OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–12
https://doi.org/10.1093/bib/bbac417
Advance access publication date 8 October 2022
Problem Solving Protocol

NetTDP: permutation-based true discovery proportions for differential co-expression network analysis

Menglan Cai, Anna Vesely, Xu Chen, Limin Li and Jelle J. Goeman

Corresponding authors: Limin Li,E-mail: liminli@mail.xjtu.edu.cn and Jelle J. Goeman, E-mail: J.J.Goeman@lumc.nl

Abstract

Existing methods for differential network analysis could only infer whether two networks of interest have differences between two groups of samples, but could not quantify and localize network differences. In this work, a novel method, permutation-based Network True Discovery Proportions (NetTDP), is proposed to quantify the number of edges (correlations) or nodes (genes) for which the co-expression networks are different. In the NetTDP method, we propose an *edge-level* statistic and a *node-level* statistic, and detect true discoveries of edges and nodes in the sense of differential co-expression network, respectively, by the permutation-based *sumSome* method. Furthermore, the NetTDP method could further localize the differences by inferring the TDPs for edge or gene subsets of interest, which can be selected post hoc. Our NetTDP method allows inference on data-driven modules or biology-driven gene sets, and remains valid even when these sub-networks are optimized using the same data. Experimental results on both simulation data sets and five real data sets show the effectiveness of the proposed method in inferring the quantification and localization of differential co-expression networks. The R code is available at https://github.com/LiminLi-xjtu/NetTDP.

Keywords: differential co-expression networks, difference quantification and localization, true discovery proportion

Introduction

Networks can characterize interaction patterns of components in complex systems and explain observed phenomena in various fields. There has been much recent attention to gene correlation networks in biological systems, which have given valuable insights into biological phenomena. With the support of next-generation sequencing technology, the measurement of high-dimensional biological data has become cheaper and more efficient, allowing large-scale networks that bring computational challenges to network analysis. Weighted gene co-expression network analysis (WGCNA) [1, 2] is a widely used network estimation method. In biological network studies, WGCNA aims to cluster thousands of genes into several synthetic groups (or modules) of densely interconnected genes. In WGCNA a pair of genes is connected in the network if their expression is correlated. Networks can be both unweighted and weighted, using correlation values as weights. Correlation distance is then used to group genes into modules, potentially reducing a large number of genes into a small number of clusters. The expression of gene clusters can be quantified by their eigengenes, defined as the first principal component of the expression values of the cluster. Highly interconnected modules can help to understand common biological processes. WGCNA has been successfully used in various computational biology problems [3–7].

Time [8], external stimuli [9] or disease status variety [10] can lead to dynamic changes in the interactions among components in biological systems. For example, the onset and progression of ovarian tumors is regulated by the network of platinum resistance-related genes [11], protein interaction networks are associated with cell proliferation in cancer cells [12] and lipid networks have been successfully used as predictive biomarkers for chronic kidney disease [13]. There is growing evidence showing that gene networks can change over time or under external stimuli, motivating the study of complex diseases from the perspective of differential networks. Differential network analysis [14] is therefore the task of comparing networks between different states and/or over time, revealing in what way these external factors affect the behavior of the system.

WGCNA has been applied to compare network topology of different co-expression networks via consensus module detection [2]. For instance, the DiffCoEx [15] method applies WGCNA to identify co-expression network differences by clustering genes into modules based on the matrix of gene correlation differences. Coxpress [16] firstly detects co-expression gene modules within one group samples and then calculates t-statistic to test whether one group of genes is differentially co-expressed. Yuan *et al.* [17] used biweight midcorrelation to measure gene similarity for gene differential co-expression analysis. GSCA[18] identified gene sets showing distinct correlation profiles across conditions by evaluating pairwise co-expression, as opposed to gene-specific expression, across a gene set. GSNCA[19] identified differentially co-expressed gene sets, which have to be defined a priori, between two sample groups. It defined weight vectors from a correlation network for each sample group. CoGA[20] is an R package for the identification of groups of differentially associated genes between two phenotypes. It utilized concepts of Information Theory applied to the spectral distributions of the gene co-expression graphs. For more differential co-expression network methods,

Menglan Cai is currently working toward the PhD degree in Xi'an Jiaotong University, China. Her research interest includes bioinformatics.
Anna Vesely is a postdoctoral fellow in University of Padova, Italy. Her research interest includes biostatistics.
Xu Chen is a researcher of Leiden University Medical Center, Netherland. Her research interest includes biostatistics.
Limin Li is a professor of Xi'an Jiaotong University, China. Her research interests include machine learning and the applications in bioinformatics and biostatistics.
Jelle J. Goeman is a professor of Leiden University Medical Center, Netherland. His research interest includes biostatistics.
Received: May 17, 2022. **Revised:** August 23, 2022. **Accepted:** August 28, 2022

please see [21]. However, existing methods fail to infer three aspects: (1) whether any difference of the gene co-expression network between groups is statistically significant; (2) how large the difference is, in terms of nodes or edges; (3) where such significant differences of the gene co-expression network are located. Quantifying and localizing the gene co-expression network differences between two groups is especially important for large gene co-expression networks, since the mere existence of a difference in a large network is not very informative. To localize the difference, we are interested in one or several selected gene co-expression sub-networks, which are defined by subsets within the gene set of interest. To correct for the multiplicity of sub-networks, a multiple testing procedure is necessary [22, 23]. This multiple testing problem becomes even more challenging if candidate sub-networks are selected post hoc, and standard multiple testing procedures cannot be directly used.

In this work, we present a novel multiple testing method, named Network True Discovery Proportions (NetTDP). Our method allows quantification and localization of co-expression modules, and is valid for modules chosen in a data-driven way. We quantify differential co-expression by defining TDP for sub-networks as the proportion of edges or nodes in the sub-network that is truly differentially co-expressed. For any selected sub-network, we find a lower $(1-\alpha)$-confidence bound for the TDP based on the statistics of the differential co-expression network. We say that network differences exist between two groups in a sub-network if the TDP lower confidence bound is non-zero, though biologically informative differences should require larger TDP values. Data-driven (post-hoc) selection of sub-networks is allowed because the confidence bounds we provide are simultaneous over all sub-networks [24, 25]. Simultaneity allows researchers to specify the subset of interest after seeing the data, while still obtaining valid confidence bounds for the TDP.

Network statistics tend to be highly correlated with each other, and analysis methods should take such correlations into account. For differential network analysis we use a permutation-based procedure to adapt to this correlation structure. In particular, we use the *sumSome* method, recently proposed by Vesely *et al.* [26], which implements the closed testing framework of [24] and [25] using sum-based permutation tests. It allows quick construction of simultaneous lower confidence bounds for the TDP of sub-networks.

The structure of the work is as follows. We first introduce WGCNA and the *sumSome* method that our novel procedure builds on before introducing the proposed method. The novel method has two variants, one for a *node-level* and one for an *edge-level* analysis. A simulation study illustrates the proper error control of the method and its power. We apply the method on five data sets to demonstrate its usefulness in practice.

Related work

In this section, we will briefly introduce two related methods, including WGCNA [2] and the *sumSome* [26] methods, since in our proposed NetTDP method, WGCNA is used for network construction and module detection, and the sumSome for TDP inference.

Network construction and module detection with WGCNA [2]

WGCNA consists of two main steps: network construction and module detection.

The nodes in the network correspond to genes. The connection strength of two genes is calculated by the similarity between them. By default, similarity between gene i and gene j is measured using their correlation. The element a_{ij} of the adjacency matrix is computed by raising the correlation absolute value to a power β:

$$a_{ij} = |\mathrm{cor}(x_i, x_j)|^{\beta}, \tag{1}$$

where $\beta = 6$ is the default value. An edge is considered to exist between genes i and j if the adjacency exceeds a threshold value. By this way WGCNA constructs large-scale networks.

Next, WGCNA defines gene modules loosely as subgraphs in which the genes are densely connected. Standard methods used to identify gene modules include hierarchical clustering based on the adjacency matrix, combined with branch cutting. For each module, the first principal component of the corresponding expression matrix is defined as the module eigengene, and genes can be reassigned to a different module if they correlate well with that module's eigengenes. Gene modules with highly correlated eigengenes can be further merged to eliminate smaller modules.

True Discovery Proportions inference with the *sumSome* [26]

The permutation-based method *sumSome* is a general closed testing procedure for sum tests, which allows to construct lower $(1-\alpha)$-confidence bounds for the TDPs simultaneously over all subsets of a family of hypotheses. The method is suitable for exploratory research, as it controls the TDP even when the subset of interest is selected after seeing the data. Furthermore, it adapts to the unknown joint distribution of the data through permutation testing. The goal and assumptions of the *sumSome* method can be formulated mathematically as follows.

Let $H_1, \ldots, H_m$ be m univariate hypotheses of interest, indexed by the set $N = \{1, 2, \ldots, m\}$. The unknown true hypotheses are collected in the subset $T \subseteq N$. Given any subset $K \subseteq N$, the corresponding intersection hypothesis

$$H_K = \bigcap_{i \in K} H_i$$

is true if and only if H_i is true for all $i \in K$, i.e. $K \subseteq T$. Let

$$S_K = \sum_{i \in K} S_i$$

be a sum test statistic for H_K, where S_i is a test statistic for the univariate hypothesis H_i. A critical value of S_K can be found using permutations, under the assumption that the joint distribution of the statistics for the true hypotheses is invariant under all permutations of data.

Subsequently, let $\mathrm{TDP}(K) = |K \setminus T|/|K|$, taken as 0 if $K = \emptyset$, be the unknown proportion of true discoveries in subset K, where $|K \setminus T|$ represents the number of elements in subset K but not in subset T. Goeman *et al.* [25] proposed a procedure to define simultaneous lower $(1-\alpha)$-confidence bounds $d(K)$ for each $\mathrm{TDP}(K)$, so that

$$P\big(\mathrm{TDP}(K) \geq d(K) \text{ for all} K \subseteq N\big) \geq 1 - \alpha.$$

In this way, $[d(K), 1]$ is a simultaneous $(1-\alpha)$-confidence interval for $\mathrm{TDP}(K)$. In general the confidence bounds $d(K)$ can be computed using closed testing, based on the idea of testing many subsets of hypotheses by means of a valid α-level test, which in this case is the permutation test. However, if used as it is the procedure has exponential complexity in the number m of hypotheses, and becomes infeasible when m is large [25].

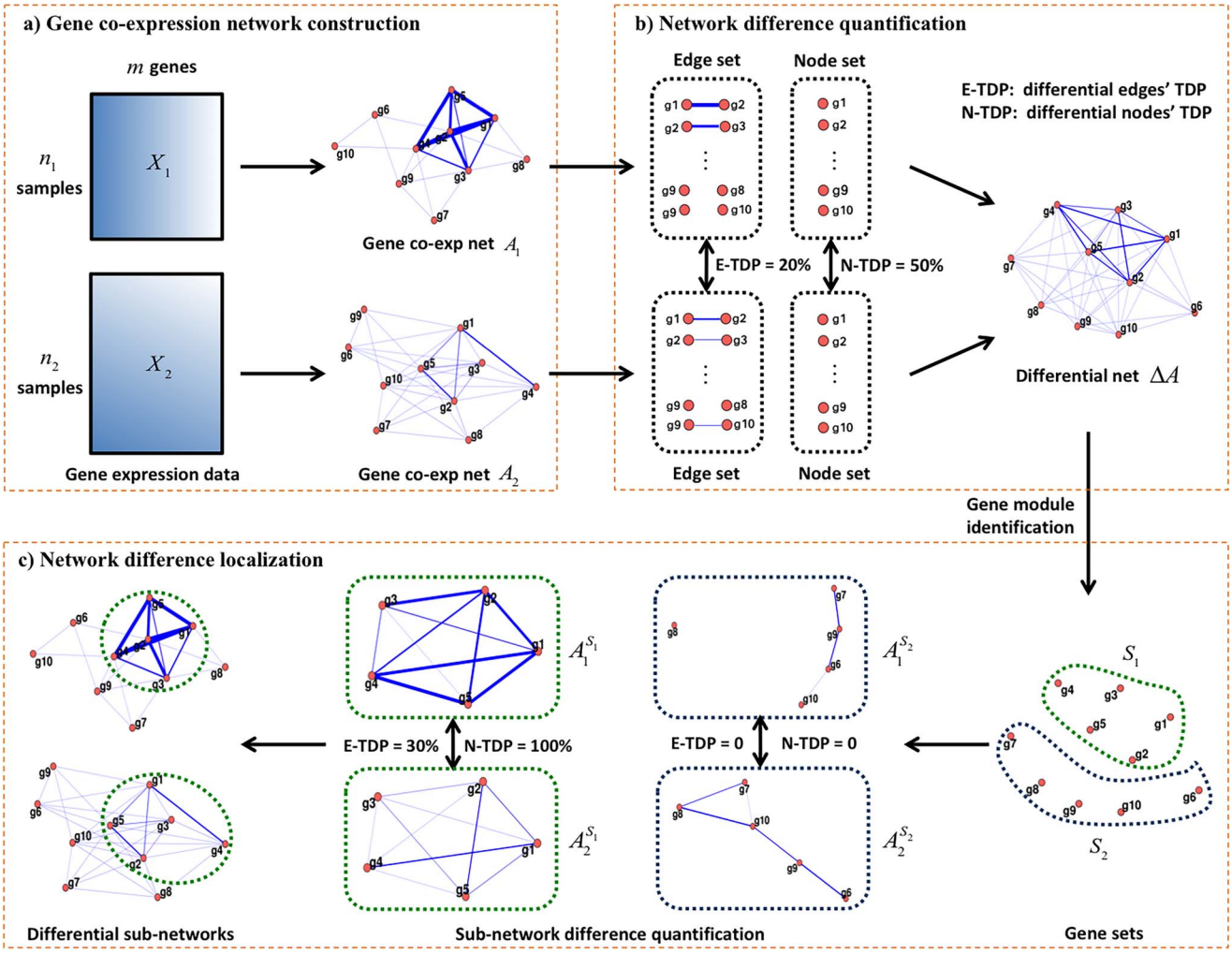

Figure 1. Flowchart of NetTDP method for differential co-expression network analysis. There are three main parts in our NetTDP, including network construction, network difference quantification and network difference localization, shown in **(a)**, **(b)** and **(c)**, respectively. **(a)** Gene co-expression network construction: given expression data for two groups of samples with the same m genes, one can construct weighted co-expression networks A_1 and A_2 based on correlation coefficients. **(b)** Network difference quantification: we quantify network difference by detecting true discoveries of edges or nodes in the sets of interest. *Edge-level* and *node-level* statistics are proposed and permutation-based *sumSome* method is used for true discovery inference. **(c)** Network difference localization: we further localize network difference by inferring the TDPs in gene sets of interest which can be selected after seeing the data. Subsets with non-zero TDPs can be identified as differential sub-networks in a network. Post hoc inference on data-driven modules and biology-driven gene sets is allowed in our proposed method.

The *sumSome* is a fast algorithm that allows quick calculation of the confidence bounds for TDP(K) by approximating the value $d(K)$ from below. For large size problems, the method can be sped up using truncation-based statistics, i.e. cutting down the statistics smaller than a given threshold. The method is suitable for testing problems up to about 10^6 hypotheses. For details, see [26].

Methods

In this section, we propose the NetTDP method for differential co-expression network analysis, which includes three steps: co-expression network construction, network difference quantification and network difference localization. The flowchart of the proposed method NetTDP is shown in Figure 1.

Gene co-expression Network construction

In differential gene co-expression network analysis, we have expression data for the same m genes in two groups of samples. The expression data are represented by $X_1 = \{x_{1,1}, x_{1,2}, \cdots, x_{1,n_1}\} \in \mathcal{R}^{m\times n_1}$ and $X_2 = \{x_{2,1}, x_{2,2}, \cdots, x_{2,n_2}\} \in \mathcal{R}^{m\times n_2}$, with n_1 and n_2 samples in two groups, respectively.

We will assume that the columns $x_{1,1}, \ldots, x_{1,n_1}$ are independently drawn from a zero mean normal distribution with unknown $m \times m$ correlation matrix Σ_1. Analogously, we assume that $x_{2,1}, \ldots, x_{2,n_2}$ are independently drawn from a zero mean normal distribution with unknown $m \times m$ correlation matrix Σ_2. If the means of given expression data are not zero, i.e. there is differential expression between the groups, we subtract the mean per group to fulfill this assumption asymptotically. For all pairs $i, j \in \{1, 2, \ldots, m\}$, let $\rho_{1,ij}$ and $\rho_{2,ij}$ be the true correlation coefficients between genes i and j in groups 1 and 2, and $\hat{\rho}_{1,ij}$ and $\hat{\rho}_{2,ij}$ be their estimates.

From these expression data, we can use WGCNA to obtain two estimated networks, G_1 and G_2, with the same m nodes (i.e. genes). The corresponding adjacency matrices are $A_1 = [a_{1,ij}] = [(\hat{\rho}_{1,ij})^\beta]$ and $A_2 = [a_{2,ij}] = [(\hat{\rho}_{2,ij})^\beta]$, with $i, j \in \{1, 2, \cdots, m\}$.

In this work, our first goal is to test whether the co-expression network is significantly differential between groups. Secondly, we

want to quantify the network difference, and in the third place to localize the difference in the network.

Network difference quantification

For the first and second goals, we find the lower $(1-\alpha)$-confidence bound of the TDP of the full network. This quantifies the network difference, and if the lower bound is positive, indicates the presence of a significant difference. For the third goal, to further identify locations of the difference, we find the TDP in sub-networks and modules of interest.

The TDP may be defined either at the node or at the edge level. The *edge-level* null hypothesis for the edge between node i and node j, $i \neq j$, may be defined as

$$H_{ij} : \rho_{1,ij} = \rho_{2,ij},$$

i.e. the co-expression pattern between genes i and j is not different between the two groups. A sensible test statistic to test this hypothesis is

$$T_{ij} = \begin{cases} |a_{1,ij} - a_{2,ij}| & \text{if } \text{sign}(\hat{\rho}_{1,ij}) = \text{sign}(\hat{\rho}_{2,ij}) \\ |a_{1,ij} + a_{2,ij}| & \text{otherwise} \end{cases}. \quad (2)$$

We can use these null hypotheses and test statistics directly in *sumSome*. We define a permutation of the data as a random reallocation of the $n_1 + n_2$ samples over the two groups, and remarking that under our model the joint distribution of the null genes is invariant to such a permutation. If the true means are not zero, the invariance is asymptotic. Alternatively, we can use rotations instead of permutations [27, 28].

The use of *sumSome* controls the *edge-level TDP*, i.e. the proportion of correcty identified differential edges. The number of hypotheses in this testing problem is $m(m-1)/2$.

Alternatively, we can look at the *node-level TDP*. We can define the intersection hypothesis of all *edge-level* hypotheses corresponding to all edges involving node i, as follows

$$H_i = \bigcap_{j \neq i} H_{ij} : \rho_{1,ij} = \rho_{2,ij}.$$

This hypothesis is false if the correlation between gene i and at least one other gene is different between the groups, i.e. if the node is differentially connected between the two groups. A suitable test statistic for this formulation of the problem is

$$T_i = \sum_{j \neq i} T_{ij}. \quad (3)$$

We can also use *sumSome* based on this choice of hypotheses and test statistics, since the same permutations apply. This will give lower confidence bounds to the *node-level TDP*, the proportion of correctly identified differentially connected nodes. The number of hypotheses in this testing problem is m.

Note that our NetTDP is not a simple application of *sumSome* method. It provides a novel procedure to quantify and localize the differences between two networks, and the *sumSome* test is only one step in the procedure. Furthermore, we proposed novel edge-level and node-level statistics for two networks, to calculate differences in genes or correlation of gene co-expression networks. Based on the two statistics, the TDP of node (gene) or edge (correlation) in the subset of interest can be inferred by *sumSome*.

Network difference localization

For detection and quantification of the network differences we will simply look at the TDP for the entire network. For localization, however, we need to look at potentially interesting subnetworks. Depending on whether a *node-level* or *edge-level* analysis is done, we need to look at subsets of the nodes or subsets of the edges.

Many strategies can be used to detect node subsets, i.e. gene sets, of interest. For example, one can collect gene sets from the existing gene set databases, e.g. Gene Ontology (GO) categories or Kyoto Encyclopedia of Genes and Genomes (KEGG). Alternatively, data-driven methods can be adopted to detect gene clusters. A natural choice is WGCNA's module detection method, identifying gene clusters for each group alone with the calculated adjacency matrix. Inspired by this method, we input the difference of adjacency matrices over two groups instead, referred to as ΔA, with entries formulated as T_{ij}. Then we use hierarchical clustering to identify gene modules with the network difference matrix ΔA.

Edge subsets can be defined from node subsets in two ways. One is collecting edges between nodes that appear in that node subset. The other is considering edges between nodes in that node subset and any other nodes.

Note that our NetTDP is a further analysis tool based on the WGCNA method, but not a competitor, since the WGCNA method is to analyze single-gene co-expression networks (to identify gene modules) or to identify gene modules shared between two gene co-expression networks, while our NetTDP is proposed for quantifying and localizing differences in the co-expression networks generated by two groups of samples.

Simulation study

In this section, Monte Carlo simulation is used to examine the performance of our proposed method. Different co-expression networks between two groups, characterized by having different gene correlations or including different number of differential genes, are constructed.

Data generation and experimental settings

A total of 100 genes and 50 samples from each of the two groups are considered. We also consider selecting subsets K from the 100 genes, with the size varying within $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$.

For the first group, the data are simulated from a standard multivariate normal distribution $\text{MVN}(0, \Sigma(\eta, \rho))$, where the first η genes are correlated to each other with an equal correlation ρ, while the other genes are assumed to be mutually independent. Here, η varies between 10, 50 and 90, and ρ is selected within $\{0.2, 0.3, 0.4, 0.5\}$. The simulated data for the second group also follow a multivariate normal distribution $\text{MVN}(0, \Sigma(\eta, w\rho))$, where w is within $\{0, 0.3, 0.6, 0.9, 2\}$. Since zero mean is assumed for both groups, we do not need to subtract per-group means to guarantee that the identified network differences are unaffected by differential gene expression.

The η co-expressed genes indicates all are differentially co-expressed between the two groups. For the gene subsets of interest we used the following selection rule: if $|K| \leq \eta$, we randomly select $|K|$ genes from η dependent genes; if $|K| > \eta$, $|K| - \eta$ genes are randomly picked from the independent genes in addition to the η dependent genes. For instance, for a chosen subset of 50 genes, we first include the $\eta = 10$ genes, and then randomly choose 40 genes from the remaining 90 independent genes. This defines the

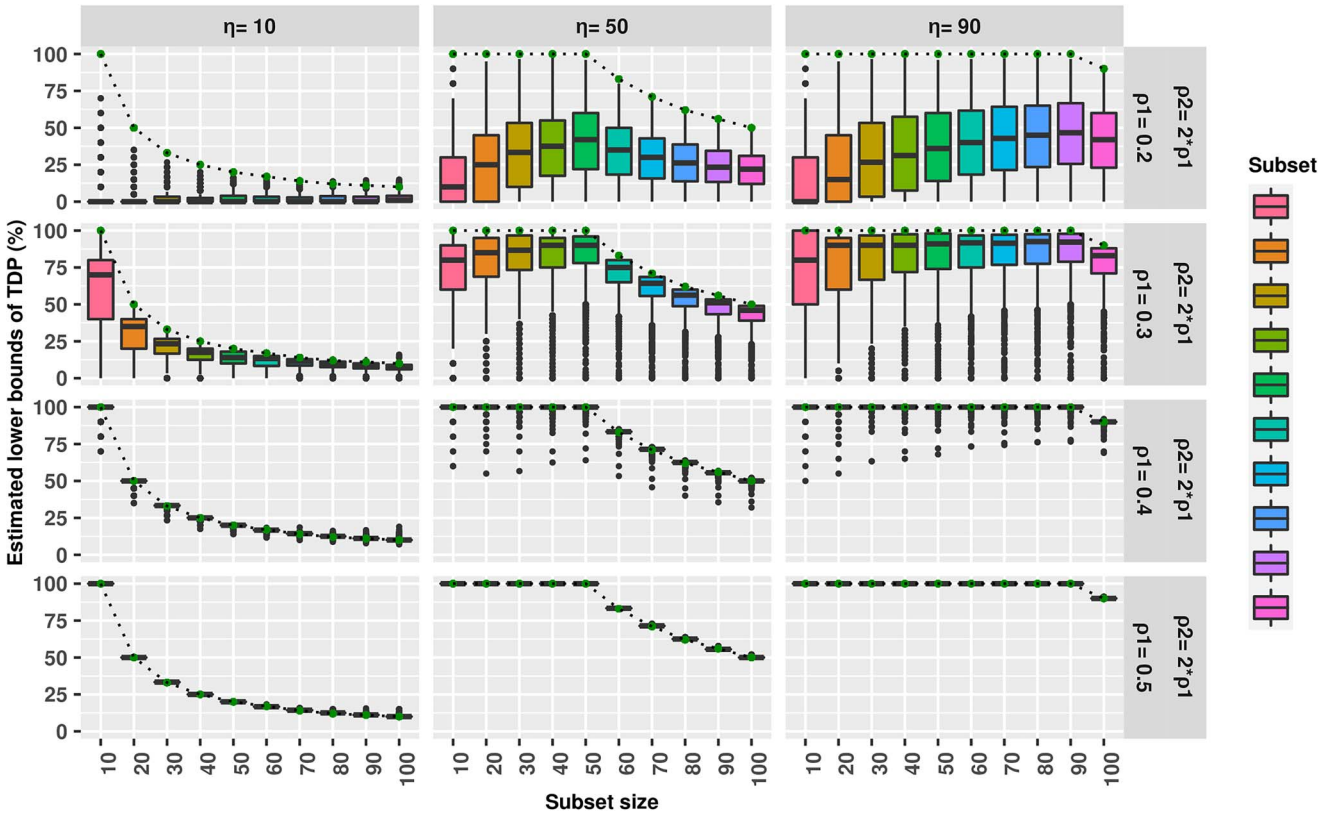

Figure 2. NetTDP *node-level* results for 1000 simulations with $w = 2$. The box plots are estimated TDP bounds. The green dots mark the ground-truth TDPs in each gene set, and the dotted lines depict the changes of these TDPs. Different gene sets are presented in different colors. The x-axis provides the number of genes in the chosen gene set. Each row corresponds to results for each setting of ρ that measures the correlation strength among differentially related gene pairs. Each column contains results for each η setting measuring how many genes are differentially related.

node-level subsets. For the *edge-level* sets, we tested edges collected by the first way, as described in Section *Methods*.

We conducted 1000 simulations, and the significance level is specified as 0.05 and for each simulation we performed 200 permutations, as advised in [26].

Results on node-level inference

The *node-level* estimated TDP bounds are shown in Figure 2, where w is fixed as 2. Each row corresponds to the results obtained for each ρ setting, and each column for each η value. The green dots are the true TDP in each gene set of interest. The dotted lines depict trends in true TDPs across gene sets. It can be observed that the estimated TDPs are all lower than the dotted line, that is, the estimated results are all less than or equal to the ground-truth TDPs. This shows that the proposed method of NetTDP can control false positives well. In other words, while the gene set is defined after seeing the data, NetTDP can also control the probability of making type I error below a given significance level α.

We observed that the signal strength of co-expression network differences increases with the increased number of related genes or enhanced correlation between genes. That is, the larger η and ρ are, the higher the estimated TDP bounds would be. We also observed that when all genes in a gene set are network-differentiated, the estimated TDP bound increases with the size of the gene set. In other words, when a network differs everywhere between two groups, increasing the chosen gene set leads to a higher TDP confidence bound.

Secondly, simulation results for the other w settings are shown in Figures S1, S3, S5 and S7 in the Supplementary Material. As the value of parameter w increases from 0.3 to 0.9, one can see a decreasing trend in the TDP estimated by our method for the gene set of interest. This is because as parameter $w \in [0, 1]$ increases, the covariance matrix difference between groups shrinks, resulting in a smaller network difference between groups. In extreme cases, such as when $w = 0.9$, the network differences between groups are very difficult to distinguish, and the TDP estimated by the proposed method of NetTDP is low. We also check the power of NetTDP in the case where genes in one group are independent of each other, while some genes in the other group are related. The results of Figure S7 show that the performance of NetTDP for this case is between the performance in the cases of $w = 2$ and $w = 0.3$.

Results on edge-level inference

Similar to *node-level* inference, we also reported corresponding simulation results for edges in Figure 3 and Figures S2, S4, S6 and S8 in the Supplementary Material. In these figures, the dotted lines with green points show ground-truth TDPs for edges within gene sets. At edge level we see qualitatively the same conclusions as at node level. In most cases, we obtain lower TDPs at the edge level than at the node level (see Figures 2 and 3). This can be explained by the statistics defined here. The *node-level* statistic, defined in equation (3.2), integrates all information of genes co-expressed with one gene, while the *edge-level* statistics, defined in equation (1), looks into each pair of co-expressed genes separately. That is to say the *node-level* statistic makes it easier to identify genes with differences by gathering more edge differences, while the *edge-level* statistic value depends on the information of one edge, which greatly increases the difficulty of identification.

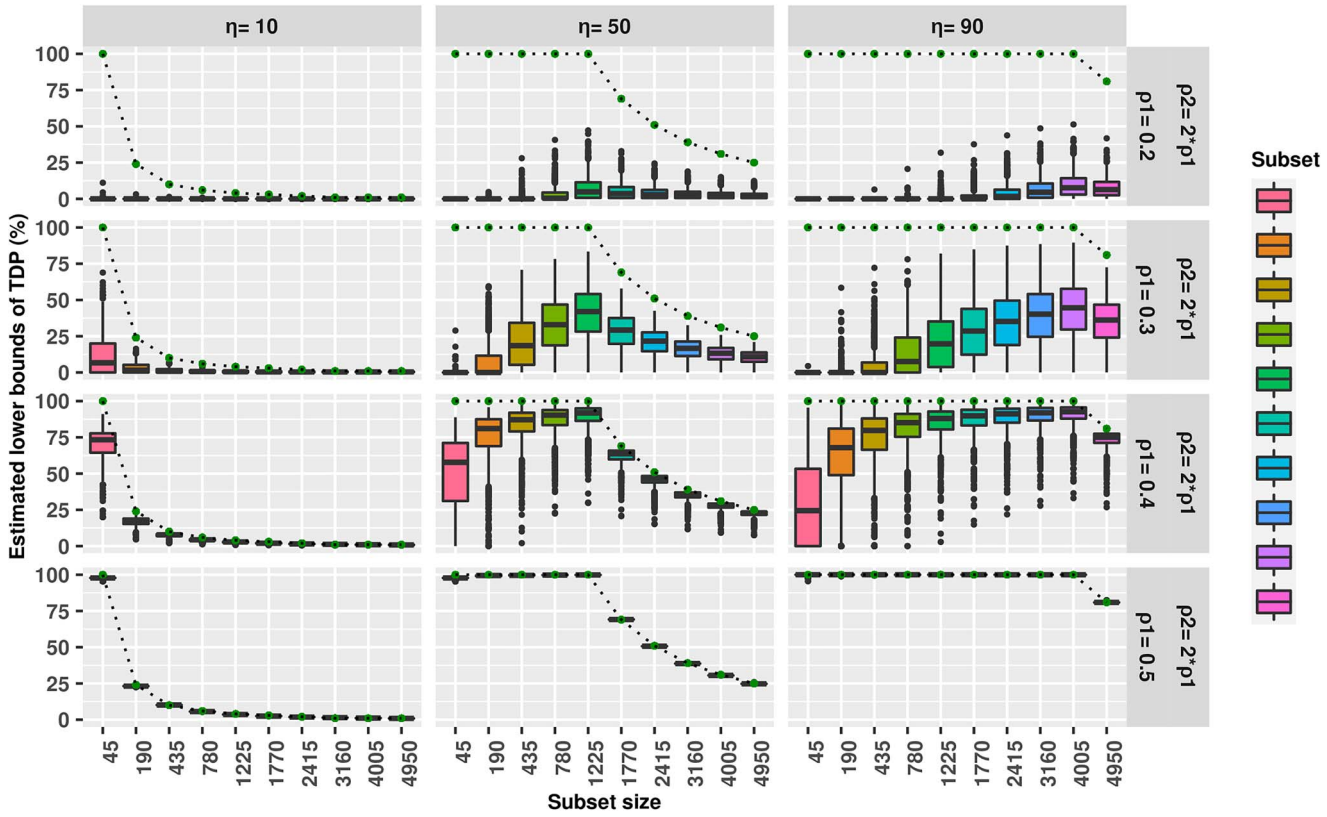

Figure 3. NetTDP *edge-level* results for simulation data with $w = 2$. The box plots are estimated TDP bounds for 1000 simulations. The green dots mark the ground-truth TDPs in each gene set, and the dotted lines depict the changes of these TDPs. Different gene sets are presented in different colors. The x-axis provides the number of genes in the chosen gene set. Each row corresponds to results for each setting of ρ that measures the correlation strength among differentially related gene pairs. Each column contains results for each η setting measuring how many genes are differentially related.

Real data study

Data collection and processing

This section focuses on the collection of real data, including gene expression data and gene sets of interest, and also describes data preprocessing.

Gene expression data collection. To evaluate our method of NetTDP with real biological data, we collected three human disease gene-expression data sets. These data sets were also applied in [29] to detect differentially co-expressed gene sets through a novel probabilistic score. The details of these data sets are shown in Table 1. As described in [29], only 3000 probes with maximum variation are firstly selected in each data set. In this way, they could cut down noise and focus on genes varying across the study. Then the probes are mapped to Entrez IDs to merge them further.

We also collected two mouse single-cell data sets from brain and embryo stem cells, shown in Table 1. Mouse brain data [30] have seven cell types, including astrocytes, endothelial, ependymal, microglia, neurons, oligos and vsm. Mouse embryo stem cells [31] are grouped into four stages by differentiation days, including day0, day2, day4 and day7, where day0 means embryo stem cells have not yet started to differentiate and day2 represents 2 days after cell differentiation.

Gene sets collection. We also collected 25 891 human gene sets downloaded through http://baderlab.org/GeneSets. It contains all pathway resources and all three divisions of GO (biological process, molecular function, cellular component) excluding annotations that have evidence code IEA (inferred from electronic annotation), ND (no biological data available) and RCA (inferred from reviewed computational analysis). Then we made an intersection of these gene sets and genes in the expression data. To reduce the number of gene subsets of interest in our experiments, we focused on sets with a size of more than 30. Consequently, we obtained 1537, 1464 and 1691 gene subsets for the AD, LC and NDD data sets, respectively. In addition to the biology-derived gene sets, we also defined gene sets with a data-driven method, i.e. using the gene module detection described in Section *Methods* above. Results are reported in the following subsection, including two ways to construct gene sets of interest. Similarly, we tested edge sets defined by the same way in our simulation study.

Mean removal. This work focuses on the identification of genes with network differences, i.e. genes whose gene–gene relationships are differentially expressed between groups. Notice that these do not coincide with differentially expressed genes, which could have the same gene–gene relationships but different means between groups. In order to eliminate the interference of differential expression on the goal of this work and further ensure that the identified genes have network differences, we removed the mean in the gene expression data to ensure that the mean within the group is zero, while preserving the inter-gene relationship. One can also normalize data to have standard deviation and zero mean before using our method.

Single-cell data preprocessing. The pipeline for analyzing single-cell data is similar to bulk data in our method, except a data pre-processing step for single-cell data with high noise. In the preprocessing step, aggregated expression profiles of single-cell data are used to conduct 'pseudocells' from specific cell populations. This strategy was also used in the single-cell co-expression network analysis method in [32]. Specifically, we applied k-nearest neighbors to compute the mean expression from 10 neighboring

Table 1. Summary of five data sets used in real data study

Technology	Species	Data set	#Genes	#Samples	#Classes	GEO ID
Bulk RNA-seq	Human	Alzheimer's disease (AD)	2698	363	2	GSE15222
		Lung cancer (LC)	2890	187	2	GSE4115
		Neuro-degenerative disorders (NDD)	2375	118	2	GSE26927
Single cell RNA-seq	Mouse	Brain	24341	2881	7	GSE74672
		Embryo stem cells	24175	2717	4	GSE65525

cells. We further reduced the dimension of features with Seurat toolbox [32] by picking the top 2000 highly variable genes. After performing the above steps, we obtained 2000 features with 576 cells and 543 cells for mouse brain and mouse embryo stem cells, respectively. Note that we only reduced the number of cells, the cell types remained the same as the original single-cell data.

To take over the challenge of various cell types, in turn, we take one cell type as a case group and the other cell types as a control group, which is a common strategy in differentially expressed gene identification. In this way, we could identify specific sub-networks that distinguish one cell type from the other cell types.

Gene module detection methods

The identification of gene modules is a critical step in differential gene co-expression analysis. Its goal is to cluster genes on a gene co-expression network. To illustrate the robustness of our NetTDP method, we compared other three gene module identification methods, including DiffCoEx [15], Coxpress [16] and lmQCM [33].

DiffCoEx: similarly to WGCNA, DiffCoEx identified and groups differentially co-expressed genes that have different partners between different samples by hierarchical clustering with the following adjacency difference:

$$D : d_{ij} = \left(\sqrt{0.5|sign(\hat{\rho}_{1,ij}) * (\hat{\rho}_{1,ij})^2 - sign(\hat{\rho}_{2,ij}) * (\hat{\rho}_{2,ij})^2|}\right)^{\beta}$$

Coxpress: It identifies co-expression modules in each sample group and tests whether the genes within these modules are also co-expressed in other groups. Coxpress method uses hierarchical cluster analysis to explore the relationship between genes, cutting the tree to form groups of genes that are co-expressed.

lmQCM: unlike DiffCoEx and Coxpress, which use hierarchical clustering and do not allow overlap between modules, lmQCM is a greedy approach allowing genes to be shared among multiple modules, consistent with the fact that the genes often participate in multiple biological processes. Same to Coxpress, only one group of samples are used to detect gene modules in this method.

Note that NetTDP, DiffCoEx and Coxpress methods all use hierarchical clusters for gene module detection. The difference is that Coxpress only inputs a co-expression network from one group, such as G_1 or G_2, while NetTDP and DiffCoEx require a co-expression network difference between two different groups. In our experiment, for Coxpress and lmQCM, gene modules were identified by only inputting case samples of AD data set, that is, disease patients on AD.

Results

In this section, we will present the results of the proposed NetTDP method on five real datasets, including bulk RNA-seq and single-cell data, taking as subsets of interest both gene sets clustered on the data and biologically meaningful gene sets obtained from databases. In bulk RNA-seq data AD, we also compared gene modules identified by three other methods with modules detected by our NetTDP.

Performance on sets defined as modules. After obtaining the network adjacency matrix within the groups, we used the WGCNA package to identify gene modules. The input is the difference between the adjacency matrices corresponding to the two sets of samples. When using the WGCNA package to identify gene modules, different parameter selections will result in different gene clustering results, but our method of NetTDP remains valid even if many such clustering settings are tried. Here, we choose Ward's hierarchical agglomerative clustering [34] method in WGCNA and we set the size of the module to be no less than 30 for all three real data sets. Only the results of our method under one clustering are shown here.

Firstly, we reported the node-level and edge-level results for gene modules in Figure 4 and Figure S9 in Supplementary Material, respectively. For each data set, we present the results in a pie chart. Each sector presents the results for one gene module. The size of the sector corresponds to the number of genes in the module. The larger the sector, the larger the gene module. The color of the sector corresponds to the estimated TDP: the darker the color, the larger the TDP value. White indicates that the TDP is 0. The TDP and the number of genes across the entire data set are listed in the title of each subplot.

For node-level results in Figure 4, the global TDP is quite different between the three data sets. It exceeds 70% in AD, is close to 50% in LC and is less than 10% in NDD. The higher the global TDP, the easier it is to find locally differentiated networks with differences. In the gene module division of the AD data set, modules with a size of more than 200 genes have non-zero TDPs; the module with 382 genes has the highest TDP, although it is not the largest of all modules. For the LC data set, the estimated TDP for the largest gene module (48.48%) is the highest, and even higher than the global TDP (42.44%). It indicates that most genes with network differences are likely to be clustered in this module. Interestingly, in the NDD data set, the TDPs of the modules are all 0, which means that the local differences are not in these sub-networks. Another way of grouping genes may be able to find different local network structures.

For *edge-level* results in Figure S9 in Supplementary Material, we explored differences in edges between genes in each gene module. Compared with the node level, the global TDP of the three data sets is greatly reduced, and the LC data set has the highest TDP of 7.53%. In the module division discussed here, only the LC data set finds a set with a non-zero TDP.

Technically, we compared different module detection methods used in the framework of our NetTDP to show the robustness of

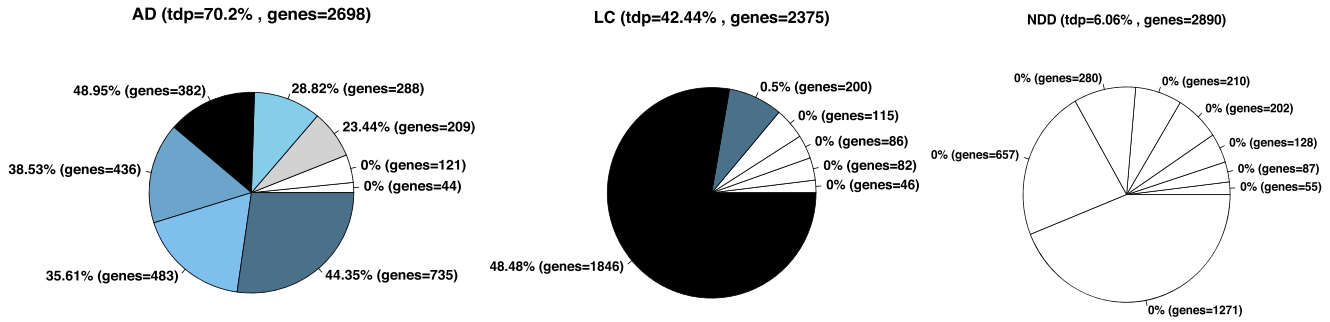

Figure 4. NetTDP *node-level* results for bulk RNA-seq data in gene modules. The TDP and the number of genes across the entire data set are listed in the title of each column. Each pie chart presents results for each data set.Each sector presents the results for one gene module. The size of the sector corresponds to the number of genes in the module: the larger the sector, the larger the gene module. The color of the sector corresponds to the estimated TDP: the darker the color, the larger the TDP value. White indicates that the TDP is 0.

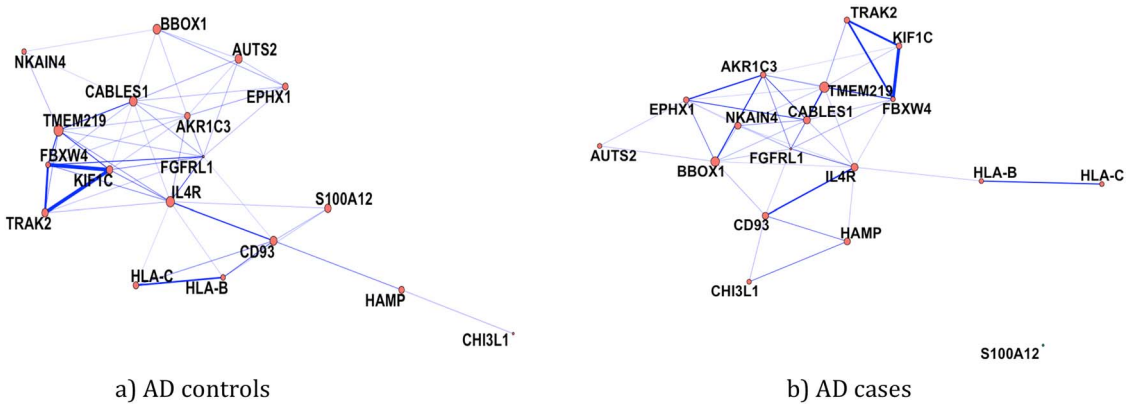

Figure 5. Co-expression networks for 18 shared genes identified by all gene module detection methods discussed here. **(a)** and **(b)** are networks for control and case samples in AD data set, respectively. Nodes size represents degree of node. Edge thickness represents weight scale: the thicker the edge, the stronger the correlation. These genes differ in degree of node or weight of edge. For example, correlation among FBXW4, KIF1C and TRAK2 genes are different between groups. And degree of S100A12 gene in cases is zero, significantly different from controls, where S100A12 are linked to two other genes.

our NetTDP in network difference quantification. We analyzed the consistency of NetTDP results using different gene modules detected by different module-based methods on AD data set. Specifically, we first inferred TDPs by our NetTDP for modules identified by each method, and further calculated true discoveries of node. Then for each method we selected the top module with the most node true discovery (223, 352, 326 and 618 genes, respectively, for NetTDP, DiffCoEx, Coxpress and lmQCM). We reported the Venn chart (shown in Figure S10) of these top modules, and over 200 genes are shared between the top modules, showing that our NetTDP could obtain consistent results by different module detection methods. The results further imply the robustness of our method with regard to module detection methods. Intuitively, in Figure 5, we visualized the co-expression network of AD and control samples, where all 18 genes share among all modules and many differences exist between these two networks (degree of node or weight of edge). This figure shows that our NetTDP method can indeed find differences between two co-expression networks. Based on observations above, we verified the robustness and rationality of our NetTDP method.

Next, in Figure S11, we reported *node-level* results of our NetTDP in gene modules for single-cell data sets, where sub-figures a) and b) present results for mouse brain data, and sub-figures c), d) and e) show results for mouse embryo stem cells data, respectively. For both these two single-cell data, their absent cell types had zero TDPs of the whole networks, that is, no statistically significant network difference existed between the two groups. It is likely that the co-expression network of the 2000 genes we selected is not different between groups. If more genes are analyzed here it may help us to discover network differences in these single-cell data. For cell types shown in this figure, we believe that their co-expression network difference could be found in current modules with the largest TDPs at least. In Figure 6 and Figure S12 in Supplementary Material, we also visualized the network differences found for specific cell types, including mouse brain neurons and mouse d4 embryo stem cells, respectively. In the detected gene module with the largest TDP value, we randomly selected 20 genes to exhibit co-expression networks constructed in our method. As we can see in these two figures, many obvious differences exist between the two compared co-expression networks, showing that our method can also handle differential gene co-expression network analysis on single-cell data well.

Performance on gene sets defined biologically. We also tested the efficacy of our method of NetTDP on biologically meaningful gene sets collected in Section *Data collection and processing*. Results are reported for *node-level* statistic in Figure 7 (left-hand side) and *edge-level* statistic in Figure S13 (left-hand side) in Supplementary Material. For both figures, the x-axis shows the size of gene sets and different colors represent different gene expression data sets. As we can see, the trends of all curves are all upwards, that is, the estimated TDP increases with the size of the gene set. This phenomenon is similar to a pattern observed in simulation experiments when the network differences between groups are small but widespread. This suggests that the three data sets could

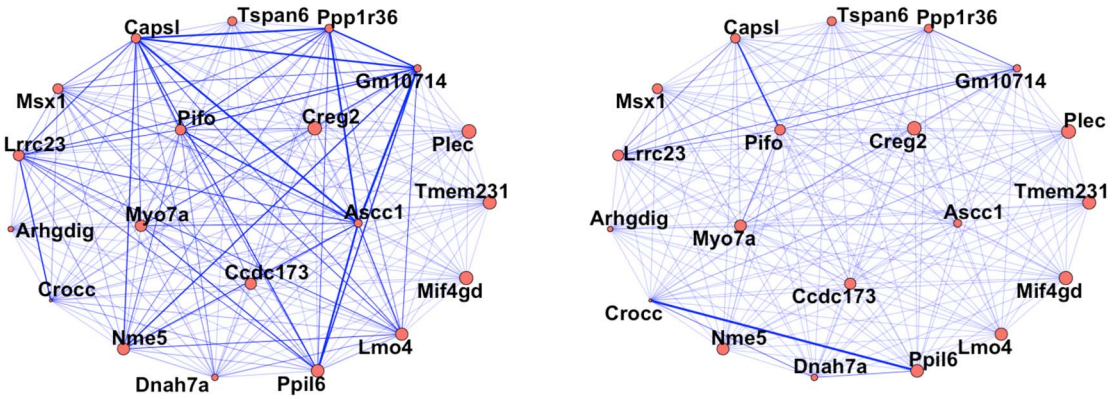

Figure 6. Co-expression networks for 20 genes identified by NetTDP for mouse brain data. **(a)** and **(b)** are networks for non-neurons and neurons, respectively. Nodes size represents the degree of the node. Edge thickness represents weight scale: the thicker the edge, the stronger the correlation. These genes differ in the weight of edge linked to the other genes. For example, the correlation between Crocc and Ppil6 is stronger in neurons than that in non-neurons.

also have widespread small differences in their networks, rather than sparsely focused differences. As in the simulation results, the TDP estimated at the edge level is much lower than the TDP estimated at the node level.

To further demonstrate the reliability of our NetTDP method, we applied NetTDP, CoGA and GSNCA, on AD gene sets collected in Data collection and processing section with less than 500 genes (totally 1451 gene sets). The overlap of the identified differential gene sets by NetTDP (non-zero TDP), CoGA (*P*-value less than 0.05) and GSNCA (*P*-value less than 0.05) methods is shown in Figure S14 in Supplementary Material. Over 88% of gene sets identified by our method can be also identified by GSNCA method, while over 13% by CoGA method. There are 51 common gene sets detected by all the three methods. Since our method could further quantify the differential networks by TDP, we also reported the gene sets with more than 20% TDP in Table S1 in Supplementary Material. Most of the reported 15 gene sets are related to Alzheimer's disease. For example, the immune system is a major factor in Alzheimer Disease (AD)[35], showing that positive regulation of immune system process (GO:0002684) may be associated with AD. Grant *et al.* [36] proved topographic regulation of kinase activity in Alzheimer's disease brains, indicating that positive regulation of kinase activity (GO:0033674) has close relationship with AD. Cell-substrate junction (GO:0030055) is also discovered by John *et al.* [37] related to AD. Mariana *et al.* [38] claimed that Alzheimer's disease is a result of stimulus reduction in a GABA-A-deficient brain, which supports positive regulation of response to stimulus (GO:0048584) and response to extracellular stimulus (GO:0009991) may play important roles in Alzheimer's disease.

Discussion

In this section, we will discuss the proposed method, NetTDP, in depth from four aspects, including network construction, statistic calculation, biological interpretation and batch effects.

Multiple network construction ways

The method of NetTDP proposed in this work is based on a weighted, signed gene correlation network for analyzing differential co-expression networks. Next, we discuss and compare some other networks, including distance networks, unweighted networks and unsigned networks.

Correlation network versus Distance network. The gene co-expression network can also be characterized by the distance of the spatial location, e.g. Euclidean distance. We call this type of network a distance network, and its adjacency matrix can be expressed as the normalized distances among genes by

$$A = 1 - (\mathrm{dist}/max(\mathrm{dist}))^2,$$

where dist is a distance matrix of genes.

Distance network-based methods are more sensitive to differentially expressed genes. If a gene is not differentially expressed between groups, distance network-based methods cannot identify the gene, even if the gene's co-expression network is different between groups. In real data studies, in order to eliminate the effect of differential expression, the within-group means were pre-processed to 0. Interestingly, the TDP estimated by the distance network-based method is everywhere 0. It shows that the method based on distance network cannot capture the fundamental co-expression relationship between genes. When the differential expression between groups is weakened, the genes identified by the method based on the correlation network are more reliable, that is, there is a higher confidence that the network in which the gene is located is different between groups.

Weighted network versus Unweighted network. When a weighted network is constructed, given a threshold, it is easy to obtain an unweighted network. Weights greater than the threshold are set to 1, otherwise 0. However, the binary network will lose a lot of useful information, which determines that the performance of the weighted network will be better than the corresponding unweighted network. Therefore, we preserve all edge information and input a weighted network in the proposed method.

Signed network versus Unsigned network. The Pearson correlation coefficient is signed, with a positive sign representing a positive correlation and a negative sign representing a negative correlation. The statistic in this work takes the sign information of the correlation coefficient into account, and captures the differences between the networks more finely. Of course, readers can also ignore the sign and simply consider the difference between

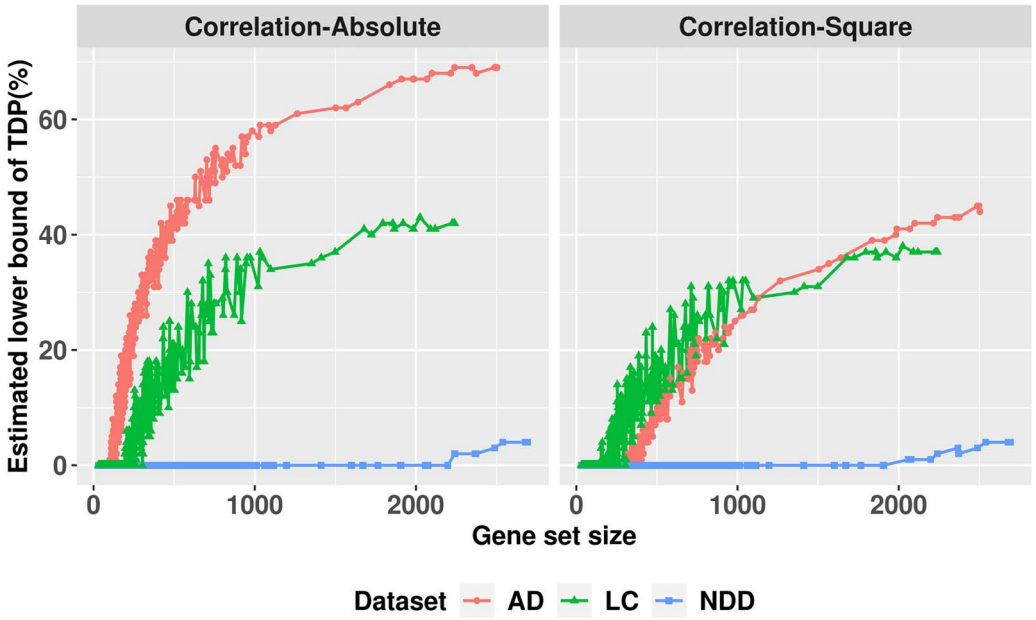

Figure 7. NetTDP *node-level* results for real data with absolute statistic (left) and square statistic (right) in gene sets defined biologically. Absolute (square) statistic means absolute (squared) difference is calculated in our method. The x-axis represents the size of these gene sets. The different colors are different data sets.

the absolute values of the correlation coefficients, although this will reduce the difference between networks to a certain extent.

Two statistic calculation ways

The statistic proposed in this work takes the absolute difference of the network, but the squared difference can also be considered. We compared these two definitions of the statistics on real data. For the square statistics, both *node-level* and *edge-level* results for gene modules and biological sets are reported in Figure 7 (right-hand side) and Figures S13 (right-hand side), S15 and S16 in Supplementary Material. We found that the absolute statistics tend to give higher TDPs than the squared statistics. For example, in the AD data set, the global TDP given by absolute statistics is 25% higher than that given by square statistics.

Biological interpretation

In this section, we firstly discuss non-localized network difference case, and then explain our findings in AD and LC data sets, respectively.

Non-localized network difference case. In differential co-expression network analysis, our goal is to identify local differences in the network. It would be most informative, biologically, if the true biological differences between two comparable networks are concentrated in a few local sub-networks. In practice, it may happen that the true biological differences between networks are widespread, that is, differences exist almost everywhere in the network. We investigated the behavior of our method in this situation in a simulation study (e.g. Figure S7 in Supplemental Material). Taking $\eta = 90$ as an example, it can be seen that in this situation of widespread effects, the estimated TDP will be larger for larger gene sets. In contrast, if the true biological differences are local, there is much less association between gene set size and TDP. Our results therefore suggest that the biological differences in the networks in the three real data sets are widespread, rather than local.

Identified genes related to AD and LC. In AD and LC data sets, we explained our findings by discussing relationships between specific diseases and genes in the differential co-expression modules identified by out NetTDP method.

For the AD data set, we focused on 17 gene sets defined biologically with TDPs estimated by our method over 60%. Based on existing literature, we validated the association between Alzheimer's disease (AD) and six genes, including APOE, BIN1, CLU, FERMT2, PTK2B and ABCA1. These six genes are included in all 17 gene sets we focused on here. Van Cauwenberghe *et al.* [39] identified that APOE, BIN1, CLU, FERMT2 and PTK2B could influence the risk of late onset AD (LOAD). Specially, APOE, BIN1 and CLU are widely known as the three greatest genetic risk factors for LOAD. Jiao *et al.* [40] used the LMR analysis method to identify three SNP pairs significantly associated with LOAD risk, including PTK2B-CR1, PTK2B-CD33 CD33-CR1. The ATP-binding cassette, sub-family A, member 1 gene (ABCA1) is a candidate risk gene for LOAD as a consequence of its role in cholesterol transport and metabolism, which is implicated in LOAD risk [41]. All evidence above strongly supports the relationship between these six genes and Alzheimer's disease.

For the LC data set, 13 gene sets biologically defined, whose TDPs are over 40%, caught our attention. We found seven genes linked to LC disease according in literature, including PIK3CA, PRKAR1A, DDX3Y, CXCR4, MARCKS, IDO1 and EGR1. All of these seven genes are contained in the 13 gene sets we are interested in. Most of the primary lung malignancies can be grouped into four major subtypes including adenocarcinoma, squamous cell carcinoma, large cell carcinoma and non-small cell lung cancer (NSCLC) [42]. Zhang *et al.* [43] found a significant role of PIK3CA in lung adenocarcinoma. They showed that PIK3CA mutations H1047R and H1047L are significant genetic alterations in lung adenocarcinoma and related to survival for lung adenocarcinoma patients who underwent curative resection. Protein Kinase cAMP-Dependent Regulatory Type I Alpha (PRKAR1A) is a tissue-specific extinguisher that transduces a signal through phosphorylation of different target proteins. Wang *et al.* [44] first found that PRKAR1A

was downregulated in lung adenocarcinoma patients. Lin [45] reported that patients expressing high DDX3Y associated with poor outcomes in multiple cancers, including lung cancer. Liu *et al.* [46] claimed that ERβ can promote NSCLC cell invasion via altering the ERβ/circ-TMX4/miR-622/CXCR4 signaling. Sekhar *et al.* [47] concluded that MARCKS is marked in combating lung cancer growthand acquired resistan. Tang *et al.* said that [48] IDO1 expression increased more in lung cancer compared with their corresponding non-tumor tissues. The over-expression of IDO1 significantly encouraged the metastasis and invasion of lung cancer cells, and IDO1 could promote metastasis formation *in vivo*. Zinc finger transcription factor early growth response gene 1 (EGR1) is underexpressed in NSCLC compared with normal lung. EGR1 expression has been linked to tumor suppression. Lower levels of EGR1 correlate with poor postoperative NSCLC outcomes. The patient's risk of disease recurrence can be diagnosed based on the expression of EGR1 [49]. The evidence above clearly shows the close relationship between these seven genes and lung cancer.

In total, the discussion here on the AD and LC data sets indicates the reliability of the results of our NetTDP method.

Batch effects

A differential analysis method could give false positives in the case of batch effects, if the experimental design is not well enough designed to counter such batch effects. If batches were not well distributed over cases and controls, there could be a difference in network between cases and controls. This is a problem for differential co-expression as for differential expression. In all analyses batch effects need to be corrected for if they are not properly balanced between cases and controls.

Conclusion

In this work, we promote the TDP as an important mathematical concept in differential co-expression network analysis. It helps researchers measure and localize the difference for a correlation based co-expression network, extending existing differential gene co-expression network analysis methods, such as WCGNA. We realize TDP inference in networks with *sumSome*, a permutation-based true discovery proportion estimation method in a closed testing framework. The method gives lower (1-α)-confidence bounds for the TDP simultaneously over all gene subsets, so that the confidence bounds remain valid even when the gene subsets of interest are selected post hoc.

Our method of NetTDP performs well in both simulation study and real data sets for inferring differential co-expression network. In biological applications, we validated that the proposed method has advantages when the differences are localized in the co-expression network.

The NetTDP method has a certain scope for use. Overall, although our NetTDP is technically applicable for any data type which WGCNA method could analyze, it may lose power for two extreme cases when the differences between two networks are too weak or spread everywhere. When the differences between the two networks are very weak, e.g. around 6% for the NDD data set, it is a great challenge to figure out where the differences are and it is reasonable that most of the annotated gene sets have zero TDPs. On the other hand, when the differences are widespread between two networks, that is, differences can not be localized in a small area, our NetTDP tend to obtain higher TDPs for larger gene sets.

Various methods have been proposed for inferring a network from the data, e.g. linear regression-based method [50–52]. Interesting future work could be to generalize the proposed method for linear regression coefficient-based networks.

Key Points

- We improve differential co-expression network analysis based on the WGCNA method by using true discovery proportions (TDP) to quantify and localize network differences. Especially in large networks it is useful to infer the strength and location of differences found.
- Our method of NetTDP allows inference on data-driven modules or biology-driven gene sets, and remains valid even when these sub-networks are optimized using the same data.
- In the proposed method, we test the difference with *edge-level* and *node-level* statistics to detect true discoveries of edges and nodes in the sense of differential co-expression network.
- Both simulation study and real data sets indicate that the proposed method has great power to analyze network differential structures for correlation based co-expression networks.

Funding

The work was funded by the National Natural Science Foundation of China project under Grant No. 12222115.

References

1. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;**4**(1):Article17.
2. Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics* 2008;**9**(1):1–13.
3. Niemira M, Collin F, Szalkowska A, *et al.* Molecular signature of subtypes of non-small-cell lung cancer by large-scale transcriptional profiling: identification of key modules and genes by weighted gene co-expression network analysis (wgcna). *Cancer* 2020;**12**(1):37.
4. Wang M, Wang L, Liyuan P, *et al.* Lncrnas related key pathways and genes in ischemic stroke by weighted gene co-expression network analysis (wgcna). *Genomics* 2020;**112**(3):2302–8.
5. Nangraj AS, Selvaraj G, Kaliamurthi S, *et al.* Integrated ppi-and wgcna-retrieval of hub gene signatures shared between barrett's esophagus and esophageal adenocarcinoma. *Front Pharmacol* 2020;**11**:881.
6. Bai K-H, He S-Y, Shu L-L, *et al.* Identification of cancer stem cell characteristics in liver hepatocellular carcinoma by wgcna analysis of transcriptome stemness index. *Cancer Med* 2020;**9**(12):4290–8.
7. DiLeo MV, Strahan GD, den Bakker, *et al.* Weighted correlation network analysis (wgcna) applied to the tomato fruit metabolome. *PLoS One* 2011;**6**(10):e26683.
8. Schmidt D, Wilson MD, Ballester B, *et al.* Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 2010;**328**(5981):1036–40.
9. Bar-Yam Y, Epstein IR. Response of complex networks to stimuli. *Proc Natl Acad Sci* 2004;**101**(13):4341–5.

10. Goh K, Cusick ME, Valle D, *et al.* (eds). The human disease network. *Proc Natl Acad Sci* 2007;**104**(21):8685–90.
11. Zhang X-F, Ou-Yang L, Zhao X-M, *et al.* Differential network analysis from cross-platform gene expression data. *Sci Rep* 2016;**6**(1):1–12.
12. West J, Bianconi G, Severini S, *et al.* Differential network entropy reveals cancer system hallmarks. *Sci Rep* 2012;**2**(1):1–8.
13. Ma J, Karnovsky A, Afshinnia F, *et al.* Differential network enrichment analysis reveals novel lipid pathways in chronic kidney disease. *Bioinformatics* 2019;**35**(18):3441–52.
14. Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol* 2012;**8**(1):565.
15. Tesson BM, Breitling R, Jansen RC. Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics* 2010;**11**(1):1–9.
16. Watson M. Coxpress: differential co-expression in gene expression data. *BMC Bioinformatics* 2006;**7**:509.
17. Yuan L, Sha W, Zhang J, *et al.* Differential coexpression analysis in gene modules level and its application to type 2 diabetes. In: *IEEE International Conference on Bioinformatics and Biomedicine*. Shanghai, China, IEEE, 2014.
18. Choi Y, Kendziorski C. Statistical methods for gene set co-expression analysis. *Bioinformatics* 2009;**25**(21):2780–6.
19. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets. *Bioinformatics* 2013;**30**(3):360–8.
20. Santos S, Galatro T, Watanabe RA, *et al.* Coga: An r package to identify differentially co-expressed gene sets by analyzing the graph spectra. *PLoS ONE* 2015;**10**(8):e0135831.
21. Sipko VD, Urmo V, Adriaan VDG, *et al.* Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* 2018;**19**(4):575–92.
22. Meijer RJ, Goeman JJ. Multiple testing of gene sets from gene ontology: possibilities and pitfalls. *Brief Bioinform* 2016;**17**(5):808–18.
23. Nichols TE. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage* 2012;**62**(2):811–5.
24. Genovese CR, Wasserman L. Exceedance control of the false discovery proportion. *J Am Stat Assoc* 2006;**101**(476):1408–17.
25. Goeman JJ, Solari A, *et al.* Multiple testing for exploratory research. *Statistical Science* 2011;**26**(4):584–97.
26. Anna V, Livio F, Goeman Jelle J. Permutation-based true discovery guarantee by sum tests. *arXiv:2102.1179v3*. 2020.
27. Langsrud Ø. *Rotation tests Statistics and computing* 2005;**15**(1):53–60.
28. Solari A, Finos L, Goeman JJ. Rotation-based multiple testing in the multivariate linear model. *Biometrics* 2014;**70**(4):954–61.
29. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol* 2013;**9**(3):e1002955.
30. Romanov RA. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* 2017;**20**:176–88.
31. Klein AM. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201.
32. Butler A, Hoffman P. Smibert. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
33. Jie Z, Kun H. Normalized imqcm: An algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. *Cancer Informatics* 2016;**CIN-S14021**:13s3.
34. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification* 2014;**31**:274–95.
35. Jevtic S, Sengar AS, Salter MW, *et al.* The role of the immune system in alzheimer disease: Etiology and treatment. *Ageing Res Rev* 2017;**40**:84–94.
36. Grant P, Pant HC. Topographic regulation of kinase activity in alzheimer's disease brains. *J Alzheimers Dis* 2002;**4**(4):269–81.
37. Siavelis John C, Bourdakou Marilena M, Athanasiadis Emmanouil I, *et al.* Bioinformatics methods in drug repurposing for alzheimer's disease. *Brief Bioinform* 2016;**17**(2):322–35.
38. Mariana A, Francisco J. Alzheimer's disease as a result of stimulus reduction in a gaba-a-deficient brain: A neurocomputational model. *Neural Plast* 2020;**2020**:26.
39. Van C. The genetic landscape of alzheimer disease: clinical implications and perspectives. *Genet Med* 2016;**18**:421–30.
40. Jiao B, Liu X, Zhou LF. Polygenic analysis of late-onset alzheimer's disease from mainland china. *PLoS One* 2015;**10**(12):e0144898.
41. Lupton MK, Proitsi P, Lin K, *et al.* The role of abca1 gene sequence variants on risk of alzheimer's disease. *J Alzheimers Dis* 2014;**38**(4):897–906.
42. Collins LG, Haines C, Perkel R, *et al.* Lung cancer: diagnosis and management. *Am Fam Physician* 2007s;**75**:56–63.
43. Jiao B, Liu X, Zhou LF. Pik3ca gene mutation associated with poor prognosis of lung adenocarcinoma. *Onco Targets Ther* 2013;**7**(6):497–502.
44. Wang S, Cheng Y, Zheng Y, *et al.* Prkar1a is a functional tumor suppressor inhibiting erk/snail/e-cadherin pathway in lung adenocarcinoma. *Rep* 2016;**6**(1):39630.
45. Lin TC. Ddx3x multifunctionally modulates tumor progression and serves as a prognostic indicator to predict cancer outcomes. *Int J Mol Sci* 2019;**21**(1):281.
46. Liu S, Hu C, Li M, *et al.* Estrogen receptor beta promotes lung cancer invasion via increasing cxcr4 expression. *Cell Death Dis* 2022;**13**(70):1–12.
47. Reddy SP, Natarajan V, Dudek AZ. Marcks is marked in combating lung cancer growth and acquired resistance. *Am J Respir Crit Care Med* 2014;**190**(10):1084–6.
48. Dongfang Tang L, Yue RY, Zhou L, *et al.* P53 prevent tumor invasion and metastasis by down-regulating ido in lung cancer. *Oncotarget* 2017;**8**(33):54548–57.
49. Ferraro B. Egr1 predicts pten and survival in patients with non-small-cell lung cancer. *Journal of Clinical Oncology Official Journal of the American Society of Clinical Oncology* 2005;**23**(9):1921–6.
50. Haury AC, Mordelet F, Vera-Licona P, *et al.* Tigress: Trustful inference of gene regulation using stability selection. *BMC Syst Biol* 2012;**6**:145.
51. Huynh-Thu VA, Irrthum A, Wehenkel L, *et al.* Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 2010;**5**(9):e12776.
52. Anh H, Guido S. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics* 2015;**31**(10):1614–22.