Briefings in Bioinformatics, 2022, **23(6)**, 1–11
https://doi.org/10.1093/bib/bbac466
Advance access publication date 21 October 2022
Problem Solving Protocol

Interaction-based transcriptome analysis via differential network inference

Jiacheng Leng and Ling-Yun Wu

Corresponding author. Ling-Yun Wu, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: lywu@amss.ac.cn

Abstract

Gene-based transcriptome analysis, such as differential expression analysis, can identify the key factors causing disease production, cell differentiation and other biological processes. However, this is not enough because basic life activities are mainly driven by the interactions between genes. Although there have been already many differential network inference methods for identifying the differential gene interactions, currently, most studies still only use the information of nodes in the network for downstream analyses. To investigate the insight into differential gene interactions, we should perform interaction-based transcriptome analysis (IBTA) instead of gene-based analysis after obtaining the differential networks. In this paper, we illustrated a workflow of IBTA by developing a Co-hub Differential Network inference (CDN) algorithm, and a novel interaction-based metric, pivot APC2. We confirmed the superior performance of CDN through simulation experiments compared with other popular differential network inference algorithms. Furthermore, three case studies are given using colorectal cancer, COVID-19 and triple-negative breast cancer datasets to demonstrate the ability of our interaction-based analytical process to uncover causative mechanisms.

Keywords: differential network inference, differential expression analysis, interaction-based transcriptome analysis

Introduction

Transcriptome analysis techniques based on gene expression are well studied, especially differential expression analysis [1–4]. However, traditional differential expression analysis has encountered a bottleneck. Differential expression analysis usually focuses on testing whether the summary statistics (e.g. mean and variance) of the distributions are equal in two groups of samples [5]. However, this kind of significance test is not enough to identify differences between groups with high intra-group heterogeneity [6]. Furthermore, life activities are mainly driven by the interactions between genes. For a pair of interacted genes, their co-expression pattern may change drastically, while the expression distributions of each gene in the two groups are the same [7–9]. Therefore, it is important to investigate the relationship or interaction between genes.

A natural model for studying interactions between genes is the graphical model, which considers genes as nodes and interactions between genes as edges, and together these nodes and edges form a graph [10–12]. We therefore expect to know which interactions vary between the two groups of samples. The graph constructed by these differential interactions is called a differential network. There have been many studies on differential network inference methods. For example, the fused graphical lasso (FGL) proposed by Danaher *et al.* [13] introduced a similarity penalty based on the Gaussian graphical lasso model; Mohan *et al.* [14] proposed PNJGL by adding a symmetric decomposition constraint of the differential part based on FGL, making it more likely to infer a differential network with a hub structure; D-trace obtained the differential network directly as a variable in the optimization model, which can reduce the variable dimensions [15]; NetDiff used variational inference for the inference of Gaussian graphical models [16]. However, all the above methods are designed for analyzing data from a single source. With the rapid development of sequencing technology, there is an increasing amount of multi-source data, such as single-cell data from multiple individuals [17, 18]. Fortunately, several differential network inference algorithms have been developed for multiple sources. For example, SIMULE focuses more on inferring the common network shared by multiple sources [19]; pDNA is a multi-source differential network inference method that considers a priori pathway constraint [20]; TDJGL assumes the existence of the shared common interactions of differential networks across multiple sources [21]; JEGN supposes that the networks from multiple sources have both the same shared common network structures and similar idiosyncratic network structures [22]. However, all these methods have a common problem, that is, they do not consider the heterogeneity

Jiacheng Leng is a PhD student at the Academy of Mathematics and System Sciences, Chinese Academy of Sciences and School of Mathematical Sciences, University of Chinese Academy of Sciences, China. His research interests include differential network inference, Gaussian graphical models, bioinformatics and deep learning.
Ling-Yun Wu is a Professor and the Director of the Bioinformatics Center at the Academy of Mathematics and System Sciences, Chinese Academy of Sciences, and a Professor at School of Mathematical Sciences, University of Chinese Academy of Sciences, China. His research interests include bioinformatics, optimization and machine learning.
Received: July 8, 2022. **Revised:** September 13, 2022. **Accepted:** September 28, 2022

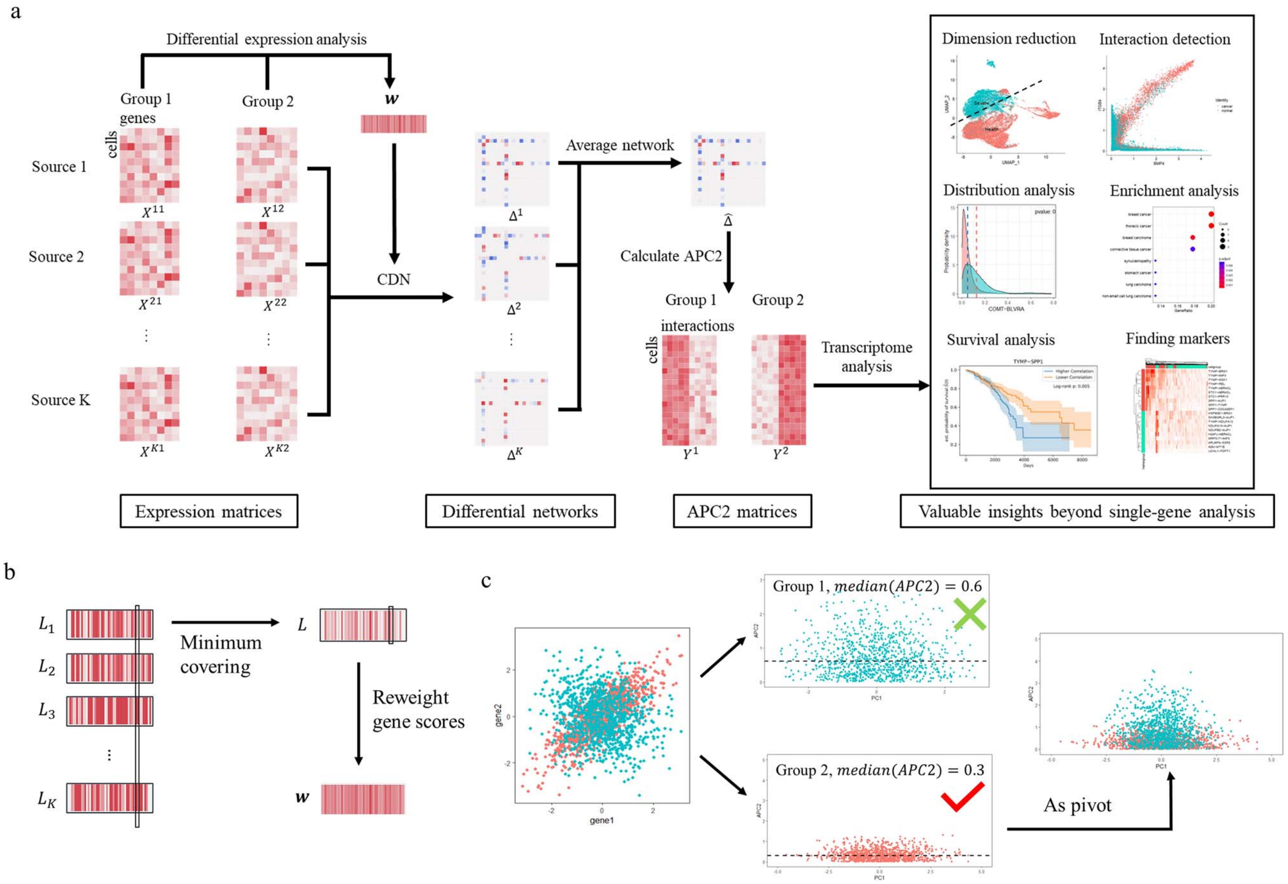


Figure 1. The workflow of IBTA. **(A)** The overall workflow of transcriptome analysis based on differential interactions. **(B)** The details about calculating gene score w. L_k is the P-value list of source k, and L is the integrated P-value list. **(C)** The diagram of calculating pivot APC2. Given expression values of two groups about gene 1 and gene 2, we calculate APC2 for each group separately and select one which has a smaller median as the pivot group. Then, we use the PCs of the pivot group to recalculate the pseudo expression values of this interaction in both two groups.

among genes. Different genes inherently have different levels of importance in the networks; therefore, in the models, genes should not be equally treated [23].

Even with differential networks, currently, most studies still only use the information of nodes in the network for downstream analyses, e.g. selecting and analyzing the genes by the degree of nodes. Few methods are available to perform transcriptome analysis directly at the level of interactions. This is mainly because at the level of genes, each gene has an expression value for each sample, which can be used in downstream analyses and visualization. However, at the level of interactions, each interaction usually has only one signed weight (e.g. Pearson's correlation coefficients) for a group of samples, rather than a value for each sample. Therefore, it becomes a challenge to design a scientifically sound transcriptome analysis method from the perspective of the interactions.

To address these issues, we first developed an algorithm that combines both differential gene expression analysis and differential network inference, called the Co-hub Differential Network inference (CDN) method. Then a metric at the level of interactions, pivot APC2 (pAPC2), was designed to measure the contribution of each sample to the differential interaction. Based on these tools, we proposed an interaction-based transcriptome analysis (IBTA) workflow. Specifically, CDN can use expression data from multiple sources, or individuals, to infer differential networks with common hub genes, and pAPC2 can assign pseudo-expression values of each interaction to each sample based on the contribution of the sample to this interaction. The source code and the demo of the workflow are accessible at https://github.com/Wu-Lab/IBTA.

Materials and methods

Workflow

Downstream analysis of differential network inference is usually vague and difficult to be performed systematically. In this study, we present a novel approach to conduct a systematical interaction-based transcriptome analysis. The overall workflow is illustrated in Figure 1. Given the gene expression matrices from K sources (e.g. single-cell sequencing samples per individual, bulk sequencing samples per subtype of disease or subgroup of individuals with similar clinical phenotypes) under two different conditions with the shared genes of interest, $X^{k1} = (x_{ij}^{k1})_{n_{k1}\times d}$ and $X^{k2} = (x_{ij}^{k2})_{n_{k2}\times d}$, $k = 1, 2., \ldots, K$, where d is the number of genes, and n_{k1} and n_{k2} are the number of samples from group (condition) 1 and group (condition) 2, respectively. We first compute the gene weights based on differential expression analysis (DEA). Combining these weights, we apply the CDN method to infer the differential networks, $\Delta^k = (\delta_{ij}^k)_{d\times d}$, $k = 1, \ldots, K$, of different sources and gain the average differential network $\overline{\Delta} = \frac{\sum_k \Delta^k}{K}$ for downstream analysis. We then calculate the pAPC2 of each interaction of interest in $\overline{\Delta}$ and obtain the pAPC2 matrices of two groups: $Y^1 = (y_{ij}^1)_{n_1\times q}$ and $Y^2 = (y_{ij}^2)_{n_2\times q}$, where $n_1 = \sum_{k=1}^{K} n_{k1}$, $n_2 = \sum_{k=1}^{K} n_{k2}$, and q is the number of the candidate interactions of interest. Finally, we use pAPC2 matrices as the features of gene interactions, to conduct downstream analyses such as dimension reduction, finding interaction markers, survival analysis, enrichment analysis and so on, which reveal many valuable insights beyond the transcriptome analysis at the level of single gene.

Co-hub differential network inference method

Given $X^{k1}, X^{k2}, k = 1, 2., \ldots, K$, CDN aims to infer the differential networks Δ^k, where $\Delta^k_{ij} = 0$ indicates that there is no change from group 1 to group 2 in the source k regarding the interaction between gene i and gene j. Under the assumption that X^{k1} and X^{k2} follow the multivariate Gaussian distribution, we infer $\Delta^k = \Theta^{k1} - \Theta^{k2}$ using Gaussian graphical models, which can measure the conditional independence between genes [12]. Specifically, the model is as follows:

$$\min_{\Theta^{k1},\Theta^{k2} \in \mathcal{S}^d_{++}} \sum_{k=1}^{K} \left[-L\left(\Theta^{k1}, \Theta^{k2}\right) + \lambda_1 \left(\left\| \Theta^{k1} \right\|_1 + \left\| \Theta^{k2} \right\|_1 \right) \right] + \lambda_2 \sum_{j=1}^{d} w_j \left\| \begin{bmatrix} \left(\Theta^{11} - \Theta^{12}\right)_j \\ \left(\Theta^{21} - \Theta^{22}\right)_j \\ \vdots \\ \left(\Theta^{K1} - \Theta^{K2}\right)_j \end{bmatrix} \right\|_2 ,$$

where $L(\Theta^{k1}, \Theta^{k2}) = [- n_{k1} \mathbf{log\,det}\,(\Theta^{k1}) + n_{k1} \mathbf{trace}(\Theta^{k1} S^{k1}) - n_{k2} \mathbf{log\,det}\,(\Theta^{k2}) + n_{k2} \mathbf{trace}(\Theta^{k2} S^{k2})]$ is the joint log-likelihood function, and S^{k1}, S^{k2} are the sample covariance matrices of the centered expression matrices, X^{k1}, X^{k2}, respectively. Θ^{k1}, Θ^{k2} are called precision matrices and taken as the variables of our optimization model. $\|\Theta\|_1 = \sum_{i,j} |\Theta_{ij}|$ is the l_1-norm which is the sum of the absolute values of the matrix elements. $|v|_2 = \sqrt{\sum_j v_j^2}$ is the l_2 vector norm, and $(\Theta)_j$ is the jth column of matrix Θ. $\mathcal{S}^d_{++}$ represents the symmetric positive definite matrix space. $\Theta^{kc}_{ij} = 0$ means gene i and gene j are conditionally independent in group c of source k, given all information of else genes. Therefore, the position of nonzero elements in Θ corresponds to the edge distribution in the network.

The first penalty, $\lambda_1(\|\Theta^{k1}\|_1 + \|\Theta^{k2}\|_1)$, is the sparsity penalty which encourages the networks of group 1 and group 2 to be sparse. The second penalty is the similarity penalty, which encourages the same gene to have a similar neighborhood structure in the differential networks of multiple sources by group lasso. λ_1, λ_2 are two hyperparameters in this model. The bigger λ_1 is, the sparser Θ^{k1} and Θ^{k2} are. The bigger λ_2 is, the more similar $\Delta^k = \Theta^{k1} - \Theta^{k2}, k = 1, 2., \ldots, K$ are. w_j is the DEA weight of gene j calculated based on P-values of differential expression analysis across K groups. In this study, we use R package 'EMDomics' to calculate p-values, which measures the overall difference between the distributions of a gene's expression in two groups [24]. Specifically, w_j is a normalized integrated p-value score. For each gene j, we first calculate DEA p-values for two groups in each source and then obtain the minimum of these p-values across K sources, denoted as p_j^*. Finally, we utilize softmax to reweight the p-values of all genes, i.e. $w_j = d \frac{e^{p_j^*}}{\sum_j e^{p_j^*}}$, which can make the mean of $\{w_j\}_{j=1,..,d}$ be d (Figure 1B). The smaller w_j is, the more differential the gene j is expressed in two groups, as a result, gene j is penalized less to encourage it to have more interactions in differential networks.

Following PNJGL [14], we introduce the symmetric decomposition of $\Theta^{k1} - \Theta^{k2} = V^k + (V^k)^T$. This kind of decomposition combined with group lasso encourages the differential networks to generate more hub genes, which interact with many other genes. It is more fit for the realistic situation since there exist many hub genes (such as housekeeping genes) in the real biological networks. With stacking penalties from several sources, the second penalty can encourage the co-hub structure in multi-source differential networks. Now, the CDN model can be rewritten as follows:

$$\min_{\Theta^{k1},\Theta^{k2} \in \mathcal{S}^d_{++}, V^k} \sum_{k=1}^{K} \left[-L\left(\Theta^{k1}, \Theta^{k2}\right) + \lambda_1 \left(\left\| \Theta^{k1} \right\|_1 + \left\| \Theta^{k2} \right\|_1 \right) \right] + \lambda_2 \sum_{j=1}^{d} w_j \left\| \begin{bmatrix} V_j^1 \\ V_j^2 \\ \vdots \\ V_j^K \end{bmatrix} \right\|_2$$

$$s.t. \Theta^{k1} - \Theta^{k2} = V^k + \left(V^k\right)^T, k = 1, \ldots, K.$$

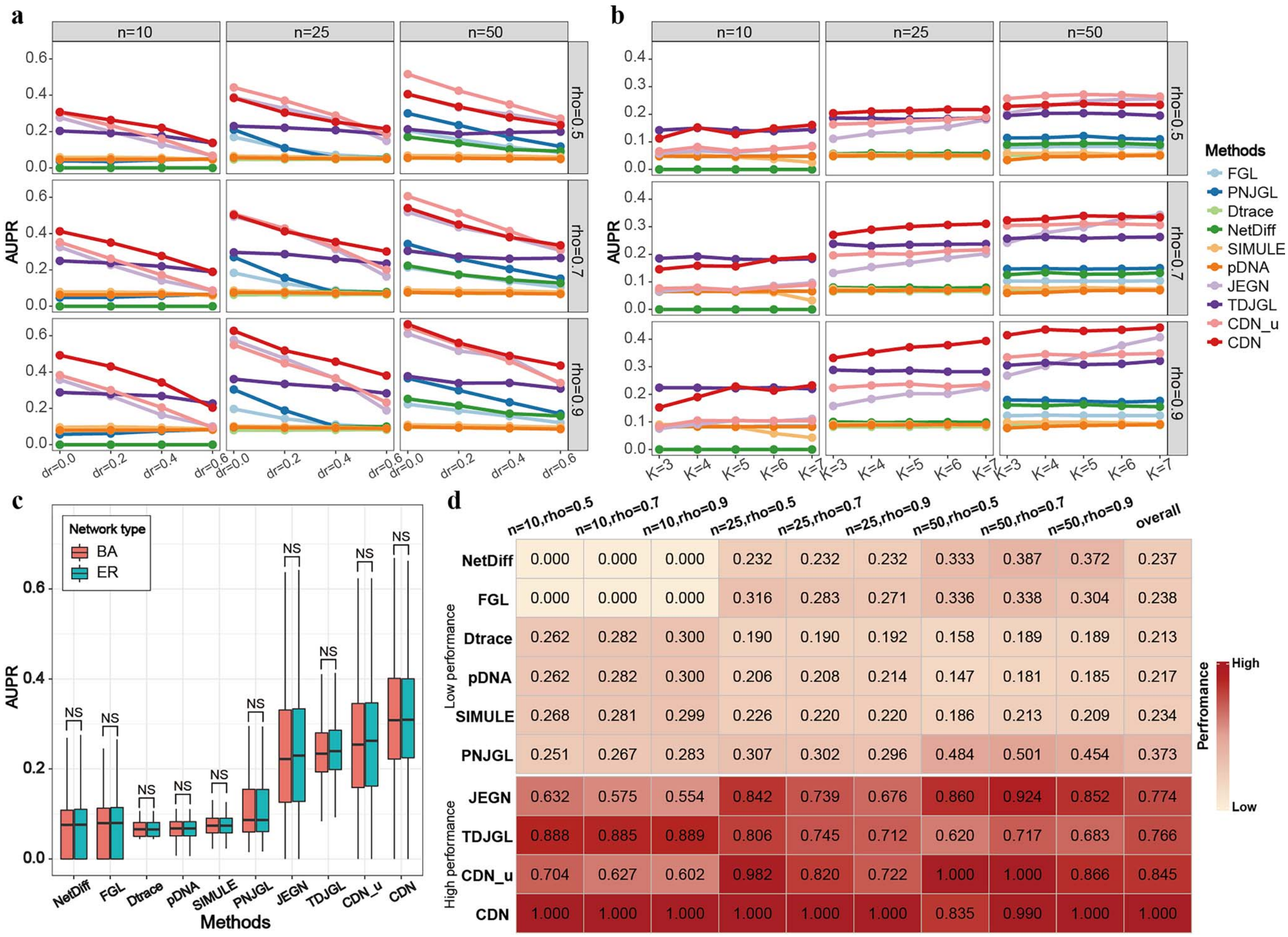
We use a variant of ADMM without fixed penalty parameter method [25] to solve this optimization problem. The details of the algorithms are shown in supplementary materials.

Pivot APC2

To exploit and analyze the results of differential network inference, we use APC2 [23] to measure the contribution of each sample to each interaction. APC2 is the absolute value of the 2nd principal component calculated by applying the Principal Component Analysis (PCA) to the expression data of two genes. PCA identifies a set of directions (technically, a linear subspace) that maximizes the variance of the data. It turns out that this is identical to minimizing the projection distance in a least-squares sense by finding a linear subspace. In our case, the linear subspace is a straight line, i.e. the 1st principal component (PC1), because our data is two-dimensional (gene1 against gene2). At this point, this line represents the trend of co-expression throughout the major cell population and the distance from one cell to this line (APC2) represents the deviation of this cell from the trend of major cell population. The original APC2 is defined for analyzing a single group of samples and is not suitable for the problem in this study. Therefore, we extend it to pivot APC2 for two groups of samples. We first separately calculate APC2 for cells from group 1 and group 2 and then select a pivot group to determine the final APC2 values (Figure 1C). Specifically, given a gene interaction j whose two genes are j_1, j_2 and their expression matrices, $\left(X^1_{\{j_1,j_2\}}\right)_{n_1 \times 2}, \left(X^2_{\{j_1,j_2\}}\right)_{n_2 \times 2}$ of group 1 and group 2 centered on $(0, 0)$, where n_c is the cell number of group $c = 1, 2$. Suppose that $X^c_{\{j_1,j_2\}} = V_c \boldsymbol{\Sigma}_\mathbf{c} U_c^T$ is the singular value decomposition. The APC2 value of the interaction j in cell i from group c is $z^c_{ij} = \left|\left(\boldsymbol{\Sigma}_\mathbf{c} U_c^T\right)_{i2}\right|$, and the median APC2 value of group c is $m^c_j = \mathbf{median}_i(z^c_{ij})$. The group $c^* = \mathbf{argmin}_c(m^c_j)$ is selected as the pivot group for the interaction j (Figure 1C). We then calculate the final APC2 values using the pivot eigen matrix $U^T_{c^*}$, and obtain the pivot APC2 value of the interaction j for cell i from group c, $y^c_{ij} = \left|\left(\boldsymbol{\Sigma}_\mathbf{c} U^T_{c^*}\right)_{i2}\right|$. We can use the pivot APC2 $Y^c = \left(y^c_{ij}\right)_{n_c \times q}$ as the features of interactions to perform downstream analysis, where q is the number of the candidate interactions. For example, we can determine whether there is a significant difference between two groups of cells by comparing the distributions of pAPC2 in different groups of cells.

Simulation experiments

To exam the performance of the new network inference method CDN, we compared it with several differential network inference methods, including single-source methods: FGL, PNJGL, D-trace and NetDiff [13–16], and multi-source methods: SIMULE, pDNA, TDJGL and JEGN [19–22] and also a simplified version of CDN without DEA weights $\{w_j\}$, called unweighted CDN (CDN_u). For fair comparison, we selected the best parameters of each method by grid search.



	n=10,rho=0.5	n=10,rho=0.7	n=10,rho=0.9	n=25,rho=0.5	n=25,rho=0.7	n=25,rho=0.9	n=50,rho=0.5	n=50,rho=0.7	n=50,rho=0.9	overall
NetDiff	0.000	0.000	0.000	0.232	0.232	0.232	0.333	0.387	0.372	0.237
FGL	0.000	0.000	0.000	0.316	0.283	0.271	0.336	0.338	0.304	0.238
Dtrace	0.262	0.282	0.300	0.190	0.190	0.192	0.158	0.189	0.189	0.213
pDNA	0.262	0.282	0.300	0.206	0.208	0.214	0.147	0.181	0.185	0.217
SIMULE	0.268	0.281	0.299	0.226	0.220	0.220	0.186	0.213	0.209	0.234
PNJGL	0.251	0.267	0.283	0.307	0.302	0.296	0.484	0.501	0.454	0.373
JEGN	0.632	0.575	0.554	0.842	0.739	0.676	0.860	0.924	0.852	0.774
TDJGL	0.888	0.885	0.889	0.806	0.745	0.712	0.620	0.717	0.683	0.766
CDN_u	0.704	0.627	0.602	0.982	0.820	0.722	1.000	1.000	0.866	0.845
CDN	1.000	1.000	1.000	1.000	1.000	1.000	0.835	0.990	1.000	1.000

Figure 2. Performance comparison of differential network inference methods. **(A)** The variation of AUPR along the dropout rate parameter $dr \in \{0, 0.2, 0.4, 0.6\}$ for different combinations of other data parameters. **(B)** The variation of AUPR along the source number parameter $K \in \{3, 4, 5, 6, 7\}$. **(C)** The overall AUPR boxplot which compares different underlying network structures. Red is BA network and green is ER network. **(D)** The overall relative performance of each combination of data parameters, 1 is the best, 0 is the worst.

We compared these methods by generating synthetic expression data from multi-Gaussian distribution on both Barabási–Albert (BA) and Erdős–Rényi (ER) network models [26, 27]. Many biological networks such as gene regulatory networks and protein–protein interaction networks have the well-known property of scale-free (i.e. a small number of genes have most of the interactions, while most genes have only a small number of interactions), which is often modeled using the BA network. On the opposite, ER network is the representative model of general random networks, in which the edges are connected independently (i.e. the emergence of edges does not depend on the degrees of nodes). We set $n_{k1} = n_{k2} = n \in \{10, 25, 50\}, k = 1, \dots, K$ and $d = 50$. Here, $n_{kc}, c = 1, 2$ is the number of samples in group c, source k. d is the number of genes. The difficulty of network inference is proportion to d/n. If $n \gg d$, the real network structure can be recovered completely in theory. Therefore, we only need to fix d and vary n to generate datasets of different difficulty and complexity for testing the performance of different methods. Considering the computational time cost, we finally chose $d = 50$ and selected the cases with $n \leq d$ to simulate the challenging real-world data as much as possible. Another three data parameters are K, dr, rho. $K \in \{3, 4, 5, 6, 7\}$ is the number of sources; $dr \in \{0, 0.2, 0.4, 0.6\}$ is the dropout rate of single-cell expressions; and $rho \in \{0.5, 0.7, 0.9\}$ controls how many genes a hub gene has interactions with. For example, $rho = 0.5$ means that a hub gene is expected to have interactions with 50% of genes. The details of the data generation are shown in supplementary materials.

The evaluation metric used in this part is the area under the precision-recall curve (AUPR). Specifically, we normalized the precision matrix as $P_{ij}^{kc} = \begin{cases} -\frac{\Theta_{ij}^{kc}}{\sqrt{\Theta_{ii}^{kc}\Theta_{jj}^{kc}}}, i \neq j \\ 1, i = j \end{cases}, k = 1, \dots, K, c = 1, 2,$ which can measure the strength of partial correlations. We use $\Delta_{\text{true}}^{k} = P_{\text{true}}^{k1} - P_{\text{true}}^{k2}$ as label matrix, $I_{\{\Delta_{\text{true}}^{k} \neq 0\}}$ as the label, where I is the indicator function. The score matrix $\Delta^{k} = P^{k1} - P^{k2}$ is used to calculate AUPR. Moreover, we calculate AUPR on positive values and negative values separately to eliminate the effect of the sign.

Comparison results

We first conducted simulation experiments on Barabási–Albert (BA) networks, which have a scale-free property. We fixed $K = 5$ and observed the variation of AUPR along the dropout rate parameter $dr \in \{0, 0.2, 0.4, 0.6\}$ for different combinations of other data parameters (Figure 2A). A lower n means a smaller number of available samples; a higher dr means the poorer sample quality; a lower rho implies lower significant differences between the two states. All three of these situations mean that the problem becomes more difficult. As expected, the AUPR of all methods decreases as the dropout rate increases, the sample size n decreases and the differential rate rho decreases.

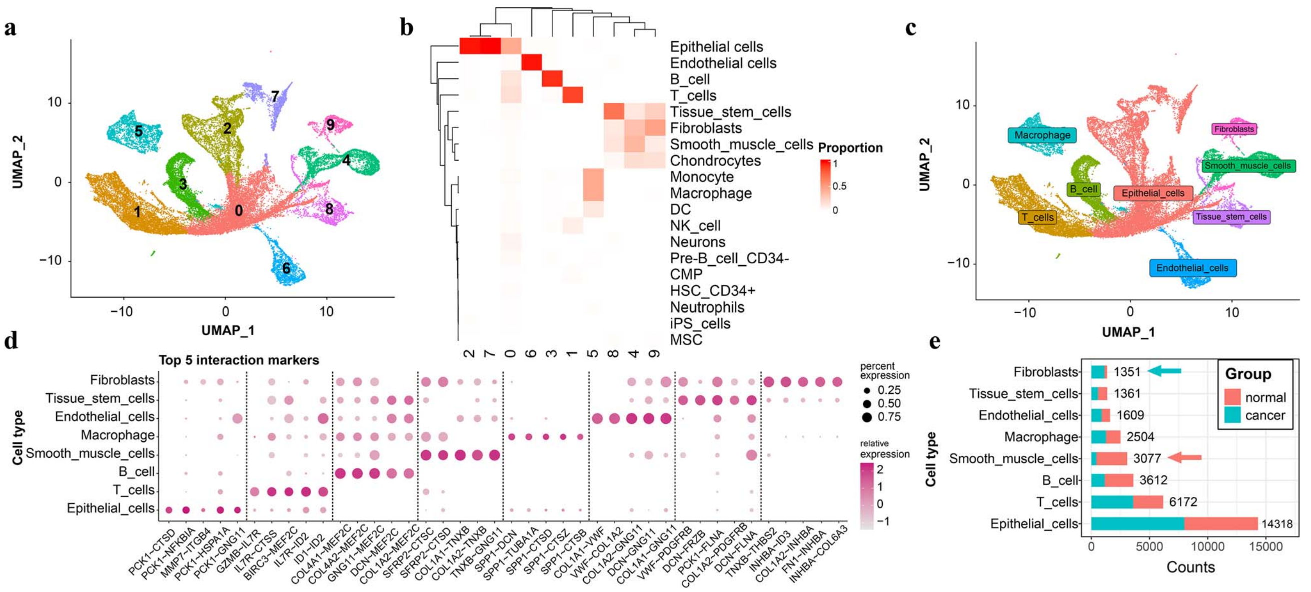


Figure 3. The IBTA results of the CRC dataset. **(A)** The UMAP visualization of 34,004 cells. **(B)** The cell annotation heatmap of 34,004 cells. The *x*-axis is the cluster obtained using pAPC2 features, the *y*-axis is the cell type. The color exhibits the cell number proportion of each cell type in a cluster. If there is only one cell type in a cluster, the proportion is 1. **(C)** The cell type annotation UMAP. **(D)** The top 5 interaction markers of each cell type. **(E)** The cell number of different groups in each cell type. Red is the normal group and green is the cancer group.

All methods achieve the highest AUPR at the simplest setting ($n = 50, rho = 0.9, dr = 0$) and the lowest AUPR at the most difficult setting ($n = 10, rho = 0.5, dr = 0.6$). As expected, almost all the multi-source methods (JEGN, TDJGL, CDN_u, CDN) outperform the single-source methods (FGL, PNJGL, D-trace, NetDiff). Meanwhile, the performance of the three methods (Dtrace, SIMULE, pDNA) is relatively poor. D-trace and pDNA both regard the differential network as a variable to solve directly rather than the subtraction of the two precision networks in different groups. Therefore, they miss the importance of the interactions in each group. SIMULE assumes that different individuals from the same group share a common network structure, and does not consider the differences between individuals. CDN and CDN_u outperform other methods in most cases, which suggests that the co-hub assumption is reasonable. CDN significantly outperforms other methods including CDN_u when d/n is relatively large. Because in this case, algorithms based solely on network inference do not have enough information to infer a reliable network. This result also confirms the necessity and importance of combining the information of differential expression analysis.

We then fixed $dr = 0.6$ and observed the variation of AUPR along the source number parameter $K \in \{3, 4, 5, 6, 7\}$ (Figure 2B). Almost, all the multi-source methods perform better as K increases (pDNA, JEGN, TDJGL, CDN_u, CDN), while the single-source methods maintain the performance as K increases. This result indicates that multi-source methods can well utilize the information shared among multiple sources. Naturally, if there are more available sources for the same groups, the better the inferred results. In all cases, CDN outperforms almost all methods, especially when the hub structure is more significant. In case $n = 10$, the AUPR increase of CDN along K is significantly larger than other methods. This is because each source has less valid information due to low n and high dr. Despite the number of sources increases, the information shared among sources is not sufficient to significantly improve the accuracy of the results. However, by integrating the DEA weights, CDN can amplify the information shared among sources, therefore, increasing the performance of the algorithm. Except for CDN, the AUPR increase of JEGN is also large in the case of $n > 10$, while that of TDJGL is not significant. This is because TDJGL assumes the differential networks share the common interactions among sources, while JEGN does not. When K increases, although the information of common interactions increases, the number of common interactions decreases, which is unfavorable in the assumptions for TDJGL.

Finally, we conduct all the experiments with the same settings on Erdős–Rényi (ER) networks. Performance-wise, ER and BA networks are very similar (Figure 2C). All the 10 methods are non-significant (P-value> 0.05). This implies the topology of the network does not affect the differential network inference significantly. Figure 2D shows the overall relative performance on both BA and ER networks under all combinations of data parameter settings. The performance of the best method is 1, and the worst is 0. Figure 2D again shows that almost all multi-source methods outperform the single-source methods consistently and CDN performs the best in seven out of nine circumstances.

Case study

Transcriptome analysis based on interactions is difficult because of the lack of expression data from an interaction perspective. To demonstrate how our IBTA workflow works, we conduct experiments on three real datasets, colorectal cancer (CRC) dataset E-MTAB-8410 [28] from ArrayExpress database, COVID-19 dataset GSE145926 [29] and triple-negative breast cancer (TNBC) dataset GSE161529 [30] from GEO database. All the datasets are processed and available on the public websites. To improve the performance of network inference, we further preprocessed three datasets using the dropout imputation method 'MAGIC' [31]. We use the CRC dataset to demonstrate the ability of our workflow to uncover pathogenic mechanisms between cell types and the COVID-19 and TNBC datasets to reveal pathogenic mechanisms within cell types, particularly those results that are virtually impossible to obtain from the single gene perspective.

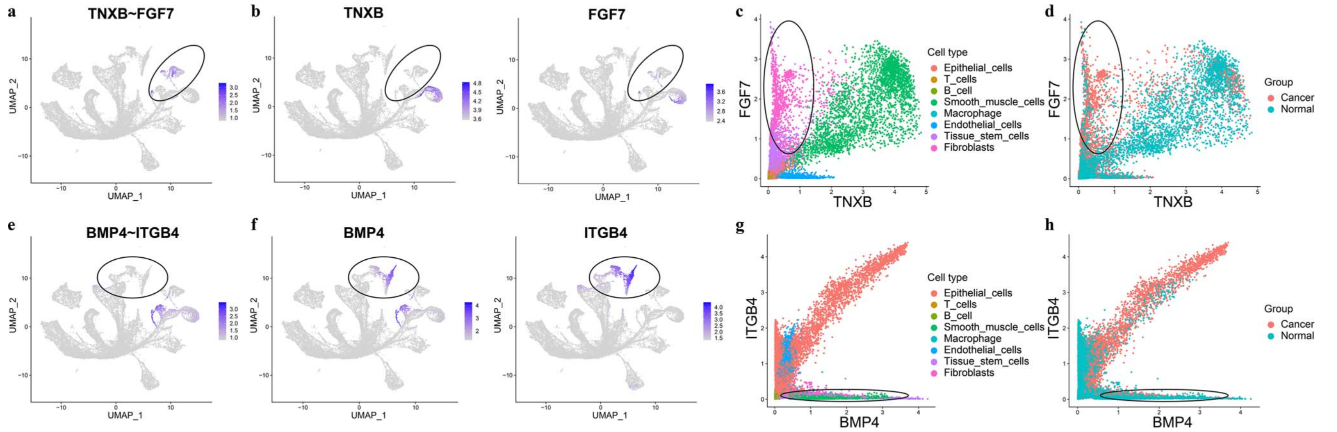


Figure 4. The expression visualization of TNXB∼FGF7 and BMP4∼ITGB4 and corresponding genes. **(A)** the UMAP visualization TNXB∼FGF7 pAPC2 values. **(B)** The UMAP visualization of TNXB and FGF7 expression values. **(C)** The joint distribution of TNXB and FGF7 expression values colored by cell types. **(D)** The joint distribution of TNXB and FGF7 expression values colored by groups. **(E)** The UMAP visualization BMP4∼ITGB4 pAPC2 values. **(F)** The UMAP visualization of BMP4 and ITGB4 expression values. **(G)** The joint distribution of BMP4 and ITGB4 expression values colored by cell types. **(H)** The joint distribution of BMP4 and ITGB4 expression values colored by groups.

Colorectal cancer

We conducted our workflow on the CRC dataset with 34,004 cells, of which 16,986 are from the tumor core, and 17,018 are from the normal tissue adjacent to the neoplasm after quality control. We first used the expression data as the input of CDN to infer the average differential network, $\overline{\Delta}$. Then we extracted the top 500 interactions according to the absolute values in $\overline{\Delta}$ as our candidate gene interactions. Finally, we calculated pAPC2 on these interactions and get a 34,004×500 pAPC2 matrix as our features to perform downstream analyses.

In our analysis, these cells were divided into 10 clusters and had clear boundaries using our pAPC2 features (Figure 3A). After annotation using the R package *SingleR* [32] based on gene features, the proportion of cell types in each cluster is shown in Figure 3B. There is a high agreement between our clustering results and cell annotations, i.e. the proportions of cell types are very close to 1 in most clusters. It implies that the pAPC2 features retain considerable biological information. We annotated the cluster as the cell type with the highest number of cells within it (Figure 3C). We also calculated the top 5 interaction markers of each cell type using the pAPC2 matrix (Figure 3D). Almost all interaction markers have a significant expression advantage in their corresponding cell types, which once again verifies that pAPC2 has captured meaningful biological information. Meanwhile, we observed that there exist some common genes in the interaction markers, such as TNXB in smooth muscle cells and INHBA in the fibroblasts, which implies they may act as hub regulators in biological activities. The cell numbers of two groups in each cell type are shown in Figure 3E. We observed two relatively pure cell types, fibroblasts and smooth muscle cells. Fibroblasts are mostly cancer cells, while smooth muscle cells are dominated by normal cells. Next, we utilized two instances to illustrate a more thorough investigation, TNXB∼FGF7 for fibroblasts and BMP4 ∼ ITGB4 for smooth muscle cells (Figure 4).

For TNXB∼FGF7, the pAPC2 in fibroblasts is higher than in other cell types, while neither of the single genes is highly expressed (circled in Figure 4A,B). We further investigated the joint distribution of TNXB and FGF7 expression (Figure 4C,D). Figure 4C shows there is a relatively significant correlation between TNBX and FGF7 expression in smooth muscle cells (Pearson’s correlation = 0.6999), which are dominated by normal cells. Meanwhile, this correlation is broken in fibroblasts (Pearson’s correlation = 0.1901, circled in Figure 4C), which are dominated by cancer cells. High expression of TNXB mRNA in cancer is associated with a good survival prognosis [30], and FGF7 is elevated in mucosal regions of cancer patients, supporting its potential as a biomarker of regional oncogenesis [33]. At the same time, studies have shown that TNC (the protein encoded by TNXB) has a high affinity for many fibroblast growth factor (FGF) families, including FGF7 [33]. Therefore, the interaction of TNXB and FGF7 may be the reason that the high expression of FGF7 loses most of its oncogenic effects, thus preventing cell carcinogenesis.

For BMP4 ∼ ITGB4, we can observe the pAPC2 in smooth muscle cells is higher than in other areas (Figure 4E). BMP4 is also highly expressed in smooth muscle cells, but both BMP4 and ITGB4 are highly expressed in cluster 7 (epithelial cells). This implies that neither gene can be a marker of smooth muscle cells. We further investigated the joint distribution of BMP4 and ITGB4 expression (Figure 4G,H). Figure 4G shows there is a relatively significant correlation between BMP4 and ITGB4 expression in epithelial cells (Pearson’s correlation = 0.8839), which are dominated by cancer cells. Meanwhile, this correlation is broken in smooth muscle cells (Pearson’s correlation = −0.078, circled in Figure 4H), which are dominated by normal cells. BMP4 is universally upregulated in human CRC cells and has been considered as a candidate treatment strategy for CRC [34]. However, CRC appears to be associated with more than just high expression of BMP4. ITGB4 has also been predicted as a diagnostic serum biomarker and a potential therapeutic target for CRC [35, 36]. Combined with Figure 4H, we think that breaking some interaction between these two genes seems to play a role in suppressing CRC.

COVID-19

We next conducted experiments on the COVID-19 dataset using our IBTA workflow. After annotation, the cell type composition is shown in Figure 5A, the most abundant of which are macrophages. The disease group composition in each cell type is also shown in Figure 5A. Considering both cell type abundance and disease group ratios, we selected macrophages as our target cell type to explore the differences between ‘severe’ and ‘health’ samples. To explore useful information which usually cannot be obtained from the gene-based analysis, we first excluded genes with adjusted *t*-test *P*-value ≤ 0.05, i.e. differential expressed genes (DEGs), then selected the top 500 genes with the highest

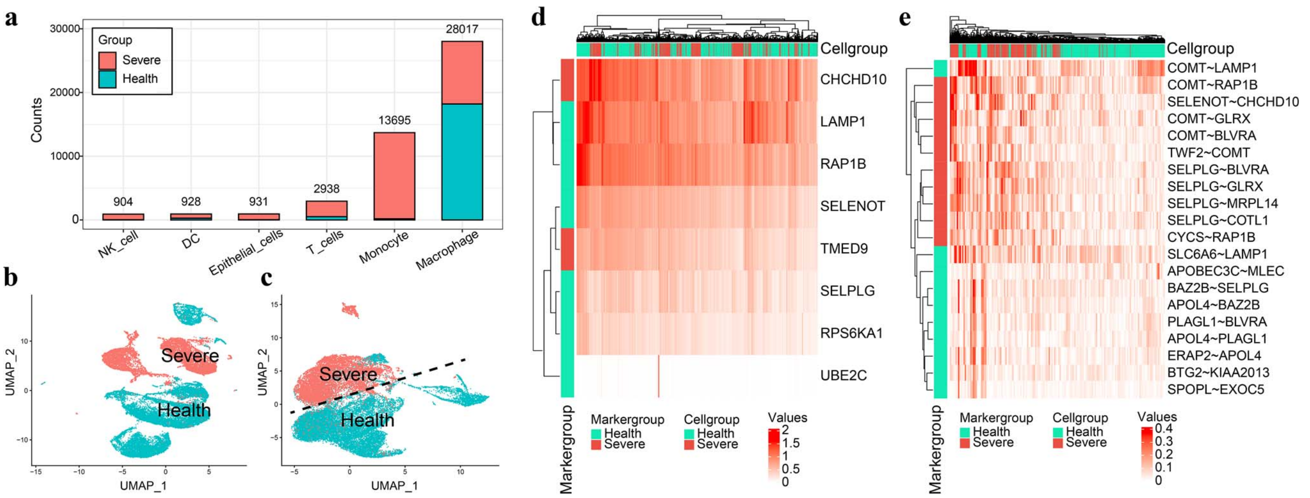


Figure 5. The IBTA results of the COVID-19 dataset. **(A)** The proportion of groups in each cell type. **(B)** The UMAP visualization of cells by gene expression features. **(C)** The UMAP visualization of cells by pAPC2 features. The straight dashed line can separate two groups well. **(D)** The heatmap of the top 10 gene markers of each group (only 8 available). (**E**) The heatmap of the top 10 interaction markers of each group.

variance as candidate genes. We applied the CDN method and calculated the pAPC2 matrix as the input features of downstream analyses. The dimension of the pAPC2 matrix is 28,017 cells × 500 interactions. For comparison, we also conducted the gene-based differential expression analysis.

The UMAP visualization of cells by gene features and pAPC2 features colored by the group are shown in Figures 5B, C, respectively. Notably, the pAPC2 features can separate 'severe' and 'health' cells into two different half-planes by a straight dashed line (Figure 5C), while we cannot find such a straight line when using gene features (Figure 5B). This implies that the pAPC2 features are more friendly for classification. We further used the top 10 markers of each group as the features of hierarchical clustering and conducted the heatmap visualization (Figure 5D, E). We performed hierarchical clustering on both cell levels and feature levels. After filtering out the DEGs, it is almost impossible to find marker genes (only eight markers in Figure 5D), while there retains a large amount of useful information in the interaction markers based on pAPC2. The hierarchical clustering results on interaction markers show that the markers representing the same group are closer and the hierarchical clustering results on cells show that the cells from the same group are closer (Figure 5E), while these phenomena cannot be observed in the heatmap based on gene features (Figure 5D). It is interesting to note that there are some hub genes shared by several interaction markers, such as COMT and SELPLG. COMT has been identified as interacting with NSP7 proteins encoded in the SARS-CoV-2 genome related to COVID-19 severity, prognosis or outcome [37]. SELPLG, also called PSGL-1, has been reported to impair the incorporation of SARS-CoV-2 spike (S) glycoproteins into pseudo virions and may inhibit coronavirus replication [38].

Furthermore, an example COMT∼BLVRA is provided here to highlight the capabilities of our framework. The UMAP, probability density, box line plot and scatter plot of the interaction and its genes are shown in Figure 6. As a marker of the severe group, COMT∼BLVRA is significantly highly expressed on one side of the dashed line (severe side), while both single genes are highly expressed on two sides of the dashed line (Figure 6A). Also, both the density plot and the box plot show that the pAPC2 value has a significant difference between the two groups with a *t*-test *P*-value smaller than 1e−16, but the *P*-values of genes are 0.570 (COMT) and 0.221 (BLVRA), i.e. not significant (Figure 6B and C). This suggests that although the single genes are not differentially expressed, their interactions may have changed significantly. From the joint distribution (Figure 6D), we can see that in healthy cells there is a strong correlation between COMT and BLVRA expression (Pearson's correlation = 0.903), while in severe cells this correlation is broken (Pearson's correlation = 0.450). The protein encoded by BLVRA can catalyze the conversion of biliverdin (BLV) to bilirubin (BR) in the presence of NADPH or NADH. Furthermore, BLV can inhibit the binding of SARS-Cov-2 to immune serum [39]. Perhaps studying the relationship between COMT and BLVRA can be helpful for revealing the pathogenesis of COVID-19 disease.

Breast cancer

We finally experiment on the TNBC dataset using our IBTA workflow. After annotation, the cell type composition is shown in Figure 7A, the most abundant of which are epithelial cells. The disease group composition in each cell type is also shown in Figure 7A. We selected epithelial cells as our target cell type to explore the differences between cancer and normal samples. We screened genes the same as in the experiment on COVID-19 and finally obtained a pAPC2 matrix with a dimension of 22,998 cells × 500 interactions.

We obtained similar results as before in the UMAP visualization (Figure 7B, C) as well as in the hierarchical clustering (Figure 7D, E). Once again, the pAPC2-based UMAP can be separated by a straight dashed line. Similar markers and cells are also closer to each other in the pAPC2 heatmap in the results of hierarchical clustering. It is worth noting that we also observed some hub genes, such as TYMP and SPP1. Capecitabine is a drug that can treat breast cancer and works by blocking DNA, RNA and protein synthesis and inhibiting cell division. At the same time, it has been shown that TYMP expression correlates with the efficacy of capecitabine in TNBC [40, 41]. SPP1, also called OPN, has been reported as a potential biomarker for anti-EGFR therapy in TNBC and the increased mutation of SPP1 in TNBC may also help for tailoring treatment in TNBC [42, 43].

We further conducted survival analysis based on both APC2 and gene expression. Because there is no survival data in the single-cell dataset, we here used bulk data of the GDC TCGA BRCA project and acquired expression profiles and survival data from the UCSC Xena platform [44]. Through the above analysis by our IBTA workflow, we observed 248 significant interaction

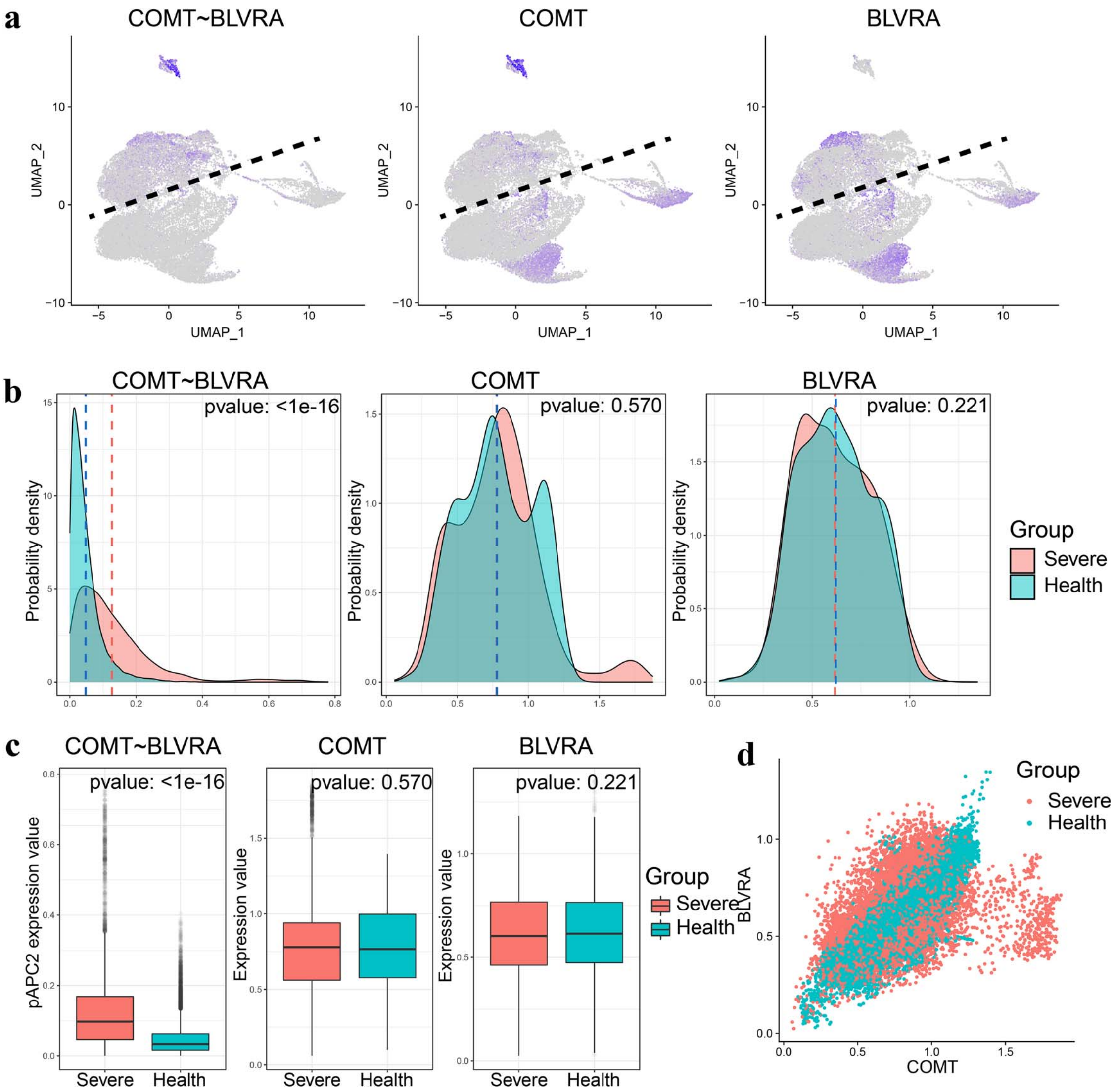


Figure 6. The expression visualization of COMT ∼ BLVRA and corresponding genes. **(A)** The UMAP visualization of COMT∼BLVRA pAPC2 values, COMT and BLVRA expression values. **(B)** The distribution of COMT∼BLVRA pAPC2 values, COMT and BLVRA expression values. **(C)** The boxplot of COMT∼BLVRA pAPC2 values, COMT and BLVRA expression values. **(D)** The joint distribution of COMT and BLVRA expression values colored by groups.

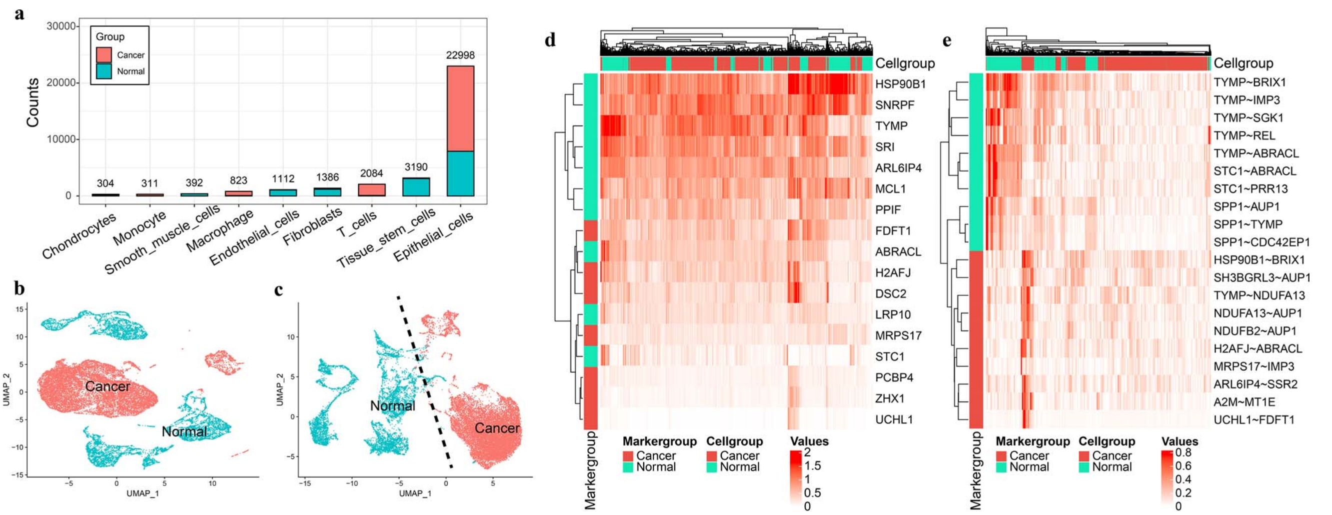


Figure 7. The IBTA results of the TNBC dataset. **(A)** The proportion of groups in each cell type. **(B)** The UMAP visualization of cells by gene features. **(C)** The UMAP visualization of cells by pAPC2 features. The straight dashed line can separate two groups well. **(D)** The heatmap of the top 10 gene markers of each group (only 17 available). **(E)** The heatmap of the top 10 interaction markers of each group.

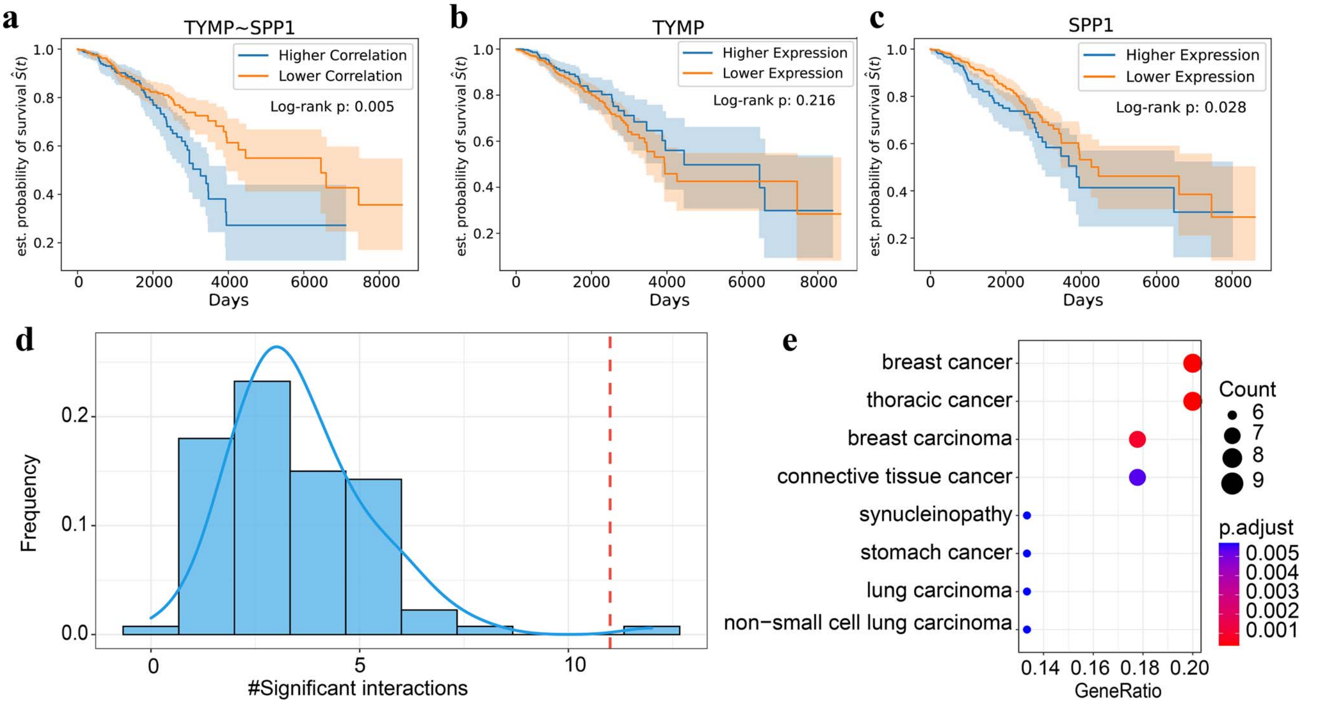


Figure 8. The survival analysis and DO analysis results of the TNBC dataset. **(A)** Kaplan–Meier survival curves of interaction TYMP∼SPP1. The x-axis is time in days. The y-axis is survival probability. **(B)** Kaplan–Meier survival curves of SPP1. **(C)** Kaplan–Meier survival curves of TYMP. **(D)** The distribution of the permutation test for interaction markers with significant survival analysis results. The red dashed line is the number of significant interactions in our markers. **(E)** The DO enrichment analysis of genes appeared in our interaction markers. 'Breast cancer' appears in the first position.

markers. Therefore, we calculated APC2 of these 248 interactions for all bulk samples. The top 33% of the samples with higher APC2 were divided into the low correlation group, and the bottom 33% of the samples with lower APC2 were divided into the high correlation group. We also conducted survival analysis based on single gene expression of all candidate genes. For example, the paired-gene survival analysis based on APC2 of TYMP∼SPP1 is shown in Figure 8A, and the single-gene survival analysis based on gene expression of TYMP and SPP1 is shown in Figure 8B and C, respectively. The results showed that the expression level of SPP1 is significant for overall survival time (log-rank *P*-value = 0.028), while TYMP is not significant (log-rank *P*-value = 0.216). But for paired-gene survival analysis, APC2 of TYMP∼SPP1 is much more significant (log-rank *P*-value = 0.005), which indicates we can easily find useful information from the interaction perspective. We also examined the enrichment levels of significant interactions and genes in our interaction markers. For the single-gene level, there are 101 genes in 248 interaction markers, and 7 out of 101 are significantly associated with patients' overall survival time. The *P*-value of the corrected chi-squared test is 0.04428. This means that the genes presented in our interaction markers are strongly associated with the significance level of survival analysis. For the interaction level, 11 of 248 interactions are significant. We conducted a permutation test by randomly selecting 248 interactions from 500 candidate interactions and repeated 100 times. The distribution of significant interaction numbers is shown in Figure 8D, and the *P*-value (the probability of the significant interaction number greater than or equal to 11 in the permutation test) is 0.0099. This implies that the interaction markers found by our method are strongly associated with the significance level of the paired-gene survival analysis. Finally, the disease ontology (DO) [45] analysis shows the 101 genes in 248 interaction markers are successfully enriched in the breast cancer term with an adjusted *P*-value of 0.001, ranking first (Figure 8E). This underlines once more how important IBTA is.

Conclusion and discussion

In this study, we proposed the CDN method for multi-source differential network inference and compared it with other popular differential network inference methods. The results of the simulation experiments show the superiority of the new method. We also proposed a novel metric, pivot APC2, which can assign values to each sample for each interaction. The main contribution of this paper is that we introduced the novel interaction-based transcriptome analysis (IBTA) workflow. By transforming the gene expression data to pivot APC2, many standard approaches for gene-based transcriptome analysis can be straightforwardly applied to conduct interaction-based analysis.

As a simple and straightforward method, APC2 may be affected by batch effect. It is worth noting that the global batch effect is spread out over every gene, while APC2 is separately calculated on each gene pair. The distance between batches is greatly reduced if only two genes are considered; therefore, the influence of batch effect is minor when calculating APC2 for most gene pairs. Of course, if possible, it is better to utilize batch correction methods to preprocess the data. Additionally, the way of APC2 to identify differential interactions is linear; however, the interactions between gene expression are usually nonlinear. Therefore, a more robust method capable of distinguishing gene co-expression patterns is urgently needed.

Overall, this workflow allows researchers to explore the key gene interactions in their datasets and obtain many meaningful and important marker interactions. These markers may help researchers design further experiments and perform validation efficiently.

Key Points

- IBTA is a transcriptome analysis framework based on gene interactions that can re-explain transcriptome data and reveal many valuables that single-gene analysis methods cannot. It consists of two components: CDN and pAPC2.
- CDN is a multi-source differential network inference method. Extensive simulation experiments demonstrated that CDNs are more capable of inferring differential network structures with hubs from multi-source data.
- pAPC2 is a novel metric for calculating the pseudo-expression value of gene interactions. It can assign each sample a pseudo-expression value according to the contribution to the interaction.
- IBTA has shown strong analytical capabilities in the case studies of all three datasets from the CRC, COVID-19 and TNBC, both between and within cell types.

Data available statement

The CRC data in this article are available in ArrayExpress repository at https://www.ebi.ac.uk/biostudies/arrayexpress, and can be accessed with E-MATB-8410. The TNBC and COVID-19 data are available in Gene Expression Omnibus repository at https://www.ncbi.nlm.nih.gov/geo, and can be accessed with GSE145926 and GSE161529. The bulk data about GDC TCGA BRCA project are available at https://xenabrowser.net/datapages.

Funding

National Key Research and Development Program of China (grant no. 2020YFA0712402); National Natural Science Foundation of China (grant no. 12231018).

References

1. Anders S, Huber W. Differential expression analysis for sequence count data. *Nat Preced* 2010;**2010**:1–1.
2. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
3. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;**14**:1–18.
4. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* 2018;**15**: 255–61.
5. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003;**4**:1–10.
6. Meng Y, Huang Y, Chang X, *et al.* Transcriptome analysis method based on differential distribution evaluation. *Brief Bioinform* 2022;**23**:bbab608.
7. de la Fuente A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet* 2010;**26**:326–33.
8. Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes. *Bioinformatics* 2004;**20**:i194–9.
9. Lai Y, Wu B, Chen L, *et al.* A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics* 2004;**20**:3146–55.
10. Zhao H, Duan Z-H. Cancer genetic network inference using Gaussian graphical models. *Bioinform Biol Insights* 2019;**13**:1177932219839402.
11. Wang J, Cheung LWK, Delabie J. New probabilistic graphical models for genetic regulatory networks studies. *J Biomed Inform* 2005;**38**:443–55.
12. Højsgaard S, Edwards D, Lauritzen S. Gaussian graphical models. In: Højsgaard S, Edwards D, Lauritzen S, eds. *Graphical Models with* R. Boston, MA: Springer US; 2012: 77–116.
13. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc Ser B Stat Methodol* 2014;**76**:373–97.
14. Mohan K, London P, Fazel M, *et al.* Node-based learning of multiple Gaussian graphical models. *J Mach Learn Res* 2014;**15**: 445–88.
15. Yuan H, Xi R, Chen C, *et al.* Differential network analysis via lasso penalized D-trace loss. *Biometrika* 2017;**104**:755–70.
16. Thorne T. NetDiff – Bayesian model selection for differential gene regulatory network inference. *Sci Reports* 2016;**6**:1–9.
17. Nawy T. Single-cell sequencing. *Nat Methods* 2013;**11**: 18–18.
18. Kashima Y, Sakamoto Y, Kaneko K, *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* 2020;**52**:1419–27.
19. Wang B, Singh R, Qi Y. A constrained ℓ 1 minimization approach for estimating multiple sparse Gaussian or nonparanormal graphical models. *Mach Learn* 2017;**106**:1381–417.
20. Zhang XF, Ou-Yang L, Yan H. Incorporating prior information into differential network analysis using non-paranormal graphical models. *Bioinformatics* 2017;**33**:2436–45.
21. Zhang X-F, Ou-Yang L, Zhao X-M, *et al.* Differential network analysis from cross-platform gene expression data. *Sci Reports* 2016;**6**:1–12.
22. Zhang XF, Ou-Yang L, Yan T, *et al.* A joint graphical model for inferring gene networks across multiple subpopulations and data types. *IEEE Trans Cybern* 2021;**51**:1043–55.
23. Leng J, Wu L-Y. Importance-penalized joint graphical lasso (IPJGL): Differential network inference via GGMs. *Bioinformatics* 2022;**38**:770–77.
24. Nabavi S, Beck AH. Earth mover's distance for differential analysis of heterogeneous genomics data. *2015 IEEE Glob Conf Signal Inf Process Glob* 2015;**2016**:963–6.
25. Xu Y, Liu M, Lin Q, *et al.* ADMM without a fixed penalty parameter: Faster convergence with new adaptive penalization. *Adv Neural Inf Process Syst* 2017;**30**:1268–78.
26. Albert R, Barabási AL. Statistical mechanics of complex networks. *Rev Mod Phys* 2002;**74**:47–97.
27. Bollobás B. *Random Graphs*. In: *Mod. Graph Theory*. New York, NY: Springer; 1998: 215–52.
28. Lee HO, Hong Y, Etlioglu HE, *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet* 2020;**52**:594–603.
29. Liao M, Liu Y, Yuan J, *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* 2020;**26**: 842–4.
30. Pal B, Chen Y, Vaillant F, *et al.* A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J* 2021;**40**:e107333.
31. van Dijk D, Sharma R, Nainys J, *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716–729.e27.

32. Aran D, Looney AP, Liu L, *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**:163–72.
33. Patel A, Tripathi G, McTernan P, *et al.* Fibroblast growth factor 7 signalling is disrupted in colorectal cancer and is a potential marker of field cancerisation. *J Gastrointest Oncol* 2019;**10**: 429.
34. Yokoyama Y, Watanabe T, Tamura Y, *et al.* Autocrine BMP-4 signaling is a therapeutic target in colorectal cancer. *Cancer Res* 2017;**77**:4026–38.
35. Jiang X, Wang J, Wang M, *et al.* ITGB4 as a novel serum diagnosis biomarker and potential therapeutic target for colorectal cancer. *Cancer Med* 2021;**10**:6823–34.
36. Li M, Jiang X, Wang G, *et al.* ITGB4 is a novel prognostic factor in colon cancer. *J Cancer* 2019;**10**:5223–33.
37. Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nat* 2020;**583**:459–68.
38. He S, Waheed AA, Hetrick B, *et al.* PSGL-1 inhibits the incorporation of SARS-CoV and SARS-CoV-2 Spike glycoproteins into pseudovirions and impairs pseudovirus attachment and infectivity. *Viruses* 2021;**13**:46.
39. Rosa A, Pye VE, Graham C, *et al.* SARS-CoV-2 can recruit a heme metabolite to evade antibody immunity. *Sci Adv* 2021;**7**:17.
40. Marangoni E, Laurent C, Coussy F, *et al.* Capecitabine efficacy is correlated with tymp and rb1 expression in pdx established from triple-negative breast cancers. *Clin Cancer Res* 2018;**24**: 2605–15.
41. Asleh K, Brauer HA, Sullivan A, *et al.* Predictive biomarkers for adjuvant capecitabine benefit in early-stage triple-negative breast cancer in the FinXX clinical trial. *Clin Cancer Res* 2020;**26**: 2603–14.
42. Anborgh PH, Lee DJ, Stam PF, *et al.* Role of osteopontin as a predictive biomarker for anti-EGFR therapy in triple-negative breast cancer. *Expert Opin Ther* 2018;**22**:727–34.
43. Elbaiomy MA, El-Ghonemy MS, Elhelaly R, *et al.* Osteopontin level and promoter polymorphism is associated with aggressiveness in breast cancer. *Ann Oncol* 2018;**29**:ix15.
44. Goldman MJ, Craft B, Hastie M, *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020;**38**:675–8.
45. Yu G, Wang LG, Yan GR, *et al.* DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2015;**31**:608–9.