Jin *et al. BMC Genomics* 2022, **23**(1):562
https://doi.org/10.1186/s12864-022-08809-w

BMC Genomics

RESEARCH

Open Access

Network regression analysis in transcriptome-wide association studies

Xiuyuan Jin[1,2], Liye Zhang[1,2], Jiadong Ji[3], Tao Ju[1,2], Jinghua Zhao[4*] and Zhongshang Yuan[1,2*]

Abstract

Background: Transcriptome-wide association studies (TWASs) have shown great promise in interpreting the findings from genome-wide association studies (GWASs) and exploring the disease mechanisms, by integrating GWAS and eQTL mapping studies. Almost all TWAS methods only focus on one gene at a time, with exception of only two published multiple-gene methods nevertheless failing to account for the inter-dependence as well as the network structure among multiple genes, which may lead to power loss in TWAS analysis as complex disease often owe to multiple genes that interact with each other as a biological network. We therefore developed a Network Regression method in a two-stage TWAS framework (NeRiT) to detect whether a given network is associated with the traits of interest. NeRiT adopts the flexible Bayesian Dirichlet process regression to obtain the gene expression prediction weights in the first stage, uses pointwise mutual information to represent the general between-node correlation in the second stage and can effectively take the network structure among different gene nodes into account.

Results: Comprehensive and realistic simulations indicated NeRiT had calibrated type I error control for testing both the node effect and edge effect, and yields higher power than the existed methods, especially in testing the edge effect. The results were consistent regardless of the GWAS sample size, the gene expression prediction model in the first step of TWAS, the network structure as well as the correlation pattern among different gene nodes. Real data applications through analyzing systolic blood pressure and diastolic blood pressure from UK Biobank showed that NeRiT can simultaneously identify the trait-related nodes as well as the trait-related edges.

Conclusions: NeRiT is a powerful and efficient network regression method in TWAS.

Keywords: TWAS, Biological networks, Dirichlet process regression, Pointwise mutual information, Blood pressure

Background

Transcriptome-wide association studies (TWASs) bridge genome-wide association studies (GWASs) and eQTL studies to make inference about the association between the genetically predicted gene expression and the phenotypes [1]. It has shown great promise in interpretation of the GWAS findings and revelation of the underlying mechanisms for disease susceptibility. It is typically done in a two-stage framework where genotype and expression data from an eQTL study are associated as the first stage to obtain the expression prediction weights, followed by the association analysis between the predicted gene expression derived from the weights from the first stage and the outcome GWAS trait. So far many statistical methods have been developed involving both stages, including for the first stage appropriate modeling of SNP effects on gene expression to improve the imputation accuracy (sparse effect as in PrediXcan [2], Bayesian sparse linear model as in TWAS [1], polygenic modeling as in PMR-Egger [3], moPMR-Egger [4], CoMM [5] and nonparametrics as in DPR [6] and TIGAR [7]),

*Correspondence: jhz22@medschl.cam.ac.uk; yuanzhongshang@sdu.edu.cn

[1] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China
[4] Department of Public Health and Primary Care, Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK
Full list of author information is available at the end of the article

constructing a composite instrumental variable [8], leveraging trans-eQTLs [9] or omics mediators [10] and epigenetic annotations [11]; for the second stage using kernel-type method [12, 13], aggregating multiple expression prediction models [14], multiple tissues [15, 16]. In addition, some methods adopted a joint likelihood-based inference procedure to improve the power [3, 4].

Almost all current TWAS methods are univariate in nature with focus on one gene at a time, which may be suboptimal due to its failure to account for the correlation among multiple gene expressions. To our knowledge, there are only two multiple-gene TWAS methods, FOCUS [17] and FOGS [18]. FOCUS extends probabilistic SNP fine-mapping approaches and models the correlation among TWAS signals to obtain risk region-based credible gene sets containing the causal gene at a given confidence level in a Bayesian framework [17]. FOGS conceptually transforms the gene-based fine-mapping into SNPs and performs conditional analysis of each specific *cis*-SNPs in one gene by adjusting the *cis*-SNPs of other genes in the same region [18]. Both FOCUS and FOGS exhibit the great advantage in modeling multiple genes over the TWAS method only modeling one gene at a time. Even so, they are unable to account for the interdependence as well as the network structure among multiple genes, thus may leading to loss of power.

A complex disease outcome is seldom the consequence of abnormality involving a single gene but often owing to multiple genes that interact with each other as a biological network whose identification can facilitate better understanding of the pathways in disease etiology. Such a network is conveniently described as a graph in which the nodes and edges are used to represent genes, and physiological interactions between nodes, respectively, so that both the node effects and edge effects can contribute to the diseases [19–22]. It is nontrivial to develop statistical methods in TWAS to detect whether a given biological network is associated with complex disease. One needs to summarize the information underlying the network, to determine a suitable measure to represent the link or connection between two nodes. It should be noted that the link may be nonlinear. We have previously proposed PMINR [22] for efficient network regression analysis, where pointwise mutual information (PMI) is used to measure the strength of the connection between a pair of nodes. PMINR has shown better performance in capturing the general relationship among different nodes in a biological network than other methods including PMNR [22], DGCA [23] and RANK [24]. Specifically, PMNR uses the common linear correlation to represent the between-node connection strength for network regression [22], DGCA is differential gene correlation analysis (i.e., edge effect) to assess the difference in gene regulatory relationships under multiple conditions [23], while RANK can detect the whole pathway due to either correlation or mean changes [24]. However, the modeling framework of PMINR requires all the gene network nodes to be observed, thus cannot be directly implemented for network regression in TWAS analysis, where the gene expression are commonly unobserved in the GWAS.

In this investigation, we developed a Network Regression method in TWAS framework, NeRiT, to detect the association between a given network and phenotypes of interest. It builds upon the two-stage analysis framework that is commonly used in TWAS, first adopts the nonparametric Bayesian Dirichlet process regression (DPR) model in the eQTL study to obtain the SNP effect size estimate on each gene within the network, given that DPR method is robust against the mis-specified distribution of SNP effect size [6]. In addition, we parallelized with Bayesian sparse linear mixed model (BSLMM) model for sensitive analysis [25]. Then, NeRiT adopts PMI to represent the between-node correlation and performs the association analysis with both the node of predicted gene expression and the edge of their correlation among these predicted values to be included in the model. In this case, it can effectively take the network structure into account, and simultaneously identify the trait-related nodes (e.g., genes) as well as the trait-related edges (e.g., gene–gene co-association).

With extensive realistic simulations, we showed that it provides calibrate type I error control for testing either the node effect or the edge effect, yields higher power, especially in testing the edge effect, than the method with product moment representing the between-node correlation. Finally, we applied NeRiT to analyze systolic blood pressure (SBP) and diastolic blood pressure (DBP) from UK Biobank to demonstrate its benefits in real data analysis.

Results

Simulations

Shown in Fig. 1 are the estimated type I error rates and statistical power of NeRiT and PMNT with the data being simulated from renin secretion network and the gene expression prediction model being constructed from DPR model. Here, PMNT is developed by replacing PMI with product moment in the proposed NeRiT framework (i.e. PM-based Network in TWAS, details in Methods). The type I error rates of both two methods were close to the given significance level ($\alpha = 0.05$) under the four simulation scenarios, regardless of the sample size, the linear or nonlinear (quadratic, sine or the combination of quadratic and sine) pattern of correlation. As expected, the power of both two methods increased with

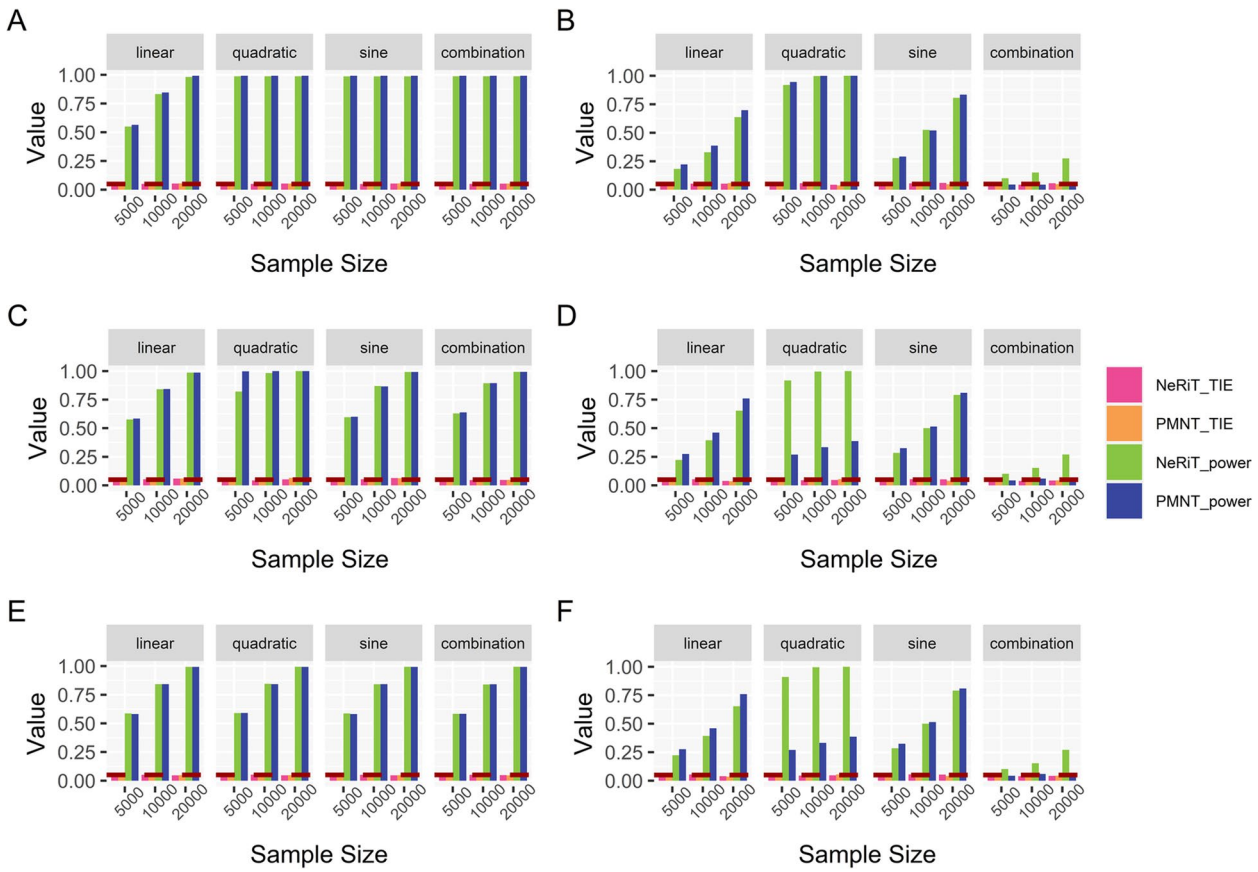

Fig. 1 Simulation results of renin secretion network under fixed effecting nodes or edges. Type I error and power of both NeRiT and PMNT with data simulated based on renin secretion network under fixed effecting nodes or edges and four different between-node correlation patterns using DPR as the imputation model in TWAS. The red dotted line represents the significance level ($\alpha = 0.05$). **A** Only node has effect; (**B**) Only edge has effect; the results for effecting node (**C**) or for effecting edge (**D**) when both node and edge change with changing node hanging on the edge; the results for effecting node (**E**) or for effecting edge (**F**) when both node and edge change with changing node not hanging on the edge

sample size under all simulation settings. In addition, both NeRiT and PMNT had comparable power to detect the effecting nodes in the settings when only node has the effect (Fig. 1A) or both node and edge have the effect (Fig. 1C and E). In detecting the effecting edge, the power of NeRiT was a little lower than that of PMNT when the inter-node correlation is linear, which was not surprising as the product moment is the gold standard to describe the inter-node relationship in this case. However, the power of NeRiT was much higher, or at least comparable, than that of PMNT when the inter-node relationship is nonlinear including quadratic, sine, as well as the combination of quadratic and sine (Fig. 1B, D, F). All results were consistent when the effecting node or edges are randomly selected (Figure S1), which illustrated that NeRiT can be reliable and robust against the specific network. In addition, similar conclusions could be drawn when the gene expression prediction model was constructed from BSLMM (Figure S2 and S3).

Such findings were also made when the data are simulated from lipid and atherosclerosis network (Fig. 2). In addition, all the results were consistent either when the effecting node or edges are randomly selected (Figure S4) or when the gene expression prediction model was constructed from BSLMM (Figure S5 and S6). Therefore, all the simulation results illustrated that NeRiT was robust against both the network size and network structure.

Applications

Shown in Table 1 are the results of the network regression in detecting the association between renin secretion network and the blood pressure traits in TWAS framework by integrating GEUVADIS data and UK Biobank GWAS. Consistent with simulations, both NeRiT and PMNT successfully detected the same node genes at a significance level of 0.05, including *CREB1* ($p = 0.001$ and 0.002 for NeRiT and PMNT, respectively) and *ADRB1* ($p = 0.029$ and 0.025 for

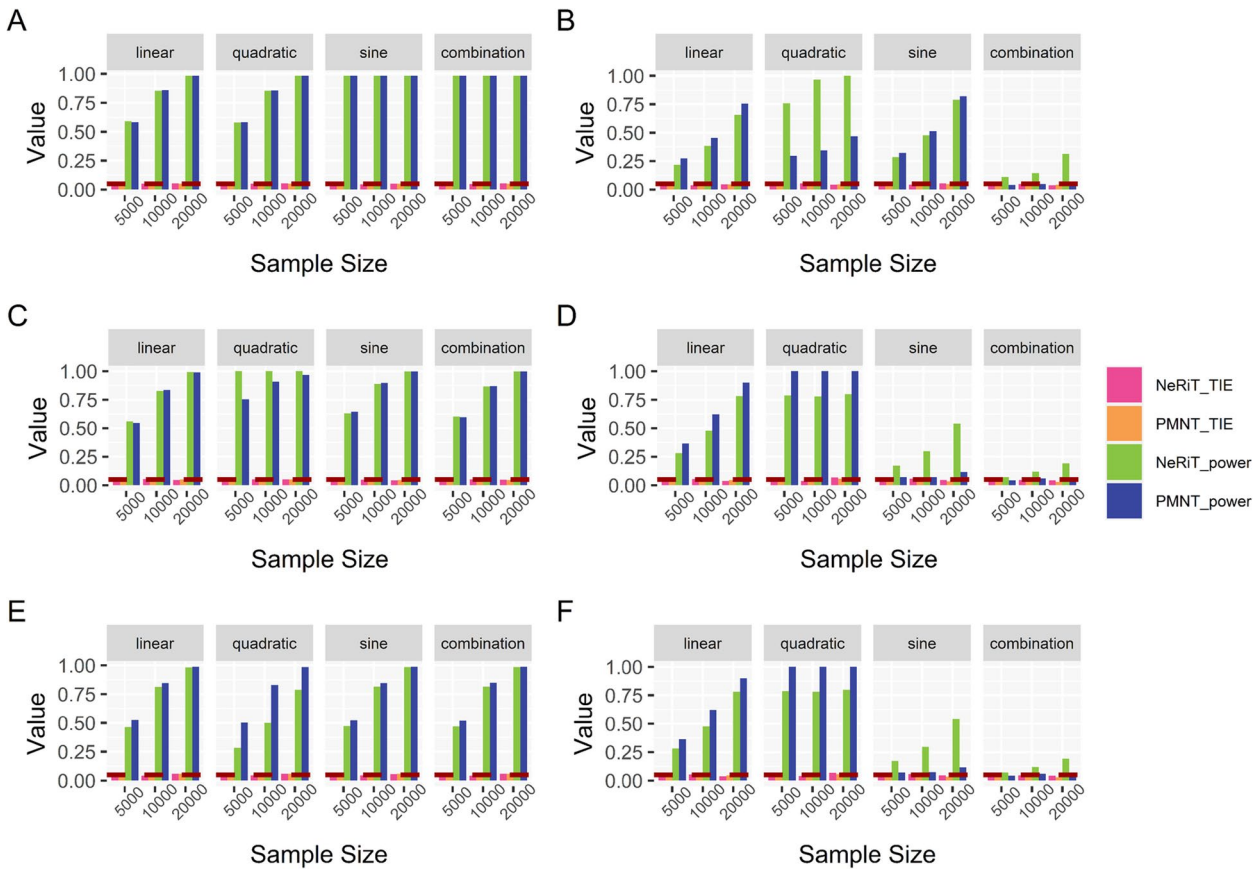

Fig. 2 Simulation results of lipid and atherosclerosis network under fixed effecting nodes or edges. Type I error and power of both NeRiT and PMNT with data simulated based on lipid and atherosclerosis network under random effecting nodes or edges and four different between-node correlation patterns using DPR as the imputation model in TWAS. The red dotted line represents the significance level ($\alpha = 0.05$). **A** Only node has effect; (**B**) Only edge has effect; the results for effecting node (**C**) or for effecting edge (**D**) when both node and edge change with changing node hanging on the edge; the results for effect node (**E**) or for effecting edge (**F**) when both node and edge change with changing node not hanging on the edge

Table 1 Renin secretion network regression of both methods with *p* values in parenthesis

Trait		NeRiT	PMNT
SBP	**Nodes**	*CREB1* (0.001) *ADRB1* (0.029)	*CREB1* (0.002) *ADRB1* (0.025)
	Edges	*GNAS-ADCY5* (0.055)	*GNAS-ADCY5* (0.025)
DBP	**Nodes**	*ADRB1* (2.540×10^{-5})	*ADRB1* (2.565×10^{-5})
	Edges	*GNAS-ADCY5* (0.024) *GNAS-PTGER2* (0.049)	

NeRiT and PMNT, respectively) for SBP and *ADRB1* ($p = 2.540 \times 10^{-5}$and 2.565×10^{-5} for NeRiT and PMNT, respectively) for DBP. For SBP, PMNT successfully identified an effecting edge *GNAS-ADCY5* ($p = 0.025$), which, to a lesser extent, has also been detected by NeRiT ($p = 0.055$). For DBP, PMNT failed to detect any effecting edges, while NeRiT successfully identify the effecting edge *GNAS-ADCY5* ($p = 0.024$) and, to a lesser extent, *GNAS-PTGER2* ($p = 0.049$). The scatter plots describing the relationship between the expression of GNAS and ADCY5 in eQTL study, as well as GNAS and PTGER2, are displayed as supplementary Figure S7 and S8, which indicates that there is no linear relationship between these genes. This highlights the important feature of PMI in capturing the nonlinear relationship and the power advantage of NeRiT.

Shown in Table 2 are the results of the network regression in detecting the association between the aldosterone-regulated sodium reabsorption network and the blood pressure traits. Consistent with simulations, both NeRiT and PMNT successfully identified the same genes at a significance level of 0.05, including *IGF1* ($p = 0.020$ and 0.021 for NeRiT and PMNT, respectively), *MAPK1* ($p = 0.028$ and 0.027 for NeRiT and

Table 2 Aldosterone-regulated sodium reabsorption network regression of both methods with *p* values in parenthesis

Trait		NeRiT	PMNT
SBP	**Nodes**	*IGF1* (0.020) *MAPK1* (0.028) *SLC9A3R2* (0.039) *IRS1* (0.047)	*IGF1* (0.021) *MAPK1* (0.027) *SLC9A3R2* (0.040) *IRS1* (0.048)
	Edges		
DBP	**Nodes**	*NEDD4L* (0.013)	*NEDD4L* (0.013)
	Edges	*SGK1-NR3C2* (0.044)	

PMNT, respectively), SLC9A3R2 ($p = 0.039$ and 0.040 for NeRiT and PMNT, respectively) and *IRS1* ($p = 0.047$ and 0.048 for NeRiT and PMNT, respectively) for SBP, and *NEDD4L* ($p = 0.013$ for both NeRiT and PMNT) for DBP. For DBP, NeRiT successfully identified the effecting edge *SGK1-NR3C2* ($p = 0.044$), while PMNT failed to detect any effecting edges. Again, Figure S9 shows the scatter plot of gene expression relationship between *SGK1* and *NR3C2* in eQTL study, which also shows the inter-node correlation is nonlinear. Again, all results were similar when using BSLMM as the gene expression prediction model (Table S1 and Table S2).

Discussion

We have presented NeRiT, a novel network regression method that detects the association between a given network and the phenotypes of interest in TWAS. It is a key step in TWAS analysis to choose the appropriate prior distribution of genotype effect size to predict gene expression, and it is often hard to determine the appropriate prior distribution of the genotype effect size since the real genetic structure is scarcely known. For network regression in TWAS, NeRiT relies on DPR to obtain the gene expression prediction weights with PMI to measure the between-node correlation and can simultaneously identify the specific gene nodes as well as edges related to the outcome traits. Comprehensive simulations illustrated that PMI can capture the general relationship among different gene nodes, and NeRiT has better performance than other competing methods.

One may be tempting to first get the PMI estimates among the network nodes of gene expression in the eQTL study, rather than among the network nodes of predicted gene expression in GWAS, given that the gene expression data are available in the eQTL study. Then, the estimate of PMI can be considered as a new exposure and the standard TWAS analysis can be conducted. However, there would be large prediction error due to the limited sample size in the eQTL study (e.g. only 465 samples in the GEUVADIS data). In addition, different from traditional TWAS analysis naturally choosing the cis-SNPs of each gene as the genotypes, it is hard to determine, both biologically and statistically, which SNPs can be chosen for the PMI between two genes as the genotypes.

Findings in our real data analysis were consistent with previous work. Loss of CREB content and function is a common, pathogenic vascular smooth muscle cells response to cardiovascular risk factors [26]. Hypertension is a multifactorial disease with a substantial genetic component. *ADRB1* is important in the regulation of blood pressure, cardiovascular function and lipid metabolism [27], and it was found that individuals with higher expression of the *ADRB1* receptor gene are at increased risk of hypertension [28]. *GNAS* implicated in variable blood pressure lowering of drug therapy in cardiovascular medicine [29]. Previous studies indicated that targeted disruption of *PTGER2* results in hypertension [30]. GNAS-ADCY5 plays a key role in a wide variety of interconnected pathways including PKA signaling and cAMP signaling, which have well-established roles in the control of blood pressure [31].

IGF1 implicated in essential hypertension [29]. *MAPK1* stimulates cardiac fibroblast and myofibroblast growth, thus contributing to the pathological actions of aldosterone in the myocardium [32]. *SLC9A3R2* is associated with SBP and/or DBP and with consistent directions of effect for SBP and DBP [33]. *NEDD4L* controls blood pressure by downregulating renal epithelial sodium channel (ENaC) expression and inhibiting sodium reabsorption, and some genetic variations in *NEDD4L* could influence the ability of the NEDD4L protein, which is significantly associated with an increased risk for adverse cardiovascular outcomes [34]. Moreover, *SGK1* phosphorylates and inactivates the ubiquitin ligase NEDD4L to reduce its interaction with the epithelial sodium channel. This consequently increases cell surface expression of the *ENaC* and thus sodium reabsorption across the apical membrane, enabling regulation of blood pressure in response to aldosterone [35].

NeRiT is not without limitations. First, the gene network structure is assumed to be known. In fact, learning gene network structure requires determining every possible edge with the highest degree of data matching, and a joint probability distribution of gene network nodes can reflect more than one network structure. Indeed, most biologists can roughly describe the specific network for the corresponding biological process, and publicly available multiple databases (such as KEGG) can also be helpful to establish the network structure. Second, the inference of PMINR directly plugs the estimate of correlation among different predicted gene expression into the regression model and fails to account for the uncertainty

during such correlation estimate, such inference procedure may lead to the biased estimate and power loss, especially in smaller sample size. Meanwhile, it ignores the direction of the link between gene codes. Third, we adjusted the p values in the real data application using Bonferroni correction as well as the FDR, but almost no significant node or edge can be detected (Tables S3, S4, S5, S6, S7, S8, S9 and S10). For network regression in TWAS, the node test and the edge test are often highly correlated, with further exacerbation since the gene expression are predicted using the cis-SNP of each gene, which are often in linkage disequilibrium. It is not straightforward to correct the p value or control the FDR. It is desirable to develop methods that can calculate the effective number of independent tests, to further address the multiple testing issue. In addition, caution should be made against the interpretation of the effect of individual node and edge, given the potential for mediation effects within the network.

Conclusions

In conclusion, NeRiT is a powerful and efficient network regression method in TWAS.

Methods

An overview

NeRiT concerns about network regression analysis in TWAS to identify the trait-associated biological network involving multiple genes from a network medicine perspective. Specifically, assume that we have a biological network with m nodes (the magnitude of each gene's expression in the regulation network) and l edges (the strength of between-node connection). We denote $\boldsymbol{X}_i (i = 1, 2, \ldots, m)$ as an n_1-vector of gene expression measurements for the i-th gene, that is measured on n_1 individuals in the gene expression study and denote $\boldsymbol{g}_i$ as an n_1 by p_i matrix of genotypes for p_i *cis*-SNPs of the i-th gene in the same study; $\boldsymbol{\eta}_i$ is a p_i-vector of SNP effect sizes on the $\boldsymbol{X}_i$. We assume $\boldsymbol{X}_i$ has been standardized with a mean of zero and a variance of one. We denote $\boldsymbol{y} = \left(y_1, y_2, \ldots, y_{n_2}\right)^{\boldsymbol{T}}$ as an n_2 -vector of outcome variable (i.e. trait) that is measured on n_2 individuals in the GWAS and denote $\boldsymbol{G}_i$ as an n_2 by p_i matrix of genotypes for the same p_i SNPs of the i -th gene. $\boldsymbol{G}_{ki} (k = 1, 2, \ldots, n_2; i = 1, 2, \ldots, m)$ denotes a 1 by p_{i} matrix of genotypes of the i-th gene for the k-th individual; $Z_{ks} (s = 1, 2, \ldots, S)$ denotes the s-th covariate for the k-th individual; E_{kij} denotes the estimator of PMI between node X_i and node X_j for the k-th individual (details in below).

NeRiT considers two linear regressions to model the gene expression study and GWAS separately in TWAS,

$$\boldsymbol{X}_i = \boldsymbol{g}_i \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_{\boldsymbol{X}_i}, (i = 1, 2, \ldots, m) \tag{1}$$

for subject $k (k = 1, 2, \ldots, n_2)$,

$$y_k = \beta_0 + \sum_{s=1}^{S} Z_{ks}\alpha_{ks} + \sum_{i=1}^{m} \boldsymbol{G}_{ki}\widehat{\boldsymbol{\eta}}_i\beta_i + \sum_{i=1}^{m}\sum_{j>i}^{m} I_{ij}E_{kij}\gamma_{ij} + e_k \tag{2}$$

where

$$I_{ij} = \begin{cases} 1 & \boldsymbol{X}_i \text{ and } \boldsymbol{X}_j \text{ are connected in the network} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Equations (1) and (2) are for the gene expression data and the GWAS data, respectively. Here, β_0 is a constant of intercept,α_{ks} is the coefficients of the s-th covariates for the k-th individual; $\widehat{\boldsymbol{\eta}}_i$ is the estimate of the SNP effect size on the i-th gene estimated by the prediction model; β_i is the causal effect of the i-th gene; γ_{ij} is the effect between the i-th gene and j-th gene. $\boldsymbol{\varepsilon}_{\boldsymbol{X}_i}$ is an n_1-vector of residual error with each element independently and identically distributed from a normal distribution$N\left(0, \sigma_X^2\right)$; $e = (e_1, e_2, \ldots, e_{n_2})$ is an n_2-vector of residual error with each element independently and identically distributed from a normal distribution$N\left(0, \sigma_y^2\right)$. Since gene expression data is unobserved in GWAS, we denote $\widehat{\boldsymbol{X}}_i = \widehat{\boldsymbol{\eta}}_i\beta_i$ as an n_2-vector of predicted gene expressions for the i-th gene, where the SNP effect $\widehat{\boldsymbol{\eta}}_i$ needs to be obtained by the gene expression prediction model (details regarding the gene expression prediction model are provided below). A key feature of NeRiT is the integration of using gene expression prediction model in the first stage and using PMI in the second stage for network regression in TWAS. It should be noted that NeRiT decomposes the change of the whole biological network into the gene node and edge changes in TWAS framework, and naturally incorporated the network structure into the model. In addition, Wald test was used to identify the gene nodes or edges that are related to the outcome traits. The NeRiT is implemented in the R package NeRiT, freely available on GitHub (https://github.com/XiuyuanJin/NeRiT).

Gene expression prediction model in TWAS

As accuracy of prediction model of gene expression is quite important for the performance of TWAS, improvement in the prediction can substantially increase the power of TWAS [36]. Different prediction models essentially differ in their assumptions about the prior distribution of the SNP effect size. In theory, the accuracy of prediction model depends on how close the prior distribution is to the real genetic structure, which is often unknown. Indeed, there are many differences in heritability, minor allele frequency and effect size across different

complex traits or diseases. Therefore, most of the existing parametric models (e.g., linear mixed model), which often use a prior effect size distribution represented by several parameters, are not sufficient to capture the true distribution of SNP effect size underlying the genetic data. In this study, we chose the non-parametric Dirichlet process regression (DPR), to construct gene expression prediction models.

DPR relies on the Dirichlet process to flexibly model the effect size distribution using infinitely many parameters and is therefore able to infer the effect size distribution from the data at hand. In addition, to investigate whether the performance of NeRiT can be influenced by the gene expression models, we alternatively adopt the commonly used Bayesian sparse linear mixed model (BSLMM), which assumes the SNP effect size on gene expressions follows two mixture normal distributions, to construct the gene expression prediction model.

Pointwise mutual information with the kernel density estimator

PMI has been illustrated to have better performance than other metrics in capturing the general relationship (linear or nonlinear) among different nodes in a biological network. For any two random variables X and Y, PMI is defined as follows [37]:

$$PMI(x;y) = \log\frac{p(x;y)}{p(x)p(y)} \tag{4}$$

where $p(x;y)$ is the joint distribution of X and Y, $p(x)$ and $p(y)$ are the marginal distributions of X and Y, respectively. Statistically, PMI can extract the general non-independency of two variables. We need to estimate the two-dimensional joint density function and marginal density function for a given sample to calculate the PMI between two network nodes. To guard against the misspecification of distribution, we chose the non-parametric kernel density estimation to characterize the corresponding distribution based on the data at hand and to improve the robustness of the PMI estimator.

The two-dimensional kernel density estimation is defined as

$$\widehat{f}_H(\boldsymbol{x}) = \frac{1}{n_2}\sum_{k=1}^{n_2} K_H(\boldsymbol{X}_k - \boldsymbol{x}) \tag{5}$$

where $\boldsymbol{X}_k = (X_{ki}, X_{kj}), k = 1, 2, ..., n_2, i \neq j, i, j = 1, 2, ..., m$ is the k-th sample of the i-th and j-th node, respective; H is a 2 by 2 bandwidth matrix, which is symmetric and positive definite; K is a bivariate kernel function and $K_H(\boldsymbol{x}) = |H|^{-\frac{1}{2}} K\left(H^{-\frac{1}{2}}\boldsymbol{x}\right)$. Here we chose the commonly-used two-dimensional normal kernel as follows:

$$K_H(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} |H|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T H^{-1}\boldsymbol{x}\right) \tag{6}$$

Simulation

Given that there are no statistical methods for network TWAS analysis yet, we performed comprehensive simulations to compare the performance of NeRiT with the method that replaces PMI with product moment in NeRiT framework (term as PM-based Network in TWAS (PMNT)). We chose this method for comparison as product moment (i.e. Pearson correlation) is commonly used to describe the dependence between two network nodes [38–41]. To make our simulation more realistic, we first mimicked a TWAS analysis by integrating the GEUVADIS [42] data with GWAS from UK Biobank [43] (details regarding these two datasets are provided below). We obtained genotype data and gene expression data from GEUVADIS and standardized the genotype and expression vector of each SNP to have a zero mean and a unit standard deviation. We then applied DPR or BSLMM to obtain the estimate of the SNP effect size $\widehat{\boldsymbol{\eta}}_i$ on gene expression, respectively. Then, we obtained genotypes for the same SNPs from UK Biobank and standardized the genotype vector of each SNP to have a zero mean and a unit standard deviation. With the standardized genotype matrix and weights vectors $\widehat{\boldsymbol{\eta}}_i$ from the previous step, we obtained the predicted gene expression. In addition, to avoid the risk of pre-specifying the network structure, we selected a realistic small network of renin secretion (Entry: hsa04924) with 13 gene nodes and 8 edges (Fig. 3) and a large network of lipid and atherosclerosis (Entry: hsa05417) with 82 gene nodes and 87 edges (Fig. 4) from Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.kegg.jp/kegg/kegg1.html), respectively. Note that we overlapped these network genes with those in the above mimicking TWAS framework to re-formulate the two biological networks.

We considered the following four scenarios for simulation:

(1) only nodes of network having the effect (e.g., node X_1 in Fig. 3),
(2) only edges of network having the effect (e.g., edge $E_{2,4}$ in Fig. 3),
(3) both nodes and edges of network having effect, with the nodes hanging on the edge (e.g., node X_{13} and edge $E_{5,13}$ in Fig. 3),

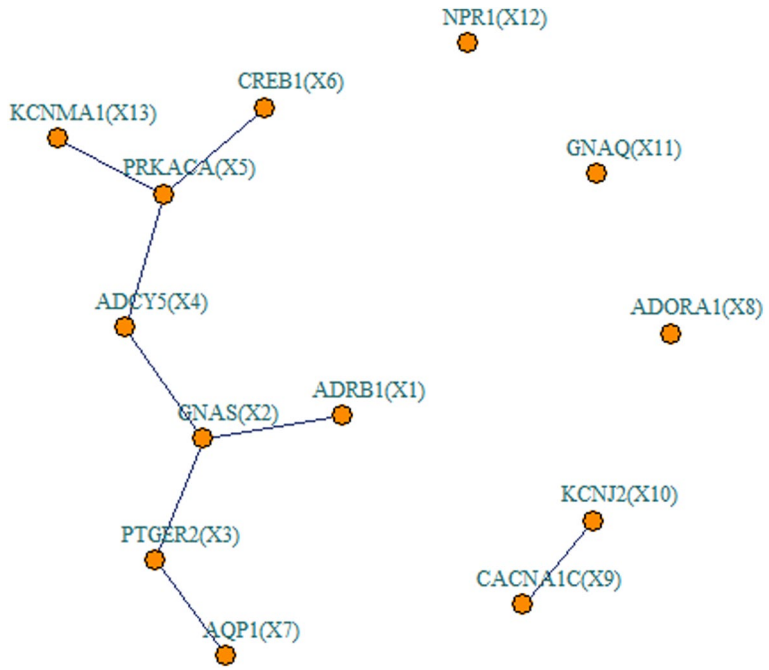

Fig. 3 The network structure based on the renin secretion pathway from KEGG

(4) both nodes and edges of network having effect, with the nodes not hanging on the edge (e.g., node X_{12} and edge $E_{5,13}$ in Fig. 3).

In each scenario, we use four inter-node relationship patterns, including the linear correlation, the quadratic relationship ($X_j = 0.5 \cdot {X_i}^2 + \varepsilon$), the sine relationship ($X_j = \sin X_i + \varepsilon$) as well as the combination of quadratic and sine relationship ($X_j = (\sin X_i)^2 + \varepsilon$), where ε is the residual error from a standard normal distribution $\varepsilon \sim N(0,1)$. For example, if we assign the quadratic relationship between node X_5 and node X_{13}, then $X_{13} = 0.5 \cdot {X_5}^2 + \varepsilon$. The nonlinear quadratic relationship between X_5 and X_{13} can be transformed to the linear relationship between ${X_5}^2$ and X_{13}, we then set $E_{5,13} = 0.5 \cdot {X_5}^2 \cdot X_{13}$ to represent the edge variable $E_{5,13}$ to simulate the traits. The type I error rate was assessed under the null hypothesis, with all node and edge effects set to be $0(\beta = 0, \gamma = 0)$, followed by the assessment of power with $\beta = 0.03, \gamma = 0.03$.

We performed 1000 simulation replicates under different sample sizes (5000, 10,000, 20,000) for each simulation replicate above. Besides pre-specifying the effecting nodes and edges, we further consider additional cases under the same above settings but randomly select the effecting nodes or edges, to eliminate the impact of network structures.

Application

We applied NeRiT through integrating gene expression data from GEUVADIS with GWASs from UK Biobank. Specifically, we obtained the GEUVADIS data as the gene expression data and examined two traits from the UK Biobank. The detailed data processing steps for the GEUVADIS data and UK Biobank data are described below.

The GEUVADIS data contains gene expression measurements for 465 individuals collected from five different populations that include CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI), and Yoruba (YRI). It performed mRNA and small RNA sequencing on 465 Epstein-Barr-virus-transformed lymphoblastoid cell line samples from five populations, and the genotype data was from the 1000 Genomes project. In the expression data, we only focused on protein coding genes and lncRNAs that are annotated in GENCODE (release 12) [44, 45]. Among these genes, we removed low-expressed genes that have zero counts in at least half of the individuals to obtain a final set of 15,810. We, following Zeng and Zhou [6], first quantile normalized the gene expression across individuals in each population to a standard normal distribution, and then normalized the gene expression to a standard normal distribution across individuals from five populations. To further remove the technical variations and batch effects, we performed PEER normalization to remove latent confounding factors for samples from five

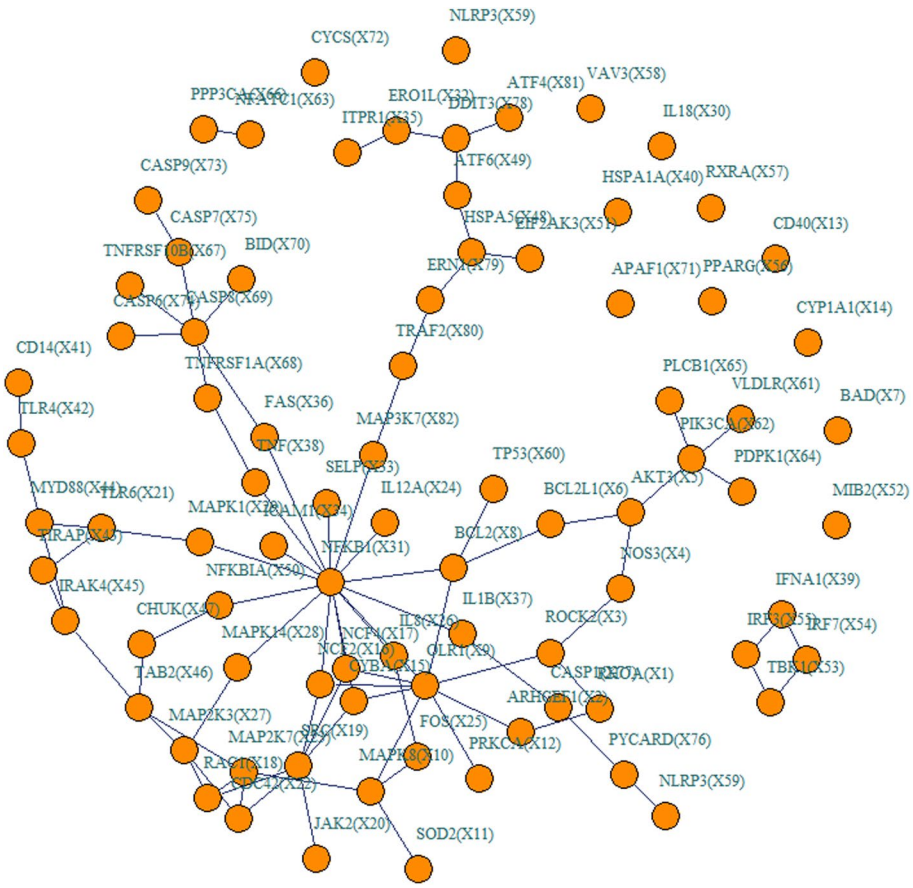

Fig. 4 The network structure based on the lipid and atherosclerosis pathway from KEGG

populations since the original gene expression measurements were read counts.

Besides the expression data, all individuals in GEUVADIS also have their genotypes sequenced in the 1000 Genomes Projects. We obtained genotype data from the 1000 Genomes Project phase 3. We filtered out SNPs that have a Hardy–Weinberg equilibrium (HWE) p value $< 10^{-4}$, a genotype call rate $< 95\%$, or a minor allele frequency (MAF) < 0.001. We retained a total of 7,072,917 SNPs for analysis.

The UK Biobank data consists of 487,298 individuals and 92,693,895 imputed SNPs [43]. We followed the same sample QC procedure in Neale lab (Web Resources) to retain a total of 337,129 individuals of European ancestry. We filtered out SNPs with an HWE p value $< 10^{-7}$, a genotype call rate $< 95\%$, or an MAF < 0.001 to obtain a total of 13,876,958 SNPs. For each trait in turn, we regressed the resulting standardized phenotypes on sex and top genotype principal components (PCs) to obtain the residuals, standardized the residuals to have a mean of zero and a standard deviation of one, and finally used these scaled residuals to conduct TWAS analysis.

We integrated the GEUVADIS data with GWAS from UK Biobank for TWAS analysis. For each gene in turn in the GEUVADIS data, we extracted *cis*-SNPs that are within either 100 kb upstream of the transcription start site (TSS) or 100 kb downstream of the transcription end site (TES). We overlapped these *cis*-SNPs of genes in GEUVADIS with the SNPs obtained from UK Biobank to obtain common sets of SNPs.

Here we focused on the UK Biobank GWAS of systolic blood pressure (SBP) and diastolic blood pressure (DBP) to investigate the association between the blood pressure and two biological networks from KEGG, one is renin secretion network (Entry: hsa04924) with 13 gene nodes and 8 edges (Fig. 3), the other is aldosterone-regulated sodium reabsorption (Entry: hsa04960) network with 12 gene nodes and 7 edges (Fig. 5). Note that we overlapped these network genes with those in the above mimicking TWAS framework to re-formulate the two biological networks. Cardiovascular diseases are a leading cause of death globally. The reason that we chose the blood pressure traits is that elevated blood pressure is a major risk factor for cardiovascular morbidity and mortality [46].

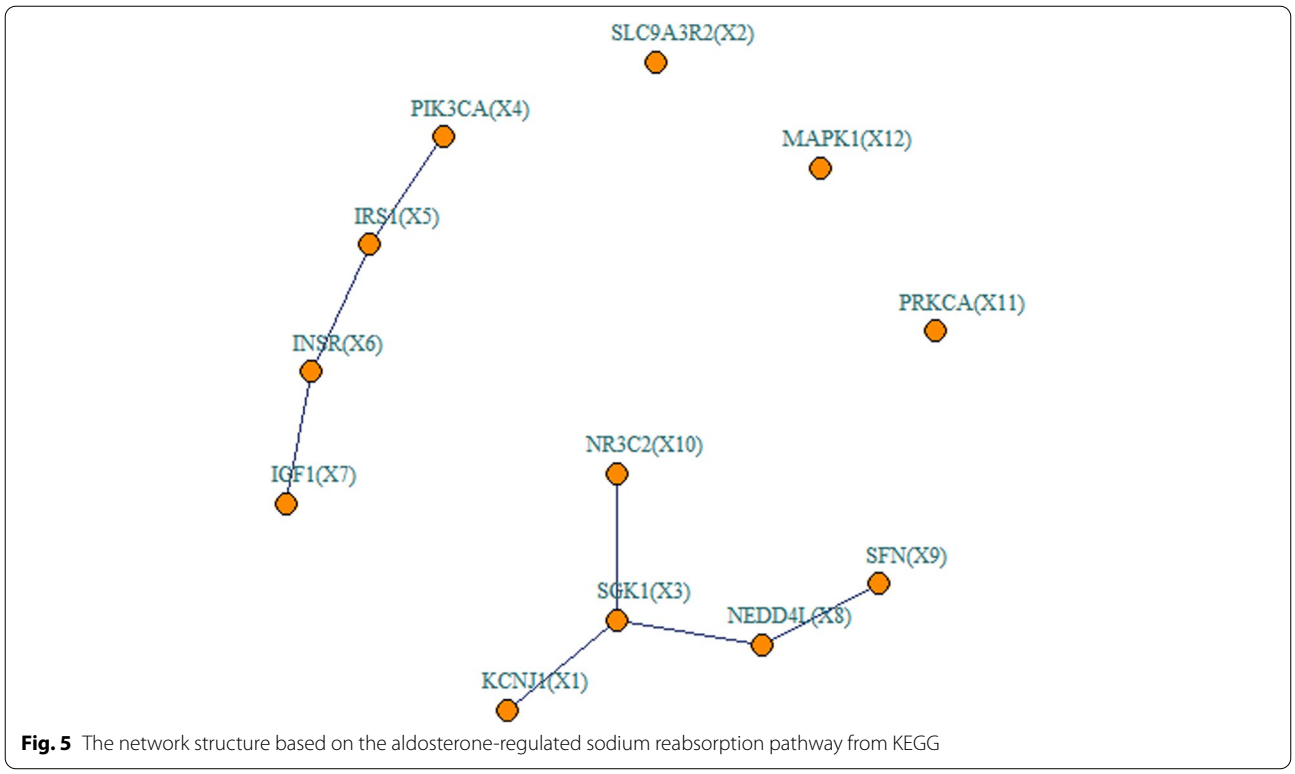

Fig. 5 The network structure based on the aldosterone-regulated sodium reabsorption pathway from KEGG

The SNP heritability of blood pressure was estimated in the range of 0.3–0.5 in previous studies [47]. In addition, the renin–angiotensin–aldosterone system (RAAS) is a critical regulator of blood volume and systemic vascular resistance. RASS is composed of renin, angiotensin and aldosterone, these three major compounds act together to elevate arterial pressure in response to decreased renal blood pressure, decreased salt delivery to the distal convoluted tubule, and/or beta-agonism. Through these mechanisms, the body can elevate blood pressure in a prolonged manner [48]. It should be noted that one important feature of NeRiT was that NeRiT can detect whether the whole network or gene or inter-gene correlation is associated with the blood pressure traits.

Web Resources

KEGG, www.kegg.jp/kegg/kegg1.html
GEUVADIS, http://www.geuvadis.org
UK Biobank, https://www.ukbiobank.ac.uk/
Sample QC procedure in Neale lab, https://github.com/Nealelab/UK_Biobank_GWAS/tree/master/imputed-v2-gwas

Abbreviations

TWAS: Transcriptome-wide association study; GWAS: Genome-wide association study; eQTL: Expression quantitative trait locus; PMI: Pointwise mutual information; DPR: Dirichlet process regression; BSLMM: Bayesian sparse linear mixed model; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; KEGG: Kyoto Encyclopedia of Genes and Genomes; SNP: Single nucleotide polypeptide; RAAS: Renin–angiotensin–aldosterone system.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08809-w.

Additional file 1: Figure S1. Simulation results of renin secretion network under random effecting nodes or edges. **Figure S2.** Simulation results of renin secretion network under fixed effecting nodes or edges. **Figure S3**. Simulation results of renin secretion network under random effecting nodes or edges. **Figure S4**. Simulation results of lipid and atherosclerosis network under random effecting nodes or edges. **Figure S5**. Simulation results of lipid and atherosclerosis network under fixed effecting nodes or edges. **Figure S6**. Simulation results of lipid and atherosclerosis network under random effecting nodes or edges. **Figure S7.** The scatter plots of relationship between the expression of *GNAS* and *ADCY5* in eQTL study. **Figure S8.** The scatter plots of relationship between the expression of *GNAS* and *PTGER2* in eQTL study. **Figure S9.** The scatter plots of relationship between the expression of *SGK1* and *NR3C2* in eQTL study. **Table S1.** Renin secretion network regression of both methods with *p* values in parenthesis. **Table S2.** Aldosterone-regulated sodium reabsorption network regression of both methods with *p* values in parenthesis. **Table S3.** Results of the renin secretion network regression on SBP using DPR as the imputation model. **Table S4.** Results of the renin secretion network regression on DBP using DPR as the imputation model. **Table S5.** Results of the renin secretion network regression on SBP using BSLMM as the imputation model. **Table S6.** Results of the renin secretion network regression on DBP using BSLMM as the imputation model. **Table S7.** Results of the aldosterone-regulated sodium reabsorption network regression on SBP using DPR as the imputation model. **Table S8.** Results of the aldosterone-regulated sodium reabsorption network regression on DBP using DPR as the imputation model. **Table S9.** Results of the

aldosterone-regulated sodium reabsorption network regression on SBP using BSLMM as the imputation model. **Table S10.** Results of the aldosterone-regulated sodium reabsorption network regression on DBP using BSLMM as the imputation model.

Acknowledgements
The authors are grateful to UK Biobank resource.

Authors' contributions
ZY conceived the study. XJ, LZ and TJ processed the data. XJ, JJ and ZY verified all the data in the study. XJ performed the analyses and interpreted the results. All authors contributed to the initial draft; XJ, ZY and JZ made numerous revisions. All authors read and approved the final manuscript.

Funding
This study was supported by the National Natural Science Foundation of China (81872712 and 82173624), the Natural Science Foundation of Shandong Province (ZR2019ZD02).

Availability of data and materials
No data were generated in the present study. The GEUVADIS gene expression data are publicly available online. The UK Biobank data is from UK Biobank resource under application number 51470.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China. [2]Institute for Medical Dataology, Shandong University, Jinan 250003, Shandong, China. [3]Institute for Financial Studies, Shandong University, Jinan 250100, Shandong, China. [4]Department of Public Health and Primary Care, Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK.

Received: 12 January 2022 Accepted: 2 August 2022
Published: 6 August 2022

References

1. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
2. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquinomichaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8.
3. Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, Liu J, Zhou X. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. Nat Commun. 2020;11(1):3861.
4. Liu L, Zeng P, Xue F, Yuan Z, Zhou X. Multi-trait transcriptome-wide association studies with probabilistic Mendelian randomization. Am J Hum Genet. 2021;108(2):240–56.
5. Yang C, Wan X, Lin X, Chen M, Zhou X, Liu J. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. Bioinformatics. 2019;35(10):1644–52.
6. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Commun. 2017;8(1):456.
7. Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, Gibson G, De Jager PL, Bennett DA, Wingo AP, Wingo TS. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. Am J Hum Genet. 2019;105:258–66.
8. Zhang Y, Quick C, Yu K, Barbeira A, Consortium GT, Luca F, Pique-Regi R, Kyung Im H, Wen X. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. Genome Biol. 2020;21(1):232.
9. Luningham JM, Chen J, Tang S, De Jager PL, Bennett DA, Buchman AS, Yang J. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. Am J Hum Genet. 2020;107(4):714–26.
10. Bhattacharya A, Li Y, Love MI. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. PLoS Genet. 2021;17(3):e1009398.
11. Zhang W, Voloudakis G, Rajagopal VM, Readhead B, Dudley JT, Schadt EE, Bjorkegren JLM, Kim Y, Fullard JF, Hoffman GE, et al. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. Nat Commun. 2019;10(1):3834.
12. Cao C, Kwok D, Edie S, Li Q, Ding B, Kossinna P, Campbell S, Wu J, Greenberg M, Long Q. kTWAS: integrating kernel machine with transcriptome-wide association studies improves statistical power and reveals novel genes. Brief Bioinform. 2021;22(4):bbaa270.
13. Tang S, Buchman AS, De Jager PL, Bennett DA, Epstein MP, Yang J. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. PLoS Genet. 2021;17(4):e1009482.
14. Zeng P, Dai J, Jin S, Zhou X. Aggregating multiple expression prediction models improves the power of transcriptome-wide association studies. Hum Mol Genet. 2021;30(10):939–51.
15. Zuber V, Colijn JM, Klaver C, Burgess S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nat Commun. 2020;11(1):29.
16. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. 2019;15(1):e1007889.
17. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, Pasaniuc B. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019;51(4):675.
18. Wu C, Pan W. A powerful fine-mapping method for transcriptome-wide association studies. Hum Genet. 2020;139(2):199–213.
19. Ji J, Yuan Z, Zhang X, Xue F. A powerful score-based statistical test for group difference in weighted biological networks. BMC Bioinformatics. 2016;17:86.
20. Yuan Z, Ji J, Zhang X, Xu J, Ma D, Xue F. A powerful weighted statistic for detecting group differences of directed biological networks. Sci Rep. 2016;6:34159.
21. Zhu Y, Ji J, Lin W, Li M, Liu L, Zhu H, Xue F, Li X, Zhou X, Yuan Z. MCC-SP: a powerful integration method for identification of causal pathways from genetic variants to complex disease. BMC Genet. 2020;21(1):90.
22. Lin W, Ji J, Zhu Y, Li M, Zhao J, Xue F, Yuan Z. PMINR: Pointwise Mutual Information-Based Network Regression - With Application to Studies of Lung Cancer and Alzheimer's Disease. Front Genet. 2020;11:556259.
23. McKenzie AT, Katsyv I, Song WM, Wang M, Zhang B. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. BMC Syst Biol. 2016;10(1):106.
24. Alvo M, Liu Z, Williams A, Yauk C. Testing for mean and correlation changes in microarray experiments: an application for pathway analysis. BMC Bioinformatics. 2010;11:60.
25. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.
26. Schauer IE, Knaub LA, Lloyd M, Watson PA, Gliwa C, Lewis KE, Chait A, Klemm DJ, Gunter JM, Bouchard R, et al. CREB downregulation in vascular disease: a common response to cardiovascular risk. Arterioscler Thromb Vasc Biol. 2010;30(4):733–41.
27. Tikhonoff V, Hasenkamp S, Kuznetsova T, Thijs L, Jin Y, Richart T, Zhang H, Brand-Herrmann SM, Brand E, Casiglia E, et al. Blood pressure and metabolic phenotypes in relation to the ADRB1 Arg389Gly and ADRA2B I/D polymorphisms in a White population. J Hum Hypertens. 2008;22(12):864–7.

28. Bengtsson K, Melander O, Orho-Melander M, Lindblad U, Ranstam J, Råstam L, Groop L. Polymorphism in the beta(1)-adrenergic receptor gene and hypertension. Circulation. 2001;104(2):187–90.
29. Arnett DK, Baird AE, Barkley RA, Basson CT, Boerwinkle E, Ganesh SK, Herrington DM, Hong Y, Jaquish C, McDermott DA, et al. Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on Epidemiology and Prevention, the Stroke Council, and the Functional Genomics and Translational Biology Interdisciplinary Working Group. Circulation. 2007;115(22):2878–901.
30. Kennedy CR, Zhang Y, Brandon S, Guan Y, Coffee K, Funk CD, Magnuson MA, Oates JA, Breyer MD, Breyer RM. Salt-sensitive hypertension and reduced fertility in mice lacking the prostaglandin EP2 receptor. Nat Med. 1999;5(2):217–20.
31. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics. 2008;92(5):265–72.
32. Stockand JD, Meszaros JG. Aldosterone stimulates proliferation of cardiac fibroblasts by activating Ki-RasA and MAPK1/2 signaling. Am J Physiol Heart Circ Physiol. 2003;284(1):H176-184.
33. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, Torstenson ES, Kovesdy CP, Sun YV, Wilson OD, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat Genet. 2019;51(1):51–62.
34. McDonough CW, Burbage SE, Duarte JD, Gong Y, Langaee TY, Turner ST, Gums JG, Chapman AB, Bailey KR, Beitelshees AL, et al. Association of variants in NEDD4L with blood pressure response and adverse cardiovascular outcomes in hypertensive patients treated with thiazide diuretics. J Hypertens. 2013;31(4):698–704.
35. Haas JG, Weber J, Gonzalez O, Zimmer R, Griffiths SJ. Antiviral activity of the mineralocorticoid receptor NR3C2 against Herpes simplex virus Type 1 (HSV-1) infection. Sci Rep. 2018;8(1):15876.
36. Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 2020;52(11):1239–46.
37. Church KW, Hanks P. Word association norms, mutual information, and lexicography. In: Proceedings of the 27th annual meeting on Association for Computational Linguistics. Vancouver, British Columbia, Canada: Association for Computational Linguistics; 1989. p. 76–83.
38 Newman ME. Assortative mixing in networks. Phys Rev Lett. 2002;89(20):208701.
39. Foster JG, Foster DV, Grassberger P, Paczuski M. Edge direction and the structure of networks. Proc Natl Acad Sci USA. 2010;107(24):10815–20.
40. Zhang WJS. Prediction of missing connections in the network:a node-similarity based algorithm. Selforganizology. 2015;2(4):91–101.
41. Barzel B, Barabási AL. Network link prediction by global silencing of indirect correlations. Nat Biotechnol. 2013;31(8):720–5.
42. Lappalainen T, Sammeth M, Friedländer MR, t Hoen PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11.
43. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9.
44. Wen X, Luca F, Pique-Regi R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS Genet. 2015;11(4):e1005176.
45. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
46. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet (London, England). 2002;360(9349):1903–13.
47. Ehret GB, Caulfield MJ. Genes for blood pressure: an opportunity to understand hypertension. Eur Heart J. 2013;34(13):951–61.
48. Hall JE, do Carmo JM, da Silva AA, Wang Z, Hall ME. Obesity, kidney dysfunction and hypertension: mechanistic links. Nat Rev Nephrol. 2019;15(6):367–85.

Publisher's Note