Methods 145 (2018) 41–50

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Leveraging multiple gene networks to prioritize GWAS candidate genes via network representation learning

Mengmeng Wu[a,d], Wanwen Zeng[b,d], Wenqiang Liu[c], Hairong Lv[b,d,*], Ting Chen[a,d,*], Rui Jiang[b,d,*]

[a] Department of Computer Science, Tsinghua University, Beijing, China
[b] Department of Automation, Tsinghua University, Beijing, China
[c] Department of Computer Science, Xi'an jiaotong University, Xi'an, China
[d] MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST, China

ABSTRACT

Genome-wide association studies (GWAS) have successfully discovered a number of disease-associated genetic variants in the past decade, providing an unprecedented opportunity for deciphering genetic basis of human inherited diseases. However, it is still a challenging task to extract biological knowledge from the GWAS data, due to such issues as missing heritability and weak interpretability. Indeed, the fact that the majority of discovered loci fall into noncoding regions without clear links to genes has been preventing the characterization of their functions and appealing for a sophisticated approach to bridge genetic and genomic studies. Towards this problem, network-based prioritization of candidate genes, which performs integrated analysis of gene networks with GWAS data, has emerged as a promising direction and attracted much attention. However, most existing methods overlook the sparse and noisy properties of gene networks and thus may lead to suboptimal performance. Motivated by this understanding, we proposed a novel method called REGENT for integrating multiple gene networks with GWAS data to prioritize candidate genes for complex diseases. We leveraged a technique called the network representation learning to embed a gene network into a compact and robust feature space, and then designed a hierarchical statistical model to integrate features of multiple gene networks with GWAS data for the effective inference of genes associated with a disease of interest. We applied our method to six complex diseases and demonstrated the superior performance of REGENT over existing approaches in recovering known disease-associated genes. We further conducted a pathway analysis and showed that the ability of REGENT to discover disease-associated pathways. We expect to see applications of our method to a broad spectrum of diseases for post-GWAS analysis. REGENT is freely available at https://github.com/wmmthu/REGENT.

1. Introduction

Genome-wide association studies (GWAS), as a major approach for deciphering genetic codes of human inherited diseases, have identified thousands of genetic variants that are possibly associated with hundreds of complex diseases over the past decade. As of July 2017, there have been over 30,000 disease-associated single nucleotide polymorphisms (SNPs) deposited the GWAS Catalog database [1]. Such a fruitful resource holds great promise to dissect genetic basis of a variety of complex phenotypes, thereby promoting genetic medicine and precision medicine. One particular problem of substantial interest is to identify disease-associated genes, which provide insights about underlying biological mechanisms as well as clues for designing appropriate drugs [2]. However, it remains difficult to identify disease-associated genes from GWAS data directly, due to such challenges as missing heritability [3], weak interpretability [4], and linkage disequilibrium [5].

Increasing sample size, though possibly improving statistical discovery power, is too costly and time-consuming to be a practical solution towards these challenges. One promising direction to improve the discovery power of GWAS without the recruitment of a large sample is network-based gene prioritization, which emerges as a powerful approach and has attracted much attention recently. For example, Lee, et al. used a network propagation algorithm to determine the risk genes from HumanNet, an integrated gene co-functional network [6]. Murat et al. designed a prix fixe strategy to formulate the problem of gene

* Corresponding authors at: Department of Computer Science and Department of Automation, Tsinghua University, Beijing, China.
E-mail addresses: wmm15@mails.tsinghua.edu.cn (M. Wu), zengww14@mails.tsinghua.edu.cn (W. Zeng), lvhairong@tsinghua.edu.cn (H. Lv), tingchen@tsinghua.edu.cn (T. Chen), ruijiang@tsinghua.edu.cn (R. Jiang).

https://doi.org/10.1016/j.ymeth.2018.06.002
Received 30 January 2018; Received in revised form 10 April 2018; Accepted 1 June 2018
Available online 03 June 2018
1046-2023/

prioritization as a combinatorial optimization problem on a gene co-functional network and used a genetic algorithm to find approximate solutions [7]. Greene et al. directly used connectivity patterns over tissue-specific gene functional networks as features and trained a support vector machine (SVM) for predicting disease-associated genes [8]. These existing methods for network-based gene prioritization directly use gene networks as inputs, holding the belief that a gene network is reliable and informative. However, gene networks are usually sparse and noisy, and thus a method overlooking this property may produce suboptimal results. Besides, most existing methods only consider a single gene network, and thus cannot make use of information stored in such a variety of networks as protein-protein interaction networks [9], gene co-expression networks [10] and gene regulatory networks [11]. Previous studies have shown that integrating multiple gene networks could significantly improve the performance of gene prioritization [12,13] and variant prioritization [14,15]. It is therefore desired to develop a novel bioinformatics method to enable an integrated analysis of GWAS data with multiple gene networks while at the same time handling sparsity and noise of these networks.

Network representation learning, with examples including DeepWalk [16] and node2vec [17], has emerged as a branch of representation learning [18] and shown successful applications in the analysis of social networks. With a network represented in a feature space of low dimensionality, such issues as sparsity and noise of the network are likely to be alleviated, and hence satisfactory performance can often be expected for such analysis tasks as node classification, graph clustering, and link prediction. A network representation learning method learns continuous representations, also called embedding, for nodes via the optimization of a carefully designed objective function for preserving structural information in original network. For example, nodes that are close to each other in the network or share many common neighbours are expected to also be close in the embedding space. The learned embedding can be naturally fed into standard machine learning algorithms such as support vector machines (SVM), k-means clustering, and principal component analysis (PCA) [19] to perform classification, clustering, and visualization, respectively. These desired characteristics motivate us to apply the network representation learning technique to network-based gene prioritization, which remains unexplored until now.

In this paper, we proposed a novel computational method called REGENT (integ**R**ating **E**mbeddings of multiple **GE**ne **NeT**works) for network-based gene prioritization. Specifically, we used the network representation learning to learn embeddings of genes from multiple gene networks and developed a hierarchical statistical model to integrate the learned embeddings of genes with GWAS summary data. We applied REGENT to GWAS summary data of six complex diseases and found our method outperformed existing methods in terms of gene prioritization. We further conducted pathway analysis to the prioritized genes for ulcerative colitis and coronary artery disease, and found that our method enhanced the significance of several disease-associated pathways. Therefore, REGENT is expected to be a useful tool for prioritizing candidate genes from GWAS and facilitating both research and practice of precision medicine.

2. Methods and materials

2.1. Schematic overview of REGENT

As illustrated in Fig. 1, REGENT takes GWAS summary data (i.e. SNP p-values) of a complex phenotype of interest and multiple gene networks as inputs, and produces inferred posterior probability of association (PPA) between a gene and the given disease as output via three computational steps. First, gene-level p-values are calculated by aggregating SNP-level p-values from GWAS summary data with the consideration of linkage disequilibrium (LD). Second, network representation learning is employed to learn embeddings of genes in each network separately in an unsupervised manner. These embeddings are learned by considering information from gene networks only and are not relevant to a phenotype, and thus the learning procedure is performed only once. Third, a hierarchical statistical model with an efficient expectation-maximization (EM) algorithm is developed to integrate learned embeddings with gene-level p-values to infer for each gene a posterior probability of association, which is in turn used for gene prioritization, with genes with larger PPAs more likely to be associated with the disease.

2.2. Calculation of gene-level p-values

We use PASCAL [20] to compute gene-level p-values from GWAS summary data. Specifically, we first perform *cis*-mapping for each gene and assign a SNP to a gene if the SNP locates within 50 kb from the gene. Then, we aggregate these SNP-level p-values into gene-level p-values. Suppose that there are a total of N genes, and there are K_i SNPs assigned to the i-th gene, with corresponding p-values for these SNPs being $p_1,\ ...,p_{K_i}$. Let Σ_i be the pairwise correlation matrix for these SNPs derived from the 1000 Genomes Project [21]. The test statistic for gene i is then defined as $T_i = \sum_{j=1}^{K_i} z_j^2 \sim \sum_{j=1}^{K_i} \lambda_j \chi_1^2$, where $z_j = \Phi^{-1}(1-p_j)$ with $\Phi(\cdot)$ being the cumulative distribution function of the standard normal distribution, λ_j the j-th eigenvalue of Σ_i, and χ_1^2 the chi-squared distribution with one degree of freedom. With this formulation, the gene-level p-value for gene i is then calculated as $p = \Pr(\sum_{j=1}^{K_i} \lambda_j \chi_1^2 \geqslant T_i)$, which can be efficiently computed by the Davies algorithm [22].

2.3. Network representation learning for extracting gene embeddings

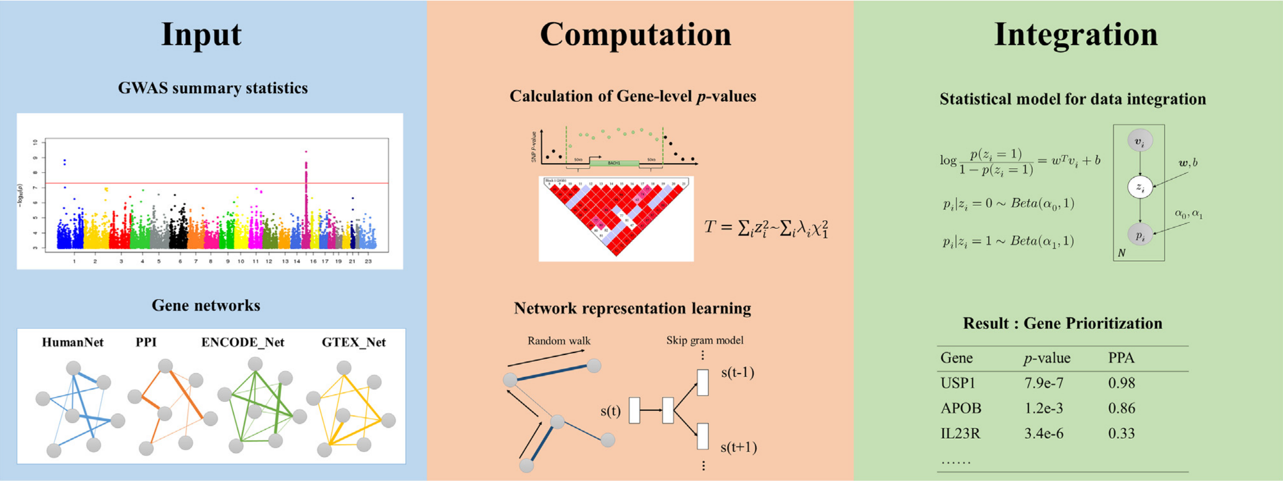
We use the framework proposed in node2vec [17] to perform network representation learning on a gene network. Specifically, a gene network is represented as $G = (V, E)$, where V is the set of nodes (genes), and E the set of edges. The weight matrix of a gene network is denoted as $\mathbf{W} = (w_{ij})_{N\times N}$, where w_{ij} is the weight of the edge between gene i and gene j. A gene network of N nodes is said to be sparse if the number of edges, $|E|$, is far less than the number of possible edges, $N(N-1)/2$. The objective of network representation learning is to learn a d-dimensional real-valued embedding vector $\boldsymbol{v}_i \in R^d$ for each gene i in the network, and the value of d is far less than N. The basic principle underlying the network representation learning is that nearby nodes in the network should have similar embedding vectors, which ensures the preservation of network structure during learning. There are two main steps in network representation learning: generating node sequences via a graph exploration algorithm (e.g., random walk) and learning embedding vectors using collected node sequences. Here, we use a random walk model to generate node sequences. For a gene u in the gene network, we simulate a l-step random walk starting from u and collect the genes appeared in this random walk as $S = \{s_0 = u, s_1, ...,s_l\}$. The i-th jump within the random walk is generated by sampling another gene according to the probabilities of transitions, defined as:

$$p(s_i = y|s_{i-1} = x) = \frac{w_{xy}}{\sum\limits_{y} w_{xy}} \tag{1}$$

Thus, the probability of transiting to the next gene relies only on the previous visited gene and edge weight between them. For each gene, we simulate the random walk procedure r times and collect corresponding gene sequences. Given a gene sequence, we use the skip-gram [23] to model the conditional distributions of the surrounding genes of each gene, as:

$$p(s_{i+t}|s_i) = \frac{\exp(\boldsymbol{v}_{s_i}^T \boldsymbol{v}_{s_{i+t}})}{\sum\limits_{k} \exp(\boldsymbol{v}_{s_i}^T \boldsymbol{v}_k)}, \quad t \in \{-k, ...,k\}\setminus\{0\} \tag{2}$$

where $\boldsymbol{v}_i \in R^d$ is the embedding vector of gene i, d the embedding dimension, and k the context size. The gene collection


Input
GWAS summary statistics
Gene networks
HumanNet
PPI
ENCODE_Net
GTEX_Net
Computation
Calculation of Gene-level p-values
$T = \sum_i z_i^2 \sim \sum_i \lambda_i \chi_1^2$
Network representation learning
Random walk
Skip gram model
s(t-1)
s(t)
s(t+1)
Integration
Statistical model for data integration
$\log \frac{p(z_i=1)}{1-p(z_i=1)} = w^T v_i + b$
$p_i|z_i = 0 \sim Beta(\alpha_0, 1)$
$p_i|z_i = 1 \sim Beta(\alpha_1, 1)$
v_i
z_i
p_i
w, b
α_0, α_1
N
Result : Gene Prioritization

Gene	p-value	PPA
USP1	7.9e-7	0.98
APOB	1.2e-3	0.86
IL23R	3.4e-6	0.33
……		

Fig. 1. Schematic overview of REGENT. The inputs consist of GWAS summary data and multiple gene networks, and the outputs are PPAs (short for posterior probability of association) of genes. There are three steps from the input to the output, including calculation of gene-level p-values, network representation learning on multiple gene networks, and statistical integration of embeddings and gene-level p-values.

$\{s_{i+t}, t \in \{-k, ..., k\} \setminus \{0\}\}$ is called the context of gene i. The likelihood of data (all collected gene sequences by random walk) is defined as:

$$L = \prod_S \prod_i \prod_t p(s_{i+t}|s_i) \tag{3}$$

To estimate $\boldsymbol{v}_i$, we adopt the maximum likelihood principle and minimize the negative logarithm of data likelihood, which is defined as:

$$\min_{\boldsymbol{v}_1,...,\boldsymbol{v}_N} \sum_{i=1}^{N} \{|N(i)|log(Z_i) - \sum_{j \in N(i)} \boldsymbol{v}_j^T \boldsymbol{v}_i\} \tag{4}$$

where $Z_i = \sum_{j=1}^{N} \exp(\boldsymbol{v}_i^T \boldsymbol{v}_j)$ is the partition function for gene i and can be efficiently approximated by negative sampling [23]. $N(i)$ collects genes that appear in the contexts of gene i within any collected random walks. Stochastic gradient descent (SGD) is used to estimate model parameters or embedding vectors. We use the default parameters $r = 10$, $l = 80$, $d = 128$, $k = 10$ because this choice is shown to achieve the-state-of-art performance in a variety of applications [17]. For each gene network, we perform the same procedure of network representation learning and obtain corresponding embedding vectors of genes in the gene network. Finally, we concatenate the embedding vectors learned from different gene networks together to form the final integrated embedding vector for each gene.

2.4. Statistical model for integrating gene embeddings and GWAS data

With the previous two steps done, we obtain a gene-level p-value p_i and an embedding vector $\boldsymbol{v}_i \in R^{md}$ for gene i ($i = 1, ..., N$), where m is the number of gene networks, d the embedding dimension, and N the number of genes. We then develop a hierarchical statistical model to infer associations between genes and the disease by integrating these two pieces of information. For gene i, we assign a binary latent variable z_i to denote its association status with a given disease, where $z_i = 1$ denotes that gene i is associated with the disease, and $z_i = 0$ denotes that gene i is not associated. As shown in Fig. 1, we specify the generative process of gene-level p-values as:

$$\log\left(\frac{p(z_i = 1)}{1 - p(z_i = 1)}\right) = \boldsymbol{w}^T \boldsymbol{v}_i + b \tag{5}$$

$$p_i|z_i = 0 \sim \mathrm{Beta}(\alpha_0, 1) \tag{6}$$

$$p_i|z_i = 1 \sim \mathrm{Beta}(\alpha_1, 1) \tag{7}$$

where $\Theta = \{\boldsymbol{w}, b, \alpha_0, \alpha_1\}$ are parameters to be estimated, and $z_i, i = 1, ..., n$ latent variables to be inferred. The usage of Beta distribution for modelling p-values is justified by previous studies [24,25]. EM algorithm is used to estimate the model parameters. In the E-step, we estimate the posterior probability of the hidden variable z_i as:

$$\gamma_i^{(t)} = p(z_i = 1|p_i, \Theta^{(t-1)}) = \frac{p(z_i = 1|\Theta^{(t-1)})p(p_i|z_i = 1)}{\sum_{j=0,1} p(z_i = j|\Theta^{(t-1)})p(p_i|z_i = j)}$$

$$= \frac{\sigma_i^{(t-1)} \alpha_1^{(t-1)} p_i^{\alpha_1^{(t-1)}-1}}{\sigma_i^{(t-1)} \alpha_1^{(t-1)} p_i^{\alpha_1^{(t-1)}-1} + (1-\sigma_i^{(t-1)}) \alpha_0^{(t-1)} p_i^{\alpha_0^{(t-1)}-1}} \tag{8}$$

where t denotes the number of iteration, $\sigma_i^{(t-1)} = \sigma((\boldsymbol{w}^{(t-1)})^T \boldsymbol{v}_i + b^{(t-1)})$ and $\sigma(x) = 1/(1 + exp(-x))$ the sigmoid function. In the M-step, we maximize the expected log-likelihood of complete data with respect to posterior distributions of latent variables, which is defined as

$$Q(\Theta) = \sum_{i=1}^{N} \gamma_i^{(t)} [\log(\sigma_i) + \log(\alpha_1) + (\alpha_1 - 1)\log(p_i)]$$

$$+ \sum_{i=1}^{N} (1 - \gamma_i^{(t)})[\log(1 - \sigma_i) + \log(\alpha_0) + (\alpha_0 - 1)\log(p_i)] \tag{9}$$

where $\sigma_i = \sigma(\boldsymbol{w}^T \boldsymbol{v}_i + b)$. Maximizing the function $Q(\Theta)$ with respect to α_0, α_1, we obtain the update formula as

$$\alpha_0^{(t)} = -\frac{\sum_{i=1}^{N} (1 - \gamma_i^{(t)})}{\sum_{i=1}^{N} (1 - \gamma_i^{(t)}) \log(p_i)} \tag{10}$$

$$\alpha_1^{(t)} = -\frac{\sum_{i=1}^{N} \gamma_i^{(t)}}{\sum_{i=1}^{N} \gamma_i^{(t)} \log(p_i)} \tag{11}$$

For $\boldsymbol{w}$, b, no closed-form update formula exist and we resort to the Newton-Raphson method [26] for updating $\boldsymbol{w}$, b:

$$\begin{pmatrix} \boldsymbol{w}^{(t)} \\ b^{(t)} \end{pmatrix} \leftarrow \begin{pmatrix} \boldsymbol{w}^{(t-1)} \\ b^{(t-1)} \end{pmatrix} - H^{-1} \begin{pmatrix} \frac{\partial Q}{\partial \boldsymbol{w}} \\ \frac{\partial W}{\partial b} \end{pmatrix} \tag{12}$$

The gradient of $Q(\Theta)$ with respect to $\boldsymbol{w}$, b and corresponding Hessian matrix are:

$$\frac{\partial Q}{\partial \boldsymbol{w}} = \sum_{i=1}^{N} (\gamma_i^{(t)} - \sigma_i^{(t)}) \boldsymbol{v}_i \tag{13}$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{N} (\gamma_i^{(t)} - \sigma_i^{(t)}) \tag{14}$$

$$H = \begin{pmatrix} \sum_{i=1}^{N} \sigma_i^{(t)}(\sigma_i^{(t)}-1)\boldsymbol{v}_i\boldsymbol{v}_i^T & \sum_{i=1}^{N} \sigma_i^{(t)}(\sigma_i^{(t)}-1)\boldsymbol{v}_i \\ \sum_{i=1}^{N} \sigma_i^{(t)}(\sigma_i^{(t)}-1)\boldsymbol{v}_i^T & \sum_{i=1}^{N} \sigma_i^{(t)}(\sigma_i^{(t)}-1) \end{pmatrix} \tag{15}$$

We alternate E-step and M-step until convergence, e.g., $|\Theta^{(t)}-\Theta^{(t-1)}|_\infty < 10^{-9}$. Empirically, the convergence can be reached within 100 iterations.

Nevertheless, direct incorporation of the learned embeddings does not work in practice due to the high dimensionality of the embedding vectors, which makes it difficult to fit the statistical model accurately. We therefore adopt a two-step strategy to solve this problem. In the first step, we perform supervised dimensionality reduction via a linear discriminant analysis (LDA) [19] to extract low-dimensional embeddings that are relevant to the disease of interest. In the second step, we fit the statistical model with the reduced embeddings. Specifically, we first fit the statistical model without embeddings and only use the intercept b in Eq. (5), from which we obtain the estimated PPAs of all genes as $\{\widehat{\gamma}_i, i = 1, ...,N\}$ via Eq. (8). Using $\{\widehat{\gamma}_i, i = 1, ...,N\}$ as soft labels, we apply LDA to embeddings, where the covariance matrices of between-class and within-class are estimated as:

$$S_b = \frac{\sum_{i=1}^{N}(1-\widehat{\gamma}_i)}{N}(\boldsymbol{\mu}_0-\boldsymbol{\mu})(\boldsymbol{\mu}_0-\boldsymbol{\mu})^T + \frac{\sum_{i=1}^{N}\widehat{\gamma}_i}{N}(\boldsymbol{\mu}_1-\boldsymbol{\mu})(\boldsymbol{\mu}_1-\boldsymbol{\mu})^T \tag{16}$$

$$S_w = \frac{1}{N}\sum_{i=1}^{N} \{\widehat{\gamma}_i(\boldsymbol{v}_i-\boldsymbol{\mu}_1)(\boldsymbol{v}_i-\boldsymbol{\mu}_1)^T + (1-\widehat{\gamma}_i)(\boldsymbol{v}_i-\boldsymbol{\mu}_0)(\boldsymbol{v}_i-\mu_0)^T\} \tag{17}$$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}_i,\ \mu_0 = \frac{\sum_{i=1}^{N}(1-\widehat{\gamma}_i)\boldsymbol{v}_i}{\sum_{i=1}^{N}(1-\widehat{\gamma}_i)},\ \boldsymbol{\mu}_1 = \frac{\sum_{i=1}^{N}\widehat{\gamma}_i\boldsymbol{v}_i}{\sum_{i=1}^{N}\widehat{\gamma}_i} \tag{18}$$

where $\boldsymbol{v}_i$ is the embedding vector of gene i. Then, we compute the projection vector $\boldsymbol{s}$ as the eigenvector of $S_w^{-1}S_b$ that corresponds to the largest eigenvalue. Finally, we compute the reduced embedding for each gene as $\widetilde{\boldsymbol{v}}_i = \boldsymbol{s}^T\boldsymbol{v}_i,\ i = 1, ...,N$. These reduced embeddings are used to fit the statistical model, in which we replace $\boldsymbol{v}_i$ with $\widetilde{\boldsymbol{v}}_i$ in Eq. (5). The fitted model gives the estimation of PPAs via Eq. (8), providing a means of gene prioritization. Genes with larger PPAs are more likely to be associated with the disease of interest and will be ranked in topper positions. In practice, it takes only several minutes to finish all computation involved in the analysis of one individual disease, making it a suitable tool for the accumulating GWAS data.

3. Results

3.1. Datasets

We collected four gene networks, including HumanNet (version v1) [6], protein interaction network (PPI) from the BioGrid database (version 3.4.147) [9], gene co-expression network from the GTEX project (GTEX_Net for short) [10] and gene regulatory network from the ENCODE project (ENCODE_Net for short) [11]. Specifically, we merged all gene co-expression networks for different tissues [10] and retained the maximum edge weight. For all gene networks, we mapped corresponding gene ID to HGNC symbol using the Ensemble Biomart [27]. As shown in Table 1, the four gene networks vary in terms of the number of nodes and edges, and all of them are sparse as the corresponding edge densities are small. Note that edge weights for these gene networks are determined by the existing studies and are not influenced by any GWAS data.

We collected GWAS summary data for six complex diseases, including Parkinson's Disease [28] (PD), Rheumatoid Arthritis [29] (RA), Crohn's Disease [30] (CD), Ulcerative Colitis [31] (UC), Coronary Artery Disease [32] (CAD) and Type 2 Diabetes [33] (T2D). These diseases belong to different types of complex diseases, such as neurological, immune-related and cardiovascular diseases. The details about GWAS data of these diseases are shown in Table 2, including the numbers of cases, controls, SNP genotyped and genes.

Table 1
Details about gene networks used in our study. The first column is the name of the gene network, and the following columns record the number of nodes and edges, average degree of nodes and edge density.

Network	#Node	#Edge	Avg. degree	Edge density
HumanNet	15,285	434,418	56.84	0.37%
PPI	16,134	305,924	37.92	0.24%
GTEX_Net	9998	1,548,622	309.79	3.09%
ENCODE_Net	19,373	532,663	54.99	0.28%

Table 2
Details about GWAS data used in our study. The first column is the disease name, and the following columns denote the numbers of cases, controls, SNP, and genes respectively.

Disease	#Case	#Control	#SNP	#Gene
PD	1713	3978	463,185	16,937
RA	5539	20,169	2,556,272	17,063
CD	6333	19,718	1,428,749	17,053
UC	6687	19,718	1,428,749	17,053
CAD	22,233	64,762	2,337,127	17,005
T2D	34,840	114,981	2,473,441	14,868

3.2. Disease-associated genes tend to be densely connected to each other in gene networks

Using the collected datasets, including the four gene networks and the six GWAS datasets, we explored whether the assumption underlying network-based gene prioritization held, that is, whether disease-associated genes tended to be densely connected to each other in the gene networks. For a gene network and a GWAS dataset, we first sorted genes according to their p-values in nondecreasing order. Then, we calculated the weighted edge density of top K (K ranges from 100 to 5000 with step size 100) genes as

$$WED_K = \frac{\sum_{1<=i<j<=K} w_{ij}}{K \times (K-1)/2} \tag{19}$$

where w_{ij} is the edge weight between gene i and gene j. A line was drawn by plotting WED_K against K for each pair of a gene network and a GWAS dataset, as shown in the Fig. 2. From this figure, we observed the clear pattern that the value of WED_K decreased as K increased, demonstrating that genes ranked higher (with smaller p-values) had stronger connections with each other. From this figure, we also found that the pattern revealed by ENCODE_Net was noisier than the others, implying that ENCODE_Net was less informative than the others. Although the ground-truth of disease-associated genes was unknown, the GWAS p-value was regarded as a good indication of the association as these GWAS datasets were well-powered by relatively large sample size (Table 2). Therefore, the assumption that disease-associated genes tend to be densely connected to each other in the gene networks was validated, making it reasonable to develop methods for network-based gene prioritization.

3.3. Gene embeddings are informative for inferring association status

We then investigated whether the learned embeddings from the four gene networks could provide information for inferring disease-associated genes. For each gene network, we collected the learned embeddings of all genes from network representation learning and evaluated

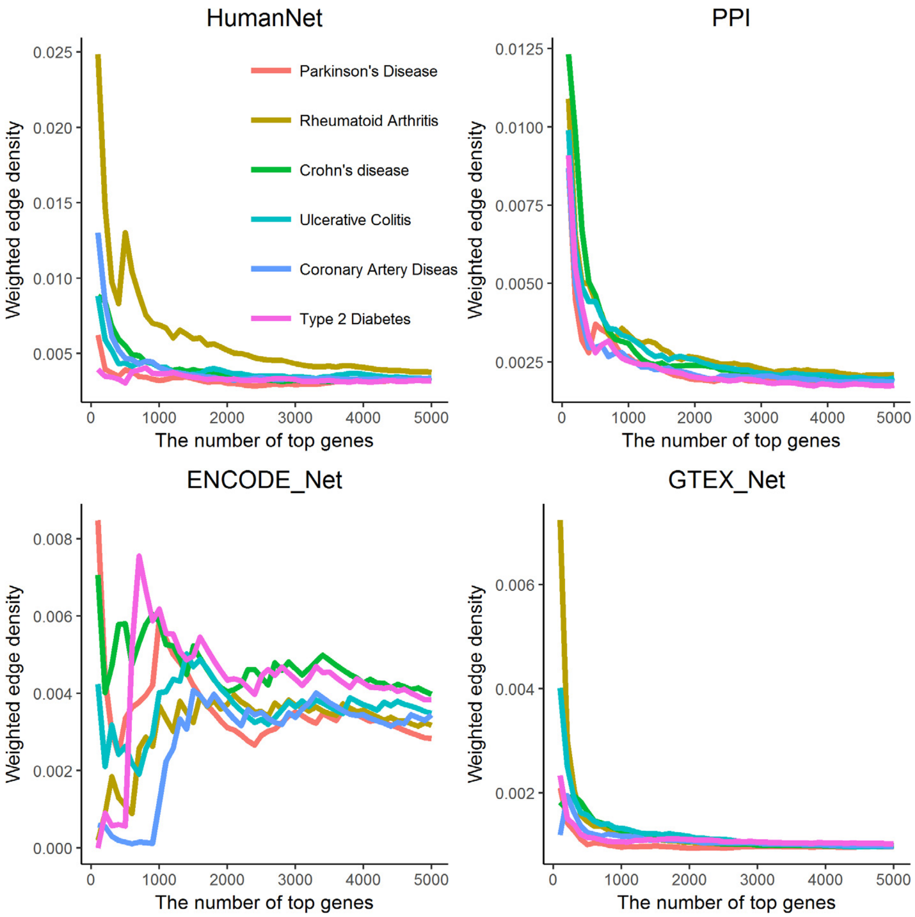


Fig. 2. Genes with smaller *p*-values connect with each other more densely in the four gene networks. In each subplot, lines with different colors represent different complex diseases. The x axis denotes the number of top genes ranked by p-values, and the y axis is the corresponding weighted edge density. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

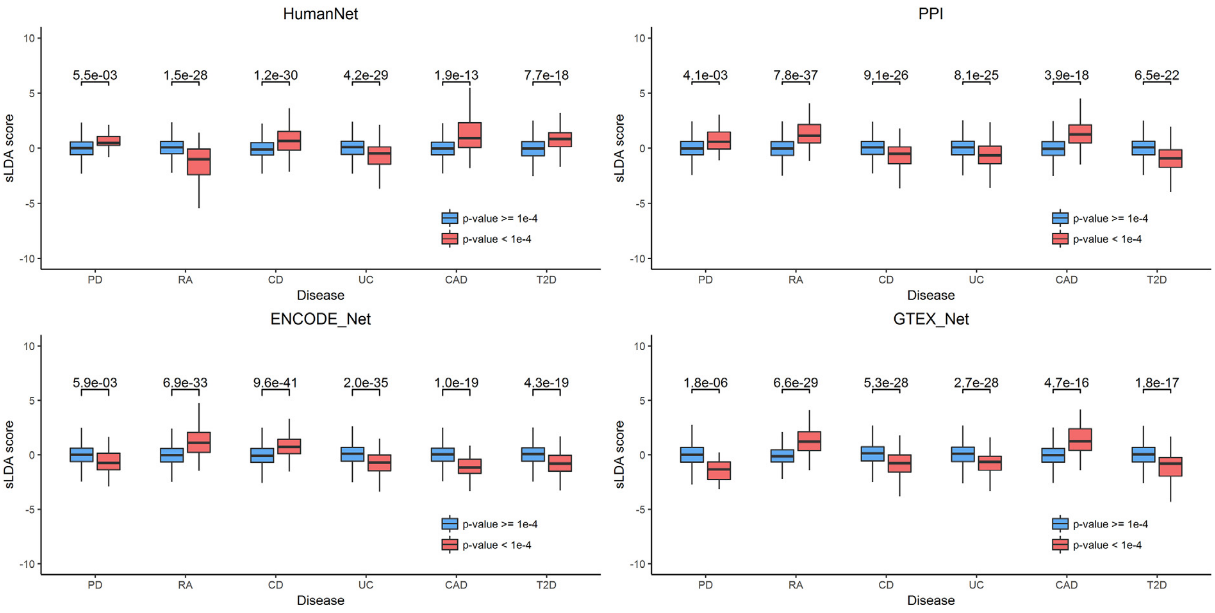


Fig. 3. Gene embeddings are informative for inferring disease-associated genes. Each subplot corresponds to one gene network. In each subplot, the x axis denotes different diseases, and the y axis denotes the compressed embedding score.

the relationship between the learned embeddings and association status of genes for a specific disease. Using the supervised dimension reduction described in Section 2.4, we obtained a one-dimensional embedding score for each gene, and this score contained the compressed information from the original embedding vector that was relevant to the disease of interest. We partitioned all genes into two groups: one group of genes with *p*-values less than 1×10^{-4} and the other one group of genes with *p*-values greater than 1×10^{-4}. We then compared the distribution of the embedding scores between the two groups of genes. As shown in Fig. 3, we found that there was obvious difference between the two groups of genes in terms of the distribution of the embedding scores and the difference was significant (Two-sided Wilcoxon rank sum test) for each pair of disease and gene network. These observations supported the claim that the learned embeddings is informative for inferring disease-associated genes.

3.4. Performance of gene prioritization on the six complex diseases

We applied REGENT to the six complex diseases and validated its effectiveness. We compared the performance of gene prioritization of REGENT with two existing the-state-of-art methods, namely GWAB [6,34] and NetWAS [8]. We used the same GWAS data of the six complex diseases and the corresponding official websites of the two methods for computation. For NetWAS, we used two versions, including NetWAS (all), which used a non tissue-specific gene network, and NetWAS (specific), which used a tissue-specific gene network. We followed the step used in GWAB [34] and used the relevant tissues of these diseases for NetWAS (specific), such as brain for Parkinson’s Disease, bone for Rheumatoid Arthritis, intestine for Crohn’s Disease and Ulcerative Colitis, heart for Coronary Artery Disease and liver for Type 2 Diabetes. Because the ground-truth of disease-associated genes was unknown, we adopted the similar strategy as GWAB and used the DisGeNet database [35] as a surrogate for the ground-truth. The DisGeNet database is a comprehensive platform for discovering disease-gene associations via integrating multiple data sources, such as OMIM [36], GAD [37] and text mining.

To evaluate the performance of gene prioritization for a specific disease, we calculated the number of genes annotated as disease-associated among the top K (ranges from 0 to 1000) prioritized genes for each method. For the same value of K, a method was said to be better than the others if it could uncover more disease-associated genes. From Fig. 4, we found that REGENT achieved better performance than the other two methods. For example, REGENT uncovered 179 genes in top 1000 when applied to Crohn's Disease, while the corresponding numbers for NetWAS (all), NetWAS (specific) and GWAB are 105, 62, and 127 respectively. This phenomenon is also apparent for Parkinson’s Disease, Rheumatoid Arthritis, and Ulcerative Colitis. For Coronary Artery Disease and Type 2 Diabetes, REGENT performed similar to GWAB, both obviously better than NetWAS (all) and NetWAS (specific). In summary, REGENT performed the best, followed by GWAB, and NetWAS was the last. As mentioned above, NetWAS directly uses adjacency matrix of a gene network as high-dimensional features for training SVM, which may explain its inferior performance. GWAB uses network propagation algorithm to propagate association evidence of genes to its neighbors, and thus it only considers local information without considering such problems as sparsity and noise in a gene network. Besides, both NetWAS and GWAB only utilize a single gene network and cannot use multiple gene networks. In contrast, we use the network representation learning to learn embeddings of genes, which can substantially solve the problem of sparsity and noise. Besides, our method can naturally integrate multiple gene networks.

3.5. Effectiveness of integrating multiple gene networks

We investigated the advantage of integrating multiple gene networks by the similar experiments described before. We applied network representation learning to learn embeddings of genes from a single gene network only, and we used REGENT to integrate these embeddings with GWAS data. We then compared the performance of gene prioritization of each individual gene network with the combined one, in which the four gene networks were all integrated by REGENT as described before. We adopted the same metric to measure the performance of gene prioritization as before. As shown in Fig. 5(A), we found that by integrating multiple gene networks, REGENT could uncover more disease-associated genes than using individual gene networks alone. For example, when applying to Crohn’s Disease, REGENT retrieved 179 disease-associated genes in top 1000 genes, while the corresponding numbers were 164, 154, 138 and 153 for the four individual gene

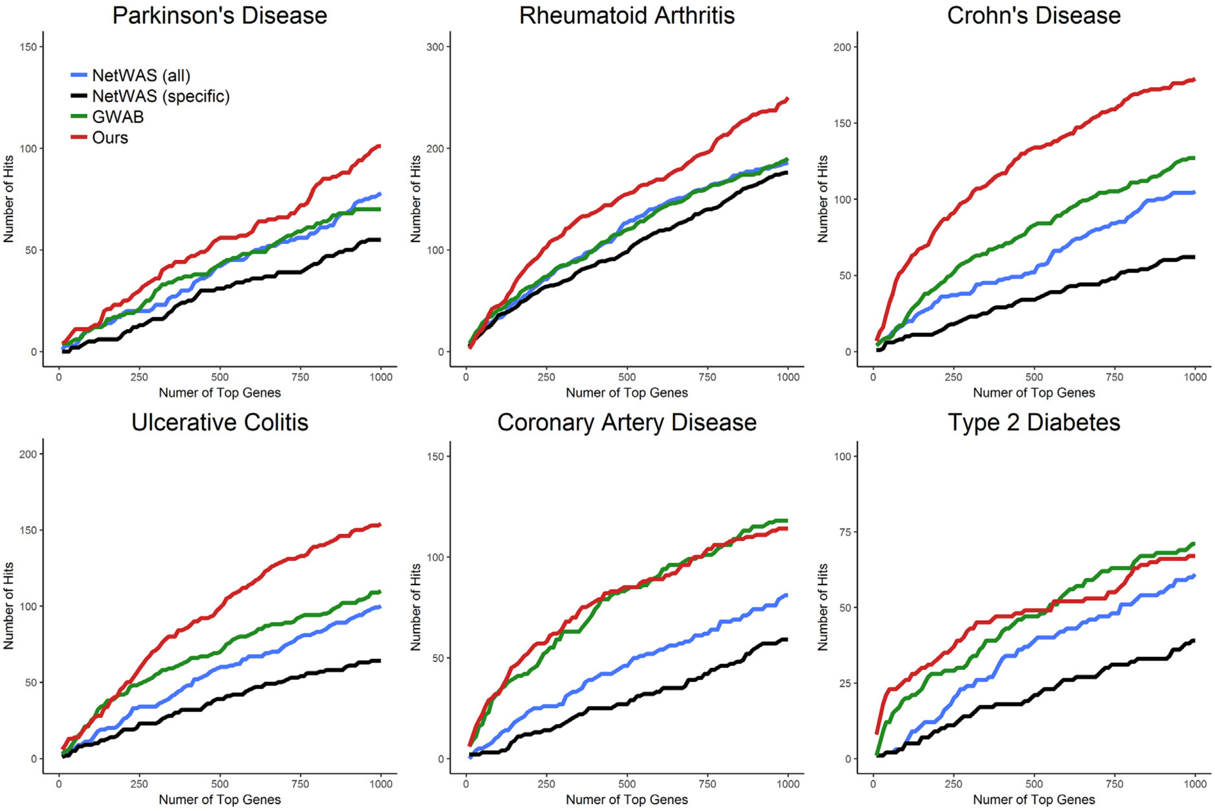


Fig. 4. Performance comparison between REGENT and two existing methods (GWAB and NetWAS). The x axis denotes the number of top ranked genes and the y axis is the corresponding number of genes annotated as disease-associated.

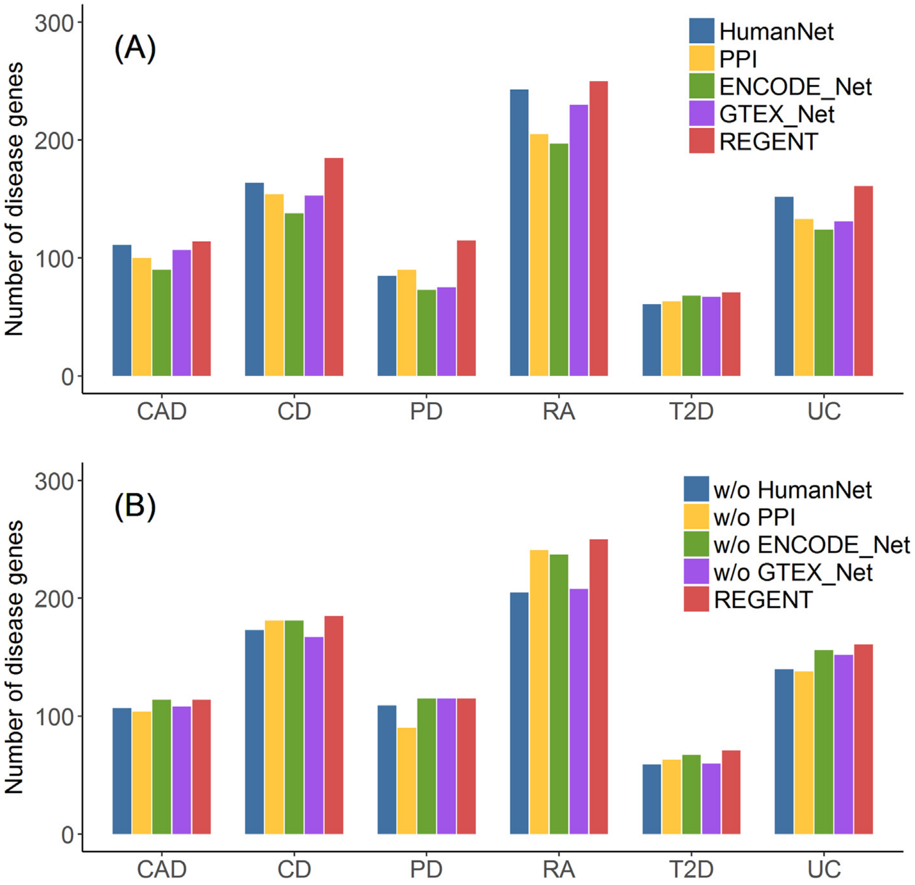


Fig. 5. The effectiveness of integrating multiple gene networks in REGENT. (A) Comparison between REGENT integrating only individual gene networks and REGENT. (B) Comparison between REGENT removing individual gene networks and REGENT.

networks respectively. Particularly, ENCODE_Net showed the worst performance for five diseases, which was consistent with previous observation that ENCODE_Net was less informative. In addition, we also evaluated the contribution of each individual gene network to the performance of the combined model. Specifically, for each gene network, we calculated the number of disease-associated genes in the top 1000 genes ranked by the REGENT that integrated the other three gene networks. From Fig. 5(B), we found that the performance dropped when any gene network was excluded, confirming the essential contribution of each gene network to the combined model.

3.6. Identification of novel susceptibility genes and functional pathways for inflammatory bowel disease

Patients affected by Inflammatory bowel diseases (IBD), such as Crohn's Disease (CD) and Ulcerative Colitis (UC), usually share symptoms like inflammation and ulcers of colon, rectum and other components of the intestine system. According to a survey [38], IBD affects approximately 1.4 million Americans, calling for molecular understanding and effective treatment for this disease. We applied REGENT to existing GWAS datasets for UC [31] and CD [30], as detailed in Table 2. From the result of UC, we found an interesting gene named GATA3, a transcription factor belonging to the GATA family. This gene had a p-value of 2.42×10^{-4}, which was not statistically significant and was ranked at 453rd by GWAS data alone and 581st by GWAB. Using our method, this gene was assigned a high PPA of 0.985 and was ranked at 140th. Recently, this gene was shown to be a key regulator of T-cell differentiation and might be involved in the disease development of UC [39]. For CD, we found a gene named STAT1, whose p-value was 1.85×10^{-4}, not statistically significant. The GWAS data alone ranked this gene at 466th while REGENT ranked it at 61st. Several evidences [40,41] supported the association between the gene STAT1 and CD.

We then investigated whether the prioritized genes revealed functional pathways for UC. We used the top 219 genes with PPA greater than 0.9 for pathway analysis with the tool ConsensusPathDB [42]. We also used the top 219 genes ranked by GWAS data alone for pathway analysis and compared REGENT with GWAS alone. As shown in Fig. 6, using prioritized genes given by REGENT, the significance of several pathways became more evident, such as immune system, Th17 cell differentiation and cytokine signaling in the immune system. These pathways are well known to be related to the immune system, which plays a vital role in UC [43,44]. On the contrary, REGENT also reduced the significance of some pathways, such as endothelins, arachidonic acid metabolism, and many others, which had no strong relevance with UC. The comparison between REGENT and the other two methods revealed the similar trend, and we omitted it due to space limitation. Therefore, our method improved the power for pathway analysis of UC, highlighting its potential for deepening our understanding of disease mechanism.

3.7. Identification of novel susceptibility genes and functional pathways for coronary artery disease

Coronary artery disease (CAD) is one of the most severe heart diseases, and it can affect the process of blood flowing through arteries and eventually result in heart failure and arrhythmias. According to World Health Organization (WHO), CAD accounts for over 15% of global deaths (about 7 millions) in 2015, appealing for molecular understanding and effective treatment for this disease. We applied REGENT to existing GWAS datasets for CAD [32], of which details were shown in

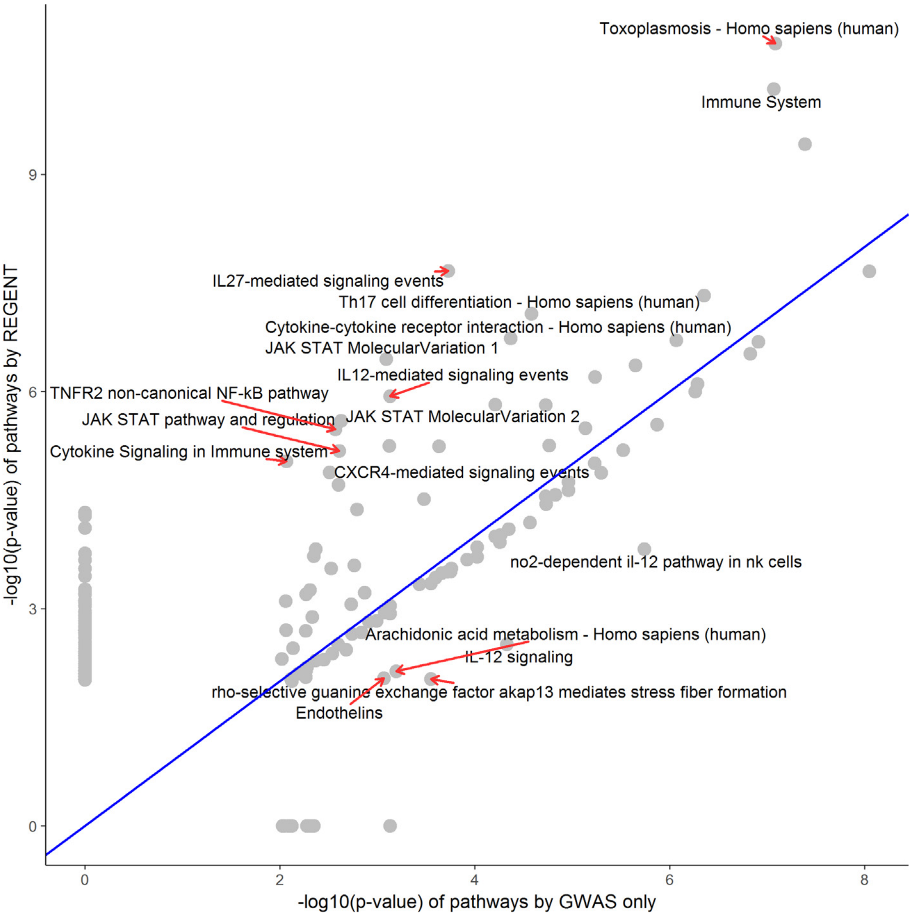


Fig. 6. REGENT improves pathway analysis of ulcerative colitis. Each point in this plot represents a pathway, and the blue line denotes the line y = x. The x axis and y axis denote the significance $-\log_{10}(p-value)$ of each pathway using prioritized genes by GWAS only and REGENT. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2. From the result of CAD, we found an interesting gene named BCAS3, a gene that was amplified and overexpressed in breast cancer cells. This gene had a p-value of 1.01×10^{-2}, which was not statistically significant and was ranked at 729th by GWAS data alone and 403rd by GWAB. Using our method, this gene received a high PPA of 0.949 and was ranked at 50th by our method. Recently, this gene was shown to be associated with CAD and showed evidence for selection and antagonistic pleiotropy [45].

Then, we investigated whether the prioritized genes revealed functional pathways for CAD. We used the top 54 genes with PPA greater than 0.9 for pathway analysis with the tool ConsensusPathDB [42]. We also used the top 54 genes ranked by GWAS data alone for pathway analysis and compared REGENT with GWAS alone. As shown in Fig. 7, using prioritized genes given by REGENT, the significance of several pathways became more evident, such as triglyceride-rich lipoprotein particle remodelling, plasma lipoprotein particle and lipid homeostasis. These pathways were related with levels of lipoprotein, plasma lipoprotein, and lipid, which were previously reported to be associated with CAD [46]. On the contrary, REGENT also reduced the significance of some pathways, such as fucosylation, regulation of growth, and neurotrophin signalling pathway, and many others, which had no strong relevance with CAD. Therefore, our method improved the power for pathway analysis of CAD, again supporting the potential of our method for deepening our understanding of disease mechanism.

4. Discussion and conclusion

We have proposed a novel method named REGENT for network-based gene prioritization. Our method uses the network representation learning to learn embeddings of genes from multiple gene networks and adopts a hierarchical statistical model to integrate the learned embeddings with GWAS data. Applications of our method to GWAS data of six complex diseases have demonstrated that our method is superior to existing methods in the identification of disease-associated genes. Further pathway analysis has shown that our method improves the significance of several disease-associated pathways.

The success of our method can be attributed to a combination of several aspects. First, we realize the importance of considering the sparse and noisy properties of gene networks and utilize the network representation learning to alleviate this issue. Second, our method has the capability to integrate multiple gene networks that provide different aspects about functional relationships between genes. Third, we design a hierarchal statistical model coupled with an efficient EM algorithm to integrate the learned network embeddings with GWAS data. Finally, we use supervised dimension reduction to extract useful signals from the learned generic network embeddings, solving the problem of dimensionality. The major shortcoming of our method is the lack of direct biological interpretations for the learned embeddings of genes. As shown in our study, the learned embeddings contain the information for inferring disease-associated genes. However, the biological meaning of a dimension in the embedding vector is not clear. This difficulty in interpretability is still a main challenge in the field of representation learning and deep learning [18]. Some efforts like matching learned convolutional filters with existing motif databases [47,48] have provided some clues for interpreting deep learning models used for DNA sequence analysis.

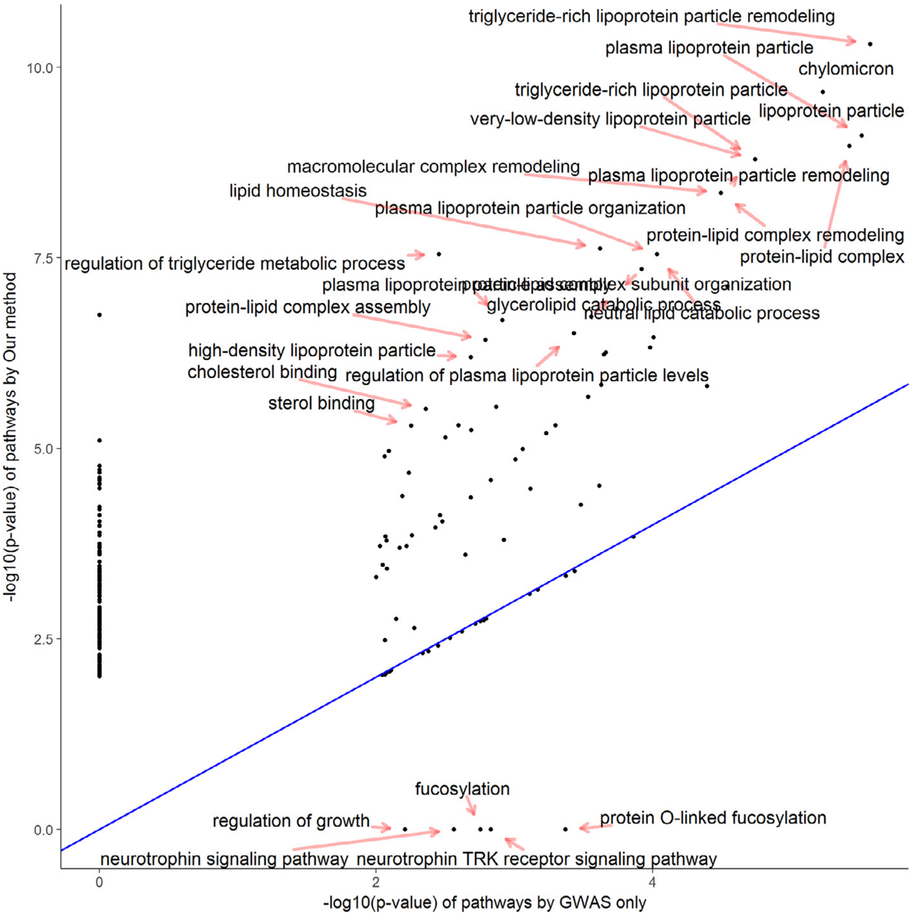


Fig. 7. REGENT improves pathway analysis of coronary heart disease. Each point in this plot represents a pathway, and the blue line denotes the line y = x. The x axis and y axis denote the significance $-\log_{10}(p{-}value)$ of each pathway using prioritized genes by GWAS only and REGENT. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

There are several directions to extend our work. First, how to integrate more gene functional networks, such as gene co-opening network [49], gene semantic similarity network [50], and tissue-specific gene regulatory network [51], would be interesting. Second, our method only considers one disease at a time, and the joint modeling of multiple diseases with pleiotropy [52] taken into consideration would be interesting and may further improve discovery power. Third, there have been some studies using traditional probabilistic models to prioritize GWAS candidate genes [12], and how to combine the strength of both traditional probabilistic models and network representation learning would be an interesting topic. Finally, how to extend the network representation learning to other biological network analysis tasks would be a valuable direction.

Funding

This research was partially supported by the National Natural Science Foundation of China (Nos. 61721003, 61573207), The Recruitment Program of Global Experts of China, and Tsinghua National Laboratory for Information Science and Technology. Rui Jiang is a RONG professor at the institute for data science, Tsinghua University.

Competing interests

The authors declare that they have no competing interests.

References

[1] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, Nucleic Acids Res. 42 (D1) (2013) D1001–D1006.
[2] P. Sanseau, P. Agarwal, M.R. Barnes, T. Pastinen, J.B. Richards, L.R. Cardon, V. Mooser, Use of genome-wide association studies for drug repositioning, Nat. Biotechnol. 30 (4) (2012) 317–320.
[3] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, Finding the missing heritability of complex diseases, Nature 461 (7265) (2009) 747–753.
[4] L.D. Ward, M. Kellis, Interpreting noncoding genetic variation in complex traits and human disease, Nat. Biotechnol. 30 (11) (2012) 1095–1106.
[5] R.M. Cantor, K. Lange, J.S. Sinsheimer, Prioritizing GWAS results: a review of statistical methods and recommendations for their application, Am. J. Human Genetics 86 (1) (2010) 6–22.
[6] I. Lee, U.M. Blom, P.I. Wang, J.E. Shim, E.M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Res. 21 (7) (2011) 1109–1121.
[7] M. Taşan, G. Musso, T. Hao, M. Vidal, C.A. MacRae, F.P. Roth, Selecting causal genes from genome-wide association studies via functionally coherent subnetworks, Nat Methods 12 (2) (2015) 154–159.
[8] C.S. Greene, A. Krishnan, A.K. Wong, E. Ricciotti, R.A. Zelaya, D.S. Himmelstein, R. Zhang, B.M. Hartmann, E. Zaslavsky, S.C. Sealfon, Understanding multicellular function and disease with human tissue-specific networks, Nat. Genetics 47 (6) (2015) 569–576.
[9] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, O'donnell L: The BioGRID interaction database: 2015 update, Nucleic Acids Res. 43 (D1) (2014) D470–D478.
[10] E. Pierson, D. Koller, A. Battle, S. Mostafavi, G. Consortium, Sharing and specificity of co-expression networks across 35 human tissues, PLoS Comput. Biol. 11 (5) (2015) e1004220.
[11] M.B. Gerstein, A. Kundaje, M. Hariharan, S.G. Landt, K.-K. Yan, C. Cheng, X.J. Mu, E. Khurana, J. Rozowsky, R. Alexander, Architecture of the human regulatory network derived from ENCODE data, Nature 489 (7414) (2012) 91–100.
[12] M. Wu, Z. Lin, S. Ma, T. Chen, R. Jiang, W.H. Wong, Simultaneous inference of

phenotype-associated genes and relevant tissues from GWAS data via Bayesian integration of multiple tissue-specific gene networks, J. Mol. Cell. Biol. 9 (6) (2017) 436–452.

[13] R. Jiang, Walking on multiple disease-gene networks to prioritize candidate genes, J. Mol. Cell Biol. 7 (3) (2015) 214–230.

[14] M. Wu, T. Chen, R. Jiang, Global inference of disease-causing single nucleotide variants from exome sequencing data, BMC Bioinform. 17 (17) (2016) 468.

[15] J. Wu, Y. Li, R. Jiang, Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies, PLoS Genetics 10 (3) (2014) e1004237.

[16] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 2014, ACM, 2014, pp. 701–710.

[17] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining: 2016, ACM, 2016, pp. 855–864.

[18] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Machine Intell. 35 (8) (2013) 1798–1828.

[19] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Springer series in statistics New York, 2001.

[20] D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, S. Bergmann, Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics, PLoS Comput. Biol. 12 (1) (2016) e1004714.

[21] Consortium GP, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (7422) (2012) 56–65.

[22] R.B. Davies, The distribution of a linear combination of x2 random variables, Appl. Stat. 29 (3) (1980) 323–333.

[23] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inform. Process. Syst. 2013 (2013) 3111–3119.

[24] D. Chung, C. Yang, C. Li, J. Gelernter, H. Zhao, GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation, PLoS Genetics 10 (11) (2014) e1004787.

[25] Y. Li, M. Kellis, Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases, Nucleic Acids Res. 44 (18) (2016) e144.

[26] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[27] R.J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, Ensembl BioMarts: a hub for data retrieval across taxonomic space, Database 2011 (2011).

[28] J. Simon-Sanchez, C. Schulte, J.M. Bras, M. Sharma, J.R. Gibbs, D. Berg, C. Paisan-Ruiz, P. Lichtner, S.W. Scholz, D.G. Hernandez, Genome-wide association study reveals genetic risk underlying Parkinson's disease, Nat. Genetics 41 (12) (2009) 1308–1312.

[29] E.A. Stahl, S. Raychaudhuri, E.F. Remmers, G. Xie, S. Eyre, B.P. Thomson, Y. Li, F.A. Kurreeman, A. Zhernakova, A. Hinks, Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci, Nat. Genetics 42 (6) (2010) 508–514.

[30] A. Franke, D.P. McGovern, J.C. Barrett, K. Wang, G.L. Radford-Smith, T. Ahmad, C.W. Lees, T. Balschun, J. Lee, R. Roberts, Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci, Nat. Genetics 42 (12) (2010) 1118–1125.

[31] C.A. Anderson, G. Boucher, C.W. Lees, A. Franke, M. D'Amato, K.D. Taylor, J.C. Lee, P. Goyette, M. Imielinski, A. Latiano, Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47, Nat. Genetics 43 (3) (2011) 246–252.

[32] H. Schunkert, I.R. König, S. Kathiresan, M.P. Reilly, T.L. Assimes, H. Holm, M. Preuss, A.F. Stewart, M. Barbalic, C. Gieger, Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease, Nat. Genetics 43 (4) (2011) 333.

[33] A.P. Morris, B.F. Voight, T.M. Teslovich, T. Ferreira, A.V. Segre, V. Steinthorsdottir, R.J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes, Nat. Genetics 44 (9) (2012) 981.

[34] J.E. Shim, C. Bang, S. Yang, T. Lee, S. Hwang, C.Y. Kim, U.M. Singh-Blom, E.M. Marcotte, I. Lee, GWAB: a web server for the network-based boosting of human genome-wide association data, Nucleic Acids Res. (2017).

[35] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L.I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. 45 (D1) (2017) D833–D839.

[36] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (suppl 1) (2005) D514–D517.

[37] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, The genetic association database, Nat. Genetics 36 (5) (2004) 431–432.

[38] C. Abraham, J.H. Cho, Inflammatory bowel disease, N. Engl. J. Med. 361 (21) (2009) 2066–2078.

[39] K. Ray, IBD: A role for GATA3 in ulcerative colitis, Nat. Rev. Gastroenterol. Hepatol. 13 (11) (2016) 624.

[40] S. Schreiber, P. Rosenstiel, J. Hampe, S. Nikolaus, B. Groessner, A. Schottelius, T. Kühbacher, J. Hämling, U. Fölsch, D. Seegert, Activation of signal transducer and activator of transcription (STAT) 1 in human chronic inflammatory bowel disease, Gut 51 (3) (2002) 379–385.

[41] J.K. Nieminen, M. Niemi, T. Sipponen, H.M. Salo, P. Klemetti, M. Färkkilä, J. Vakkila, O. Vaarala, Dendritic cells from Crohn's disease patients show aberrant STAT1 and STAT3 signaling, PloS One 8 (8) (2013) e70738.

[42] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, R. Herwig, ConsensusPathDB: toward a more complete picture of cell biology, Nucleic Acids Res. 39 (suppl_1) (2010) D712–D717.

[43] Z.-J. Liu, P.K. Yadav, J.-L. Su, J.-S. Wang, K. Fei, Potential role of Th17 cells in the pathogenesis of inflammatory bowel disease, World J. Gastroenterol. WJG 15 (46) (2009) 5784.

[44] R.C. Langan, P.B. Gotsch, M.A. Krafczyk, D.D. Skillinge, Ulcerative colitis: diagnosis and treatment, Am. Family Phys. 76 (9) (2007).

[45] S.G. Byars, Q.Q. Huang, L.-A. Gray, A. Bakshi, S. Ripatti, G. Abraham, S.C. Stearns, M. Inouye, Genetic loci associated with coronary artery disease harbor evidence of selection and antagonistic pleiotropy, PLoS Genetics 13 (6) (2017) e1006328.

[46] G.H. Dahlen, J.R. Guyton, M. Attar, J.A. Farmer, J.A. Kautz, A. Gotto, Association of levels of lipoprotein Lp (a), plasma lipids, and other lipoproteins with coronary artery disease documented by angiography, Circulation 74 (4) (1986) 758–765.

[47] X. Min, W. Zeng, N. Chen, T. Chen, R. Jiang, Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding, Bioinformatics 33 (14) (2017) i92–i101.

[48] Q. Liu, F. Xia, Q. Yin, R. Jiang, Chromatin accessibility prediction via a hybrid deep convolutional neural network, Bioinformatics (2017).

[49] W. Li, M. Wang, J. Sun, Y. Wang, R. Jiang, Gene co-opening network deciphers gene functional relationships, Mol. BioSystems 13 (11) (2017) 2428–2439.

[50] R. Jiang, M. Gan, P. He, Constructing a gene semantic similarity network for the inference of disease genes, BMC Systems Biology: 2011, BioMed Central, 2011, p. S2.

[51] Z. Duren, X. Chen, R. Jiang, Y. Wang, W.H. Wong, Modeling gene regulation from paired expression and chromatin accessibility data, Proc. Natl. Acad. Sci. (2017) 201704553.

[52] N. Solovieff, C. Cotsapas, P.H. Lee, S.M. Purcell, J.W. Smoller, Pleiotropy in complex traits: challenges and strategies, Nat. Rev. Genet. 14 (7) (2013) 483–495.