genes

MDPI

Article

CoNet: Efficient Network Regression for Survival Analysis in Transcriptome-Wide Association Studies—With Applications to Studies of Breast Cancer

Jiayi Han [1,2], Liye Zhang [1,2], Ran Yan [1,2], Tao Ju [1,2], Xiuyuan Jin [1,2], Shukang Wang [1,2], Zhongshang Yuan [1,2] and Jiadong Ji [3,*]

[1] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, China
[2] Institute for Medical Dataology, Shandong University, Jinan 250003, China
[3] Institute for Financial Studies, Shandong University, Jinan 250100, China
* Correspondence: jiadong@sdu.edu.cn

Abstract: Transcriptome-wide association studies (TWASs) aim to detect associations between genetically predicted gene expression and complex diseases or traits through integrating genome-wide association studies (GWASs) and expression quantitative trait loci (eQTL) mapping studies. Most current TWAS methods analyze one gene at a time, ignoring the correlations between multiple genes. Few of the existing TWAS methods focus on survival outcomes. Here, we propose a novel method, namely a COx proportional hazards model for NEtwork regression in TWAS (CoNet), that is applicable for identifying the association between one given network and the survival time. CoNet considers the general relationship among the predicted gene expression as edges of the network and quantifies it through pointwise mutual information (PMI), which is under a two-stage TWAS. Extensive simulation studies illustrate that CoNet can not only achieve type I error calibration control in testing both the node effect and edge effect, but it can also gain more power compared with currently available methods. In addition, it demonstrates superior performance in real data application, namely utilizing the breast cancer survival data of UK Biobank. CoNet effectively accounts for network structure and can simultaneously identify the potential effecting nodes and edges that are related to survival outcomes in TWAS.

Keywords: TWAS; biological network; breast cancer; survival analysis

Citation: Han, J.; Zhang, L.; Yan, R.; Ju, T.; Jin, X.; Wang, S.; Yuan, Z.; Ji, J. CoNet: Efficient Network Regression for Survival Analysis in Transcriptome-Wide Association Studies—With Applications to Studies of Breast Cancer. *Genes* **2023**, *14*, 586. https://doi.org/10.3390/genes14030586

Academic Editors: Jingyun Yang and Chuntao Zhao

Received: 24 December 2022
Revised: 23 February 2023
Accepted: 23 February 2023
Published: 25 February 2023

1. Introduction

Genome-wide association studies (GWASs) have detected hundreds of thousands of single nucleotide polymorphisms (SNPs) that are related with complex diseases, including various cancers [1]. However, most GWAS signals are located in non-coding regions across the genome [2], leading to difficulties in the validation and interpretation of associations, and challenges in uncovering the regulatory mechanism underlying the disease. Concurrently, expression quantitative trait loci (eQTL) mapping studies have successfully detected several genetic variants that are related to gene expression. By integrating GWAS and eQTL studies, the recently developed transcriptome-wide association study (TWAS) provides a promising technique for interpreting GWAS associations and identifying disease-related genes. TWAS has facilitated the identification of potential genes that have expression values associated with various GWAS outcome traits, such as lung cancer [3], pancreatic cancer [4], and schizophrenia [5]. In typical TWAS analysis, the effect size of the genotype on gene expression is first estimated from the eQTL study, which is further used to predict gene expression in GWAS. Then, the regression association analysis is usually conducted between the gene expression prediction and the trait in GWAS. To date, many TWAS statistical tools have been developed; some methods aim to improve the performance

of genotype effect size estimation using different models (e.g., PrediXcan [6], TWAS [7], DPR [8], and TIGAR [9]), some aim to improve the statistical power in the final association analysis of TWAS (e.g., kernel-type methods [10]), and some make the statistical inference in a likelihood framework to account for the uncertainty in the estimation of the genotype effect size (e.g., PMR-Egger [11] and moPMR-Egger [12]).

It should be noted that most currently available TWAS methods encounter important challenges. The first is that most current TWAS analyses can only focus on one gene at a time, thus ignoring the correlation structure among multiple genes. To the best of our knowledge, FOCUS [13] and FOGS [14] are the only two existing multiple gene-based TWAS methods. FOCUS constructs the multiple-gene TWAS model from a Bayesian perspective, aiming to obtain credible gene sets that contain all the associated genes at a nominal confidence level. FOGS is essentially a multiple SNP model, which identifies genes based on conditional analyses of SNPs of each gene by adjusting the other SNPs residing in the same region. However, both methods fail to take the network relationship among multiple genes into account, thus resulting in these methods possibly losing efficiency.

A complex disease can reflect the interactions among multiple genes in a biological network [15]. Identifying the specific biological network related with a complex disease can help explore the network mechanism of complex diseases. Often, in a multiple-gene network, nodes are used to represent genes and edges are used to represent the possible interactions between different genes. Correspondingly, genes and their interactions included in the network can make contributions to the development of disease. Quantifying the correlation between nodes to represent the edge is challenging: it is not easy to determine the suitable measure to capture the general between-node correlations. PMI was confirmed as an efficient measure to represent the complex relationship among different network nodes in the network regression [16].

Previously, we have developed two network regression method in TWAS: the NeRiT [17] method for continuous outcomes and PoLoNet [18] for binary or categorical outcomes, which illustrated the advantage of PMI in capturing the general relationship among different network nodes. However, these methods cannot be easily extended to the survival outcomes context given that the distribution of the survival time is often non-normal, coupled with the censoring issue. On the other hand, time-to-event data are commonly encountered in GWAS, especially in cancer genomics [19,20]. For example, in the UK Biobank, the breast cancer survival time is often the main outcome for exploring the biological network related to breast cancer progression, but some participants may be censored and are unable to experience the event by the end of the follow-up.

In this study, a COx proportional hazards model for NEtwork regression in TWAS (CoNet) was developed to detect the association between one given network and the survival time. CoNet is developed under a two-stage TWAS framework. In the first stage, the SNP effect size for each specific gene within one network is estimated in the eQTL study. In the second stage, CoNet adopts PMI to quantify the edges of the network to describe the general relationship among the nodes and conducts the association analysis with all the nodes along with all the edges in the model. With the network structure effectively accounted for, CoNet can simultaneously identify the potential nodes and edges that are associated with the survival time. Comprehensive and extensive realistic simulations are conducted to evaluate the performance of CoNet, including the type I error control and power for detecting either the node effect or the edge effect. In addition, breast cancer survival data in UK Biobank were used to highlight the advantages of CoNet in real applications.

2. Materials and Methods

2.1. eQTL Study

Suppose there are a total of m genes and $\boldsymbol{x_i}$ is an n_1-vector of gene expression measurements for the i-th gene, which is measured on n_1 individuals in the gene expression study. $\boldsymbol{G_{x_i}}$ is denoted as an n_1 by p_i matrix of genotypes for p_i cis-SNPs (1 Mb windows around

the gene) of the i-th gene. In two-stage TWAS, it is common to obtain the genotype effect size using the following model:

$$\boldsymbol{x_i} = \boldsymbol{G_{x_i}}\boldsymbol{\beta_i} + \boldsymbol{\varepsilon_{x_i}}, (i = 1, 2, \ldots, m) \quad (1)$$

where $\boldsymbol{\beta_i}$ is a p_i-vector of cis-SNP effect sizes on the i-th gene expression and $\boldsymbol{\varepsilon_{x_i}}$ is an error term with n_1 sampled values that is distributed from a normal distribution $N\left(0, \sigma_x^2\right)$.

2.2. Gene Expression Prediction

The estimator of genotype effect size $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ can be obtained from the eQTL study and the predicted gene expression of the i-th gene is derived as $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{G_{yi}}\,\widehat{\boldsymbol{\beta}}_i$, where $\boldsymbol{G_{y_i}}$ is the n_2 by p_i matrix of genotypes for the same p_i cis-SNPs of the i-th gene measured on n_2 individuals in the GWAS study.

CoNet aims to detect the association between a given network and the survival time under the two-stage TWAS framework, with the prior known network structure. The performance of TWAS depends on the estimation accuracy of the effect of cis-SNP on the gene expression [21] which is strongly associated with the extent of the consistency between the assumed prior distribution and the true distribution of the genetic effect size. However, it is usually hard to access the true distribution of SNP effect size. As BSLMM and DPR modeling assumptions are more flexible than the normality hypothesis and tend to outperform sparse models in predicting gene expression in TWAS applications [22]. Here, we choose the non-parametric Dirichlet process regression (DPR) [8] to model the SNP effect size due to its robustness to the distribution of the genetic effect size. Furthermore, the Bayesian sparse linear mixed model (BSLMM) [23] was also applied for sensitivity analysis. BSLMM is a hybrid modeling assumption between a sparse modeling assumption and the standard polygenic modeling assumption. The SNP effect size is assumed to follow a mixture of two normal distributions in the BSLMM model. We compare the performance of CoNet when using the DPR model in the first stage and when using the BSLMM model in the first stage. Evaluation of the performance of CoNet cannot be substantially influenced by the model assumption of the genetic effect size.

For subject j in the GWAS study, $j = 1, \ldots, n_2$, the CoNet model based on the Cox Proportional Hazard Model is defined as:

$$h\left(t; Z_{ij}, \widetilde{x}_{ij}, E_{jlk}\right) = h_0(t) exp\left(\mu_0 + \sum_{i=1}^{s} Z_{ij}\alpha_i + \sum_{i=1}^{m} \widetilde{x}_{ij}\eta_i + \sum_{l=1}^{m}\sum_{k>l}^{m} I_{lk}E_{jlk}\gamma_{lk}\right) \quad (2)$$

where

$$I_{lk} = \begin{cases} 1 & l-th\ gene\ and\ k-th\ gene\ are\ connected\ in\ the\ network \\ 0 & otherwise \end{cases}$$

and $h_0(t)$ is the baseline hazard function, $\tilde{\boldsymbol{x}}_{\boldsymbol{ij}}$ denotes the predicted gene expression of the i-th gene derived from the above model (4), E_{jlk} is the estimator of PMI between the l-th and k-th node estimated from the BKDE method for the j-th individual, $\boldsymbol{Z_j} = \left(Z_{1j}, Z_{2j}, \ldots, Z_{sj}\right)^T$ denotes an s-vector of the covariates for the j-th individual (e.g., top five genotype principle components (PCs)), $\alpha_i (i = 1, \ldots, s)$ represents the coefficient of the corresponding covariate, $\eta_i (i = 1, \ldots, m)$ is the effect of the i-th node, and γ_{lk} indicates the effect of the edge linking the l-th and the k-th node.

CoNet obtains the predicted gene expression included in the target network with the DPR model, then uses PMI to describe the correlations between nodes. After directly plugging the predicted gene expression and PMI estimator among different network nodes, CoNet is regarded as a Cox proportional model. The main goal is to estimate and test the node effects η and edge effects γ. The partial likelihood framework can be utilized to infer the parameters of interest and obtain estimates of $\hat{\eta}$ and $\hat{\gamma}$, as well as their standard error values $se(\hat{\eta})$ and $se(\hat{\gamma})$. Subsequently, the corresponding Wald test can be constructed to obtain a p-value for hypothesis testing. CoNet is computationally scalable and imple-

mented in an R package, which is available at https://github.com/hanjiayi626/CoNet (accessed on 20 April 2022).

2.3. PMI between Two Predicted Gene Expressions in TWAS Framework

PMI is commonly defined for discrete variables, whereas here we used it for continuous variables. The PMI between $\widetilde{x}_i$ and $\widetilde{x}_j$ is defined as the log ratio between their joint distribution $\boldsymbol{p}(\widetilde{x}_i, \widetilde{x}_j)$ and the product of their marginal distribution $p(\widetilde{x}_i)\boldsymbol{p}(\widetilde{x}_j)$ [24]:

$$PMI(\widetilde{x}_i, \widetilde{x}_j) = log\frac{p(\widetilde{x}_i, \widetilde{x}_j)}{p(\widetilde{x}_i)p(\widetilde{x}_j)} \tag{3}$$

The marginal distribution $p(\widetilde{x}_i)$ and $p(\widetilde{x}_j)$ can be estimated using the kernel density estimation method. The joint distribution $p(\widetilde{x}_i, \widetilde{x}_j)$ can be estimated using the bivariate kernel density estimation (BKDE) method, which is non-parametric and robust against the misspecification of data distribution [25,26]. Assume $Z_i = (X_i, Y_i)^T, i = 1, 2, \ldots, n$, is a bivariate sample from a bivariate distribution p. The BKDE is:

$$\hat{p}_H(z; H) = n^{-1} \sum_{i=1}^{n} K_H(z - Z_i) \tag{4}$$

where $z = (x, y)^T$ and H is the bandwidth (or smoothing) matrix which is symmetric and positive-definite.

$$H = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}$$

Bivariate kernel function K is symmetric, with $K_H(z) = |H|^{-\frac{1}{2}} K\left(H^{-\frac{1}{2}} z\right)$. This study used the bivariate normal kernel:

$$K_H(z) = (2\pi)^{-\frac{d}{2}} |H|^{-\frac{1}{2}} exp\left(-\frac{1}{2} z^T H^{-1} z\right) \tag{5}$$

with d = 2.

2.4. Simulation Study

No other methods have been developed so far for network regression for failure time data in TWAS. Therefore, extensive simulations were performed to compare CoNet with the CPNT method, which was intuitively developed under the CoNet framework but replaced PMI with the product moment (PM) to describe the edges of the network. The expectation of the product of the two scaled node variables (PM) is the linear correlation coefficient, and the PM can be regarded as the individual observed value of the correlation coefficient. To make these simulations more realistic, a realistic TWAS setting was mimicked by integrating data from the GEUVADIS study and GWAS from the UK Biobank. Data from GEUVADIS ($n_1 = 465$) were obtained, then each SNP was standardized along with the gene expression vector (the expression of a specific gene for all the individuals) to obtain a zero mean and a unit standard deviation. For each gene, we chose two models, DPR or BSLMM, to estimate the effects of cis-SNPs on gene expression. The same SNPs were obtained from the UK Biobank, and the genotype vector of each SNP was also standardized. Next, the predicted gene expression was obtained using the standardized genotype matrix. The survival phenotype was also simulated using the above gene expression prediction. Additionally, the PI3K-AKT signaling pathway (hsa04151-nt06214) from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database was selected as the network. All the genes in the pathway overlapped with those in the UK Biobank, containing a total of ten nodes and ten edges (Figure 1).

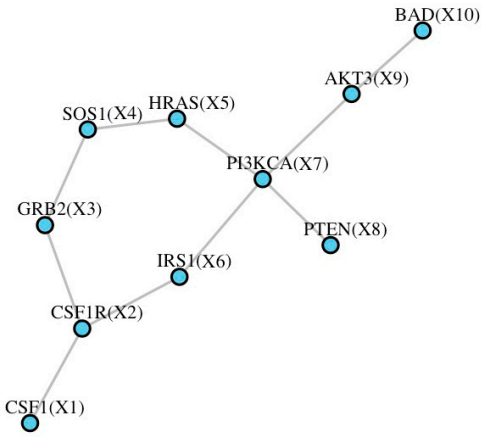

Figure 1. The simulated network for the PI3K-AKT signaling pathway from KEGG.

For subject j ($j = 1, \ldots, n_2$), according to Bender's method [27] the complete survival time T_j^* was simulated from a Cox model with a Weibull baseline hazard function as $T_j^* = -\left(\frac{log(U_j)}{\lambda exp(w_j)}\right)^{\frac{1}{v}}$, where the scale parameter v was set as 0.5 and shape parameter λ was set as 1. Furthermore, U_j was simulated from a uniform (0,1) distribution and $w_j = 0.5z_{1j} + 0.5z_{2j} + \sum_{i=1}^{m} \widetilde{x}_{ij}\eta_i + \sum_{l=1}^{m}\sum_{k>l}^{m} I_{lk}E_{jlk}\gamma_{lk}$. The covariate z_{1j} was simulated from a standard normal distribution $N(0,1)$ and z_{2j} from a Bernoulli (0.5) distribution. Various GWAS sample sizes ($n_2 = 5000,\ 10,000,\ 20,000$) were randomly selected from the 337,129 individuals from the UK Biobank. In addition, censoring time C_j was randomly simulated on a uniform distribution U (0, T_j^*) and n_2q censored observations were randomly selected based on the pre-specified censoring rate q. Next, the observed time-to-event T_j and indicator variables δ_j (0 for censored data and 1 for event) were obtained. The censoring rate was set at 10%, 30%, and 50%. The type I error rates were assessed with $\eta = 0$ and $\gamma = 0$. Empirical power was evaluated with the effects of genes and edges were set to be $\eta = 0.05$ and $\gamma = 0.05$. These effects were calculated to be the 50% quantile from the effect estimate from the real data. Briefly, four scenarios were designed regarding different patterns of the network effect: (1) only nodes of the network have effects (e.g., X_6); (2) only edges of the network have effects (e.g., E_{7_9}); (3) both nodes and edges have effects, with the nodes hanging on the edges (e.g., X_4 and E_{4_5}); and (4) both nodes and edges have effects, with the nodes not hanging on the edges (e.g., X_2 and E_{4_5}). In each scenario, four between-node correlation patterns were considered, including a simple linear correlation, quadratic relationship ($x_k = 0.1{x_l}^2$), sine relationship ($x_k = sinx_l$), and a combination of sine and quadratic relationships ($x_k = sin^2x_l$). For example, if the correlation between node X_4 and node X_5 was set to be quadratic, then $X_5 = 0.1{X_4}^2 + \varepsilon$, where ε is the residual error and $\varepsilon \sim N(0,1)$, that is the linear correlation between ${X_4}^2$ and X_5, can be used to represent the nonlinear quadratic relationship between X_4 and X_5, setting $E_{4\_5} = 0.1 \cdot {X_4}^2 \cdot X_5$. In addition to pre-specifying the effecting nodes and effecting edges in each simulation, random selection of the effecting nodes and effecting edges was considered in each simulation to minimize the influence of the network structure. In addition, each setting of the simulations was repeated 1000 times.

2.5. Real Data Analysis

CoNet was applied to perform network regression for survival time in a TWAS framework. Specifically, the gene expression data were obtained from the GEUVADIS study, then breast cancer patient survival data from the UK Biobank were examined. The GEUVADIS data [28] include 465 individuals from CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI), and Yoruba (YRI) populations. Only protein-coding genes and long intergenic non-coding RNAs (lincRNAs) annotated in GENCODE (release 12) [29,30] are considered in this study. Low-expressed genes with zero counts in at least half of

the individuals were removed. The PEER normalization method was used to eliminate confounding effects and unwanted variations [8,31]. To remove population stratification, the gene expression measurements were quantile normalized across individuals in each population to a standard normal distribution, which was further quantile normalized to a standard normal distribution across individuals from all five of the populations. A total of 15,810 genes were finally retained. All the individuals also had their genotypes sequenced in the 1000 Genomes Project. Genotype data were therefore obtained from the 1000 Genomes Project phase 3. SNPs with a Hardy–Weinberg equilibrium (HWE) p-value $< 10^{-4}$, genotype call rate < 95%, or minor allele frequency (MAF) < 0.01 were filtered out. Overall, 7,072,917 SNPS were ultimately left for further analysis.

The UK Biobank data comprises 487,298 individuals and 92,693,895 imputed SNPs [32]. We followed the same sample QC procedure as performed in the Neala laboratory, and 337,129 individuals with European ancestry were retained. SNPs with an HWE p-value $< 10^{-7}$, genotype call rate < 95%, or MAF < 0.001 were filtered out to retain 13,876,958 SNPs. For each gene, the cis-SNPs that were either within 1Mb upstream of the transcription start site (TSS) or within 1Mb downstream of the transcription end site (TES) were extracted. The cis-SNPs of genes in GEUVADIS were overlapped with those from the UK Biobank to obtain common SNPs. The initial focus was on breast cancer survival and included 818 patients with breast cancer based on the ICD-10 code (C50) within the UK Biobank cohort. Survival time can be calculated by the age of death minus the age at cancer diagnosis. Overall, 241 patients were censored because their time of death was not recorded or there was a competing risk of death. To verify the robustness of the real data, we first searched the networks potentially related to breast cancer from KEGG and involved 7 networks (hsa04630-nt06219, hsa04115, hsa04330, hsa04960, hsa04622, hsa04623, and hsa05211). After overlapping the network genes with those from UK Biobank, we finally analyzed 7 networks, including 338 nodes and 440 edges (Table 1). For both the nodes test and the edges test, we adjusted the p values using the false discover rate (FDR) with the Benjamini-Hochberg (BH) procedure to perform multiple tests, and declared the significance at an FDR threshold of 0.05. In addition, as population stratification may have an impact on the results, the top five PCs were treated as covariates in both the CoNet and the CPNT model.

Table 1. Summary of analyzed networks.

Network	KEGG		Overlap	
	Node	Edge	Node	Edge
hsa04115	73	88	68	82
hsa04330	59	164	57	131
hsa04623	75	100	60	73
hsa04960	37	37	31	31
hsa04622	71	147	53	86
hsa05211	68	98	62	31
hsa04630-nt06219	7	6	7	6
Total	390	640	338	440

3. Results

3.1. Simulation

Since there are currently no statistical tools available for network regression for failure time data in TWAS, here we aimed to compare CoNet with the CPNT method, which was intuitively developed under the CoNet framework but replaced PMI with the PM to describe the network edge. First, when the gene expression was predicted based on the DPR model, both methods performed well in detecting the node effect. Table 2 summarizes the estimated type I error for survival phenotype with a sample size of 5000 under three scenarios. Both methods yielded calibrated type I error control, regardless of the correlation pattern among the network nodes, the censoring rate, or the sample size (Tables S1 and S2).

Figure 2 displays the power for the survival phenotype in three scenarios with different node-effecting patterns. Overall, CoNet has a similar power with CPNT: the power of both methods increases as the sample size increases and the censoring rate decreases, regardless of the correlation patterns.

Table 2. The type I error for detecting the effect of the node on the survival phenotype under three scenarios where the effecting node is pre-specified (n = 5000), with the SNP effect obtained from DPR model. Simulations were conducted with four different between-node correlation patterns (the combination of sine and quadratic, sine, quadratic, and linear) and three different censoring rates (0.1, 0.3, and 0.5).

Scenario 1: Only Node Changes				
Correlation patterns	Methods	Censoring rate		
		0.1	0.3	0.5
$x_k = 0.5x_l$	CoNet	0.051	0.050	0.046
	CPNT	0.051	0.047	0.048
$x_k = 0.1{x_l}^2$	CoNet	0.038	0.049	0.043
	CPNT	0.039	0.050	0.044
$x_k = sinx_l$	CoNet	0.037	0.047	0.043
	CPNT	0.040	0.048	0.045
$x_k = sin^2x_l$	CoNet	0.037	0.045	0.043
	CPNT	0.040	0.050	0.045
Scenario 2: Both node and edge change with node hanging on the edge				
Correlation patterns	Methods	Censoring rate		
		0.1	0.3	0.5
$x_k = 0.5x_l$	CoNet	0.044	0.046	0.047
	CPNT	0.048	0.044	0.051
$x_k = 0.1{x_l}^2$	CoNet	0.051	0.047	0.055
	CPNT	0.057	0.049	0.052
$x_k = sinx_l$	CoNet	0.046	0.048	0.054
	CPNT	0.046	0.046	0.054
$x_k = sin^2x_l$	CoNet	0.058	0.055	0.055
	CPNT	0.053	0.045	0.053
Scenario 3: Both node and edge change with node not hanging on the edge				
Correlation patterns	Methods	Censoring rate		
		0.1	0.3	0.5
$x_k = 0.5x_l$	CoNet	0.047	0.061	0.068
	CPNT	0.044	0.055	0.064
$x_k = 0.1{x_l}^2$	CoNet	0.069	0.068	0.061
	CPNT	0.060	0.071	0.054
$x_k = sinx_l$	CoNet	0.069	0.069	0.061
	CPNT	0.058	0.071	0.057
$x_k = sin^2x_l$	CoNet	0.068	0.068	0.062
	CPNT	0.059	0.072	0.058

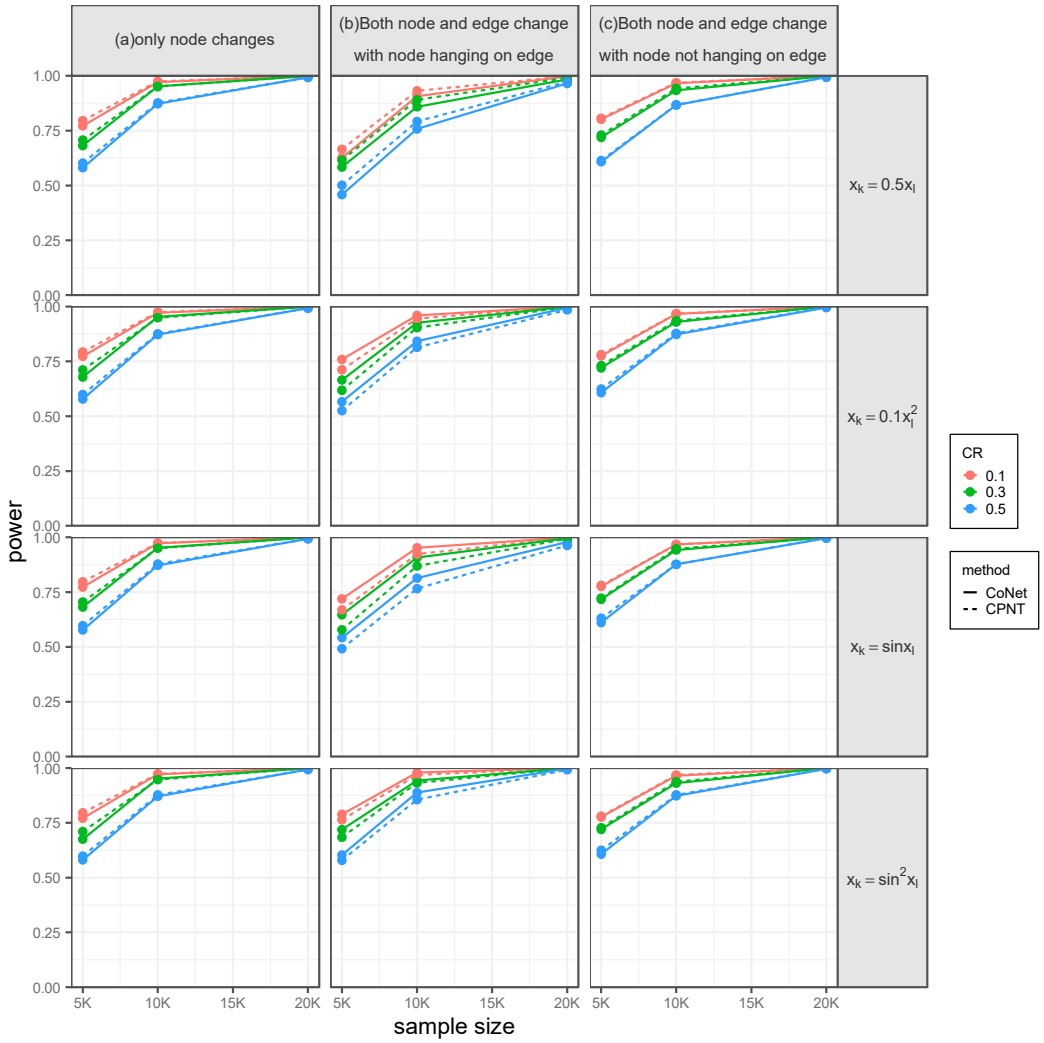

Figure 2. The power for testing the effect of the node on the survival phenotype under the setting that the effecting node is pre-specified, with the SNP effect obtained from the DPR model. Simulations were conducted with four different between-node correlation patterns (the combination of sine and quadratic, sine, quadratic, and linear) and three different censoring rates (0.1, 0.3, and 0.5). (**a**) Only a node has an effect. (**b**) Both node and edge have effects, with the effecting node hanging on the edge. (**c**) Both node and edge have effects, with the effecting node not hanging on the edge.

Then, we also evaluated the ability of both methods in identifying the significant edge. Table 3 summarizes the estimated type I error for survival phenotype with a sample size of 5000 under three scenarios. Similarly, the type I error rates of CoNet and CPNT remain calibrated with different correlation patterns among nodes, the censoring rate, and the sample size (Tables S3 and S4). Figure 3 shows the power for survival phenotype when only an edge has a pre-specified effect. With the same settings as in the detecting node effect, we can draw the same conclusions as for the power of both methods. With nonlinear correlation between the nodes inside the network, the power of CoNet has a better performance than that of CPNT. For instance, under a combination of sine and quadratic relationships, with a sample size of 20,000 and censoring rate of 0.1, the power of CoNet is 0.449 and the power of CPNT is 0.042 (Figure 3a). Although the power of CoNet decreases to 0.298 when the censoring rate increases to 0.5, it is still higher than that of CPNT (0.043) (Figure 3a).

Table 3. The type I error for testing the effect of the edge on the survival phenotype under three scenarios where the effecting edge is pre-specified (n = 5000), with the SNP effect obtained from DPR model. Simulations were conducted with four different between-node correlation patterns (the combination of sine and quadratic, sine, quadratic, and linear) and three different censoring rates (0.1, 0.3, and 0.5).

Scenario 1: Only Edge Changes				
Correlation patterns	Methods	Censoring rates		
		0.1	0.3	0.5
$x_k = 0.5x_l$	CoNet	0.049	0.039	0.043
	CPNT	0.046	0.040	0.043
$x_k = 0.1{x_l}^2$	CoNet	0.055	0.058	0.054
	CPNT	0.042	0.055	0.056
$x_k = sinx_l$	CoNet	0.040	0.048	0.049
	CPNT	0.040	0.054	0.052
$x_k = sin^2x_l$	CoNet	0.050	0.050	0.042
	CPNT	0.043	0.058	0.060
Scenario 2: Both node and edge change with node hanging on the edge				
Correlation patterns	Methods	Censoring rates		
		0.1	0.3	0.5
$x_k = 0.5x_l$	CoNet	0.051	0.053	0.049
	CPNT	0.053	0.060	0.051
$x_k = 0.1{x_l}^2$	CoNet	0.056	0.052	0.054
	CPNT	0.056	0.053	0.046
$x_k = sinx_l$	CoNet	0.058	0.060	0.053
	CPNT	0.056	0.058	0.058
$x_k = sin^2x_l$	CoNet	0.042	0.044	0.039
	CPNT	0.057	0.056	0.050
Scenario 3: Both node and edge change with node not hanging on the edge				
Correlation patterns	Methods	Censoring rates		
		0.1	0.3	0.5
$x_k = 0.5x_l$	CoNet	0.069	0.044	0.063
	CPNT	0.059	0.039	0.055
$x_k = 0.1{x_l}^2$	CoNet	0.056	0.052	0.055
	CPNT	0.056	0.053	0.052
$x_k = sinx_l$	CoNet	0.064	0.057	0.053
	CPNT	0.059	0.056	0.058
$x_k = sin^2x_l$	CoNet	0.043	0.046	0.039
	CPNT	0.066	0.054	0.050

As expected, the type of nonlinear pattern plays a key role in the difference of power for both methods. For example, the power of CoNet is much higher than that of CPNT when the relationship between nodes is a combination of both sine and quadratic, while they have a comparable performance when nodes are in the sine relationship. As the sample size increases and the censoring rate decreases, the power of CoNet increases even more dramatically. For example, if the nodes are correlated within a pattern of recombination of sine and quadratic, when the sample size increases from 5000 to 20,000 with a fixed censoring rate = 0.1, the power of CoNet increases from 0.142 to 0.449 whereas the power of CPNT increases from 0.038 to 0.042 (Figure 3a). If the nodes are correlated within a pattern of the combination of sine and quadratic, when the censoring rate decreases from 0.3 to 0.1 with a fixed sample size = 20,000, the power of CoNet increases from 0.376 to 0.449 while the power of CPNT increases from 0.037 to 0.042 (Figure 3a).

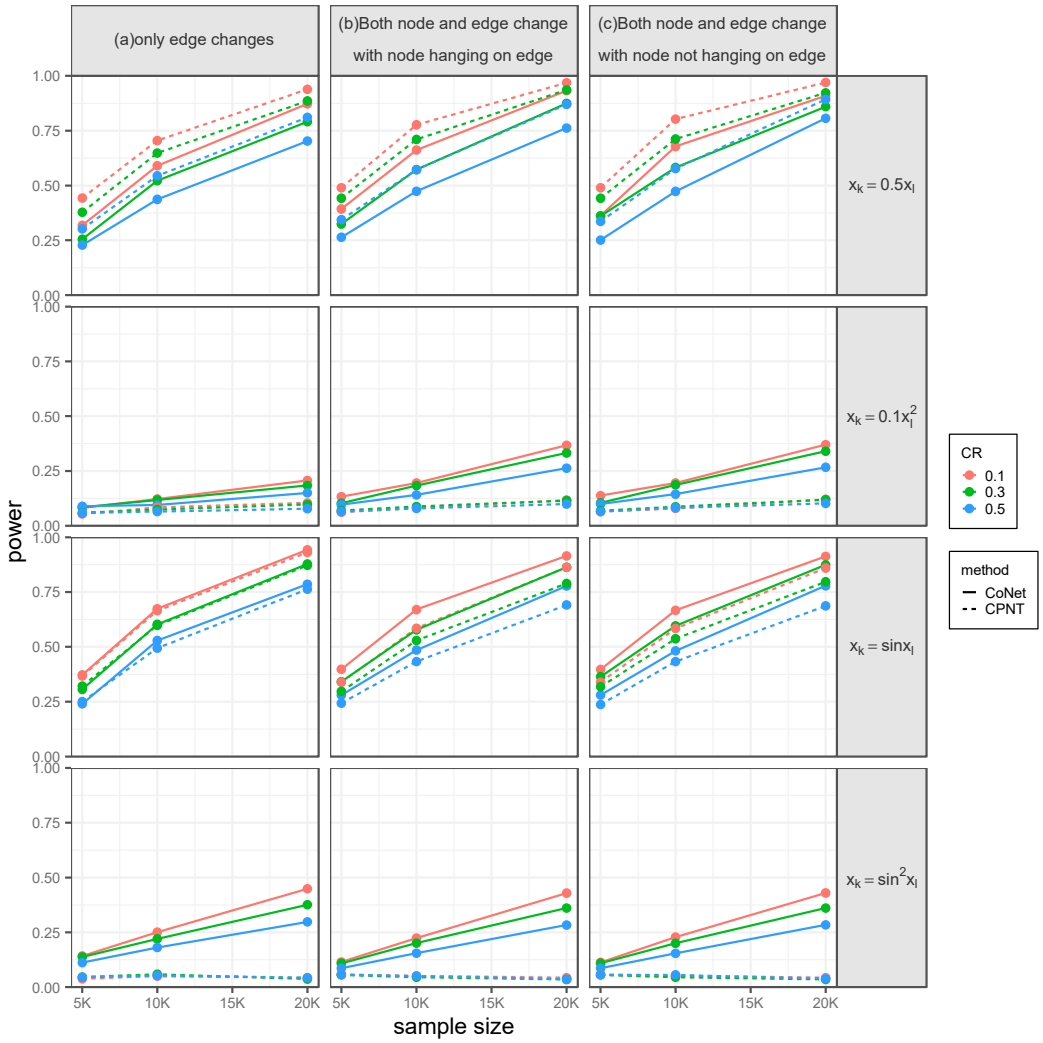

Figure 3. The power for testing the effect of the edge on the survival phenotype under pre-specified effecting edge settings, with the SNP effect obtained from the DPR model. Simulations were conducted with four different between-node correlation patterns (the combination of sine and quadratic, sine, quadratic, and linear) and three different censoring rates (0.1, 0.3, and 0.5). (**a**) Only an edge has an effect. (**b**) Both node and edge have effects, with the effecting node hanging on the edge. (**c**) Both node and edge have effects, with the effecting node not hanging on the edge.

The power advantage of CoNet over CPNT remains with nonlinear correlation patterns, regardless of the censoring rate. Together, these results illustrate that CoNet is more powerful to capture the nonlinear relationships than linear ones. Note that, CoNet only just has a slightly lower power than CPNT with the linear correlation pattern, possibly because the PM is the gold standard to capture the linear correlation in this case. We can obtain the similar findings when both the node and edge had effects, either with the effecting node hanging on the edge or with the effecting node not hanging on the edge.

Similar results can also be found when randomly selecting the effecting nodes and effecting edges (Tables S5–S10, Figures S1 and S2) and when the predicted value of gene expression was calculated by the BSLMM model (Tables S11–S22, Figures S3–S6). For example, in the setting that only an edge has a pre-specified effect, under a combination of sine and quadratic relationships with a sample size of 20,000 and censoring rate of 0.1, the power of CoNet and CPNT is 0.449 and 0.042 using the DPR model (Figure 3a) and 0.465 and 0.053 using the BSLMM model (Figure S4a).

Additional simulations also illustrated the advantage of CoNet with a high censoring rate (Tables S23 and S24, Figures S8 and S9) and the robustness of CoNet when several genes

are unavailable (Table S26–S31, Figures S10 and S11). In addition, CoNet is computationally efficient (Table S25).

3.2. Application

We completely analyzed 7 networks, including 338 nodes and 440 edges. Overall, CoNet successfully identified 3 genes and 7 edges (Tables 1 and 4), while CPNT identified 3 genes and failed to identify any edges. Consistent with the simulations that both methods have a comparable performance in detecting the node effect, both methods successfully identified 3 genes, respectively, including the common genes *CDK6* in hsa04115 (adjusted $p = 0.048$ for CoNet and 0.043 for CPNT) and *DTX3L* in hsa04330 (adjusted $p = 0.034$ for CoNet and 0.012 for CPNT). The significant node identified by CoNet rather than CPNT was *IL6* (adjusted $p = 0.012$ for CoNet and 0.156 for CPNT) in hsa04623. The significant node identified by CPNT rather than CoNet was *MAML2* in hsa04330 (adjusted $p = 0.096$ for CoNet and 0.043 for CPNT).

Table 4. Significantly affecting nodes and edge for two methods with *p*-values being corrected by FDR.

	CoNet		**CPNT**	
	Node	**Edge**	**Node**	**Edge**
hsa04115	*CDK6* ($p = 0.048$)	*GADD45A_CDK1* ($p = 0.027$)	*CDK6* ($p = 0.043$)	/
hsa04623	*IL6* ($p = 0.012$)	/	/	/
hsa04622	/	*DDX58_TRIM25* ($p = 1.15 \times 10^{-3}$) *MAP3K1_MAPK12* ($p = 0.027$)	/	/
hsa05211	/	*VHL_EP300* ($p = 5.78 \times 10^{-6}$)	/	/
hsa04960		*NEDD4L_SFN* ($p = 3.28 \times 10^{-3}$)	/	
hsa04330	*DTX3L* ($p = 0.034$) *MAML2* ($p = 0.096$)	*NOTCH2_DVL2* ($p = 0.011$) *NOTCH4_DVL2*($p = 1.26 \times 10^{-3}$)	*DTX3L* ($p = 0.012$) *MAML2*($p = 0.043$)	/

The significant edges identified only by CoNet included *GADD45A_CDK1* (adjusted $p = 0.027$) in hsa04115, *DDX58_TRIM25* (adjusted $p = 1.15 \times 10^{-3}$) in hsa04622, *MAP3K1_MAPK12* (adjusted $p = 0.027$) in hsa04622, *VHL_EP300* (adjusted $p = 5.78 \times 10^{-6}$) in hsa05211, *NEDD4L_SFN* (adjusted $p = 3.28 \times 10^{-3}$) in hsa04960, *NOTCH2_DVL2* (adjusted $p = 0.011$) in hsa04330, and *NOTCH4_DVL2* ($p = 1.26 \times 10^{-3}$, hsa04330). Two exemplary scatter plots (Figure S7a–d) further illustrated that the joint distribution of the two nodes linked by the significant edges may be different between the individuals with a short survival time (less than 25% quantile) and a long survival time (higher than the 75% quantile).

4. Discussion

It is essential to identify biological networks that are associated with complex traits to understand the network mechanism related to complex diseases. In this study, we proposed CoNet, a novel statistical method for detecting the association between one given network and the survival time. CoNet applies DPR to find the gene expression prediction weights, then implements PMI to quantify the correlations between the nodes. Moreover, CoNet can provide the significant effecting gene nodes and edges associated with the survival outcomes at once. CoNet uses nonparametric kernel density estimation to calculate PMI between two genes. Here, we demonstrated several benefits of CoNet through extensive simulations and real data analysis.

It could be argued that the PMI estimate could be obtained among the network nodes of observed gene expression from the eQTL study instead of predicted gene expression in GWAS. The standard TWAS analysis could then be performed using the Cox proportional hazards model in the second stage, using the PMI estimate as a new exposure. However, the eQTL study would have a large prediction error because of its limited sample size (e.g., only 465 samples in the GEUVADIS data). In addition, unlike traditional TWAS analysis that uses the cis-SNPs of each gene as the genotypes, it is challenging, both biologically

and statistically, to determine the SNPs that are suitable for the PMI between two genes as the genotypes.

In real data analysis, we found several genes or gene–gene interactions associated with breast cancer. As for the specific genes, *CDK6* and *DTX3L* were identified by both CoNet and CPNT, while *IL6* was identified by CoNet only. *CDK6* is a known classic cell cycle kinase that facilitates the progression of cells. Some studies have detected *CDK6* mRNA expression increases in breast cancer tissues versus that in adjacent tissues [33]. *DTX3L* is found to be overexpressed in breast cancer, which functions as a negative regulator of ATRA-induced growth inhibition in breast cancer cells [34]. *IL-6* was shown to promote or inhibit the growth of breast cancer cells. Indeed, some studies considered *IL-6* as a potential marker in the prognosis of breast cancer [35]. As for the significant edges, CoNet identified *GADD45A_CDK1*, *DDX58_TRIM25*, *MAP3K1_MAPK12*, *VHL_EP300*, *NEDD4L_SFN*, *NOTCH2_DVL2*, and *NOTCH4_DVL2*, but CPNT failed to identify any edges. Previous studies confirmed that *SFN* inhibited TGF-β1-induced migration and invasion in breast cancer cells [36] and *NEDD4L* expression significantly reduced in breast invasive carcinoma [37]. In ER^+- tumor tissues, the mRNA levels of *DDX58* were significantly higher than in adjacent tissues [38]. Similarly, *TRIM25* was reported as overexpressed in breast cancer cells [39].

CoNet is not without limitations. Firstly, only CPNT was used as a reference to evaluate the performance of CoNet. We have performed additional simulations to compare CoNet with the modified TIGAR, where we changed the linear regression in the second stage of TIGAR to be the Cox model. The results illustrated that CoNet has a higher power than TIGAR (Table S32, Figure S12). Secondly, it is assumed that the network structure is prior known. The learning network structure needs to identify all the possible edges that match the data. Often, a joint probability distribution of an all-gene network can reveal multiple network structures. In CoNet, the PMI estimation for different gene expression predictions is directly plugged into the regression model, ignoring the accuracy of the PMI estimator, particularly in eQTL studies with a small sample size. The interpretation of the regression coefficients of the edges (PMI) is correlation pattern specific. Intuitively, the positive coefficient indicates that the hazard will increase as the strength of the non-independency between the two node variables increases. The negative effect indicates that the hazard will decrease as the strength of the non-independency between the two node variables increases. Even so, the effect size should be interpreted in caution. In addition, the interaction studies in this research are statistical interactions. Indeed, it is hard to define what the specific biological interactions are. They can be proteins coded by two interaction genes working in the same pathway, wherein one gene affects the expression of the other, or both genes sharing a common regulatory mechanism. However, despite these limitations, our study highlights that CoNet is an appealing approach for simultaneously identifying the potential nodes and edges that are related to the survival time in large datasets.

5. Conclusions

The proposed method here, CoNet, effectively accounts for network structure and can capture and quantify the general relationship among different genes. It is robust against different model assumptions of genetic effect sizes and different censoring rates and can simultaneously identify the potential nodes and edges that are related to the survival time in TWAS.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14030586/s1, Figure S1: Power for the effecting node is randomly selected, using DPR as imputation model; Figure S2: Power for the effecting edge is randomly selected, using DPR as imputation model; Figure S3: Power for the effecting node is pre-specified, using BSLMM as imputation model; Figure S4: Power for the effecting edge is pre-specified, using BSLMM as imputation model; Figure S5: Power for the effecting node is randomly selected, using BSLMM as imputation model; Figure S6: Power for the effecting edge is randomly selected, using BSLMM as imputation model; Figure S7: Scatter plot of the expression of two different genes with different survival

status; Figure S8: Power for the effecting node and the effecting edge are pre-specified, using DPR as imputation model (including high censoring rate of 0.7); Figure S9: Power for the effecting node and the effecting edge are randomly selected, using DPR as imputation model (including high censoring rate of 0.7); Figure S10: Power of CoNet in a reduced gene network with some genes being unavailable, where the effecting node and effecting edge are pre-specified; Figure S11: Power of CoNet in a reduced gene network with some genes being unavailable, where the effecting node and effecting edge are randomly selected; Figure S12: Power of CoNet and TIGAR to test the effect of nodes under the effect size of between-node correlation being 0.1; Table S1: Type I error for the effecting node is pre-specified, using DPR as imputation model (n = 10,000); Table S2: Type I error for the effecting node is pre-specified, using DPR as imputation model (n = 20,000); Table S3: Type I error for the effecting edge is pre-specified, using DPR as imputation model (n = 10,000); Table S4: Type I error for the effecting edge is pre-specified, using DPR as imputation model (n = 20,000); Table S5: Type I error for the effecting node is randomly selected, using DPR as imputation model (n = 5000); Table S6: Type I error for the effecting node is randomly selected, using DPR as imputation model (n = 10,000); Table S7: Type I error for the effecting node is randomly selected, using DPR as imputation model (n = 20,000); Table S8: Type I error for the effecting edge is randomly selected, using DPR as imputation model (n = 5000); Table S9: Type I error for the effecting edge is randomly selected, using DPR as imputation model (n = 10,000); Table S10: Type I error for the effecting edge is randomly selected, using DPR as imputation model (n = 20,000); Table S11: Type I error for the effecting node is pre-specified, using BSLMM as imputation model (n = 5000); Table S12: Type I error for the effecting node is pre-specified, using BSLMM as imputation model (n = 10,000); Table S13: Type I error for the effecting node is pre-specified, using BSLMM as imputation model (n = 20,000); Table S14: Type I error for the effecting edge is pre-specified, using BSLMM as imputation model (n = 5000); Table S15: Type I error for the effecting edge is pre-specified, using BSLMM as imputation model (n = 10,000); Table S16: Type I error for the effecting edge is pre-specified, using BSLMM as imputation model (n = 20,000); Table S17: Type I error for the effecting node is randomly selected, using BSLMM as imputation model (n = 5000); Table S18: Type I error for the effecting node is randomly selected, using BSLMM as imputation model (n = 10,000); Table S19: Type I error for the effecting node is randomly selected, using BSLMM as imputation model (n = 20,000); Table S20: Type I error for the effecting edge is randomly selected, using BSLMM as imputation model (n = 5000); Table S21: Type I error for the effecting edge is randomly selected, using BSLMM as imputation model (n = 10,000); Table S22: Type I error for the effecting edge is randomly selected, using BSLMM as imputation model (n = 20,000); Table S23: Type I error for the effecting node and the effecting edge are pre-specified, using DPR as imputation model(censoring rate = 0.7); Table S24: Type I error for the effecting node and the effecting edge are randomly selected, using DPR as imputation model(censoring rate = 0.7); Table S25: Mean computational time (seconds) of both methods; Table S26: Type I error of CoNet in a reduced gene network with some proportions (0, 20%, 30%) of genes being unavailable, where the effecting node and effecting edge are pre-specified (n = 5000); Table S27: Type I error of CoNet in a reduced gene network with some proportions (0, 20%, 30%) of genes being unavailable, where the effecting node and effecting edge are pre-specified (n = 10,000); Table S28: Type I error of CoNet in a reduced gene network with some proportions (0, 20%, 30%) of genes being unavailable, where the effecting node and effecting edge are pre-specified (n = 20,000); Table S29: Type I error of CoNet in a reduced gene network with some proportions (0, 20%, 30%) of genes being unavailable, where the effecting node and effecting edge are randomly selected (n = 5000); Table S30: Type I error of CoNet in a reduced gene network with some proportions (0, 20%, 30%) of genes being unavailable, where the effecting node and effecting edge are randomly selected (n = 10,000); Table S31: Type I error of CoNet in a reduced gene network with some proportions (0, 20%, 30%) of genes being unavailable, where the effecting node and effecting edge are randomly selected (n = 20,000); Table S32: Type I error of CoNet and TIGAR to test the effect of nodes under the survival phenotype.

Author Contributions: Conceptualization, Z.Y.; methodology, J.H.; software, J.H. and L.Z.; validation, J.H.; formal analysis, J.H.; investigation, J.H., L.Z., R.Y., T.J. and X.J.; resources, J.J. and Z.Y.; data curation, J.H., L.Z., R.Y., T.J. and X.J.; writing—original draft preparation, J.H.; writing—review and editing, J.H., J.J., S.W. and Z.Y.; visualization, J.H.; supervision, J.J., Z.Y. and S.W.; project administration, J.J. and Z.Y.; funding acquisition, J.J and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [81872712, 82173624, 81673272], the Natural Science Foundation of Shandong Province [ZR2019ZD02], the National Statistical Scientific Research Project (2022LY031), and the Young Scholars Program of Shandong University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The GEUVADIS data underlying this article are available at http://www.geuvadis.org (accessed on 20 April 2022). The breast cancer data of the UK Biobank are available at https://www.ukbiobank.ac.uk/ (accessed on 15 June 2022), with application number 51470. The sample QC procedure in Neale lab is available at https://github.com/Nealelab/UK_Biobank_GWAS/tree/master/imputed-v2-gwas (accessed on 20 April 2022).

Acknowledgments: The UK Biobank was established by the Wellcome Trust medical charity, the Medical Research Council, the Department of Health, the Scottish Government, and the Northwest Regional Development Agency. The authors are grateful to the UK Biobank resource.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zeng, P.; Zhao, Y.; Qian, C.; Zhang, L.; Zhang, R.; Gou, J.; Liu, J.; Liu, L.; Chen, F. Statistical Analysis for Genome-Wide Association Study. *J. Biomed. Res.* **2015**, *29*, 285. [PubMed]
2. Li, B.; Ritchie, M.D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front. Genet.* **2021**, *12*, 713230. [CrossRef] [PubMed]
3. Bossé, Y.; Li, Z.; Xia, J.; Manem, V.; Carreras-Torres, R.; Gabriel, A.; Gaudreault, N.; Albanes, D.; Aldrich, M.C.; Andrew, A.; et al. Transcriptome-wide Association Study Reveals Candidate Causal Genes for Lung Cancer. *Int. J. Cancer* **2020**, *146*, 1862–1878. [CrossRef] [PubMed]
4. Gong, L.; Zhang, D.; Lei, Y.; Qian, Y.; Tan, X.; Han, S. Transcriptome-Wide Association Study Identifies Multiple Genes and Pathways Associated with Pancreatic Cancer. *Cancer Med.* **2018**, *7*, 5727–5732. [CrossRef]
5. Gusev, A.; Mancuso, N.; Won, H.; Kousi, M.; Finucane, H.K.; Reshef, Y.; Song, L.; Safi, A.; McCarroll, S.; Neale, B.; et al. Transcriptome-Wide Association Study of Schizophrenia and Chromatin Activity Yields Mechanistic Disease Insights. *Nat. Genet.* **2018**, *50*, 538–548. [CrossRef]
6. Gamazon, E.R.; Wheeler, H.E.; Shah, K.P.; Mozaffari, S.V.; Aquino-Michaels, K.; Carroll, R.J.; Eyler, A.E.; Denny, J.C.; Nicolae, D.L.; Cox, N.J.; et al. A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data. *Nat. Genet.* **2015**, *47*, 1091–1098. [CrossRef]
7. Gusev, A.; Ko, A.; Shi, H.; Bhatia, G.; Chung, W.; Penninx, B.W.J.H.; Jansen, R.; de Geus, E.J.C.; Boomsma, D.I.; Wright, F.A.; et al. Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies. *Nat. Genet.* **2016**, *48*, 245–252. [CrossRef]
8. Zeng, P.; Zhou, X. Non-Parametric Genetic Prediction of Complex Traits with Latent Dirichlet Process Regression Models. *Nat. Commun.* **2017**, *8*, 456. [CrossRef]
9. Nagpal, S.; Meng, X.; Epstein, M.P.; Tsoi, L.C.; Patrick, M.; Gibson, G.; De Jager, P.L.; Bennett, D.A.; Wingo, A.P.; Wingo, T.S.; et al. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am. J. Hum. Genet.* **2019**, *105*, 258–266. [CrossRef]
10. Tang, S.; Buchman, A.S.; De Jager, P.L.; Bennett, D.A.; Epstein, M.P.; Yang, J. Novel Variance-Component TWAS Method for Studying Complex Human Diseases with Applications to Alzheimer's Dementia. *PLoS Genet.* **2021**, *17*, e1009482. [CrossRef]
11. 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*, 3861. [CrossRef] [PubMed]
12. 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*, 240–256. [CrossRef] [PubMed]
13. Mancuso, N.; Freund, M.K.; Johnson, R.; Shi, H.; Kichaev, G.; Gusev, A.; Pasaniuc, B. Probabilistic Fine-Mapping of Transcriptome-Wide Association Studies. *Nat. Genet.* **2019**, *51*, 675–682. [CrossRef] [PubMed]
14. Wu, C.; Pan, W. A Powerful Fine-Mapping Method for Transcriptome-Wide Association Studies. *Hum. Genet.* **2020**, *139*, 199–213. [CrossRef] [PubMed]
15. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-Based Approach to Human Disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef] [PubMed]
16. 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. [CrossRef] [PubMed]
17. Jin, X.; Zhang, L.; Ji, J.; Ju, T.; Zhao, J.; Yuan, Z. Network Regression Analysis in Transcriptome-Wide Association Studies. *BMC Genom.* **2022**, *23*, 562. [CrossRef]

18. Zhang, L.; Ju, T.; Jin, X.; Ji, J.; Han, J.; Zhou, X.; Yuan, Z. Network Regression Analysis for Binary and Ordinal Categorical Phenotypes in Transcriptome-Wide Association Studies. *Genetics* **2022**, *222*, iyac153. [CrossRef]
19. Johnson, D.C.; Weinhold, N.; Mitchell, J.S.; Chen, B.; Kaiser, M.; Begum, D.B.; Hillengass, J.; Bertsch, U.; Gregory, W.A.; Cairns, D.; et al. Genome-Wide Association Study Identifies Variation at 6q25.1 Associated with Survival in Multiple Myeloma. *Nat. Commun.* **2016**, *7*, 10290. [CrossRef]
20. Labadie, J.D.; Savas, S.; Harrison, T.A.; Banbury, B.; Huang, Y.; Buchanan, D.D.; Campbell, P.T.; Gallinger, S.J.; Giles, G.G.; Gunter, M.J.; et al. Genome-wide Association Study Identifies Tumor Anatomical Site-specific Risk Variants for Colorectal Cancer Survival. *Sci. Rep.* **2022**, *12*, 127. [CrossRef]
21. Cao, C.; Ding, B.; Li, Q.; Kwok, D.; Wu, J.; Long, Q. Power Analysis of Transcriptome-Wide Association Study: Implications for Practical Protocol Choice. *PLoS Genet.* **2021**, *17*, e1009405. [CrossRef]
22. 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*, 939–951. [CrossRef] [PubMed]
23. Zhou, X.; Carbonetto, P.; Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* **2013**, *9*, e1003264. [CrossRef] [PubMed]
24. Church, K.W.; Hanks, P. Word Association Norms, Mutual Information and Lexicography. In Proceedings of the 27th Annual Meeting on Association for Computational Linguistics; Association for Computational Linguistics, Vancouver, BC, Canada, 26–29 June 1989; pp. 76–83.
25. Duong, T.; Hazelton, M. Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation. *J. Nonparametric Stat.* **2003**, *15*, 17–30. [CrossRef]
26. Bilock, A.; Jidling, C.; Rydin, Y. *Modelling Bivariate Distributions Using Kernel Density Estimation*; Uppsala University: Uppsala, Sweden, 2016.
27. Bender, R.; Augustin, T.; Blettner, M. Generating Survival Times to Simulate Cox Proportional Hazards Models. *Stat. Med.* **2005**, *24*, 1713–1723. [CrossRef]
28. Lappalainen, T.; Sammeth, M.; Friedländer, M.R.; 't Hoen, P.A.; Monlong, J.; Rivas, M.A.; Gonzàlez-Porta, M.; Kurbatova, N.; Griebel, T.; Ferreira, P.G.; et al. Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans. *Nature* **2013**, *501*, 506–511. [CrossRef] [PubMed]
29. Wen, X.; Luca, F.; Pique-Regi, R. Cross-Population Joint Analysis of EQTLs: Fine Mapping and Functional Annotation. *PLoS Genet.* **2015**, *11*, e1005176. [CrossRef] [PubMed]
30. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S. GENCODE: The Reference Human Genome Annotation for The ENCODE Project. *Genome Res.* **2012**, *22*, 1760–1774. [CrossRef] [PubMed]
31. Stegle, O.; Parts, L.; Piipari, M.; Winn, J.; Durbin, R. Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses. *Nat. Protoc.* **2012**, *7*, 500–507. [CrossRef]
32. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J. The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **2018**, *562*, 203–209. [CrossRef]
33. Zhang, H.; Zhao, B.; Wang, X.; Zhang, F.; Yu, W. LINC00511 Knockdown Enhances Paclitaxel Cytotoxicity in Breast Cancer via Regulating MiR-29c/CDK6 Axis. *Life Sci.* **2019**, *228*, 135–144. [CrossRef] [PubMed]
34. Hu, W.; Hu, Y.; Pei, Y.; Li, R.; Xu, F.; Chi, X.; Mi, J.; Bergquist, J.; Lu, L.; Zhang, L. Silencing DTX3L Inhibits the Progression of Cervical Carcinoma by Regulating PI3K/AKT/MTOR Signaling Pathway. *Int. J. Mol. Sci.* **2023**, *24*, 861. [CrossRef] [PubMed]
35. Knüpfer, H.; Preiß, R. Significance of Interleukin-6 (IL-6) in Breast Cancer. *Breast Cancer Res. Treat.* **2007**, *102*, 129–135. [CrossRef] [PubMed]
36. Zhang, Y.; Lu, Q.; Li, N.; Xu, M.; Miyamoto, T.; Liu, J. Sulforaphane Suppresses Metastasis of Triple-Negative Breast Cancer Cells by Targeting the RAF/MEK/ERK Pathway. *NPJ Breast Cancer* **2022**, *8*, 40. [CrossRef]
37. Dong, H.; Zhu, L.; Sun, J.; Zhang, Y.; Cui, Q.; Wu, L.; Chen, S.; Lu, J. Pan-Cancer Analysis of NEDD4L and Its Tumor Suppressor Effects in Clear Cell Renal Cell Carcinoma. *J. Cancer* **2021**, *12*, 6242. [CrossRef]
38. Fang, Q.; Yao, S.; Luo, G.; Zhang, X. Identification of Differentially Expressed Genes in Human Breast Cancer Cells Induced by 4-Hydroxyltamoxifen and Elucidation of Their Pathophysiological Relevance and Mechanisms. *Oncotarget* **2018**, *9*, 2475. [CrossRef]
39. Urano, T.; Saito, T.; Tsukui, T.; Fujita, M.; Hosoi, T.; Muramatsu, M.; Ouchi, Y.; Inoue, S. Efp Targets 14-3-3σ for Proteolysis and Promotes Breast Tumour Growth. *Nature* **2002**, *417*, 871–875. [CrossRef]