Bioinformatics, 37(23), 2021, 4414–4423
doi: 10.1093/bioinformatics/btab502
Advance Access Publication Date: 10 July 2021
Original Paper

Gene expression

Differential network analysis by simultaneously considering changes in gene interactions and gene expression

Jia-Juan Tu[1], Le Ou-Yang[2], Yuan Zhu [3,4], Hong Yan[5], Hong Qin[6] and Xiao-Fei Zhang [1,*]

[1]School of Mathematics and Statistics and Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan 430079, China, [2]College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China, [3]School of Automation, China University of Geosciences, Wuhan 430074, China, [4]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, China University of Geosciences, Wuhan 430074, China, [5]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China and [6]Department of Statistics, Zhongnan University of Economics and Law, Wuhan 430073, China

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on January 14, 2021; revised on June 13, 2021; editorial decision on July 2, 2021; accepted on July 5, 2021

Abstract

Motivation: Differential network analysis is an important tool to investigate the rewiring of gene interactions under different conditions. Several computational methods have been developed to estimate differential networks from gene expression data, but most of them do not consider that gene network rewiring may be driven by the differential expression of individual genes. New differential network analysis methods that simultaneously take account of the changes in gene interactions and changes in expression levels are needed.

Results: : In this article, we propose a differential network analysis method that considers the differential expression of individual genes when identifying differential edges. First, two hypothesis test statistics are used to quantify changes in partial correlations between gene pairs and changes in expression levels for individual genes. Then, an optimization framework is proposed to combine the two test statistics so that the resulting differential network has a hierarchical property, where a differential edge can be considered only if at least one of the two involved genes is differentially expressed. Simulation results indicate that our method outperforms current state-of-the-art methods. We apply our method to identify the differential networks between the luminal A and basal-like subtypes of breast cancer and those between acute myeloid leukemia and normal samples. Hub nodes in the differential networks estimated by our method, including both differentially and nondifferentially expressed genes, have important biological functions.

Availability and implementation: All the datasets underlying this article are publicly available. Processed data and source code can be accessed through the Github repository at https://github.com/Zhangxf-ccnu/chNet.

Contact: zhangxf@mail.ccnu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological processes (BPs) often involve complicated, network-like interactions between genes. Gene dependency networks, where nodes represent genes and edges represent their functional dependencies, can provide key insights into the mechanisms of diseases (Barabási *et al.*, 2011; Schadt *et al.*, 2009). These networks often change under different conditions (Ha *et al.*, 2015; Ou-Yang *et al.*, 2014). As the development and progression of complex diseases are associated with the rewiring of gene networks (Schadt *et al.*, 2009), identifying the rewiring pattern of gene networks between two conditions could reveal the underlying biological mechanisms (Liu *et al.*, 2017; Yuan *et al.*, 2017; Zhang *et al.*, 2018, 2019a). In previous studies, a variety of differential expression analysis methods have been developed to identify a set of genes of which the expression levels are changed between two different conditions (Robinson *et al.*, 2010). However, differential expression analysis considers each gene separately and cannot characterize the change of gene interactions under different conditions. Recently, the focus of differential analysis research has shifted from differential expression analysis to differential network analysis that characterize gene network rewiring patterns. Unlike differential expression analysis that identifies a set of differentially expressed genes, differential network analysis aims to infer a differential network from gene expression data, where nodes represent genes and edges represent gene interactions that are changed between two conditions (Ha *et al.*, 2015; Ideker *et al.*, 2012; Liu *et al.*, 2017; Tan *et al.*, 2020; Yuan *et al.*, 2017; Zhang *et al.*, 2016, 2017, 2018).

Several Gaussian Graphical Model (GGM)-based methods have been proposed to infer differential networks from gene expression data, on the assumption that gene expression measurements follow multivariate Gaussian distributions (Ha *et al.*, 2015; Liu *et al.*, 2017; Yuan *et al.*, 2017; Zhang *et al.*, 2016, 2017, 2018). GGMs have an advantage over marginal correlation-based methods in distinguishing direct and indirect interactions by borrowing from the strength of conditional dependencies between genes (Meinshausen *et al.*, 2006; Yuan *et al.*, 2007), which can be captured directly by the precision matrix (the inverse of the covariance matrix). Partial correlations are determined by precision matrices and conditional variances, and, even if they do not vary across conditions, precision matrices may change if the conditional variances of individual genes change. This could lead to differential networks defined by differences in precision matrices that include false differential edges caused by variants of conditional variances (Liu *et al.*, 2017; Zhang *et al.*, 2019a). To deal with this problem, several methods have defined differential networks as the difference of partial correlations across different conditions (Liu *et al.*, 2017; Tan *et al.*, 2020; Zhang *et al.*, 2019a).

In biology, mutations of a gene can alter its expression level and functional relationships with other genes (Bashashati *et al.*, 2012; Grechkin *et al.*, 2016). Several recent studies have shown that gene network rewiring may be driven by certain perturbed genes that are mutated or differentially expressed across conditions (Bashashati *et al.*, 2012; Mohan *et al.*, 2014). It is reasonable to assume that differentially expressed genes are more likely to lead to differential interactions than nondifferentially expressed genes, and also that the interactions corresponding to differentially expressed genes may be in some sense of more practical importance. In fact, if an interaction is detected as a differential interaction by a method but neither of the two involved genes is differentially expressed, it will be difficult to interpret why the interaction is rewired. Therefore, a natural assumption is that a change in the interaction between two genes is derived by a change in the expression level of at least one of the two involved genes (e.g. genes 3 and 5, or genes 4 and 6 in Fig. 1B). As a result, a reasonable differential network should satisfy the hierarchical constraints that a differential edge can be considered only if at least one of the two involved genes is differentially expressed. From a statistical perspective, when the number of genes is large, many potential pairwise interactions make it difficult to identify the true differential interactions. Under the hierarchical constraints, the procedure only analyzes interactions involved with differentially expressed genes instead of all possible interactions, which may improve the statistical power (Bien *et al.*, 2013, 2015; Lim *et al.*, 2015). Most previously developed differential network analysis methods focus on searching all possible interactions, and do not take into account changes in expression levels of individual genes (Liu *et al.*, 2017; Tan *et al.*, 2020; Zhang *et al.*, 2016, 2017, 2018, 2019a). Therefore, the differential networks identified by these methods may not satisfy the hierarchical constraints, and may be less interpretable and powerful (Bien *et al.*, 2013, 2015; Lim *et al.*, 2015).

In this paper, we propose a new hierarCHical differential NETwork analysis model (chNet) to estimate differential networks that satisfy the hierarchical constraints (Fig. 1). A differential network is defined as the difference of partial correlations between two conditions (Fig. 1A), and a new test statistic to quantify the change of partial correlations between gene pairs is developed (Fig. 1B). The Student's *t*-test statistic is used to identify changes in the expression levels of genes between two conditions (Fig. 1C). An optimization model is proposed to combine the two types of test statistics to produce differential networks that exhibit the hierarchical structures (Fig. 1D). A closed-formed solution is derived to solve the optimization model. In addition, based on a subsampling approach, a weighted hierarchical differential network can also be inferred. Simulation experiments show that our method outperforms the state-of-the-art ones. We also apply chNet to gene expression data from breast cancer and acute myeloid leukemia (AML), and investigate the differential networks estimated by our method by assessing the biological significance of the hub nodes. The results show that the hub nodes in the estimated differential networks, including both differentially and nondifferentially expressed genes, play critical roles in cancers.

2 Materials and methods

2.1 Problem formulation

Supposed that we are given gene expression datasets corresponding to two different conditions, $X^{(1)} = \left(\left(x_1^{(1)}\right)^T, \ldots, \left(x_{n_1}^{(1)}\right)^T\right)^T$ and $X^{(2)} = \left(\left(x_1^{(2)}\right)^T, \ldots, \left(x_{n_2}^{(2)}\right)^T\right)^T$, where $x_\ell^{(c)} = \left(x_{\ell 1}^{(c)}, \ldots, x_{\ell p}^{(c)}\right)^T$ is gene expression level of sample ℓ on p genes for $\ell = 1, \ldots, n_c$, and $c = 1, 2$ represents the condition (Fig. 1A). To estimate a differential network between two conditions, we assume that the gene expression datasets are sampled from two different multivariate normal distributions, i.e. $x_1^{(c)}, \ldots, x_{n_c}^{(c)} \sim N_p\left(\mu^{(c)}, \Sigma^{(c)}\right)$, where $\mu^{(c)}$ is the mean vector and $\Sigma^{(c)}$ is the covariance matrix for $c = 1, 2$. Let $\Omega^{(c)} = \left(\omega_{ij}^{(c)}\right) = \left(\Sigma^{(c)}\right)^{-1}$ be the precision matrix, and $\rho_{ij}^{(c)}$ be the partial correlation between two genes i and j given the other $p - 2$ genes, where $\rho_{ij}^{(c)} = -\omega_{ij}^{(c)} / \sqrt{\omega_{ii}^{(c)} \omega_{jj}^{(c)}}$. Then the differential network is defined as the difference of partial correlations between the two conditions, e.g. $\theta_{ij} = \rho_{ij}^{(1)} - \rho_{ij}^{(2)}$. There is a differential edge between genes i and j if and only if $\theta_{ij} \neq 0$ (Fig. 1B).

As mentioned above, rewiring of gene interactions may be driven by the differential expression of the involved genes. In this study, we define the differentially expressed genes based on the differences in mean expression levels between two conditions, e.g. $\phi_i = \mu_i^{(1)} - \mu_i^{(2)}$. Gene i is considered as a differentially expressed gene if $\phi_i \neq 0$ (Fig. 1C). The hierarchical constraints are attempted to impose on the resulting differential network so that an edge is allowed in the resulting differential network only if at least one of the two involved genes is differentially expressed. That is, $\theta_{ij} \neq 0$ implies $\phi_i \neq 0$ or $\phi_j \neq 0$. Therefore, given gene expression data corresponding to two conditions, the goal of this study is to estimate θ_{ij} and ϕ_i that satisfy the hierarchical constraints, for which at least one of ϕ_i and ϕ_j is nonzero if θ_{ij} is nonzero (Fig. 1D).

2.2 Test statistics for quantifying changes in gene interactions

Following the method of our previous study (Zhang *et al.*, 2019a), identification of differential edges is cast as a statistical hypothesis test of partial correlations:

$$H_{0,ij}^{(1)} : \rho_{ij}^{(1)} = \rho_{ij}^{(2)} \quad \text{versus} \quad H_{1,ij}^{(1)} : \rho_{ij}^{(1)} \neq \rho_{ij}^{(2)}, \text{for } 1 \leq i < j \leq p. \tag{1}$$

There will be a differential edge between genes i and j if $H_{0,ij}^{(1)}$ is rejected. The test statistic proposed by Liu *et al.* (2017) is used to quantify changes in partial correlations between two genes across the conditions. A linear regression method is used to compute the estimator of partial correlation for each condition as

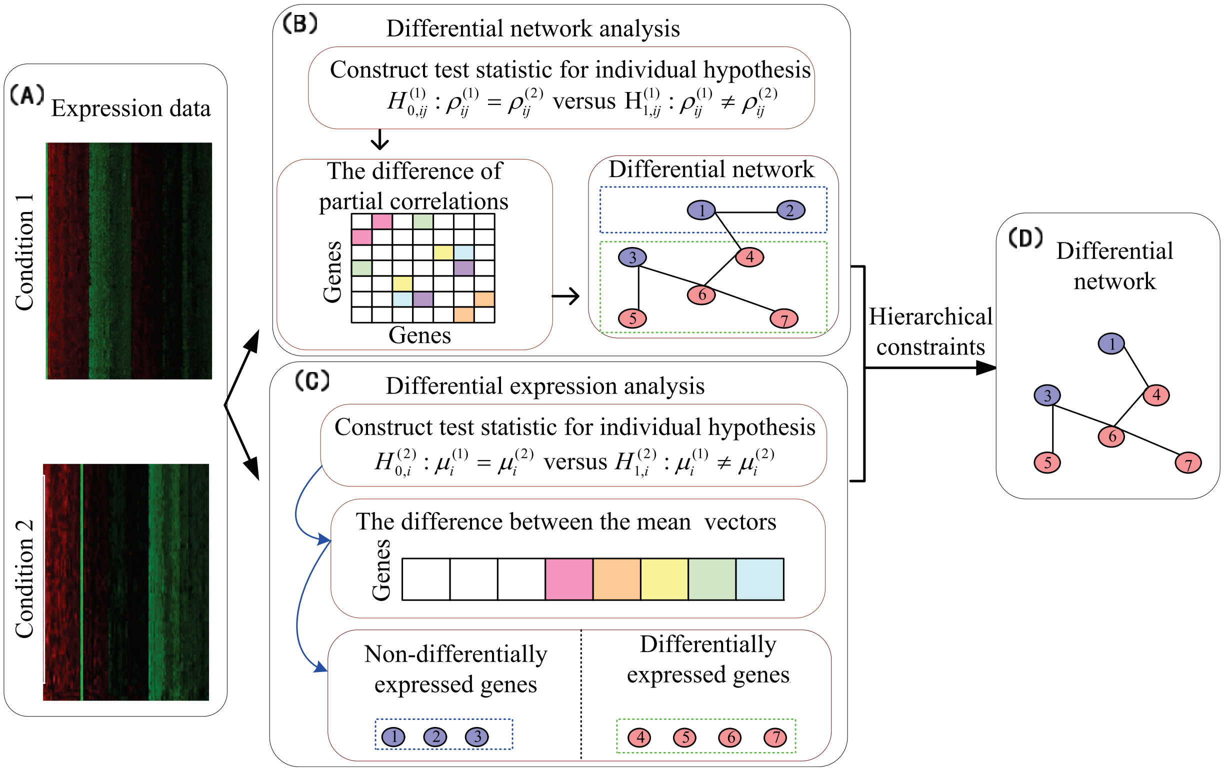
(A) Expression data
Condition 1
Condition 2
(B) Differential network analysis
Construct test statistic for individual hypothesis
$H_{0,ij}^{(1)}: \rho_{ij}^{(1)} = \rho_{ij}^{(2)}$ versus $H_{1,ij}^{(1)}: \rho_{ij}^{(1)} \neq \rho_{ij}^{(2)}$
The difference of partial correlations
Genes
Genes
Differential network
1
2
3
4
5
6
7
Hierarchical constraints
(C) Differential expression analysis
Construct test statistic for individual hypothesis
$H_{0,i}^{(2)}: \mu_i^{(1)} = \mu_i^{(2)}$ versus $H_{1,i}^{(2)}: \mu_i^{(1)} \neq \mu_i^{(2)}$
The difference between the mean vectors
Genes
Non-differentially expressed genes
Differentially expressed genes
1 2 3
4 5 6 7
(D) Differential network
1
3
4
5
6
7

Fig. 1. Overview of chNet. (**A**) The input of chNet is gene expression data from two different conditions. (**B**) The differential network is defined as the changes in partial correlations between gene pairs under different conditions and use a statistical hypothesis test over partial correlations to identify the differential edges. There may exist differential edges connected by two nondifferentially expressed genes if the changes in expression levels of individual genes are not taken into account (e.g. the detected differential edge between genes 1 and 2 highlighted in the purple box). (**C**) The differentially expressed genes are defined as the changes in expression levels of genes. Identification of differentially expressed genes is equivalent to be a statistical hypothesis testing problem on sample means. Nondifferentially and differentially expressed genes are shown in purple and red, respectively. (**D**) We propose a differential network analysis model by simultaneously considering changes in gene interaction and gene expression. Specially, an optimization model is used to combine the test statistics that quantify the changes in partial correlations between genes and the test statistics that quantify the changes in expression levels of individual genes to produce a differential network that satisfies the hierarchical constraints. In doing so, at least one of the two involved genes is differentially expressed for all differential edges in the resulting networks, and significant differences in partial correlations that do not obey the hierarchical constraints [e.g. the detected interaction between genes 1 and 2 presented in (B)] will be excluded

$$t_{ij}^{(c)} = \sqrt{\frac{1}{\tilde{r}_{ii}^{(c)} \tilde{r}_{jj}^{(c)}}}\hat{r}_{ij}^{(c)}, 1 \le i < j \le p,$$

where $\tilde{r}_{ii}^{(c)}$ and $\tilde{r}_{jj}^{(c)}$ are the sample variances of the residuals of the regression models, and $\hat{r}_{ij}^{(c)}$ is a bias-corrected estimator of the covariance of the error terms. The variance of $t_{ij}^{(c)}$ is calculated as

$$var\left(t_{ij}^{(c)}\right) = \frac{1}{n_c}\left(1 - \hat{\rho}_{ij}^{(c)}\right)^2,$$

where $\hat{\rho}_{ij}^{(c)}$ is a partial correlation estimated by a threshold operation, $\hat{\rho}_{ij}^{(c)} = t_{ij}^{(c)} 1\left\{|t_{ij}^{(c)}| \ge 2\sqrt{\frac{\log p}{n_c}}\right\}$, and 1 is the indicator function. The test statistic to quantify changes in partial correlations between two conditions is

$$t_{ij} = \frac{t_{ij}^{(1)} - t_{ij}^{(2)}}{\sqrt{var\left(t_{ij}^{(1)}\right) + var\left(t_{ij}^{(2)}\right)}}, 1 \le i < j \le p. \qquad (2)$$

The details to compute this test statistics t_{ij} ($1 \le i < j \le p$) are provided in Supplementary Section S3.1.

Under $H_{0,ij}^{(1)}$, the test statistic t_{ij} approximately follows a standard normal distribution, i.e. $t_{ij} \to N(0, 1)$ (Liu *et al.*, 2017). The null hypotheses $H_{0,ij}^{(1)}$ can be rejected if $|t_{ij}| > \lambda$ at a given threshold $\lambda > 0$. The set of differential edges can be determined by the set $\{(i, j) : |t_{ij}| > \lambda\}$.

2.3 Test statistics to quantify changes in gene expression levels

Changes in gene expression levels are evaluated by the statistical hypothesis test

$$H_{0,i}^{(2)} : \mu_i^{(1)} = \mu_i^{(2)} \quad \text{versus} \quad H_{1,i}^{(2)} : \mu_i^{(1)} \neq \mu_i^{(2)}, \quad \text{for } i = 1, \ldots, p. \quad (3)$$

Gene i is considered to be differentially expressed if $H_{0,i}^{(2)}$ is rejected. We use the Student's t-test statistics to test $H_{0,i}^{(2)}$

$$z_i = \frac{\bar{x}_i^{(1)} - \bar{x}_i^{(2)}}{\sqrt{(S_i^{(1)})^2/n_1 + (S_i^{(2)})^2/n_2}}, 1 \le i \le p, \qquad (4)$$

where $\bar{x}_i^{(c)} = \frac{1}{n_c}\sum_{\ell=1}^{n_c} x_{\ell i}^{(c)}$ and $S_i^{(c)} = \frac{1}{n_c - 1}\sum_{\ell=1}^{n_c}\left(x_{\ell i}^{(c)} - \bar{x}_i^{(c)}\right)^2$ are the sample mean and variance for gene i in condition c.

Under $H_{0,i}^{(2)}$, the test statistic z_i approximately follows a standard normal distribution, i.e. $z_i \to N(0, 1)$. The null hypotheses $H_{0,i}^{(2)}$ can be rejected if $|z_i| > \lambda$ for a given threshold $\lambda > 0$. The set of differential genes can be determined by the set $\{i : |z_i| > \lambda\}$.

2.4 Hierarchical differential network analysis model

Indeed, we can determine the set of differential edges only using the test statistics t_{ij} following previous studies (Liu *et al.*, 2017; Zhang *et al.*, 2019a). However, the differences in expression levels of individual genes will be ignored and the resulting differential network will not satisfy the hierarchical constraints. To improve the interpretability and accuracy of the resulting differential networks, we propose a testing procedure, through an optimization problem involving both t_{ij} and z_i, to produce a differential network that obey the hierarchical constraints. Let $\theta = (\theta_{ij})$ and $\phi = (\phi_i)$ be optimization variables that are associated with t_{ij} and z_i, respectively. Based on the hierarchical network model proposed by Bien *et al.* (2015), we develop a hierarCHical differential NETwork analysis model (chNet) to impose hierarchical constraints on the resulting differential network:

$$\begin{aligned} \min\nolimits_{\theta,\phi} \quad & \frac{1}{2}\sum_{i=1}^{p}\sum_{j\neq i}(t_{ij}-\theta_{ij})^2 + \frac{1}{2}\sum_{i=1}^{p}(z_i-\phi_i)^2 + \lambda\sum_{i=1}^{p}\sum_{j\neq i}|\theta_{ij}| + \lambda\sum_{i=1}^{p}|\phi_i| \\ s.t. \quad & \sum_{j=1}^{p}|\theta_{ij}| \le |\phi_i|, \text{ for } i = 1,\ldots,p. \end{aligned} \tag{5}$$

Here, the first term of the objective function quantifies the loss between the optimization variables (θ_{ij} and ϕ_i) and the corresponding statistics (t_{ij} and z_i). The second term imposes sparsity on the resulting differential network and the set of differentially expressed genes, with λ being a tuning parameter that controls the level of sparsity. λ is associated with the threshold parameter used to reject the null hypotheses $H_{0,ij}^{(1)}$ and $H_{0,i}^{(2)}$. A large value of λ will produce a sparse differential network and a few differentially expressed genes. The constraint $\sum_{j=1}^{p}|\theta_{ij}| \le |\phi_i|$ is used to relate θ_{ij} and ϕ_i and to exploit hierarchy. Since $|\theta_{ij}| \le \sum_{j=1}^{p}|\theta_{ij}|$, the constraint $\sum_{j=1}^{p}|\theta_{ij}| \le |\phi_i|$ implies that ϕ_i must be nonzero in order for θ_{ij} to be nonzero. In other words, $\theta_{ij} \neq 0$ implies $\phi_i \neq 0$, and similarly for θ_{ji}. Thus, $\theta_{ij} \neq 0$ implies at least one of ϕ_i and ϕ_j is nonzero, indicating that if there is a differential edge between genes *i* and *j*, at least one of genes *i* and *j* is a differentially expressed gene. Thus, chNet simultaneously considers changes in partial correlations between gene pairs and changes in gene expression levels with a new constraint to produce a differential network that satisfies the hierarchy constraints.

Due to the constraints $\sum_{j=1}^{p}|\theta_{ij}| \le |\phi_i|$, the optimization problem (5) is not convex. Following (Bien *et al.*, 2015), we represent ϕ_i as the difference of two nonnegative quantities, ϕ_i^+ and ϕ_i^-, i.e. $\phi_i = \phi_i^+ - \phi_i^-$, in the optimization problem (5). Equation (5) is then reformulated as a convex optimization problem that can be solved by the method of Bien *et al.* (2015). Let the solution to problem (5) be $\hat{\phi}_i(\lambda)$and $\hat{\theta}_{ij}(\lambda)$, which are functions of λ. Let $\hat{\lambda}_i$ and $\hat{\lambda}_{ij}$ are the maximal value of λ values at which $\hat{\phi}_i(\lambda)$ and $\hat{\theta}_{ij}(\lambda)$ become nonzero, that is $\hat{\lambda}_i = \sup\{\lambda \ge 0 : \hat{\phi}_i(\lambda) \neq 0\}$ and $\hat{\lambda}_{ij} = \sup\{\lambda \ge 0 : \hat{\theta}_{ij}(\lambda) \neq 0\}$. Bien *et al.* (2015) have shown that $\hat{\lambda}_i$ and $\hat{\lambda}_{ij}$ can be computed as

$$\hat{\lambda}_i = \max\left\{|z_i|, \frac{|z_i| + \max_j\{|t_{ij}| : j \neq i\}}{2}\right\}, \tag{6}$$

and

$$\hat{\lambda}_{ij} = \min\left\{|t_{ij}|, \frac{|t_{ij}|}{2} + \frac{[|z_i| - \sum_{j':|t_{ij'}|>|t_{ij}|}(|t_{ij'}| - |t_{ij}|)]_+}{2}\right\}, \tag{7}$$

where $[x]_+ = \max(x, 0)$. According to the definitions of $\hat{\lambda}_i$ and $\hat{\lambda}_{ij}$, for a given tuning parameter $\overline{\lambda}$, gene *i* will be considered as a differentially expressed gene if $\hat{\lambda}_i \ge \overline{\lambda}$, and the interaction between genes *i* and *j* will be considered as a differential edge if $\hat{\lambda}_{ij} \ge \overline{\lambda}$. Please note that $\hat{\lambda}_{ij}$ is not symmetric according to Equation (7). To produce a symmetric differential network, we define $\hat{\lambda}'_{ij} = \max\{\hat{\lambda}_{ij}, \hat{\lambda}_{ji}\}$ and determine the differential network based on $\hat{\lambda}'_{ij}$. That is, we determine the set of differential edges by $\{(i,j) : \hat{\lambda}'_{ij} \ge \overline{\lambda}\}$ and the set of differentially expressed genes by $\{i : \hat{\lambda}_i \ge \overline{\lambda}\}$. The estimated differential network can be represented as $\hat{\theta}_{ij} = 1(\hat{\lambda}'_{ij} \ge \overline{\lambda})$ and the estimated differentially expressed genes can be represented as $\hat{\phi}_i = 1(\hat{\lambda}_i \ge \overline{\lambda})$. The complete procedure for estimating the set of differential edges and the set of differentially expressed genes using chNet is presented in Algorithm 1.Note that our method is similar to that proposed by Bien *et al.* (2015) in the sense that methods first use test statistics to quantify the changes in correlations between genes and expression levels of individual genes across different conditions, and then combine the two types of test statistics to produce a differential network that satisfies the hierarchical constraints. The key difference between the two methods is that Bien *et al.* (2015) focus on changes in marginal correlations while our method attempts to identify changes in partial correlations. It is known that partial correlations have an advantage over marginal correlations in distinguishing direct interactions from indirect effects (Liu *et al.*, 2017; Zhang *et al.*, 2019a). Therefore, the differential networks estimated by our method may have fewer false differential edges caused by indirect effects than those estimated by the method proposed by Bien *et al.* (2015).

Algorithm 1 Complete procedure of chNet

- **Inputs:** Gene expression datasets corresponding to two different conditions, $X^{(1)}$ and $X^{(2)}$, the tuning (threshold) parameter $\overline{\lambda}$.
- **Output:** Estimated set of differential edges and set of differentially expressed genes.
- Compute test statistics t_{ij} according to Equation (2);
- Compute test statistics z_i according to Equation (4);
- Compute test statistics $\hat{\lambda}_i$ according to Equation (6);
- Compute test statistics $\hat{\lambda}_{ij}$ and $\hat{\lambda}'_{ij}$ according to Equation (7);
- Compute the set of differential edges by $\{(i,j) : \hat{\lambda}'_{ij} \ge \overline{\lambda}\}$ [i.e. $\hat{\theta}_{ij} = 1(\hat{\lambda}'_{ij} \ge \overline{\lambda})$] and the set of differentially expressed genes by $\{i : \hat{\lambda}_i \ge \overline{\lambda}\}$ [i.e. $\hat{\phi}_i = 1(\hat{\lambda}_i \ge \overline{\lambda})$].

2.5 Construction of weighted differential networks

The estimated differential network depends on the datasets used, and different sets of differential edges may be detected if the datasets are varied. Therefore, it is useful to generate a weighted differential

Algorithm 2 Complete procedure for estimating weighted differential networks

```
Step-1: Use Algorithm 1 to estimate a differential network $\hat{\theta}^0$ from $X^{(1)}$ and $X^{(2)}$.
Step-2: Determine the significance of each edge under random sampling:
    2-1: Subsampling $R$ datasets $D_1, \ldots, D_R$ from $X^{(1)}$ and $X^{(2)}$ without replacement, each with size $0.8n$.
    2-2: For each $D_r$, $r = 1, \ldots, R$, use Algorithm 1 to estimate a differential network $\hat{\theta}^r$.
    2-3: Compute the selection frequency of a pair of genes $(i, j)$ being connected in the estimated differential networks,
        $b_{ij} = \frac{1}{R}\sum_{r=1}^{R} 1(\hat{\theta}^r_{ij} \neq 0)$.
Step-3: Compute the final weighted differential network $\hat{\theta}$,
    $\hat{\theta}_{ij} = \hat{\theta}^0_{ij} * b_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, p$.
```

network, for which the edge weights can quantify the selection probabilities of differential edges with varied datasets, and the edges with high selection probabilities would be more likely to be true differential edges. We estimate the selection probabilities by the selection frequencies over differential networks estimated from different subsampled datasets (Li *et al.*, 2013; Liu *et al.*, 2010). We first sample R subsampled datasets from the total datasets with sample size $0.8n$, and then estimate the differential networks from each subsampled dataset, and finally compute the selection frequency for each edge over different subsampled datasets and use the frequencies to quantify the selection probabilities of edges in the differential network estimated from the total datasets. The complete procedure for estimate a weighted differential network is presented in Algorithm 2. In this study, we set the number of subsampled datasets, R, to 20 as the default value.

3 Simulation studies

Due to the lack of gold standard that represents the true differential network, the evaluation of the performance of differential network analysis methods on real data is a common challenge (Tian *et al.*, 2016). Therefore, we first carry out simulation experiments to assess the empirical performance of chNet in this section, and then apply it to real datasets and analyze the biological significance of the inferred differential networks in the next section. In simulation studies, we conduct two different experiments to assess the generalization of our method. In the first experiment, we simulate the scenario where the true differential network is driven by several hub genes that can be both differentially expressed and nondifferentially expressed, and evaluate the performance in terms of the accuracy of the estimated differential networks and the identified hub nodes. In the second one, we simulate the scenario where the differential edges in the true differential network are arisen separately and there is no significantly hub node in the true differential network, and we evaluate the performance only in terms of the accuracy of the estimated differential networks (Supplementary Section S3.3.2).

3.1 Data generation

We briefly present the procedure used to generate the simulated data in the first experiment. The sample sizes are set as $n_1 = n_2 = 200$ for each condition and the number of genes is set as $p = 100$, 200, 400. The top $0.1p$ genes are modeled as differentially expressed genes and the remains are modeled as nondifferentially expressed. To model genes which may play key roles in the changes of networks, four of the p genes, including three differentially expressed genes and one nondifferentially expressed gene, are generated as hub nodes in the differential network following the strategy illustrated by (Mohan *et al.*, 2014). That is, the true differential network is determined by the four hub genes. Conditional variances of the top $0.2p$ genes are changed between the two conditions (Zhang *et al.*, 2019a). Therefore, there are more changes in the precision matrices than those in the partial correlations. The true differential network is generated to guarantee that the hierarchical constraints are satisfied. Details of the generation of the simulation data are provided in Supplementary Section S3.3.1.

3.2 Simulation results

The empirical performance of chNet is evaluated by comparison with several competing methods: CHT (Bien *et al.*, 2015), Pcor (Zhang *et al.*, 2019a), Pmat (Zhang *et al.*, 2019a) and Dtrace (Yuan *et al.*, 2017). CHT is a hierarchical test-based method designed to estimate a differential network based on marginal correlations. The direct and indirect effects between genes may not be distinguished in the resulting differential networks. A comparison to CHT can show the advantage of our chNet in excluding false differential edges owing to indirect effects by using partial correlations. Pcor is a hypothesis test-based method that defines the differential network as the difference of partial correlations. However, Pcor does not consider changes in expression levels of individual genes and impose the hierarchical constraints on the resulting networks. A comparison to Pcor can show the gain of our chNet by taking into account the hierarchical constraints. Pmat and Dtrace are two methods that define the differential network as the difference of precision matrices and do not consider the hierarchical constraints. A comparison with Pmat and Dtrace can show the advantage of chNet by defining differential networks based on partial correlations and taking into account the hierarchical constraints.

Two types of metrics are used to measure the performance in this experiment. We first evaluate the accuracy of differential networks estimated by different methods by comparing them with the true differential network in terms of precision–recall tradeoff. Since we cannot obtain the true differential network in reality and we often investigate the estimated differential networks by analyzing the biological significance of the hub nodes, so we also measure the performance in terms of the successful detection of hub nodes in the true differential networks. Let $\hat{\theta}_{ij}$ be the (i, j)th entry of a given estimator $\hat{\theta}$ and θ_{ij} be the (i, j)th entry of the true θ. For a given method, precision, recall, the positive hub columns (PHC), and the true positive hub columns (TPHC) are calculated as

$$\text{precision} = \frac{\sum_{i<j} 1(\theta_{ij} \neq 0, \hat{\theta}_{ij} \neq 0)}{\sum_{i<j} 1(\hat{\theta}_{ij} \neq 0)},$$

$$\text{recall} = \frac{\sum_{i<j} 1(\theta_{ij} \neq 0, \hat{\theta}_{ij} \neq 0)}{\sum_{i<j} 1(\theta_{ij} \neq 0)},$$

$$\text{PHC} = \sum_{i=1}^{p} 1(w_i \geq t_s),$$

$$\text{TPHC} = \textstyle\sum_{i \in I_p} 1(w_i \geq t_s).$$

Here, $w_i = \sum_{j=1}^{p} \hat{\theta}_{ij}$ can be considered as the weighted degree of nodes in the estimated networks and $\hat{\theta}_{ij}$ is the edge weight computed using Algorithm 2, and t_s is a threshold to define a node as hub, and we set $t_s = \mu + 3\sigma$ in a similar way to Mohan *et al.* (2014), where μ is the mean and σ is the standard deviation of $\{w_i\}_{i=1}^{p}$, and I_p is the set of the true hub genes in the true differential network. A node i in the estimated differential network will be considered as a hub if $w_i \geq t_s$. For a fixed number of detected differential edges, the ratio of PHC to TPHC (TPHC/PHC) can be used to measure the proportion of correctly identified hub nodes in the estimated differential network. In addition, the precision–recall curve is used to quantify the recovery of differential edges as a function of the tuning parameter.

Figure 2 shows precision against recall in the first row and TPHC/PHC against the number of detected differential edges in the second row (averaged over 20 random generations of the data). Within each plot, each line corresponds to the results obtained from different values of tuning parameters (e.g. $\bar{\lambda}$ for chNet) that control the sparsity level of the estimated differential networks. chNet obtains a higher precision than the other methods for a fixed value of recall (the first row), and it correctly identifies more hub nodes than the compared methods for a fixed number of detected differential edges (the second row) (Fig. 2). The advantage of chNet over CHT may be partially due to the fact chNet can distinguish between direct and indirect effects. Pcor, Pmat and Dtrace do not perform as well as chNet since they ignore the hierarchical constraints in the estimated differential network.

In order to assess the generalization of our method, we also generate simulation datasets without hub nodes in the true differential networks to evaluate the performance. The simulation results also show that chNet outperforms the other compared methods (Supplementary Section S3.3.2).

4 Applications to real data

4.1 Breast cancer application

4.1.1 Datasets

Breast cancer is a common cause of death from cancer among women and has four main molecular subtypes: luminal A, luminal B, basal-like and HER2-enriched (Network *et al.*, 2012). The clinical and genomic characteristics of breast cancer differ across

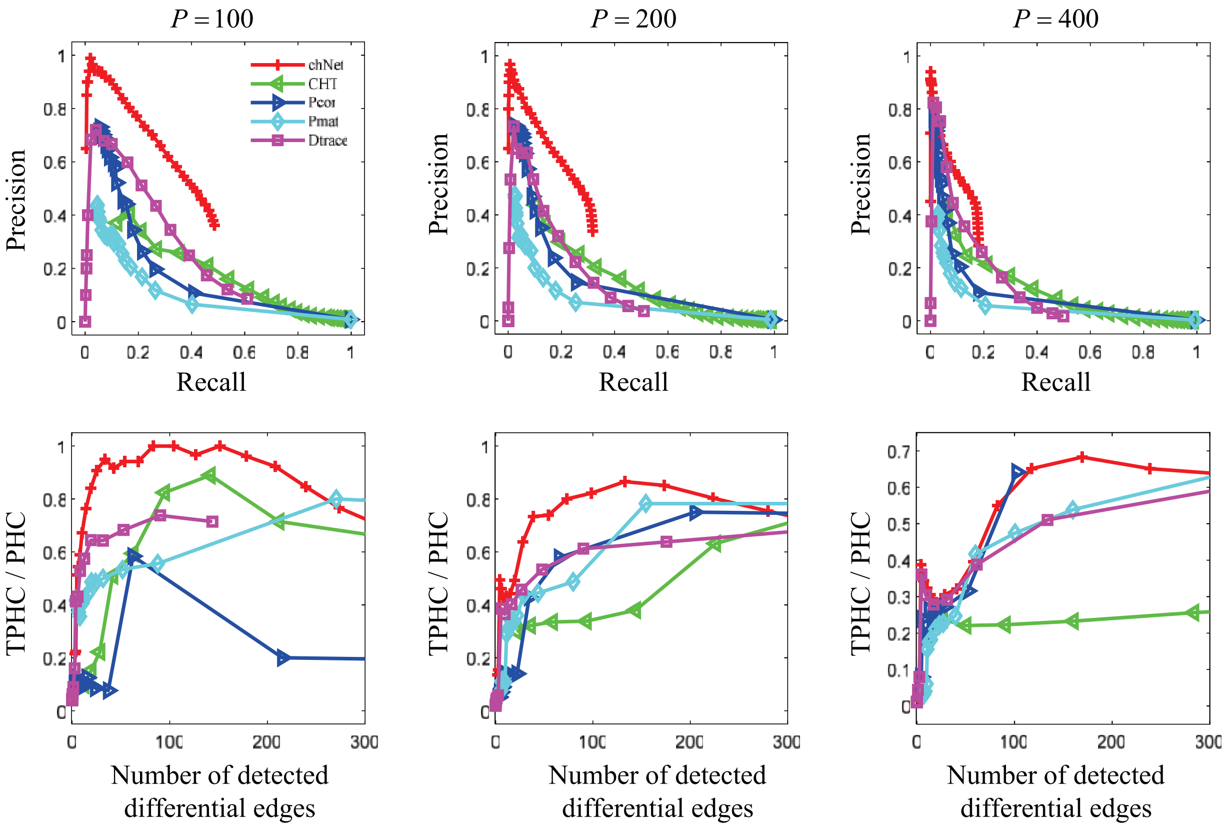

Fig. 2. Performance of the compared methods on simulation data with sample size $n_1 = n_2 = 200$. The first row presents precision against recall, and the second presents the TPHC/PHC against the number of detected differential edges. The columns correspond to the different numbers of genes, $p = 100, 200, 400$, examined. In each plot, each curve corresponds to the performance of a method as the tuning parameter that controls the sparsity level of its differential network is varied. Results are averaged over 20 random generations of the data

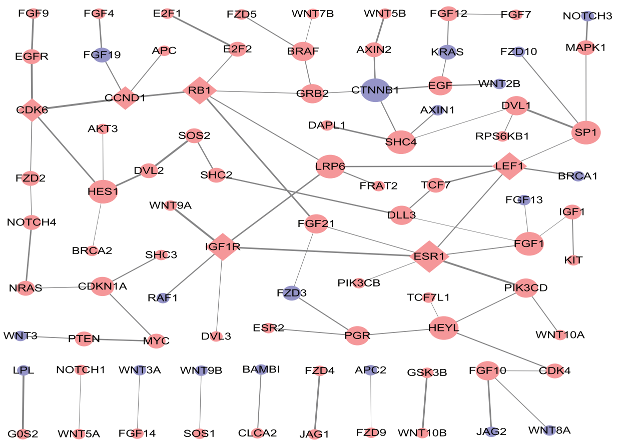

Fig. 3. Differential network estimated by chNet from breast cancer gene expression datasets. The size of each node is proportional to the node's degree. The hub nodes of the differential network are highlighted by a diamond. Differentially expressed genes are marked red, and nondifferentially expressed genes are marked purple. The thickness of lines represents the weight of edges

subtypes. Luminal A cancers are hormone-receptor positive, while basal-like cancers are hormone-receptor negative. Basal-like cancers are often aggressive and have a poorer prognosis than luminal A cancers (Schnitt, 2010). Identifying rewiring of gene networks between different cancer subtypes may give insights into the molecular mechanisms of breast cancer. We aim to estimate differential networks between the luminal A and basal-like subtypes using gene expression datasets from The Cancer Genome Atlas (level 3, Agilent G450 microarray, version: May 6, 2017) (Network *et al.*, 2012). There are 231 luminal A cancers and 95 basal-like cancers. A pathway-based analysis is used to reveal network aberrations. The breast cancer pathway (hsa05224) is downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2000). In order to investigate the role of nondifferentially expressed genes in the differential networks, we also consider 50 nondifferentially expressed genes (Student's *t*-test, *P*-value >0.05) that have the highest variations but are not included in the breast cancer pathway. As a result, there are 189 genes in the gene expression datasets.

4.1.2 Analysis of gene network rewiring

The tuning parameter $\overline{\lambda}$ is selected from a set of possible values (61 values from 2 to 3.5 on an equal space) using stability selection (Liu *et al.*, 2010) (Supplementary Section S3.2). chNet is implemented with the tuning parameter $\overline{\lambda} = 2.825$. The selection frequency b_{ij} of a pair of genes (i, j) being connected in the differential networks is used to quantify the edge weight between genes i and j (Section 2.5). The weighted differential network inferred by chNet is presented in Figure 3. There are 86 differential edges between 89 genes, of which 71 are differentially expressed (red) and 18 are nondifferentially expressed (purple). The thickness of the lines represents the weights of edges. The hierarchical constraints ensure that every differential edge connects to at least one differentially expressed gene and that there is no differential edge between two nondifferentially expressed genes (Fig. 3).

Since there is no true differential network between luminal A and basal-like subtypes in reality, we cannot evaluate the accuracy of estimated differential network by comparing it with the reference. Considering the fact that gene network rewiring may be driven by hub genes in the differential network, we provide more insights into the estimated differential network by analyzing the functional significance of its hub nodes. As mentioned in the simulation studies,

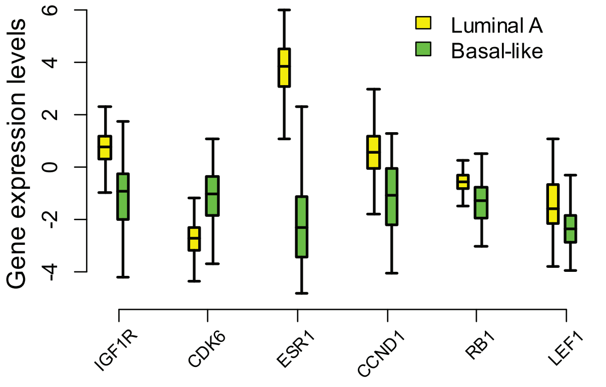

Fig. 4. Expression levels of hub nodes (as boxplots) in the estimated differential network of breast cancer. Gene expression in the luminal A cancers is in yellow, and in green for the basal-like cancers

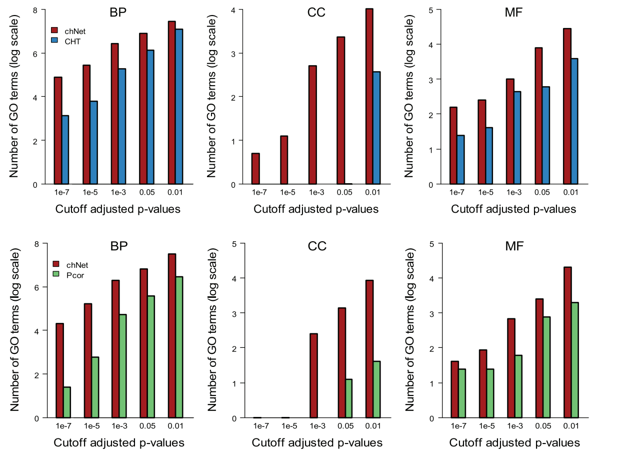

Fig. 5. Number of GO terms enriched in genes corresponding to differential edges unique to individual methods for two compared methods on breast cancer datasets. The *X*-axis represents different adjusted *P*-value cutoffs for enrichment analysis, and the *Y*-axis represents the number of GO terms (in log scale) enriched in genes involved in the subnetwork unique to individual method. (**A**) Comparison between chNet and CHT. (**B**) Comparison between chNet and Pcor

we determine the set of hub genes using the three-sigma rule. A gene will be considered as a hub (highlighted by diamonds in Fig. 3) if its weighted degree is larger than $\mu + 3\sigma$, where μ and σ are the mean and standard deviation of weighted degrees across all genes. Six hub genes are determined, all of which are differentially expressed. The expression levels of the hub genes in different subtypes are shown in Figure 4. We find that the hub genes identified by chNet are functionally significant in breast cancer. IGF1R tends to be more highly expressed in luminal A cancers and comparatively under expressed in basal-like cancers (Fig. 4). Its effects on survival vary among individual breast cancer subtypes (Yerushalmi *et al.*, 2012) and it is a therapeutic target for basal-like breast cancer (Klinakis *et al.*, 2009). CDK6 is a member of the cyclin-dependent kinase (CDK) gene family and is under expressed in breast cancer (Zinia *et al.*, 2020), which might induce poor prognosis in patients, and CDK6 could be a useful predictive biomarker for breast cancer (Nebenfuehr *et al.*, 2020; Zinia *et al.*, 2020). ESR1 is one of the two main types of estrogen receptor. Clinical risk stratification and subtype classification of breast cancer are often based on the expression level of ESR1 (Network *et al.*, 2012). CCND1 is often amplified in basal-like cancers (Elsheikh *et al.*, 2008) and its product, Cyclin D1, could be considered for routine diagnosis (Elsheikh *et al.*, 2008; Rakha *et al.*, 2006). The expression of RB1 is upregulated in luminal A cancer (Fig. 4). RB1 is a tumor suppressor that is frequently lost in basal-like cancers (Jiang *et al.*, 2011). Expression of LEF1 contributes to high canonical Wnt activity, which is a key player during normal mammary gland development and mammary tumorigenesis (Gracanin *et al.*, 2014; Hatsell *et al.*, 2003). LEF1 affects the viability, invasion and migration of breast cancer cells (Hsieh *et al.*, 2012).

4.1.3 Comparison with other methods

As chNet is a hypothesis test-based differential network analysis method with the hierarchical constraints based on partial correlations, it is only compared with the similar methods CHT and Pcor. CHT considers hierarchical constraints, and Pcor is based on partial correlations. To provide interpretable results, the tuning parameters of CHT and Pcor are selected to give a similar number of edges to that of chNet. We find that the differential networks estimated by chNet and CHT are quite different (Supplementary Fig. S1), this may be partially explained by the fact that chNet defines the differential networks based on partial correlations while CHT is based on marginal correlations. Some hub genes estimated by chNet (e.g. IGF1R, CDK6 and ESR1) have high degrees in the network estimated by CHT, while the remains (e.g. CCND1, RB1 and LEF1) are not (Supplementary Table S1), which indicates that some functionally significant genes (e.g. CCND1, RB1 and LEF1) detected by chNet may be neglected by CHT. Since a network which includes indirect effects tends to contain many triangles (Zalesky *et al.*, 2012), we assess whether a method can distinguish direct and indirect interactions by counting the number of triangles in the estimated networks. There are 11 triangles in the network estimated by CHT (Supplementary Fig. S2), which may be due to CHT estimating differential networks based on marginal correlations that cannot distinguish between direct and indirect effects. Forty-four of the 86 edges in the differential network estimated by chNet are in the network estimated by Pcor (Supplementary Fig. S1). The differences between differential networks estimated by the two methods may be due to the fact chNet have imposed the hierarchical constraints on the resulting networks. Most hub genes estimated by chNet also have high degrees in the network estimated by Pcor (Supplementary Table S1), which may be owing to the fact both methods define the differential networks based on partial correlations. Although no triangles are in the differential network estimated by Pcor, edges between two nondifferentially expressed genes (purple) are found (Supplementary Fig. S3). Even though Pcor can distinguish between direct and indirect effects, it does not take account of the hierarchical constraints.

To further evaluate the functional significance of differential networks estimated by different methods, for two compared methods (e.g. chNet versus CHT, chNet versus Pcor), the R package clusterProfiler (Yu *et al.*, 2012) is used to perform Gene Ontology (GO) enrichment analysis on genes involved in the subnetworks unique to an individual method. For chNet versus CHT, there are 78 genes corresponding the edges unique to the differential network estimated by chNet and 59 genes corresponding the edges unique to the differential network resulting by CHT. For all the three complementary biological concepts [BP, cellular component (CC) and molecular function (MF)], the genes unique to our method are enriched with more GO terms than those unique to CHT at different adjusted *P*-value cutoffs (Fig. 5A). This result indicates that the differential networks estimated by chNet can deliver more biological insights than those estimated by CHT, and that distinguishing direct interactions from indirect effect is important in differential network analysis. For chNet versus Pcor, there are 61 genes corresponding to the subnetwork unique to the differential network estimated by chNet, and 69 genes corresponding to the subnetwork unique to that by Pcor. We find that the genes unique to chNet are enriched with a larger number of GO terms than those unique to Pcor at different adjusted *P*-value cutoffs (Fig. 5B). These results indicate that the differential networks estimated by our method are more biologically significant than those estimated by Pcor, and that imposing the hierarchical constraints may be reasonable and valuable in reality.

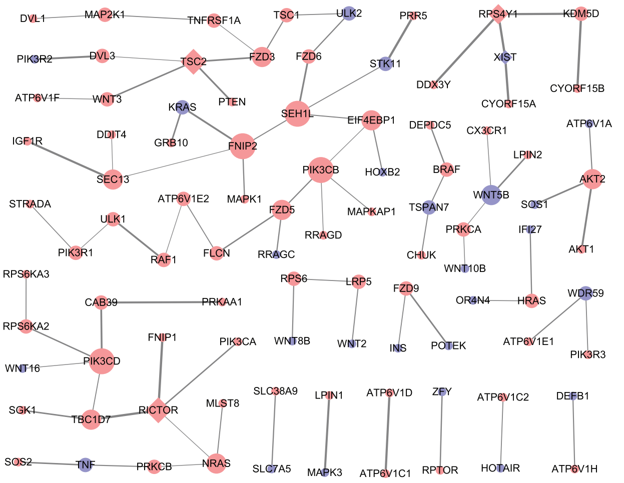

Fig. 6. Differential network estimated by chNet from the AML gene expression datasets. The size of each node is proportional to the node's degree. The hub nodes of the differential network are highlighted by a diamond. The differentially and non-differentially expressed genes are red and purple, respectively. The thickness of lines represents the weight of edges

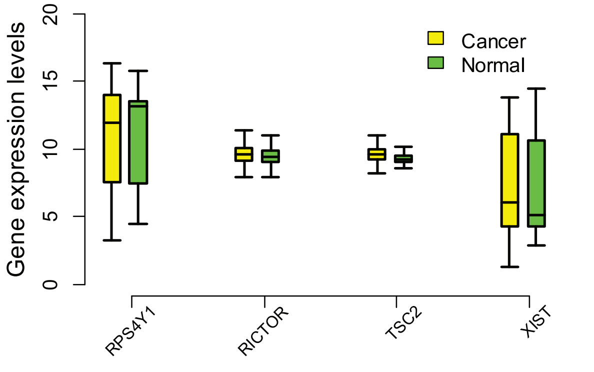

Fig. 7. Expression level of hub nodes in the estimated differential networks of AML. The yellow and green boxplots represent the gene expression in cancer and normal patients, respectively

4.2 AML application

4.2.1 Datasets

AML is characterized by abnormal, rapid growth of cells in the bone marrow and blood. In this experiment, rewiring of gene networks between AML and normal samples is examined using gene expression data retrieved from GEO (GSE13159) (Haferlach *et al.*, 2010). Data for 541 AML cancers and 73 normal samples are downloaded from http://discem-leelab.cs.washington.edu/ (Grechkin *et al.*, 2016). The mTOR signaling pathway is often activated in AML and the inhibition of mTOR signaling has growth-inhibitory effects (Park *et al.*, 2010; Tabe *et al.*, 2017). Therefore, a pathway-based analysis is used in this study and the mTOR signaling pathway (hsa04150) is obtained from the KEGG database. Besides the 138 genes that overlap with the mTOR signaling pathway, we select 50 nondifferentially expressed genes (Student's *t*-test, *P*-value >0.05) with the highest variations to explore the role of nondifferentially expressed genes in the differential network. As a result, 188 genes are considered.

4.2.2 Analysis of gene network rewiring

chNet is applied to the AML datasets to construct a differential network. The tuning parameter $\bar{\lambda}$ is selected from 31 values in the range of 2–5 on a log scale, using stability selection (Liu *et al.*, 2010) (Supplementary Section S3.2) and giving $\bar{\lambda} = 2.8$. The selection frequency, b_{ij}, of a pair of genes (i, j) being connected is used to define the edge weight (Section 2.5). The differential network estimated by chNet is presented in Figure 6. There are 80 differential edges between 96 genes, including 70 differentially expressed genes (red) and 26 nondifferentially expressed genes (purple). The thickness of lines indicates the weight of edges. There is no connection between two nondifferentially expressed genes.

We evaluate the estimated differential network by analyzing the biological function of the hub nodes. Based on the weighted degrees, we also use the three-sigma rule to determine hub nodes in the estimated differential network (diamond in Fig. 6). Three differentially expressed hub genes (RPS4Y1, RICTOR and TSC2) and one nondifferentially expressed hub gene (XIST) are identified. The expression levels of the hub genes in the different conditions are shown in Figure 7. Most hub genes identified by chNet are functionally significant in AML. AML shows a diverse spectrum of chromosomal aberrations and some AML patients relate to Y chromosome loss (Group *et al.*, 1992; Kim *et al.*, 2001). RPS4Y1 is a member of human Y chromosome harbors genes, which may be involved in the makeup of the marker chromosomes (Kim *et al.*, 2001). RICTOR is a component of the protein complex that integrates nutrient and growth factor-derived signals to regulate cell growth. As a subunit of the mammalian target of rapamycin complex 2 (mTORC2), RICTOR may participate in regulating hematopoietic stem cell self-renewal and suppressing leukemia (Magee *et al.*, 2012). Disrupting this balance by inhibiting RICTOR could be significant in the development of novel therapies to eliminate high-risk AML (Fang *et al.*, 2017). TSC2, a putative tumor suppressor gene, could affect cell growth, differentiation and proliferation and its aberrant expression might be associated with the pathogenesis of leukemogenesis (Xu *et al.*, 2009). Hypermethylation of the promoter of TSC2 may be a biomarker for AML therapy (Xu *et al.*, 2009). XIST is identified as a nondifferentially expressed (Student's *t*-test, *P*-value = 0.106) hub node in the differential network estimated by chNet. Silencing XIST inhibits cellular activity and drug resistance, and can also inhibit tumorigenesis of AML cells *in vivo* (Wang *et al.*, 2020).

4.2.3 Comparison with other methods

The network estimated by chNet is compared with those estimated by CHT and Pcor for the AML datasets. The tuning parameters of CHT and Pcor are chosen to provide a similar number of differential edges to those identified by chNet. Similar to the results obtained from the breast cancer datasets, there are a few edges in common between the differential networks estimated by chNet and CHT (Supplementary Fig. S4). The hub genes identified by chNet do not have a high degree in the differential network estimated by CHT (Supplementary Table S2), indicating that the functionally significant genes identified by chNet may be missed by chNet. Six triangles are found in the network estimated by CHT (Supplementary Fig. S5), which indicate that the differential network estimated by CHT may include indirect effects. Most edges in the differential network estimated by chNet are in common with that estimated by Pcor (Supplementary Figs S4 and S6) and most hub genes estimated by chNet also have a high degree in the differential network estimated by Pcor (Supplementary Table S2). There is no triangle in the differential network identified by Pcor (Supplementary Fig. S4). However, edges between two nondifferentially expressed genes (purple) are found in the differential network estimated by Pcor, indicating that the differential network estimated by Pcor does not obey the hierarchical constraints (Supplementary Fig. S6).

For two compared methods, we also perform GO enrichment analysis on genes corresponding to differential edges unique to each individual methods. For chNet versus CHT, there are 88 genes corresponding to the subnetwork unique to chNet and 62 genes corresponding to the subnetwork unique to that CHT. We find that the number of GO terms enriched in genes unique to our method is comparable to that of CHT at different adjusted *P*-value cutoffs, and that our method often enriches more GO terms at a small adjusted *P*-value cutoff (Supplementary Fig. S7A). For chNet versus Pcor, there are 41 genes unique to chNet and 44 genes unique to Pcor. The genes unique to our chNet are enriched with more GO terms on the BP and MF subontologies than those unique to Pcor

(Supplementary Fig. S7B). These results indicate that the differential network estimated by our method may be more biologically significant than those estimated by the two compared methods.

5 Discussion

In this paper, we have proposed a new differential network analysis method by simultaneously considering the change of partial correlations between gene pairs and the expression levels of genes. The competitive performance of our method is demonstrated by both simulation studies and real data analysis. The novelty of our method lies in taking advantage of the hierarchical constraints to borrow information from differentially expressed genes to infer the differential networks. Unlike most previous differential network analysis methods that only consider changes in gene interactions, our method also takes account of changes in the expression levels of genes. The differential network is defined by our method as the difference of partial correlations from two different conditions, whereas some other methods define the differential network as the difference of marginal correlations, allowing our method to be more powerful in excluding spurious differential edges caused by indirect effects. For methods that define the differential network as the difference of precision matrices, our method performs better by removing false differential edges produced by variants of the conditional variances of individual genes. Differential networks estimated by our method have good interpretability and high accuracy.

Please note that the hierarchical constraints considered in this study may be violated in reality. That is, there may be a differential interaction for which neither of the two involved genes is differentially expressed. However, it will be difficult to interpret which factor drives the change of the interaction if the two involved genes are not differentially expressed. In this study, we assume that the change of an interaction is driven by at least one of the two involved genes, and that at least one of the two involved genes needs to be differentially expressed to form a differential edge. In addition, differentially expressed genes may be more likely to produce differential interactions than nondifferentially expressed genes, and there may be more differential interactions associated with differentially expressed genes than those not associated with any differentially expressed genes. Thus, we use the hierarchical constraints to reduce the search space of possible differential interactions to improve the statistical power. Even though our method may the certain true differential edges between nondifferentially expressed genes, it can still exclude many false positive differential edges thanks to the constraints of search space, and improve the overall accuracy of the estimated differential networks (Bien *et al.*, 2015). Simulation studies have shown that our method performs better than methods that do not consider the hierarchical constraints if the true differential networks exhibits the hierarchical structures. Furthermore, the real data experiments show that the differential networks estimated by our method are more biologically significant than those estimated by the compared methods. These results indicate that it is reasonable and meaningful to impose the hierarchical constraints on the differential networks.

We consider the hierarchical constraint that at least one of the two involved genes must be differentially expressed to derive a differential edge, which has been called a weak hierarchical constraint in previous studies (Bien *et al.*, 2015). A strong hierarchical constraint, which requires both involved genes to be differentially expressed to derive a differential edge, has also been considered (Bien *et al.*, 2015). To obtain a differential network which satisfies a strong hierarchical constraint, one can first identify a set of differentially expressed genes using a differential expression analysis method and then estimate a differential network among the identified differentially expressed genes. In doing so, genes which are biologically significant, but not differentially expressed, will be ignored (Chuang *et al.*, 2007). To avoid this, we only considered a weak hierarchical constraint in this study.

Our method uses the partial correlation-based statistics to quantify the change of gene interactions and uses the Student's *t*-test statistics to quantify the changes in expression levels of individual genes. Other statistics based on marginal correlations or precision matrices can also be used. Here, we consider partial correlation-based statistics to remove false differential edges caused by indirect effects and variants of conditional variances of individual genes. Statistics other than the Student's *t*-test could be used to quantify changes in the expression levels of genes. The Student's *t*-test statistics is used as they are simple and widely used in differential expression analysis and, under the null hypothesis, both the Student's *t*-test statistics and the partial correlations-based statistics, approximately follow a standard normal distribution. As a result, we can easily combine these two statistics to obtain a differential network that satisfies the hierarchical constraint.

In general, estimating parameters of Gaussian graphical models is time consuming. To reduce the running time, previous related studies often choose a fraction of genes to perform real data analysis (e.g. genes in a considered pathway or high variable genes) (Danaher *et al.*, 2014; Deng *et al.*, 2018; Zhang *et al.*, 2016, 2018, 2017). In this paper, we also carry out a pathway-based analysis. In particular, we pay our attention to the breast cancer pathway (hsa05224) in breast cancer and the mTOR pathway in AML since they play an important role in related cancers. The goal of this paper is to propose a new statistical model to estimate a hierarchical differential network from gene expression data. Therefore, we do not analyze other pathways. Interested readers can analyze other pathways by using our R package.

Thanks to recent breakthroughs in technology, a considerable amount of single-cell RNA-sequencing (scRNA-seq) data has been accumulated, which will allow molecular mechanisms underlying BPs and complex diseases to be elucidated at single-cell resolution. To understand heterogeneity in complex samples (e.g. cancers) at a network level, differential networks between different cell types or cell states need to be estimated from scRNA-seq data. However, scRNA-seq data do not follow the normal distributions considered in this study and the biological signal in scRNA-seq data is often corrupted by dropout events (Jin *et al.*, 2020; Zhang *et al.*, 2019b). Therefore, extending our method to deal with scRNA-seq data is a future challenge.

Funding

This work was supported by the National Natural Science Foundation of China [11871026, 11871237 and 61602309], Hubei Provincial Science and Technology Innovation Base (Platform) Special Project [2020DFH002], Shenzhen Fundamental Research Program [JCYJ20170817095210760], Guangdong Basic and Applied Basic Research Foundation [2019A1515011384], Hong Kong Research Grants Council [Project 11200818] and the Hong Kong Innovation and Technology Commission.

Conflict of Interest: none declared.

References

Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Bashashati,A. *et al.* (2012) Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.*, **13**, R124–R124.

Bien,J. *et al.* (2013) A lasso for hierarchical interactions. *Ann. Stat.*, **41**, 1111–1141.

Bien,J. *et al.* (2015) Convex hierarchical testing of interactions. *Ann. Appl. Stat.*, **9**, 27–42.

Chuang,H.Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Danaher,P. *et al.* (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc.*, **76**, 373–397.

Deng,W. *et al.* (2018) JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions. *Bioinformatics*, **34**, 3470–3478.

Elsheikh,S. *et al.* (2008) Ccnd1 amplification and cyclin d1 expression in breast cancer and their relation with proteomic subgroups and patient outcome. *Breast Cancer Res. Treat.*, **109**, 325–335.

Fang,Y. *et al.* (2017) Rictor has a pivotal role in maintaining quiescence as well as stemness of leukemia stem cells in MLL-driven leukemia. *Leukemia*, **31**, 414–422.

Gracanin,A. *et al.* (2014) Ligand-independent canonical WNT activity in canine mammary tumor cell lines associated with aberrant lef1 expression. *PLoS One*, **9**, e98698.

Grechkin,M. *et al.* (2016) Identifying network perturbation in cancer. *PLoS Comput. Biol.*, **12**, e1004888.

Group,U.K.C.C. *et al.* (1992) Loss of the y chromosome from normal and neoplastic bone marrows. *Genes Chromosomes Cancer*, **5**, 83–88.

Ha,M.J. *et al.* (2015) Dingo: differential network analysis in genomics. *Bioinformatics*, **31**, 3413–3420.

Haferlach,T. *et al.* (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. *J. Clin. Oncol.*, **28**, 2529–2537.

Hatsell,S. *et al.* (2003) β-catenin and TCFS in mammary development and cancer. *J. Mammary Gland Biol. Neoplasia*, **8**, 145–158.

Hsieh,T.H. *et al.* (2012) n-butyl benzyl phthalate promotes breast cancer progression by inducing expression of lymphoid enhancer factor 1. *PLoS One*, **7**, e42750.

Ideker,T. *et al.* (2012) Differential network biology. *Mol. Syst. Biol.*, **8**, 565.

Jiang,Z. *et al.* (2011) Rb1 and p53 at the crossroad of EMT and triple-negative breast cancer. *Cell Cycle*, **10**, 1563–1570.

Jin,K. *et al.* (2020) sctssr: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation. *Bioinformatics*, **36**, 3131–3138.

Kanehisa,M. *et al.* (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kim,M.H. *et al.* (2001) The application of comparative genomic hybridization as an additional tool in the chromosome analysis of acute myeloid leukemia and myelodysplastic syndromes. *Cancer Genet. Cytogenet.*, **126**, 26–33.

Klinakis,A. *et al.* (2009) Igf1r as a therapeutic target in a mouse model of basal-like breast cancer. *Proc. Natl. Acad. Sci. USA*, **106**, 2359–2364.

Li,S. *et al.* (2013) Bootstrap inference for network construction with an application to a breast cancer microarray study. *Ann. Appl. Stat.*, **7**, 391–417.

Lim,M. *et al.* (2015) Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Stat.*, **24**, 627–654.

Liu,H. *et al.* (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst.*, **24**, 1432–1440.

Liu,W. *et al.* (2017) Structural similarity and difference testing on multiple sparse Gaussian graphical models. *Ann. Appl. Stat.*, **45**, 2680–2707.

Magee,J.A. *et al.* (2012) Temporal changes in PTEN and mTORC2 regulation of hematopoietic stem cell self-renewal and leukemia suppression. *Cell Stem Cell*, **11**, 415–428.

Meinshausen,N. *et al.* (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Appl. Stat.*, **34**, 1436–1462.

Mohan,K. *et al.* (2014) Node-based learning of multiple Gaussian graphical models. *J. Mach. Learn. Res.*, **15**, 445–488.

Nebenfuehr,S. *et al.* (2020) The role of CDK6 in cancer. *Int. J. Cancer*, **147**, 2988–2995.

Network,C.G.A. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.

Ou-Yang,L. *et al.* (2014) Detecting temporal protein complexes from dynamic protein-protein interaction networks. *BMC Bioinf.*, **15**, 335.

Park,S. *et al.* (2010) Role of the PI3K/AKT and mTOR signaling pathways in acute myeloid leukemia. *Haematologica*, **95**, 819–828.

Rakha,E.A. *et al.* (2006) Chromosome 16 tumor-suppressor genes in breast cancer. *Genes Chromosomes Cancer*, **45**, 527–535.

Robinson,M.D. *et al.* (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Schadt,E.E. *et al.* (2009) Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218–223.

Schnitt,S.J. (2010) Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod. Pathol.*, **23**, S60–S64.

Tabe,Y. *et al.* (2017) Inhibition of mTOR kinase as a therapeutic target for acute myeloid leukemia. *Expert Opin. Ther. Targets*, **21**, 705–714.

Tan,Y.T. *et al.* (2020) Identifying gene network rewiring based on partial correlation. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **99**, 1.

Tian,D. *et al.* (2016) Identifying gene regulatory network rewiring using latent differential graphical models. *Nucleic Acids Res.*, **44**, e140.

Wang,C. *et al.* (2020) Silencing long non-coding RNA XIST suppresses drug resistance in acute myeloid leukemia through down-regulation of MYC by elevating microrna-29a expression. *Mol. Med.*, **26**, 1–11.

Xu,Z. *et al.* (2009) Aberrant expression of tsc2 gene in the newly diagnosed acute leukemia. *Leuk. Res.*, **33**, 891–897.

Yerushalmi,R. *et al.* (2012) Insulin-like growth factor receptor (igf-1r) in breast cancer subtypes. *Breast Cancer Res. Treat.*, **132**, 131–142.

Yu,G. *et al.* (2012) clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS: J. Integr. Biol.*, **16**, 284–287.

Yuan,H. *et al.* (2017) Differential network analysis via lasso penalized d-trace loss. *Biometrika*, **104**, 755–770.

Yuan,M. *et al.* (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.

Zalesky,A. *et al.* (2012) On the use of correlation as a measure of network connectivity. *Neuroimage*, **60**, 2096–2106.

Zhang,X.F. *et al.* (2016) Differential network analysis from cross-platform gene expression data. *Sci. Rep.*, **6**, 34112.

Zhang,X.F. *et al.* (2017) Incorporating prior information into differential network analysis using non-paranormal graphical models. *Bioinformatics*, **33**, 2436–2445.

Zhang,X.F. *et al.* (2018) Diffgraph: an R package for identifying gene network rewiring using differential graphical models. *Bioinformatics*, **34**, 1571–1573.

Zhang,X.F. *et al.* (2019a) Diffnetfdr: differential network analysis with false discovery rate control. *Bioinformatics*, **35**, 3184–3186.

Zhang,X.F. *et al.* (2019b) Enimpute: imputing dropout events in single-cell RNA-sequencing data via ensemble learning. *Bioinformatics*, **35**, 4827–4829.

Zinia,J.A. *et al.* (2020) Evaluation of the prognostic significance of cdk6 in breast cancer. *Netw. Model Anal. Health Inform. Bioinform.*, **9**, 1–9.