


METHODOLOGY ARTICLE

Open Access



# Enhanced construction of gene regulatory networks using hub gene information

Donghyeon Yu<sup>1</sup>, Johan Lim<sup>2</sup>, Xinlei Wang<sup>3</sup>, Faming Liang<sup>4</sup> and Guanghua Xiao<sup>5\*</sup> 

## Abstract

**Background:** Gene regulatory networks reveal how genes work together to carry out their biological functions. Reconstructions of gene networks from gene expression data greatly facilitate our understanding of underlying biological mechanisms and provide new opportunities for biomarker and drug discoveries. In gene networks, a gene that has many interactions with other genes is called a hub gene, which usually plays an essential role in gene regulation and biological processes. In this study, we developed a method for reconstructing gene networks using a partial correlation-based approach that incorporates prior information about hub genes. Through simulation studies and two real-data examples, we compare the performance in estimating the network structures between the existing methods and the proposed method.

**Results:** In simulation studies, we show that the proposed strategy reduces errors in estimating network structures compared to the existing methods. When applied to *Escherichia coli*, the regulation network constructed by our proposed ESPACE method is more consistent with current biological knowledge than the SPACE method. Furthermore, application of the proposed method in lung cancer has identified hub genes whose mRNA expression predicts cancer progress and patient response to treatment.

**Conclusions:** We have demonstrated that incorporating hub gene information in estimating network structures can improve the performance of the existing methods.

**Keywords:** Gene regulatory network, Hub gene, Partial correlation, Sparse partial correlation estimation, *Escherichia coli*, Lung cancer

## Background

A gene regulatory network (GRN) describes interactions and regulatory relationships among genes. It provides a systematic understanding of the molecular mechanisms underlying biological processes by revealing how genes work together to form modules that carry out cell functions [1–4]. In addition, the visualization of genetic dependencies through the GRN facilitates the systematic interpretation and comprehension of analysis results from genome-wide studies using high-throughput data. GRNs have proven valuable in a variety of contexts, including identifying druggable targets [5], detecting driver genes in diseases [6], and even optimizing prognostic and predictive signatures [7].

Gene expression microarrays monitor the transcription activities of thousands of genes simultaneously, which provides a great opportunity to study the “relationships” among genes on a large scale. However, challenges lie in constructing large-scale GRNs from gene expression microarray data due to the small sample sizes of microarray studies and the extremely large solution space. Computational techniques and algorithms have been proposed to reconstruct GRNs from gene expression data, including probability-based approaches such as Bayesian networks [8–12], correlation-based approaches [13], likelihood-based approaches [14–16], partial-correlation-based approaches [17, 18], and information-theory-based approaches [19–22]. The existing methods are briefly reviewed in the Methods Section. Readers can also refer to Bansal et al. [23] and Allan et al. [24] for a more detailed review of network construction methods.

\*Correspondence: guanghua.xiao@utsouthwestern.edu

<sup>5</sup>Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA

Full list of author information is available at the end of the article

The sparse partial correlation estimation (SPACE) method, proposed by Peng et al. [18], considers a penalized regression approach to estimate edges in the GRN, which utilizes the sparse feature of the GRN. Comparative studies have shown that the SPACE method performs well in estimating sparse networks with high accuracy [24]. Peng et al. [18] also showed that the method was able to identify functional relevant molecular networks. In addition, recent studies of network analysis have revealed its advantage in detecting genes or modules associated with phenotypes [25–27].

In gene networks, genes that have many interactions with other genes are defined as hub genes. Because of these interactions, hub genes usually play an important role in a biological system. For example, transcription factor (TF), a protein that binds to specific DNA sequences, can regulate a given set of genes. In humans, approximately 10% of genes in the genome code for around 2600 TFs [28]. The combinatorial human TFs account for most of the regulation activities in the human genome, especially during the development stage. As a result, the genes that code TFs, called TF-encoding genes, are usually regarded as hub genes. Furthermore, in cancer research, cancer genes (oncogenes or tumor suppressor genes) take part in tumor genesis and are likely to be hub genes in the genetic networks of tumors [29, 30]. Through decades of biological studies, knowledge on important genes (such as TFs or cancer genes) has been accumulated. Our hypothesis is that incorporating prior knowledge about hub genes can improve accuracy in estimating the gene network structure. It is worth noting that there is a reweighted  $\ell_1$  regularization method [31] that repeatedly estimates the structures and modifies the weights of the penalties by using the information on degrees from the previous estimation to encourage the appearance of the hubs. This method does not use the prior information obtainable from the other resources while our method uses additional information not contained in the observed dataset.

To explicitly account for the information on hub genes, we propose an extension of the SPACE method, which introduces an additional tuning parameter to open up the possibility of reducing penalization and increasing the likelihood of selecting the edges connected to such genes. We numerically show that the proposed method reduces errors in estimating network structures. Although we focus on extending the SPACE method in this paper, the idea can also be applied to penalized likelihood methods as well as to other penalized regression methods. Note that there is no rigorous definition of a hub in the context of a network; the definition of a hub varies depending on the sparsity of the network. For sparse protein networks, a hub is defined in [32] as a protein whose degree lies over the 0.95 quantile of the degree distribution or in

[33] and [7] as a protein whose degree is greater than 7. In this paper, we conservatively define a hub as a node whose degree is both greater than 7 and above the 0.95 quantile of the degree distribution, because most nodes in sparse networks have relatively small degrees between 0 and 3.

In this study, we briefly introduce seven existing methods, including the SPACE and the graphical lasso, and propose the extended SPACE (ESPACE) method to incorporate the biological knowledge about important genes, i.e. network hubs. Moreover, it is worth noting that the ESPACE only incorporates the previously known biological information not contained in the observed dataset compared to the other existing methods. Through simulation studies, we show that the proposed approach reduces error in estimating the network structures compared to the seven other existing methods that we reviewed in the “Methods” section. Finally, we demonstrate the improvement of the ESPACE method compared to the SPACE method with two real-data examples.

## Methods

### Review of existing methods

Here, we briefly review the existing methods; the GeneNet [34], the NS [17], the GLASSO [15], the GLASSO-SF [31], the PCACMI [21], the CMI2NI [22], and the SPACE [18]. Let  $X_i^k$  be the expression level of the  $i$ th gene of the  $k$ th array for  $i = 1, 2, \dots, p$  and  $k = 1, 2, \dots, n$ . Let  $\mathbf{X}_i = (X_i^1, X_i^2, \dots, X_i^n)^T$  so that observed gene expression data can be denoted by an  $n \times p$  matrix  $\mathbf{X} = (\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_p)$  whose rows and columns denote arrays and genes, respectively. Suppose row vectors  $\mathbf{X}^k = (X_1^k, X_2^k, \dots, X_p^k)$  for  $k = 1, 2, \dots, n$  are independently and identically distributed random vectors from the multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\Sigma$ . We assume that  $\Sigma$  is positive definite, and let  $\Omega \equiv \Sigma^{-1} = (\omega_{ij})_{1 \leq i, j \leq p}$  be the inverse of the covariance matrix  $\Sigma$ , which is referred to as a concentration matrix or a precision matrix.

### GeneNet

Schäfer and Strimmer [34] propose the linear shrinkage estimator for a covariance matrix and the Gaussian graphical model (GGM) selection based on the partial correlation obtained from their shrinkage estimator. With multiple testing procedure using the local false discovery rate [35], the GGM selection controls the false discovery rate under a pre-determined level  $\alpha$ . Since Schäfer and Strimmer [34] provide their GGM selection procedure in the R package GeneNet, we denote their GGM selection procedure as GeneNet in this paper. To be specific, one of the most commonly used linear shrinkage estimators  $S^*$  for the covariance matrix  $\Sigma$  is

$$S^* = \lambda^* T + (1 - \lambda^*) S,$$

where  $S = (s_{ij})_{1 \leq i, j \leq p}$  is the sample covariance matrix,  $T = \text{diag}(s_{11}, s_{22}, \dots, s_{pp})$  is the shrinkage target matrix, and  $\lambda^* = \sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) / \left( \sum_{i \neq j} s_{ij}^2 \right)$  is the optimal shrinkage intensity. With this estimator  $S^*$ , the matrix of the partial correlations  $P = (\hat{\rho}^{ij})_{1 \leq i, j \leq p}$  is defined as  $\hat{\rho}^{ij} = -\hat{\omega}_{ij} / \sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}$ , where  $\hat{\Omega} = (\hat{\omega}_{ij})_{1 \leq i, j \leq p} = (S^*)^{-1}$ .

To identify the significant edges, Schäfer and Strimmer [34] suppose the distribution of the partial correlations as the mixture

$$f(\rho) = \eta_0 f_0(\rho, \nu) + (1 - \eta_0) f_1(\rho),$$

where  $f_0$  is the null distribution,  $f_1$  is the alternative distribution corresponding to the true edges, and  $\eta_0$  is the unknown mixing parameter. Using the algorithm in [35], GeneNet identifies significant edges that have the local false positive rate

$$\text{fdr}(\rho) = \frac{\hat{\eta}_0 f_0(\rho, \hat{\nu})}{\hat{f}(\rho)}$$

smaller than the pre-determined level  $\alpha$ , where  $f_0(\rho, \nu) = |\rho| \text{Be}(\rho^2; 0.5, (\nu - 1)/2)$ ,  $\text{Be}(x; a, b)$  is the density of the Beta distribution and  $\nu$  is the reciprocal variance of the null  $\rho$ .

**Neighborhood selection (NS)**

Meinshausen and Bühlmann [17] propose the neighborhood selection (NS) method, which separately solves the lasso [36] problem and identifies edges with nonzero estimated regression coefficients for each node. Meinshausen and Bühlmann [17] prove that the NS method is asymptotically consistent in identifying the neighborhood of each node when the neighborhood stability condition is satisfied. Note that the neighborhood stability condition is related to the irrerepresentable condition in linear model literature [37].

To be specific, for each node  $i \in V = \{1, 2, \dots, p\}$ , NS solves the following lasso problem

$$\hat{\beta}^{i,\lambda} = \arg \min_{\beta \in \mathbb{R}^p: \beta_i = 0} \frac{1}{2} \|\mathbf{X}_i - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1,$$

where  $\|\mathbf{x}\|_2^2 = \sum_{i=1}^p x_i^2$  and  $\|\mathbf{x}\|_1 = \sum_{i=1}^p |x_i|$  for  $\mathbf{x} \in \mathbb{R}^p$ . With the estimate  $\hat{\beta}^{i,\lambda}$ , NS identifies the neighborhood of the node  $i$  as  $N_i(\lambda) = \{k \mid \hat{\beta}_k^{i,\lambda} \neq 0\}$ , which defines an edge set  $E_i^\lambda = \{(i, j) \mid j \in N_i(\lambda)\}$ . Since NS separately solves  $p$  lasso problems, contradictory edges may occur when we define the total edge set  $E^\lambda = \cup_{i=1}^p E_i^\lambda$ , i.e.,  $\hat{\beta}_i^{i,\lambda} \neq 0$  and  $\hat{\beta}_i^{k,\lambda} = 0$ . To avoid these contradictory edges, NS suggests two types of edge sets  $E^{\lambda,\wedge}$  and  $E^{\lambda,\vee}$  defined as follows:

$$E^{\lambda,\wedge} = \{(i, j) \mid i \in N_j(\lambda) \text{ and } j \in N_i(\lambda)\},$$

$$E^{\lambda,\vee} = \{(i, j) \mid i \in N_j(\lambda) \text{ or } j \in N_i(\lambda)\}.$$

Meinshausen and Bühlmann [17] mentioned these two edge sets have only small differences in their experience

and the differences vanish asymptotically. Meinshausen and Bühlmann [17] also propose the choice of the tuning parameter  $\lambda_i(\alpha)$  for the  $i$ th node

$$\lambda_i(\alpha) = \|\mathbf{X}_i\|_2 \tilde{\Phi}^{-1} \left( \frac{\alpha}{2p^2} \right),$$

where  $\tilde{\Phi} = 1 - \Phi$  and  $\Phi$  is the distribution function of the standard normal distribution. With this choice of  $\lambda_i(\alpha)$  for  $i = 1, 2, \dots, p$ , the probability of falsely identifying edges in the network is bounded by the level  $\alpha$ . Note that we estimate the edge set with  $E^{\lambda,\wedge}$  and solve the lasso problems using the R package CDLasso proposed by [38] in this paper.

**Graphical lasso (GLASSO)**

Friedman et al. [15] propose the graphical lasso method that estimates a sparse inverse covariance matrix  $\Omega$  by maximizing the  $\ell_1$  penalized log-likelihood

$$l(\Omega) = \log |\Omega| - \text{tr}(S\Omega) - \lambda \|\Omega\|_1, \tag{1}$$

where  $S$  is the sample covariance matrix,  $\text{tr}(A)$  is the trace of  $A$  and  $\|A\|_1$  is the  $\ell_1$  norm of  $A$  for  $A \in \mathbb{R}^{p \times p}$ .

To be specific, let  $W$  be the estimate of the covariance matrix  $\Sigma$  and consider partitioning  $W$  and  $S$

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix}, \Omega = \begin{pmatrix} \Omega_{11} & \omega_{12} \\ \omega_{12}^T & \omega_{22} \end{pmatrix}$$

Motivated by [39], Friedman et al. [15] show that the solution  $\hat{\Omega}$  of (1) is equivalent to the inverse of  $W$  whose partitioned entity  $w_{12}$  satisfies  $w_{12} = W_{11}\beta^*$ , where  $\beta^*$  is the solution of the lasso problem

$$\min_{\beta} \frac{1}{2} \left\| W_{11}^{1/2} \beta - W_{11}^{-1/2} s_{12} \right\|_2^2 + \lambda \|\beta\|_1. \tag{2}$$

Based on the above property, the graphical lasso sets the diagonal elements  $w_{ii} = s_{ii} + \rho$  and obtains the off-diagonal elements of  $W$  by repeatedly applying the following two steps:

1. Permuting the columns and rows to locate the target elements at the position of  $w_{12}$ .
2. Finding the solution  $w_{12} = W_{11}\beta^*$  by solving the lasso problem (2).

until convergence occurs. After finding  $W$ , the estimate  $\hat{\Omega}$  is obtained from the relationships  $\omega_{12} = -\hat{\beta}\hat{\omega}_{22}$  and  $\hat{\omega}_{22} = 1/(w_{22} - w_{12}^T \hat{\beta})$ , where  $\hat{\beta} = W_{11}^{-1} w_{12}$ . This graphical lasso algorithm was proposed in [15] and had its computational efficiency improved in [16] and [40]. Witten et al. [16] provide the R package `glasso` version 1.7.

**GLASSO with reweighted strategy for scale-free network (GLASSO-SF)**

Liu and Ihler [31] propose the reweighted  $\ell_1$  regularization method to improve the performance of the estimation for the scale-free network whose degrees follows the

power law distribution. Motivated by the fact that the existing methods work poorly for the scale-free networks, Liu and Ihler [31] consider changing the  $\ell_1$  norm penalty in the existing methods to the power law regularization

$$p_{\lambda,\gamma}(\Omega) = \lambda \sum_{i=1}^p \log(\|\omega_{-i}\|_1 + \epsilon_i) + \gamma \sum_{i=1}^p |\omega_{ii}|, \quad (3)$$

where  $\lambda$  and  $\gamma$  are nonnegative tuning parameters,  $\omega_{-i} = \{\omega_{ij} \mid j \neq i\}$ ,  $\|\omega_{-i}\|_1 = \sum_{j \neq i} |\omega_{ij}|$ , and  $\epsilon_i$  is a small positive number for  $i = 1, 2, \dots, p$ . Thus, Liu and Ihler [31] consider optimizing the following objective function

$$f(\Omega; \mathbf{X}, \lambda, \gamma) = L(\mathbf{X}, \Omega) + u_L \cdot p_{\lambda,\gamma}(\Omega), \quad (4)$$

where  $L(\mathbf{X}, \Omega)$  denotes the objective function of the existing method without its penalty terms,  $u_L = 1$  if  $L$  is convex and  $u_L = -1$  if  $L$  is concave for  $\Omega$ . Note that the choice of  $L$  is flexible. For instance,  $L(\mathbf{X}, \Omega)$  can be the log-likelihood function of  $\Omega$  as in the graphical lasso or the squared loss function as in the NS and the SPACE. In this section, we suppose that  $L$  is concave for the purpose of notational simplicity.

To obtain the maximizer of  $f(\Omega; \mathbf{X}, \lambda, \gamma)$ , Liu and Ihler [31] propose the iteratively reweighted  $\ell_1$  regularization procedure based on the minorization-maximization (MM) algorithm [41]. The reweighted procedure iteratively solves the following problem:

$$\Omega^{(k+1)} = \arg \max_{\Omega} L(\mathbf{X}, \Omega) - \sum_{i=1}^p \sum_{j \neq i} \eta_{ij}^{(k)} |\omega_{ij}| - \gamma \sum_{i=1}^p |\omega_{ii}|, \quad (5)$$

where  $\Omega^{(k)} = (\omega_{ij}^{(k)})$  is the estimate at the  $k$ th iteration,  $\|\omega_{-i}^{(k)}\|_1 = \sum_{l \neq i} |\omega_{il}^{(k)}|$ , and  $\eta_{ij}^{(k)} = \lambda \left( 1 / (\|\omega_{-i}^{(k)}\|_1 + \epsilon_i) + 1 / (\|\omega_{-j}^{(k)}\|_1 + \epsilon_j) \right)$ . In practice, [31] suggest  $\epsilon_i = 1$ ,  $\gamma = 2\lambda / \epsilon_i$ , and the initial estimate  $\Omega^{(0)} = I_p$ , where  $I_p$  is the  $p$ -dimensional identity matrix. Note that this reweighted strategy facilitates to estimate the hub nodes by adjusting weights in the penalty term but weights are updated by solely using the observed dataset without previously known information from other literatures.

In this paper, we consider  $L(\mathbf{X}, \Omega) = \log |\Omega| - \text{tr}(S\Omega)$ , which is the same to the component in the objective function of the GLASSO. Thus, we call this procedure as the GLASSO with a reweighted strategy for the scale-free network (GLASSO-SF). As applied in [31], we stop the reweighting iteration after 5 iterations. The R package `glasso` version 1.7 is used to obtain the solution of (5) at each iteration with the penalty matrix  $E^{(k)} = (e_{ij}^{(k)})$ , where  $e_{ij}^{(k)} = \eta_{ij}^{(k)}$  for  $i \neq j$  and  $e_{ii}^{(k)} = 2\lambda$  for  $i = 1, 2, \dots, p$ .

### Path consistency algorithm based on conditional mutual information (PCACMI)

Mutual information (MI) is a widely used measure of dependency between variables in information theory. MI even measures non-linear dependency between variables and can be considered as a generalization of the correlation. Several mutual information (MI) based methods have been developed such as ARACNE [20], CLR [42], and minet [43]. However, similar to the correlation, MI only measures pairwise dependency between two variables. Thus, it usually identifies many undirected interactions between variables. To resolve this difficulty, Zhang et al. [21] propose the information theoretic method for reconstruction of the gene regulatory networks based on the conditional mutual information (CMI).

To be specific, let  $H(X)$  and  $H(X, Y)$  be the entropy of a random variable  $X$  and the joint entropy of random variables  $X$  and  $Y$ , respectively. For two random variables  $X$  and  $Y$ ,  $H(X)$  and  $H(X, Y)$  can be expressed as

$$H(X) = E(-\log f_X(X)), \quad H(X, Y) = E(-\log f_{XY}(X, Y)),$$

where  $f_X(x)$  is the marginal probability density function (PDF) of  $X$  and  $f_{XY}(x, y)$  is the joint PDF of  $X$  and  $Y$ . With these notations, MI is defined as

$$\begin{aligned} I(X, Y) &= E\left(-\log \frac{f_{XY}(X, Y)}{f_X(X)f_Y(Y)}\right) \\ &= H(X) + H(Y) - H(X, Y). \end{aligned} \quad (6)$$

It is known that MI measures dependency between two variables that contain both directed dependency and undirected dependency through other variables. While MI can not distinguish directed and undirected dependency, CMI can measure directed dependency between two variables by conditioning on other variables. CMI for  $X$  and  $Y$  given  $Z$  is defined as

$$I(X, Y|Z) = H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z). \quad (7)$$

To estimate the entropies in (7), Zhang et al. [21] consider the Gaussian kernel density estimator used in [19]. Using the Gaussian kernel density estimator, MI and CMI are defined as

$$\begin{aligned} \hat{I}(X, Y) &= \frac{1}{2} \log \frac{|C(X)| |C(Y)|}{|C(X, Y)|}, \\ \hat{I}(X, Y|Z) &= \frac{1}{2} \log \frac{|C(X, Z)| |C(Y, Z)|}{|C(Z)| |C(X, Y, Z)|}, \end{aligned} \quad (8)$$

where  $|A|$  is the determinant of a matrix  $A$ ,  $C(X)$ ,  $C(Y)$  and  $C(Z)$  are the variances of  $X$ ,  $Y$  and  $Z$ , respectively, and  $C(X, Z)$ ,  $C(Y, Z)$  and  $C(X, Y, Z)$  are the covariance matrices of  $(X, Z)$ ,  $(Y, Z)$  and  $(X, Y, Z)$ , respectively.

To efficiently identify dependent pairs of variables, Zhang et al. [21] adopt the path consistency algorithm (PCA) in [44]. Thus, the authors called their procedure as PCA based on CMI (PCACMI). The PCACMI

method sets  $L = 0$  and calculates with  $L$ -order CMI, which is equivalent to MI if  $L = 0$ . Then, PCACMI removes the pairs of variables such that the maximal CMI of two variables given  $L + 1$  adjacent variables is less than a given threshold  $\alpha$ , where  $\alpha$  determines whether two variables are independent or not and adjacent variables denote variables connected to the two target variables in PCACMI at the previous step. PCACMI repeats the above steps until there is no higher order connection. The MATLAB code for PCACMI is provided by [21] at the author's website <https://sites.google.com/site/xiujunzhangcsb/software/pca-cmi>.

**Conditional mutual inclusive information-based network inference (CMI2NI)**

Recently, Zhang et al. [22] proposed the conditional mutual inclusive information-based network inference (CMI2NI) method that improves the PCACMI method [21]. CMI2NI considers the Kullback-Leibler divergences from the joint probability density function (PDF) of target variables to the interventional PDFs removing the dependency between two variables of interest. Instead of using CMI, CMI2NI uses the conditional mutual inclusive information (CMI2) as the measure of dependency between two variables of interest given other variables. To be specific, we consider three random variables  $X, Y$  and  $Z$ . For these three random variables, the CMI2 between  $X$  and  $Y$  given  $Z$  is defined as

$$CMI2(X, Y|Z) = (D_{KL}(P||P_{X \rightarrow Y}) + D_{KL}(P||P_{Y \rightarrow X})) / 2, \tag{9}$$

where  $D_{KL}(f||g)$  is the Kullback-Leibler divergence from  $f$  to  $g$ ,  $P$  is the joint PDF of  $X, Y$  and  $Z$ , and  $P_{X \rightarrow Y}$  is the interventional probability of  $X, Y$  and  $Z$  for removing the connection from  $X$  to  $Y$ .

With Gaussian assumption on the observed data, the CMI2 for two random variables  $X$  and  $Y$  given  $m$ -dimensional vector  $Z$  can be expressed as

$$CMI2(X, Y|Z) = \frac{1}{4} \left( \text{tr}(C^{-1}\Sigma) + \text{tr}(\tilde{C}^{-1}\tilde{\Sigma}) + \log C_0 + \log \tilde{C}_0 - 2n \right), \tag{10}$$

where  $\Sigma$  is the covariance matrix of  $(X, Y, Z^T)^T$ ,  $\tilde{\Sigma}$  is the covariance matrix of  $(Y, X, Z^T)^T$ ,  $\Sigma_{XZ}$  is the covariance matrix of  $(X, Z^T)^T$ ,  $\Sigma_{YZ}$  is the covariance matrix of  $(Y, Z^T)^T$ ,  $n = m + 2$ , and  $C, \tilde{C}, C_0$  and  $\tilde{C}_0$  are defined with the elements of  $\Sigma, \Sigma_{XZ}, \Sigma_{YZ}, \Sigma^{-1}, \Sigma_{XZ}^{-1}$  and  $\Sigma_{YZ}^{-1}$  (see Theorem 1 in [22] for details). As applied in PCACMI, CMI2NI adopts the path consistency algorithm (PCA) to efficiently calculate the CMI2 estimates. All steps of the PCA in CMI2NI are the same as one of PCACMI if we change the CMI to the CMI2. In the PCA steps

of CMI2NI, two variables are regarded as independent if the corresponding CMI2 estimate is less than a given threshold  $\alpha$ . The MATLAB code for CMI2NI is available at the author's website <https://sites.google.com/site/xiujunzhangcsb/software/cmi2ni>.

**Sparse partial correlation estimation (SPACE)**

In the Gaussian graphical models [45], the conditional dependencies among  $p$  variables can be represented by a graph  $\mathcal{G} = (V, E)$ , where  $V = \{1, 2, \dots, p\}$  is a set of nodes representing  $p$  variables and  $E = \{(i, j) \mid \omega_{ij} \neq 0, 1 \leq i \neq j \leq p\}$  is a set of edges corresponding to the nonzero off-diagonal elements of  $\Omega$ .

To describe the SPACE method, we consider linear models such that for  $i = 1, 2, \dots, p$ ,

$$X_i = \sum_{j \neq i} \beta_{ij} X_j + \epsilon_i \tag{11}$$

where  $\epsilon_i$  is an  $n$ -dimensional random vector from the multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $(1/\omega_{ii})I_n$ , and  $I_n$  is an identity matrix with size of  $n \times n$ . Under normality, the regression coefficients  $\beta_{ij}$ s can be replaced with the partial correlations  $\rho^{ij}$ s by the relationship

$$\beta_{ij} = -\frac{\omega_{ij}}{\omega_{ii}} = \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}}, \tag{12}$$

where  $\rho^{ij} = \text{corr}(X_i, X_j \mid X_k, k \neq i, j) = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}}$  is a partial correlation between  $X_i$  and  $X_j$ . Motivated by the relationship (12), Peng et al. [18] propose the SPACE method for solving the following  $\ell_1$ -regularized problem:

$$\min_{\rho} \frac{1}{2} \sum_{i=1}^p \left\{ w_i \sum_{k=1}^n \left( X_i^k - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} X_j^k \right)^2 \right\} + \lambda \sum_{1 \leq i < j \leq p} |\rho^{ij}|, \tag{13}$$

where  $w_i$  is a nonnegative weight for the  $i$ -th squared error loss.

**Proposed approach incorporating previously known hub information**

**Extended sparse partial correlation estimation (ESPACE)**

In this paper, we assume that some genes (or nodes), which are referred to as hub genes (or hub nodes), regulate many other genes, and we also assume that many of these hub genes were identified from previous experiments. To incorporate information about hub nodes, we propose the extended SPACE (ESPACE) method, which extends the model space by using an additional tuning parameter  $\alpha$  on edges connected to the given hub nodes. This additional tuning parameter can reflect the hub gene information by reducing the penalty on edges connected to hub nodes. To be specific, let  $\mathcal{H}$  be the set of hub nodes

that were previously identified. The ESPACE method we propose solves

$$\min_{\rho} \frac{1}{2} \sum_{i=1}^p \left\{ w_i \sum_{k=1}^n \left( X_i^k - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} X_j^k \right)^2 \right\} + \alpha \lambda \sum_{\substack{i < j \\ (i \in \mathcal{H}) \cup \{j \in \mathcal{H}\}}} |\rho^{ij}| + \lambda \sum_{\substack{i < j \\ i, j \in \mathcal{H}^c}} |\rho^{ij}|, \quad (14)$$

where  $0 < \alpha \leq 1$ . Note that we consider the weights  $w_i$  for the squared error loss as one in this paper. To summarize the process of the proposed method, we depict the flowchart of the ESPACE method in Fig. 1. As described in Fig. 1, the ESPACE has the prior knowledge about hub genes as an additional input, which is the novelty of the proposed method compared to the other existing methods.

### Extended graphical lasso (EGLASSO)

In the Background, we mentioned the proposed procedure is applicable to other methods such as the graphical lasso. For the purpose of fair comparison and the investigation of the performance, we also applied the proposed strategy to the GLASSO, which is the GLASSO incorporating the hub gene information. We call this procedure the extended graphical lasso (EGLASSO). Similar to the ESPACE, the EGLASSO maximizes

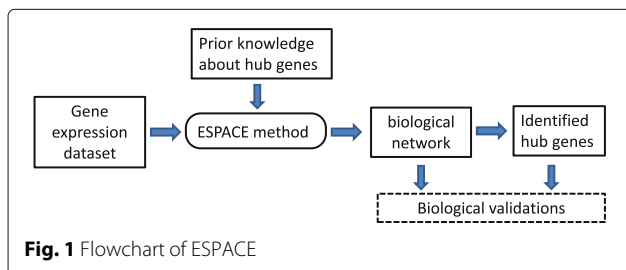
$$\log |\Omega| - \text{tr}(S\Omega) - \alpha \lambda \sum_{\substack{i < j \\ (i \in \mathcal{H}) \cup \{j \in \mathcal{H}\}}} |\omega_{ij}| - \lambda \sum_{\substack{i < j \\ i, j \in \mathcal{H}^c}} |\omega_{ij}|, \quad (15)$$

where  $\lambda \geq 0$  and  $0 < \alpha \leq 1$  are two tuning parameters,  $S$  is the sample covariance matrix,  $\text{tr}(A)$  is the trace of  $A$  and  $\mathcal{H}$  is the set of hub nodes that were previously identified. Note that we can use the R package `glasso` version 1.7 for the EGLASSO by defining the penalty matrix corresponding to the penalty term in (15).

### Active shooting algorithm for ESPACE

To solve (14), we adopt the active shooting algorithm introduced in [18]. We rewrite the problem (14) as

$$\min_{\rho} \frac{1}{2} \left\| \mathbf{Y} - \tilde{\mathbf{X}}\rho \right\|_2^2 + \alpha \lambda \sum_{\substack{i < j \\ (i \in \mathcal{H}) \cup \{j \in \mathcal{H}\}}} |\rho^{ij}| + \lambda \sum_{\substack{i < j \\ i, j \in \mathcal{H}^c}} |\rho^{ij}|, \quad (16)$$



where  $\mathbf{Y} = (\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_p^T)^T$  is an  $n \times p$  column vector;  $\mathbf{X}^{k,l} = \left( \mathbf{0}_{n(k-1) \times 1}^T, \mathbf{X}_{l(k)}^T, \mathbf{0}_{n(l-k-1) \times 1}^T, \mathbf{X}_{k(l)}^T, \mathbf{0}_{n(p-l) \times 1}^T \right)^T$  is an  $n \times p$  column vector as well, with  $\mathbf{X}_{k(l)} = \sqrt{\frac{\omega_{kk}}{\omega_{ll}}} \mathbf{X}_k$ ;

$$\tilde{\mathbf{X}} = (\mathbf{X}^{1,2}; \mathbf{X}^{1,3}; \dots; \mathbf{X}^{1,p}; \mathbf{X}^{2,3}; \mathbf{X}^{2,4}; \dots; \mathbf{X}^{(p-1),p}),$$

and  $\rho = (\rho^{12}; \rho^{13}; \dots; \rho^{1p}; \rho^{23}; \rho^{24}; \dots; \rho^{(p-1)p})^T$ . Let  $\hat{\rho}^{(m)}$  and  $\hat{\omega}_{ii}^{(m)}$  be estimates of  $\rho$  and  $\omega_{ii}$  at the  $m$ -th iteration, respectively. Then, the steps of the modified algorithm are outlined below:

- Step 1: (Initialization of  $\hat{\omega}_{ii}$ ) For  $i = 1, 2, \dots, p$ ,  $\hat{\omega}_{ii}^{(0)} = 1$  and  $s = 0$ .
- Step 2: (Initialization of  $\hat{\rho}$ ) For  $1 \leq i < j \leq p$  and  $m = 0$ ,

$$\hat{\rho}^{ij,(0)} = \text{sign}(\mathbf{Y}^T \mathbf{X}^{ij}) \frac{(|\mathbf{Y}^T \mathbf{X}^{ij}| - \alpha \lambda)_+}{(\mathbf{X}^{ij})^T \mathbf{X}^{ij}} \text{ for } \{i \in \mathcal{H}\} \cup \{j \in \mathcal{H}\},$$

$$\hat{\rho}^{ij,(0)} = \text{sign}(\mathbf{Y}^T \mathbf{X}^{ij}) \frac{(|\mathbf{Y}^T \mathbf{X}^{ij}| - \lambda)_+}{(\mathbf{X}^{ij})^T \mathbf{X}^{ij}} \text{ for } i, j \in \mathcal{H}^c,$$

where  $(x)_+ = \max(x, 0)$  and  $\mathbf{X}^{ij}$ s are defined in (16) with  $\hat{\omega}_{ii}^{(s)}$ .

- Step 3: Define an active set  $\Lambda = \{(i, j) \mid \hat{\rho}^{ij,(m)} \neq 0\}$ .
- Step 4: Iteratively update  $\hat{\rho}^{(m)}$  for  $(k, l) \in \Lambda$ ,

$$\hat{\rho}^{kl,(m)} = \text{sign}((\mathbf{X}^{k,l})^T \boldsymbol{\epsilon}') \frac{(|(\mathbf{X}^{k,l})^T \boldsymbol{\epsilon}'| - \alpha \lambda)_+}{(\mathbf{X}^{k,l})^T \mathbf{X}^{k,l}} \text{ for } \{k \in \mathcal{H}\} \cup \{l \in \mathcal{H}\},$$

$$\hat{\rho}^{kl,(m)} = \text{sign}((\mathbf{X}^{k,l})^T \boldsymbol{\epsilon}') \frac{(|(\mathbf{X}^{k,l})^T \boldsymbol{\epsilon}'| - \lambda)_+}{(\mathbf{X}^{k,l})^T \mathbf{X}^{k,l}} \text{ for } k, l \in \mathcal{H}^c,$$

where  $\boldsymbol{\epsilon}' = \mathbf{Y} - \sum_{(i,j) \neq (k,l)} \hat{\rho}^{ij} \mathbf{X}^{ij}$  and  $\hat{\rho}^{ij}$ s are current estimates at the step for updating the  $(k, l)$ -th partial correlation.

Step 5: Repeat Step 4 until convergence occurs on the active set  $\Lambda$ .

Step 6: Update  $\hat{\rho}^{(m+1)}$  for  $1 \leq i < j \leq p$  by using the equations in Step 4. If the maximum difference between  $\hat{\rho}^{(m+1)}$  and  $\hat{\rho}^{(m)}$  is less than a pre-determined tolerance  $\tau$ , then go to Step 7 with the estimates  $\hat{\rho}^{(m+1)}$ . Otherwise, consider  $m = m + 1$  and go back to Step 3.

Step 7: Update  $\hat{\omega}_{ii}^{(s+1)}$  for  $i = 1, 2, \dots, p$ ,

$$\frac{1}{\hat{\omega}_{ii}^{(s+1)}} = \frac{1}{n} \left\| \mathbf{X}_i - \sum_{j \neq i} \hat{\rho}^{ij,(m+1)} \sqrt{\frac{\hat{\omega}_{jj}^{(s)}}{\hat{\omega}_{ii}^{(s)}}} \mathbf{X}_j \right\|_2^2$$

for  $i = 1, 2, \dots, p$ .

Step 8: Repeat Step 2 through Step 7 with  $s = s + 1$  until convergence occurs on  $\hat{\omega}_{ii}$ s.

Note that the number of iterations of  $\widehat{\omega}_{ii}$ s is usually small for stabilization of the estimates of  $\rho$ . In our numerical study, the estimates of  $\omega_{ii}$ s converge within 10 iterations. Moreover, the inner products such as  $\mathbf{Y}^T \mathbf{X}^{ij}$ , whose complexity is  $O(np)$ , can efficiently be computed by rewriting  $\mathbf{Y}^T \mathbf{X}^{ij} = \sum_{k=1}^n (\sqrt{\omega_{jj}/\omega_{ii}} + \sqrt{\omega_{jj}/\omega_{ii}}) X_i^k X_j^k$ , whose complexity is  $O(n)$ . We implemented the R package `espace`, which is available from <https://sites.google.com/site/dhyeonyu/software>.

**Choice of tuning parameters**

We have introduced the ESPACE method, which relaxes the penalty on edges connected to the hub genes (i.e.,  $\alpha < 1$ ) but uses the same penalty on edges connected to non-hub gene (i.e.,  $\alpha = 1$ ). When no hub genes are involved in a network, ESPACE is reduced to SPACE. For a given  $\lambda$ , this modification allows us to find more edges connected to the hub genes by reducing  $\alpha$ . In practice, however, we do not know the values of  $\lambda$  and  $\alpha$ . In this paper, we consider the GIC-type criterion used in [46] for the Gaussian graphical model to choose the optimal tuning parameters  $(\lambda^*, \alpha^*)$ . Let  $\widehat{\rho}_{(\lambda, \alpha)}^{ij}$  be the  $(i, j)$ -th estimate of partial correlation for given  $\lambda$  and  $\alpha$ . The GIC-type criterion is defined as

$$\text{GIC}(\lambda, \alpha) = \sum_{i=1}^p \left\{ n \cdot \log \text{RSS}_i + \log \log n \log(p - 1) \times \left| \left\{ j : j \neq i, \widehat{\rho}_{\lambda, \alpha}^{ij} \neq 0 \right\} \right| \right\},$$

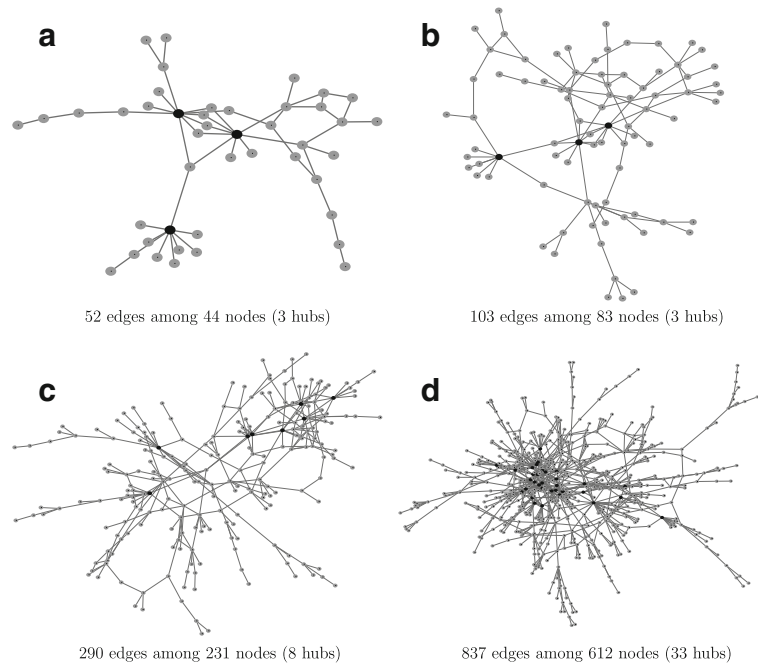
where  $\text{RSS}_i = \left\| \mathbf{X}_i - \sum_{j \neq i} \widehat{\rho}_{(\lambda, \alpha)}^{ij} \mathbf{X}_{j(i)} \right\|_2^2$  and  $|A|$  denotes a cardinality of a set  $A$ . We choose the tuning parameters which minimize the GIC-type criterion,

$$(\lambda^*, \alpha^*) = \underset{\lambda, \alpha}{\text{argmin}} \text{GIC}(\lambda, \alpha).$$

**Simulation studies**

**Simulation settings**

In this simulation, we consider four real protein-protein interaction (PPI) networks used in a comparative study [24], which were partially selected from the human protein reference database [47]. As mentioned earlier, genes whose degrees are greater than 7 and above the 0.95 quantile of the degree distribution are thought of as hub genes. Figure 2 shows the four PPI networks and their hub genes. Let  $p$  be the number of nodes in a network. We consider the number of samples as  $p/2$  and  $p$  and



**Fig. 2** The network structures of the four simulated networks. The structure of the real protein-protein interaction networks [47] were used to construct networks of different sizes by varying the number of references required to support each connection. In the degree distribution, the 0.95 quantile is 7 (connections), so the nodes with more than 7 connections were defined as hub nodes, which are represented as *black nodes* in the network structure. **a** 52 edges among 44 nodes (3 hubs), **b** 103 edges among 83 nodes (3 hubs), **c** 290 edges among 231 nodes (8 hubs) and **d** 837 edges among 612 nodes (33 hubs)

generate samples from the multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\Sigma$  defined with  $(\Sigma)_{ij} = (\Omega^{-1})_{ij} / \sqrt{(\Omega^{-1})_{ii}(\Omega^{-1})_{jj}}$ , where  $\Omega$  is a concentration matrix corresponding to a given network structure. To generate a positive definite concentration matrix, we use the following procedure as described in [18]:

Step G1: For a given edge set  $E$ , we generate an initial concentration matrix  $\tilde{\Omega} = (\tilde{\omega}_{ij})_{1 \leq i, j \leq p}$  with

$$\tilde{\omega}_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j, (i, j) \notin E \\ \sim Unif(D) & i \neq j, (i, j) \in E \end{cases}$$

where  $D = [-1, -0.5] \cup [0.5, 1]$ .

Step G2: For positive definiteness and symmetry of the concentration matrix, we define a concentration matrix  $\Omega = (\omega_{ij})_{1 \leq i, j \leq p}$  as

$$\Omega = \frac{1}{2} (A + A^T),$$

where  $A = (a_{ij})_{1 \leq i, j \leq p}$ ,  $a_{ij} = \tilde{\omega}_{ij} / (1.5 \cdot d_i)$  and  $d_i = \sum_{k \neq i} |\tilde{\omega}_{ik}|$  for  $i = 1, 2, \dots, p$ .

Step G3: Set  $\omega_{ii} = 1$  for  $i = 1, 2, \dots, p$  and  $\omega_{ij} = 0.1 \cdot \text{sign}(\omega_{ij})$  if  $0 < |\omega_{ij}| < 0.1$ .

With these four networks, we have conducted the numerical comparisons of the ESPACE and the SPACE methods, as well as seven other methods including the other reviewed existing methods and EGLASSO. For the purpose of fair comparison, we select the optimal model by the GIC for SPACE, ESPACE, GLASSO, GLASSO-SF, and EGLASSO. Since there is no specific rule for the model selection in the other methods, we set the level  $\alpha = 0.2$  for GeneNet and NS, and the threshold  $\alpha = 0.03$  for PCACMI and CMI2NI. Note that the pre-determined level  $\alpha = 0.2$  is a default of the GeneNet package and used in [35]. The pre-determined threshold  $\alpha = 0.03$  was used in [21, 22].

Note that all the existing methods need  $O(p^2)$  memory space to store and calculate values corresponding to the interactions between variables. We can reduce this memory consumption when the whole variables can be divided into several conditionally independent blocks by using the condition described in [16].

### Sensitivity analysis on random noise in the observed data

To investigate the effect of the random noise contained in the observed data, we consider sensitivity analysis for the variance of the random noise. To be specific, suppose that a random vector  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  follows the multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\Sigma$ , a vector of random noise  $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)^T$  follows the multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\sigma_\varepsilon^2 I$ , and  $\mathbf{X}$  and  $\varepsilon$  are independent,

where  $I$  is the identity matrix. Furthermore, we assume that an observed random vector  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_p)^T$  such that

$$\mathbf{Z} = \mathbf{X} + \varepsilon. \tag{17}$$

Thus, the covariance matrix of  $\mathbf{Z}$  becomes  $\Sigma + \sigma_\varepsilon^2 I$ , which may have a different conditional dependent structure to one of  $\mathbf{X}$ .

For example, if we consider  $\sigma_\varepsilon^2 = 0.5$  and the following  $\Sigma$  and  $\Sigma_Z$

$$\Sigma = \begin{pmatrix} 15/11 & -8/11 & 2/11 \\ -8/11 & 16/11 & -4/11 \\ 2/11 & -4/11 & 12/11 \end{pmatrix}, \Sigma_Z = \Sigma + \sigma_\varepsilon^2 I, \tag{18}$$

then the inverse matrices of  $\Sigma$  and  $\Sigma_Z$  are calculated as

$$\Sigma^{-1} = \begin{pmatrix} 1 & 0.5 & 0 \\ -0.5 & 1 & 0.25 \\ 0 & 0.25 & 1 \end{pmatrix} \text{ and} \tag{19}$$

$$\Sigma_Z^{-1} = \begin{pmatrix} 0.63 & 0.23 & -0.02 \\ 0.23 & 0.62 & 0.12 \\ -0.02 & 0.12 & 0.66 \end{pmatrix}, \text{ respectively.}$$

Thus, we can see that  $Z_1$  and  $Z_3$  are conditionally dependent given  $Z_2$  while  $X_1$  and  $X_3$  are conditionally independent given  $X_2$ . Moreover, the nonzero partial correlations decrease when the variance of the random noises increases. From these observations, the performance of the estimation becomes worse if the variance of the random noise increases.

In this sensitivity analysis, we consider  $\sigma_\varepsilon^2 = 0, 0.01, 0.1, 0.25, 0.5$  and  $p = 231$  and  $n = 115, 231$  with the same network structure as the one of  $p = 231$  in Fig. 2. To focus on the proposed method, we apply the SPACE and the ESPACE methods to the 50 generated datasets containing random noise having variance  $\sigma_\varepsilon^2$ .

### Performance measures

To investigate the gains from the extension, we use five performance measures: sensitivity (SEN), specificity (SPE), false discovery rate (FDR), mis-specification rate (MISR) and Matthews correlation coefficients (MCC). Note that the MCC, which lies between  $-1$  and  $+1$ , has been used to measure the performance of binary classification, where  $+1$ ,  $0$ , and  $-1$  denote a perfect classification, a random classification, and a total discordance of classification, respectively. Let  $\rho$  and  $\hat{\rho}_{\lambda, \alpha}$  be  $(p(p-1)/2)$ -dimensional vectors of the true and estimated partial



correlation, respectively. The above five measures are defined as

$$\begin{aligned} \text{SEN} &\equiv \text{TP}/(\text{TP} + \text{FN}), \quad \text{SPE} \equiv \text{TN}/(\text{TN} + \text{FP}), \\ \text{FDR} &\equiv \text{FP}/(\text{TP} + \text{FP}), \quad \text{MISR} \equiv (\text{FN} + \text{FP})/(p(p-1)/2) \quad \text{and} \\ \text{MCC} &\equiv \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \end{aligned}$$

where  $\text{TP} = \sum_{i < j} I(\rho^{ij} \neq 0)I(\hat{\rho}_{\lambda, \alpha}^{ij} \neq 0)$ ,  $\text{FP} = \sum_{i < j} I(\rho^{ij} = 0)I(\hat{\rho}_{\lambda, \alpha}^{ij} \neq 0)$ ,  $\text{FN} = \sum_{i < j} I(\rho^{ij} \neq 0)I(\hat{\rho}_{\lambda, \alpha}^{ij} = 0)$  and  $\text{TN} = \sum_{i < j} I(\rho^{ij} = 0)I(\hat{\rho}_{\lambda, \alpha}^{ij} = 0)$ .

### Application to *Escherichia coli* dataset

We applied the ESPACE method to the largest public *Escherichia coli* (*E.coli*) microarray dataset available from the Many Microbe Microarrays database (M3D) [48]. The M3D contains 907 microarrays measured under 466 experimental conditions using Affymetrix GeneChip *E.coli* genome arrays. Microarrays from the same experimental conditions were averaged to derive the mean expression. The data set (“E\_coli\_v4\_Build\_6” from the M3D) contains the expression levels of 4297 genes from 446 samples. In the *E.coli* genome, a number of studies have been conducted to identify transcriptional regulations. The RegulonDB [49] curates the largest and best-known information on the transcriptional regulation of *E.coli*. To combine the information from the above two databases, we focus on the 1623 genes reported in both the M3D and the RegulonDB. As mentioned before, the TFs are known to regulate many other genes in the genome and can be considered potential hubs. To incorporate the information about the potential hubs, we used a list of 180 known TF-encoding genes from the RegulonDB. The RegulonDB also provides 3811 transcriptional interactions among the 1623 genes, which were used as the gold standard to evaluate the accuracy of the constructed networks.

### Application to lung cancer adenocarcinoma dataset

Lung cancer is the leading cause of death from cancer, both in the United States and worldwide; it has a 5-year survival rate of approximately 15% [50]. The progression and metastasis of lung cancer varies greatly among early stage lung cancer patients. To customize treatment plans for individual patients, it is important to identify prognostic or predictive biomarkers, which allows for more precise classification of lung cancer patients. In this study, we applied the extended SPACE method to reconstruct the gene regulatory network in lung cancer. Exploring network structures can facilitate comprehension of biological mechanisms underlying lung cancer and identification of important genes that could be potential lung cancer biomarkers. We constructed the gene network using microarray data from 442 lung cancer adenocarcinoma patients in the Lung Cancer Consortium study [51]. For

detail about preprocessing this dataset, please refer to [7]. First, univariate Cox regression was used to identify the genes whose expression levels are correlated with patient survival outcomes, after adjusting for clinical factors such as study site, age, gender, and stage. The false discovery rate (FDR) was then calculated using a Beta-Uniform model [52]. By controlling the FDR to less than 10%, we identified 794 genes that were associated with the survival outcome of lung cancer patients. Among these 794 genes, 22 were found to appear among the 236 carefully curated cancer genes of the FoundationOne™ gene panel (Foundation Medicine, Inc.). Current biological knowledge indicates genes from this panel play a key role in various types of cancer. These 22 genes were then input as known hub genes to the ESPACE method.

## Results and discussion

### Simulation results

#### Comparison results for existing methods

For each network, we generated 50 datasets and reconstructed the network from each dataset using nine different network construction methods, including both the SPACE and the ESPACE methods. In addition to the five performance measures, we also measure the computation time (Time) of each method to compare the efficiency. Note that all methods are executed on R software [53] for the purpose of fair comparison. We implemented the R codes for PCACMI and CMI2NI using the authors' MATLAB codes. The computation times are measured in CPU time (seconds) by using a desktop PC (Intel Core(TM) i7-4790K CPU (4.00 GHz) and 32 GB RAM).

Tables 1, 2, 3 and 4 report the averages and standard errors of the number of the estimated edges, the five performance measures of the estimation of the network structures and computation times with the optimal tuning parameter  $\lambda^*$  for SPACE, GLASSO, GLASSO-SF; the optimal tuning parameters  $\alpha^*$  and  $\lambda^*$  for ESPACE and EGLASSO; and the pre-determined  $\alpha$  for GeneNet, NS, PCACMI, and CMI2NI.

Overall, ESPACE has the best performance in estimating network structures in terms of the MCC and the MISR except for the case  $(p, n) = (83, 41)$ , where ESPACE has the second smallest FDR while the MCC and the MISR of ESPACE show the moderate performance among all methods. In the case  $(p, n) = (83, 41)$ , the CMI-based methods have better performance than the others in terms of the MCC and the MISR, but the CMI-based methods also have the large FDRs ( $\approx 41\%$ ) more than double of those of the other methods. As we described in the Methods Section, the MCC has been used to measure the performance of binary classification and the MISR denotes the total error rate. Thus, this comparison results show that ESPACE is favorable for the identification of edges for the networks with high-dimensional data.

**Table 1** The averages of the number of estimated edges, the five performance measures and the computation time (sec.) over 50 datasets

$p$	$n$	Method	$ \hat{E} $	SEN	SPE	FDR	MISR	MCC	Time	
44 ( $ E  = 52$ )	22	GeneNet	0.80 (0.31)	1.00 (0.37)	99.97 (0.02)	4.38 (2.05)	5.47 (0.01)	3.25 (1.06)	0.02 (0.00)	
		NS	0.66 (0.13)	1.19 (0.23)	100.00 (0.00)	2.50 (2.05)	5.44 (0.01)	6.65 (1.16)	0.03 (0.00)	
		SPACE	10.48 (1.21)	12.50 (1.32)	99.55 (0.07)	33.38 (3.40)	5.23 (0.05)	24.58 (1.69)	0.01 (0.00)	
		ESPACE	12.06 (1.12)	16.50 (1.33)	99.61 (0.06)	23.62 (2.52)	4.96 (0.05)	31.46 (1.67)	0.00 (0.00)	
		GLASSO	6.64 (0.95)	6.88 (0.66)	99.66 (0.08)	33.86 (4.12)	5.44 (0.06)	17.98 (1.19)	0.00 (0.00)	
		GLASSO-SF	7.14 (1.20)	6.77 (0.70)	99.60 (0.10)	32.51 (4.11)	5.51 (0.07)	17.39 (1.12)	0.04 (0.00)	
		EGLASSO	6.34 (0.74)	8.65 (0.96)	99.79 (0.04)	26.36 (3.47)	5.22 (0.04)	22.14 (1.49)	0.00 (0.00)	
		PCACMI	55.60 (0.62)	33.15 (0.73)	95.71 (0.07)	68.93 (0.65)	7.73 (0.08)	27.98 (0.71)	0.21 (0.01)	
	44	44	GeneNet	9.28 (1.32)	13.58 (1.48)	99.75 (0.07)	12.19 (2.11)	4.99 (0.05)	29.73 (1.61)	0.02 (0.00)
			NS	4.16 (0.23)	7.65 (0.42)	99.98 (0.01)	3.48 (1.26)	5.10 (0.02)	25.92 (0.75)	0.03 (0.00)
			SPACE	27.28 (1.50)	37.38 (1.73)	99.12 (0.08)	25.79 (1.47)	4.27 (0.07)	48.78 (1.71)	0.01 (0.00)
			ESPACE	23.12 (1.20)	34.38 (1.36)	99.41 (0.07)	20.03 (1.47)	4.16 (0.06)	50.01 (0.94)	0.00 (0.00)
			GLASSO	11.32 (1.35)	13.12 (1.08)	99.50 (0.10)	26.48 (3.03)	5.25 (0.05)	27.20 (0.99)	0.00 (0.00)
			GLASSO-SF	13.62 (1.52)	14.38 (1.23)	99.31 (0.11)	30.92 (3.22)	5.36 (0.07)	26.77 (1.16)	0.04 (0.00)
			EGLASSO	14.34 (1.10)	22.38 (1.63)	99.70 (0.04)	16.31 (1.81)	4.55 (0.07)	39.94 (1.73)	0.00 (0.00)
			PCACMI	26.70 (0.44)	34.00 (0.70)	98.99 (0.05)	33.39 (1.28)	4.58 (0.07)	45.47 (0.89)	0.18 (0.00)
		CMI2NI	28.84 (0.48)	35.50 (0.67)	98.84 (0.04)	35.75 (0.96)	4.64 (0.06)	45.54 (0.74)	0.27 (0.01)	

The reported values for the SEN, SPE, FDR, MISR and MCC were multiplied by 100. Numbers in the parentheses denote the standard errors

**Table 2** The averages of the number of estimated edges, the five performance measures and the computation time (sec.) over 50 datasets

$\rho$	$n$	Method	$ \hat{E} $	SEN	SPE	FDR	MISR	MCC	Time
83 ( $ E  = 103$ )	41	GeneNet	6.04 (0.61)	5.32 (0.52)	99.98 (0.00)	6.57 (1.64)	2.88 (0.01)	19.21 (1.46)	0.05 (0.00)
		NS	2.36 (0.20)	2.21 (0.19)	100.00 (0.00)	2.83 (1.42)	2.96 (0.01)	13.29 (0.82)	0.09 (0.01)
		SPACE	4.62 (0.79)	4.04 (0.64)	99.99 (0.00)	6.41 (2.36)	2.92 (0.02)	16.18 (1.40)	0.03 (0.00)
		ESPACE	7.28 (0.94)	6.37 (0.78)	99.98 (0.01)	5.75 (1.38)	2.86 (0.02)	21.04 (1.55)	0.01 (0.00)
		GLASSO	11.40 (0.87)	8.78 (0.59)	99.93 (0.01)	16.81 (2.05)	2.83 (0.01)	25.48 (0.88)	0.00 (0.00)
		GLASSO-SF	9.90 (0.81)	7.53 (0.53)	99.94 (0.01)	16.30 (2.08)	2.86 (0.01)	23.44 (0.84)	0.12 (0.00)
		EGLASSO	11.28 (0.86)	8.89 (0.59)	99.94 (0.01)	15.36 (1.86)	2.82 (0.01)	25.93 (0.88)	0.00 (0.00)
		PCACMI	48.44 (0.82)	27.24 (0.55)	99.38 (0.02)	41.86 (1.00)	2.80 (0.03)	38.52 (0.68)	0.54 (0.01)
		CMI2NI	48.72 (0.85)	27.88 (0.55)	99.39 (0.02)	40.82 (0.96)	2.77 (0.03)	39.35 (0.65)	0.54 (0.01)
		GeneNet	34.74 (0.83)	31.17 (0.61)	99.92 (0.01)	7.05 (0.74)	2.16 (0.02)	52.97 (0.50)	0.06 (0.00)
		NS	15.14 (0.43)	14.52 (0.41)	99.99 (0.00)	1.07 (0.38)	2.59 (0.01)	37.20 (0.55)	0.15 (0.01)
		SPACE	51.84 (1.95)	41.03 (1.28)	99.71 (0.02)	16.81 (1.06)	2.07 (0.03)	56.67 (1.11)	0.05 (0.00)
		ESPACE	52.34 (1.46)	42.45 (0.86)	99.74 (0.02)	15.42 (0.96)	2.00 (0.02)	58.82 (0.54)	0.03 (0.00)
		GLASSO	27.98 (1.33)	23.24 (0.96)	99.88 (0.02)	12.57 (1.32)	2.44 (0.03)	43.63 (0.97)	0.00 (0.00)
GLASSO-SF	28.04 (1.31)	22.82 (0.90)	99.86 (0.02)	14.53 (1.41)	2.47 (0.02)	42.86 (0.89)	0.13 (0.00)		
EGLASSO	30.06 (1.33)	26.17 (1.03)	99.91 (0.01)	8.91 (1.15)	2.33 (0.03)	47.48 (1.01)	0.00 (0.00)		
PCACMI	28.44 (0.44)	25.96 (0.49)	99.95 (0.01)	6.09 (0.79)	2.29 (0.02)	48.66 (0.61)	0.56 (0.00)		
CMI2NI	28.66 (0.45)	26.76 (0.46)	99.97 (0.01)	3.85 (0.61)	2.25 (0.02)	50.02 (0.53)	0.54 (0.01)		

The reported values for the SEN, SPE, FDR, MISR and MCC were multiplied by 100. Numbers in the parentheses denote the standard errors

**Table 3** The averages of the number of estimated edges, the five performance measures and the computation time (sec.) over 50 datasets

$p$	$n$	Method	$ \hat{E} $	SEN	SPE	FDR	MISR	MCC	Time
231 ( $ E  = 290$ )	115	GeneNet	115.98 (1.27)	38.39 (0.39)	99.98 (0.00)	3.94 (0.26)	0.69 (0.00)	60.46 (0.30)	0.56 (0.01)
		NS	54.28 (0.85)	18.71 (0.29)	100.00 (0.00)	0.04 (0.04)	0.89 (0.00)	42.99 (0.34)	0.34 (0.01)
		SPACE	160.38 (2.39)	46.90 (0.43)	99.91 (0.01)	14.78 (0.69)	0.67 (0.00)	62.85 (0.28)	0.36 (0.01)
		ESPACE	172.74 (1.58)	49.41 (0.39)	99.89 (0.00)	16.95 (0.41)	0.66 (0.00)	63.74 (0.31)	0.13 (0.00)
		GLASSO	85.78 (3.64)	28.26 (1.12)	99.99 (0.00)	3.87 (0.41)	0.80 (0.01)	51.25 (1.03)	0.03 (0.00)
		GLASSO-SF	76.64 (3.26)	25.19 (1.01)	99.99 (0.00)	4.03 (0.46)	0.83 (0.01)	48.31 (0.98)	1.06 (0.02)
		EGLASSO	86.50 (3.54)	28.63 (1.11)	99.99 (0.00)	3.51 (0.39)	0.79 (0.01)	51.71 (1.05)	0.04 (0.00)
		PCACMI	73.42 (0.94)	25.14 (0.33)	100.00 (0.00)	0.72 (0.14)	0.82 (0.00)	49.70 (0.34)	4.16 (0.05)
		CMi2NI	73.42 (0.94)	25.17 (0.33)	100.00 (0.00)	0.62 (0.13)	0.82 (0.00)	49.75 (0.34)	6.25 (0.07)
		GeneNet	173.78 (1.21)	56.66 (0.28)	99.96 (0.00)	5.38 (0.30)	0.51 (0.00)	72.99 (0.18)	0.74 (0.01)
		NS	128.10 (0.54)	44.15 (0.19)	100.00 (0.00)	0.05 (0.03)	0.61 (0.00)	66.22 (0.14)	0.97 (0.01)
		SPACE	235.54 (2.20)	68.37 (0.35)	99.86 (0.01)	15.62 (0.50)	0.49 (0.00)	75.68 (0.22)	0.60 (0.00)
		ESPACE	235.86 (1.99)	69.35 (0.32)	99.87 (0.01)	14.55 (0.49)	0.47 (0.00)	76.72 (0.23)	0.23 (0.01)
		GLASSO	222.38 (2.62)	64.97 (0.47)	99.87 (0.01)	15.00 (0.54)	0.51 (0.00)	74.00 (0.20)	0.07 (0.00)
GLASSO-SF	176.66 (2.11)	56.42 (0.35)	99.95 (0.01)	7.12 (0.52)	0.52 (0.00)	72.13 (0.24)	0.70 (0.02)		
EGLASSO	222.86 (2.57)	65.68 (0.46)	99.88 (0.01)	14.26 (0.54)	0.50 (0.00)	74.74 (0.21)	0.09 (0.01)		
PCACMI	74.28 (0.79)	25.61 (0.27)	100.00 (0.00)	0.02 (0.02)	0.81 (0.00)	50.36 (0.27)	6.23 (0.13)		
CMi2NI	74.28 (0.79)	25.61 (0.27)	100.00 (0.00)	0.02 (0.02)	0.81 (0.00)	50.36 (0.27)	7.99 (0.11)		

The reported values for the SEN, SPE, FDR, MISR and MCC were multiplied by 100. Numbers in the parentheses denote the standard errors

**Table 4** The averages of the number of estimated edges, the five performance measures and the computation time (sec.) over 50 datasets

$\rho$	$n$	Method	$ \hat{E} $	SEN	SPE	FDR	MISR	MCC	Time	
612 ( $ E  = 837$ )	306	GeneNet	597.72 (4.89)	55.42 (0.22)	99.93 (0.00)	22.24 (0.41)	0.27 (0.00)	65.49 (0.15)	3.91 (0.03)	
		NS	343.52 (0.86)	41.00 (0.10)	100.00 (0.00)	0.10 (0.03)	0.26 (0.00)	63.91 (0.08)	5.72 (0.07)	
		SPACE	781.04 (8.90)	66.72 (0.31)	99.88 (0.00)	28.21 (0.55)	0.27 (0.00)	69.02 (0.22)	16.57 (0.14)	
		ESPACE	765.50 (5.95)	67.36 (0.29)	99.89 (0.00)	26.22 (0.41)	0.25 (0.00)	70.35 (0.22)	3.69 (0.05)	
		GLASSO	1097.32 (6.08)	65.66 (0.18)	99.71 (0.00)	49.86 (0.25)	0.45 (0.00)	57.15 (0.17)	4.86 (0.37)	
		GLASSO-SF	1069.56 (14.93)	60.67 (0.64)	99.70 (0.01)	52.38 (0.29)	0.48 (0.00)	53.45 (0.27)	29.36 (1.85)	
		EGLASSO	1042.64 (8.54)	67.86 (0.34)	99.74 (0.00)	45.43 (0.28)	0.40 (0.00)	60.63 (0.18)	7.59 (0.65)	
		PCACMI	272.08 (0.99)	27.40 (0.11)	99.98 (0.00)	15.70 (0.24)	0.35 (0.00)	47.94 (0.14)	42.09 (1.41)	
	612	612	GeneNet	727.94 (4.53)	68.80 (0.20)	99.92 (0.00)	20.80 (0.32)	0.22 (0.00)	73.69 (0.13)	5.56 (0.06)
			NS	453.50 (1.19)	54.14 (0.14)	100.00 (0.00)	0.08 (0.02)	0.21 (0.00)	73.47 (0.10)	25.69 (0.38)
			SPACE	983.38 (6.38)	84.01 (0.24)	99.85 (0.00)	28.39 (0.34)	0.22 (0.00)	77.43 (0.17)	62.72 (1.41)
			ESPACE	983.84 (4.88)	84.96 (0.19)	99.85 (0.00)	27.66 (0.27)	0.21 (0.00)	78.28 (0.14)	17.86 (0.63)
			GLASSO	1467.52 (5.04)	85.61 (0.11)	99.60 (0.00)	51.15 (0.17)	0.47 (0.00)	64.47 (0.12)	29.86 (1.21)
			GLASSO-SF	1615.16 (6.56)	85.60 (0.13)	99.52 (0.00)	55.61 (0.17)	0.55 (0.00)	61.42 (0.13)	117.78 (3.68)
			EGLASSO	1385.60 (5.37)	87.87 (0.12)	99.65 (0.00)	46.89 (0.20)	0.40 (0.00)	68.14 (0.14)	38.67 (1.35)
			PCACMI	273.22 (0.98)	27.68 (0.11)	99.98 (0.00)	15.20 (0.17)	0.35 (0.00)	48.33 (0.12)	39.25 (0.59)
		CMI2NI	298.38 (0.97)	29.55 (0.09)	99.97 (0.00)	17.10 (0.11)	0.34 (0.00)	49.37 (0.09)	59.62 (0.84)	

The reported values for the SEN, SPE, FDR, MISR and MCC were multiplied by 100. Numbers in the parentheses denote the standard errors

In addition, we made several interesting observations from the results of the our simulation study. First, ESPACE and EGLASSO improve SPACE and GLASSO in terms of the FDR, the MISR, and the MCC for almost scenarios, respectively. The only exception is the case  $(p, n) = (231, 115)$  for the ESPACE and the SPACE methods. In this case, although the FDR of ESPACE increases 2.17% compared to one of SPACE, ESPACE still improves SPACE in terms of the SEN, the MISR, and the MCC. This suggests that our proposed strategy, which incorporates the previously known hub information, can reduce the errors in estimating network structures compared to the existing method without considering known hub information. Second, GeneNet controls the FDR relatively close to the given level  $\alpha$  while the FDRs of NS are controlled conservatively. For instance, the FDRs of GeneNet are measured between 3.94 and 22.24% and NS has the FDRs less than 3.48%. Note that GeneNet and NS control the FDR under 20% ( $\alpha = 0.2$ ) in this simulation study. Third, all methods except the CMI-based methods (the PCACMI and the CMI2NI) have similar efficiency for the relatively low dimensions ( $p = 44, 83$ ). The CMI-based methods are relatively slower than the other methods for all the scenarios except for the case  $(p, n) = (612, 612)$ , where GLASSO-SF is the slowest and 1.4 times slower than CMI2NI. CMI2NI is slightly slower than PCACMI for the relatively high dimensions ( $p = 231, 612$ ). Finally, even though ESPACE is not the fastest method among the nine methods we consider, there is no overall winner and ESPACE is the third best in terms of the computation time for  $p = 231, 612$  except for the case  $(p, n) = (612, 612)$  where ESPACE is faster than SPACE, GLASSO-SF, PCACMI and CMI2NI.

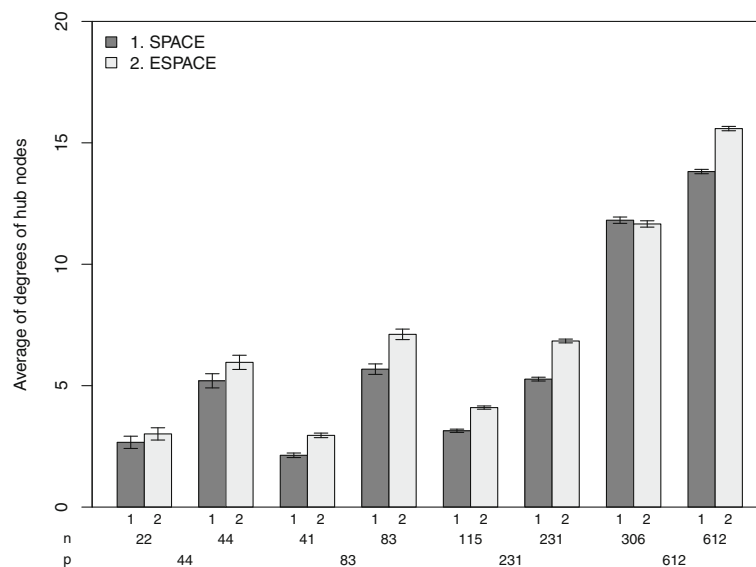
To investigate the other property of the proposed approach, we depict barplots of the averages of degrees of known hub genes over 50 datasets for ESPACE and SPACE in Fig. 3. Figure 3 shows that ESPACE tends to find more edges connected to known hub genes than SPACE. The only exception is the case  $(p, n) = (612, 306)$ , where the average by the ESPACE is 0.57 less than that by SPACE. We conjecture this is simply due the difference in the number of estimated edges, which by ESPACE is 15.54 less than that of SPACE on average. This property is due to the result that the averages of  $\alpha^*$  selected by the GIC in the ESPACE method lie between 0.76 and 0.97 for all the scenarios, which indicates that ESPACE has incorporated prior information about the hub genes and reduced the penalty on edges connected to known hub genes.

**Results of sensitivity analysis on random noise**

Table 5 reports the averages of the number of the estimated edges and the five performance measures. From the results in Table 5, we can see that the performance of estimation decreases when the variance of the random noises increases in both the SPACE and the ESPACE. For a relatively small sample size ( $n = 115$ ), both the SPACE and the ESPACE are more sensitive to the variance  $\sigma_\epsilon^2$  compared to the case of  $n = 231$ . Even though the performance of two methods decreases by similar amounts as the variance  $\sigma_\epsilon^2$  increases, the ESPACE has better performance than the SPACE in terms of the MCC and the MISR.

**Comparison of the identified GRNs in *Escherichia coli* dataset**

In this study, we compared the performance of network construction using the SPACE and the ESPACE methods



**Fig. 3** Plots of the averages of degrees of hub nodes over the simulated 50 datasets. Vertical lines denote 95% confidence intervals of the averages

**Table 5** The averages of the number of estimated edges and the five performance measures over 50 datasets

$\rho$	$n$	$\sigma_\epsilon^2$	Method	$ \hat{E} $	SEN	SPE	FDR	MISR	MCC	
231 ( $ E  = 290$ )	115	0	SPACE	160.38 (2.39)	46.90 (0.43)	99.91 (0.01)	14.78 (0.69)	0.67 (0.00)	62.85 (0.28)	
			ESPACE	172.74 (1.58)	49.41 (0.39)	99.89 (0.00)	16.95 (0.41)	0.66 (0.00)	63.74 (0.31)	
		0.01	SPACE	156.00 (2.79)	45.97 (0.57)	99.91 (0.01)	14.06 (0.66)	0.68 (0.00)	62.45 (0.33)	
			ESPACE	153.16 (2.64)	45.79 (0.55)	99.92 (0.01)	12.85 (0.67)	0.67 (0.00)	62.78 (0.34)	
		0.1	SPACE	106.08 (5.11)	32.61 (1.41)	99.96 (0.00)	9.15 (0.80)	0.78 (0.01)	53.07 (1.21)	
			ESPACE	104.30 (4.96)	32.31 (1.37)	99.96 (0.00)	8.64 (0.73)	0.78 (0.01)	53.15 (1.10)	
		0.25	SPACE	44.76 (1.75)	14.81 (0.54)	99.99 (0.00)	3.63 (0.42)	0.94 (0.01)	37.27 (0.62)	
			ESPACE	49.14 (1.47)	16.12 (0.45)	99.99 (0.00)	4.59 (0.41)	0.92 (0.00)	38.79 (0.53)	
	0.5	SPACE	55.88 (0.90)	14.97 (0.25)	99.95 (0.00)	22.20 (0.72)	0.98 (0.00)	33.81 (0.37)		
		ESPACE	57.34 (1.03)	15.32 (0.31)	99.95 (0.00)	22.49 (0.69)	0.97 (0.00)	34.14 (0.44)		
	231	231	0	SPACE	235.54 (2.20)	68.37 (0.35)	99.86 (0.01)	15.62 (0.50)	0.49 (0.00)	75.68 (0.22)
				ESPACE	235.86 (1.99)	69.35 (0.32)	99.87 (0.01)	14.55 (0.49)	0.47 (0.00)	76.72 (0.23)
			0.01	SPACE	230.90 (2.11)	67.54 (0.34)	99.87 (0.01)	15.00 (0.46)	0.49 (0.00)	75.50 (0.21)
				ESPACE	231.86 (2.21)	68.54 (0.33)	99.87 (0.01)	14.05 (0.55)	0.47 (0.01)	76.49 (0.24)
			0.1	SPACE	214.30 (2.22)	62.28 (0.37)	99.87 (0.01)	15.48 (0.53)	0.54 (0.00)	72.25 (0.25)
				ESPACE	214.04 (2.33)	63.22 (0.35)	99.88 (0.01)	14.07 (0.56)	0.52 (0.00)	73.41 (0.22)
0.25			SPACE	184.76 (2.38)	54.10 (0.43)	99.89 (0.01)	14.77 (0.60)	0.61 (0.00)	67.57 (0.27)	
			ESPACE	181.02 (1.42)	54.52 (0.27)	99.91 (0.00)	12.53 (0.42)	0.58 (0.00)	68.78 (0.21)	
0.5		SPACE	112.28 (3.23)	35.38 (0.82)	99.96 (0.00)	7.89 (0.61)	0.74 (0.01)	56.55 (0.54)		
		ESPACE	123.06 (2.80)	38.58 (0.65)	99.96 (0.00)	8.48 (0.60)	0.71 (0.00)	58.98 (0.37)		

The reported values for the SEN, SPE, FDR, MISR and MCC were multiplied by 100. Numbers in the parentheses denote the standard errors

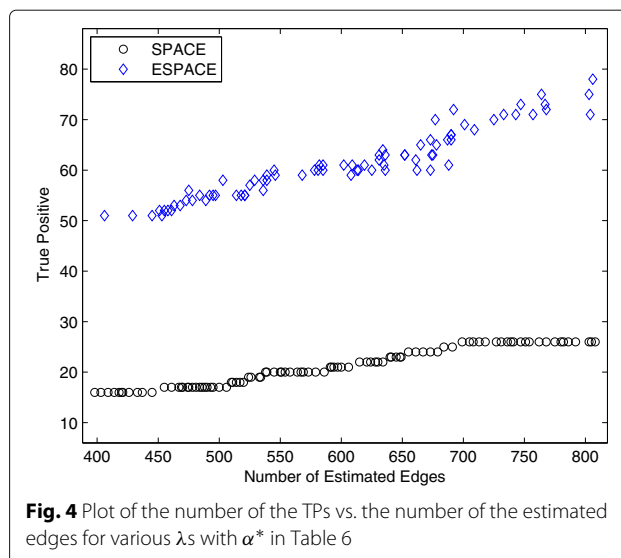
for the model selected by the GIC. We report the number of estimated edges and the true positives, which are matched to the transcriptional interactions in the RegulonDB, in Table 6. The SPACE method estimated 368 edges among 524 genes, which contain 16 TF-encoding genes, and identified 16 transcriptional interactions as true positives. In comparison, the ESPACE method estimated 349 edges among 478 genes containing 29 TF-encoding genes and found 45 transcriptional interactions in the RegulonDB. The ESPACE method found more interactions than the SPACE method and increased the ratio of the number of TPs versus the number of estimated edges as 8.55%. Figure 4 shows the number of TPs vs. the number of estimated edges for various  $\lambda$  values with  $\alpha^*$  in Table 6. The number of TPs of the ESPACE method is consistently greater than those of the SPACE method at similar sparsity. These results clearly indicate that incorporating potential hub gene information improves the accuracy of network construction.

#### Comparison of the identified GRNs in lung cancer adenocarcinoma dataset

We again compared the performances of network construction using the SPACE and the ESPACE methods. An overview of the networks constructed using both methods is shown in Fig. 5. The SPACE method estimated 234 edges between 114 genes and the ESPACE method found 272 edges between 132 genes. Although the numbers of estimated edges from both the SPACE and ESPACE methods are quite similar, 16.7 and 28.3% of the estimated edges in networks by the SPACE and the ESPACE are different, respectively. We identified hub genes using the criterion mentioned at the beginning of this paper. The lists of hub genes identified in both networks are reported in Table 7. Interestingly, all hub genes identified by the SPACE method were also found using ESPACE. Note that this is not the usual case. For instance, if we define a hub as a node whose degree is greater than 5, the set of hub genes identified by the SPACE is not a subset of the hub genes identified by the ESPACE. To investigate the gains of the ESPACE method, therefore, we focused on the hub genes identified only by ESPACE (AURKA, APC, CDKN3), among which, AURKA and APC are among the 22 pre-specified hub genes while CDKN3 is not.

**Table 6** Summary of the estimated networks using the SPACE and ESPACE methods from the *E.coli* dataset. We denote a set of estimated edges and a set of the interactions from the RegulonDB by  $\hat{E}$  and  $T$ , respectively

Method	$\alpha^*$	$\lambda^*$	$ \hat{E} $	$ \hat{E} \cap T $	$ \hat{E} \cap T  /  \hat{E} $
SPACE	1	806.6	368	16	4.35%
ESPACE	0.85	835.2	349	45	12.89%

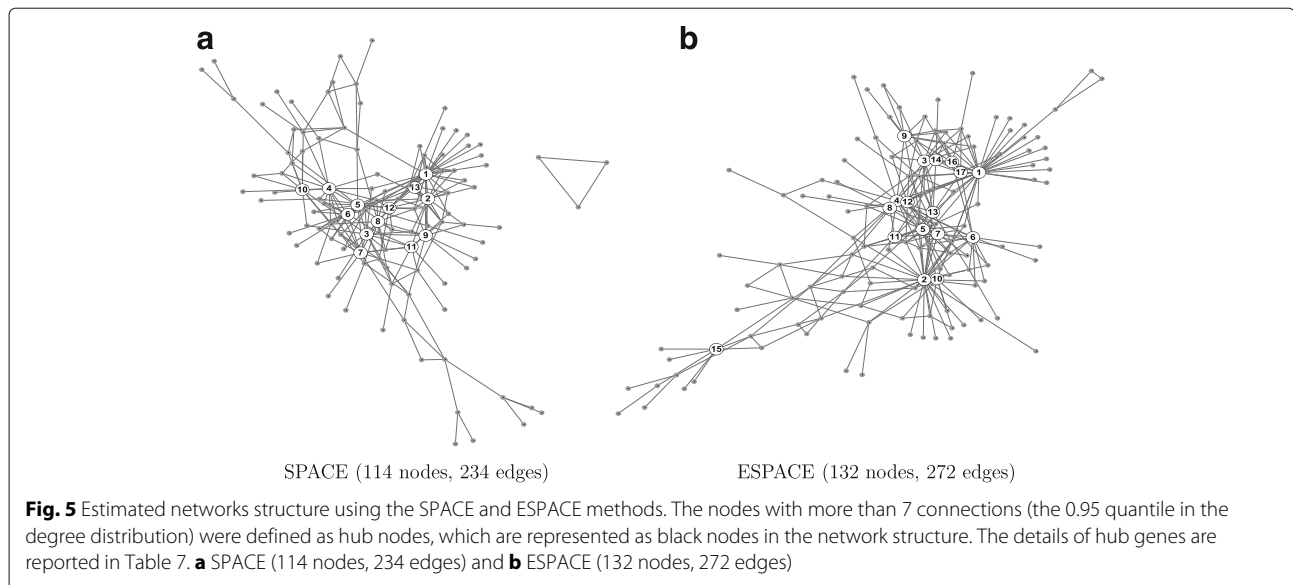


**Fig. 4** Plot of the number of the TPs vs. the number of the estimated edges for various  $\lambda$ s with  $\alpha^*$  in Table 6

The CDKN3 (Cyclin-Dependent Kinase Inhibitor 3) protein coded by the CDKN3 gene is a cyclin-dependent kinase inhibitor. Recent studies [54, 55] show that CDKN3 overexpression was associated with poorer survival outcomes in lung adenocarcinoma, but not in lung squamous cell carcinoma. We validated that CDKN3 is associated with the prognosis of lung adenocarcinoma patients in two independent datasets (see Fig. 6). The CDKN3 expression allowed us to separate the lung adenocarcinoma patients into high CDKN3 and low CDKN3 groups with significantly different survival outcomes: in the GSE13213 dataset [56] ( $n = 117$ ), hazard ratio = 2.02 (high CDKN3 vs. low CDKN3),  $p=0.0146$ ; in the GSE1037 dataset [57] ( $n = 61$ ), hazard ratio = 3.39 (high CDKN3 vs. low CDKN3),  $p=0.0126$ . Note that we divided patients into “high” and “low” groups by their gene expression levels with the K-means clustering method.

APC (Adenomatous Polyposis Coli) is a tumor suppressor gene, and is involved in the Wnt signaling pathway as a negative regulator. It has been identified as one of the key mutation genes in lung adenocarcinoma by a comprehensive study on the somatic mutations in lung adenocarcinoma [58]. AURKA (aurora kinase A) is a protein-coding gene found to be associated with many different types of cancer. Aurora kinase inhibitors have been studied as a potential cancer treatment [59]. Using the GSE42127 dataset [7] ( $n = 209$ ), we found that AURKA expression can predict lung cancer patients’ response to chemotherapy. The dataset contains expression profiles and treatment information for 209 lung cancer patients from MD Anderson Cancer Center, among whom 62 received adjuvant chemotherapy (ACT group) and the remaining 147 did not (no ACT group). The AURKA gene expression allowed us to separate the 209 patients into a low AURKA group ( $n = 104$ ) and high AURKA group





( $n = 105$ ) using the median AURKA expression as a cut-off. The patients in the low AURKA group (Fig. 7a) showed significant improvement in survival after ACT: hazard ratio = 0.289 (ACT vs. no ACT) and  $p$  value = 0.0312. The patients in the high AURKA group (Fig. 7b), on the other hand, showed no significant survival benefit after ACT: hazard ratio = 0.679 (ACT vs. no ACT) and  $p$

value = 0.241. These results indicate that AURKA expression could potentially be a predictive biomarker for lung cancer adjuvant chemotherapy, since only patients with low AURKA expression benefit from the treatment, while those with high AURKA expression are less likely to benefit. In addition, it is possible that Aurora kinase inhibitors, which suppress the expression of AURKA genes, may synergize the effect of adjuvant chemotherapy, i.e. improve the chance that a patient responds to adjuvant chemotherapy. In fact, a recent study has demonstrated that Aurora kinase inhibitors may synergize the effect of adjuvant chemotherapy in ovarian cancer, which is consistent with our results in lung cancer.

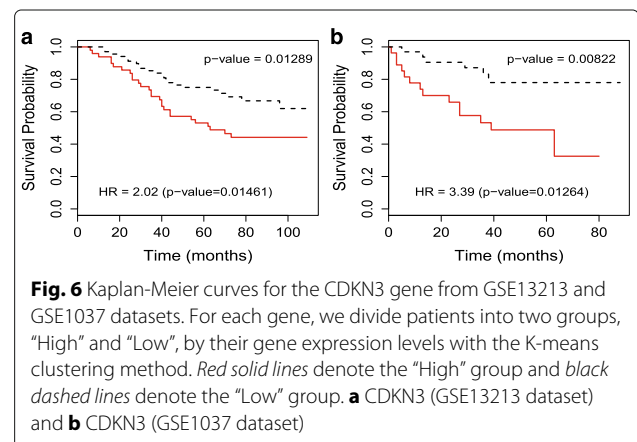
**Table 7** Hub genes from the estimated graphs using the SPACE and ESPACE methods

SPACE			ESPACE		
No.	Gene	Degree	No.	Gene	Degree
1	PRC1	26	1	<b>AURKA</b>	36
2	RRM2	17	2	NKX2-1	35
3	GPR116	16	3	RRM2	18
4	NKX2-1	16	4	CYP2B7P1	16
5	CYP2B7P1	15	5	GPR116	16
6	SFTPB	15	6	SFTPB	15
7	HOP	13	7	HOP	12
8	C1orf116	12	8	HSD17B6	11
9	HSD17B6	12	9	PRC1	11
10	TFF1	12	10	TFF1	11
11	CD302	10	11	C1orf116	10
12	FMO5	10	12	CD302	10
13	TPX2	8	13	FMO5	10
			14	UBE2C	10
			15	<b>APC</b>	8
			16	<b>CDKN3</b>	8
			17	TPX2	8

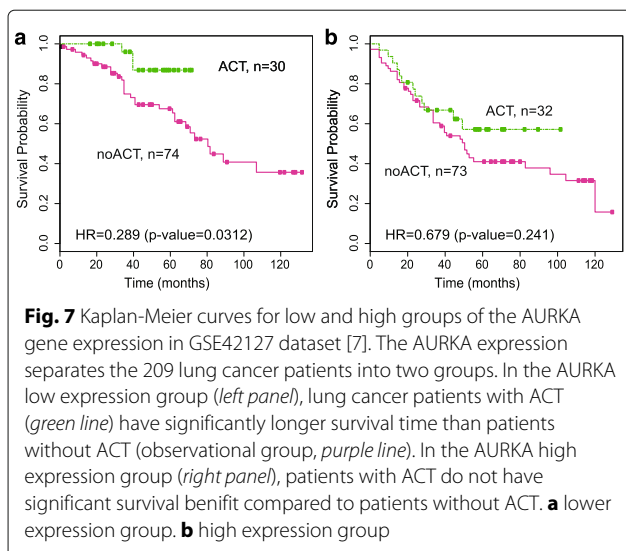
Bold font denotes the genes only identified by the modified method

### Conclusions

We have demonstrated incorporating hub gene information in estimating network structures by extending SPACE with an additional tuning parameter. Our simulation study



**Fig. 6** Kaplan-Meier curves for the CDKN3 gene from GSE13213 and GSE1037 datasets. For each gene, we divide patients into two groups, “High” and “Low”, by their gene expression levels with the K-means clustering method. Red solid lines denote the “High” group and black dashed lines denote the “Low” group. **a** CDKN3 (GSE13213 dataset) and **b** CDKN3 (GSE1037 dataset)



shows that the ESPACE method reduces errors in the construction of networks when the networks have previously-known hub nodes. Through two applications, we illustrate that the ESPACE method can improve the SPACE method by using the information about the potential hub genes. Although we adopted the GIC to select the optimal tuning parameters in this paper, the ESPACE method can directly be applied with other model selection criteria. The performance of the ESPACE method varies with the chosen criterion. However, the performance of the ESPACE method is at least comparable to the SPACE method since the ESPACE includes the SPACE as a reduced case.

#### Abbreviations

ACT: Adjuvant chemotherapy; APC: Adenomatous polyposis coli; AURKA: Aurora kinase A; CDKN3: Cyclin-dependent kinase inhibitor 3; CMI2NI: Conditional mutual inclusive information-based network inference; *E.coli*: *Escherichia coli*; EGLASSO: Extended GLASSO; ESPACE: Extended SPACE; FDR: False discovery rate; GIC: Generalized information criterion; GGM: Gaussian graphical model; GLASSO: Graphical lasso; GLASSO-SF: GLASSO with reweighted strategy for scale-free network; GRN: Gene regulatory network; M3D: Many Microbe Microarrays database; MCC: Matthews correlation coefficients; MI: Mutual information; MISR: Mis-specification rate; MM: Minorization-maximization; NS: Neighborhood selection; PCACMI: path consistency algorithm based on conditional mutual information; PPI: Protein-protein interaction; SEN: Sensitivity; SPACE: Sparse partial correlation estimation; SPE: Specificity; TF: Transcription factor

#### Acknowledgements

We gratefully thank Jessie Norris for language editing of the manuscript.

#### Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT and Future Planning (NRF-2015R1C1A1A02036312 and NRF-2011-0030810), and the National Institutes of Health grants (1R01CA17221, 1R01GM117597, R15GM113157).

#### Availability of data and materials

The *Escherichia coli* dataset analyzed during the current study is available in the Many Microbe Microarrays database (M3D) [48], <http://m3d.mssm.edu>. The RegulonDB [49] dataset used for validation of the result is available at <http://regulondb.ccg.unam.mx>. The Lung Cancer Consortium study dataset analyzed during this study is included in the published article [51]. The

information of cancer genes used during this study is available from the FoundationOne™, <http://www.foundationone.com>. The proposed method ESPACE is implemented the R package “*espace*”, which is available from <https://sites.google.com/site/dhyeonyu/software>.

#### Authors' contributions

DY, JL and GX drafted the manuscript. DY and JL formulate the proposed model and performed simulation studies. GX performed the interpretation of the results in the application to the lung cancer adenocarcinoma dataset. GX designed preprocessing procedure in the real-data applications. FL and XW helped in the verification of the proposed model and revised the manuscript. All authors read and approved the final manuscript.

#### Competing interests

We have no financial or personal relationships with other people and organizations that cause conflict of interests. The authors declare that they have no competing interests.

#### Consent for publication

Not applicable.

#### Ethics approval and consent to participate

Not applicable.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Department of Statistics, Inha University, Incheon, Korea. <sup>2</sup>Department of Statistics, Seoul National University, Seoul, Korea. <sup>3</sup>Department of Statistical Science, Southern Methodist University, 6425 Boaz Lane, Dallas, TX 75205, USA. <sup>4</sup>Department of Biostatistics, University of Florida, 2004 Mowry Road, Gainesville, FL 32611, USA. <sup>5</sup>Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA.

Received: 21 December 2016 Accepted: 3 March 2017

Published online: 23 March 2017

#### References

- Friedman N. Inferring cellular networks using probabilistic graphical models. *Science*. 2004;303(5659):799–805. doi:10.1126/science.1094068.
- Ihmels J, Friedlander G, Bergmann S, Sarg O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. *Nat Genet*. 2002;31(4):370–7. doi:10.1038/Ng941.
- Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*. 2003;34(2):166–76. doi:10.1038/ng1165 ng1165 [pii].
- Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. 2005;308(5721):523–9. doi:10.1126/science.1105809.
- Zhong R, Allen JD, Xiao G, Xie Y. Ensemble-based network aggregation improves the accuracy of gene network reconstruction. *PLoS ONE*. 2014;9(11):106319. doi:10.1371/journal.pone.0106319.
- Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. *Cell*. 2010;143(6):1005–17. doi:10.1016/j.cell.2010.11.013.
- Tang H, Xiao G, Behrens C, Schiller J, Allen J, Chow CW, Suraokar M, Corvalan A, Mao J, White MA, Wistuba I, Minna JD, Xie Y. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clin Cancer Res*. 2013;19(6):1577–86. doi:10.1158/1078-0432.CCR-12-2321.
- Cooper GF, Herskovits E. A bayesian method for the induction of probabilistic networks from data. *Mach Learn*. 1992;9(4):309–47. doi:10.1023/A:1022649401552.
- Ellis B, Wong WH. Learning causal bayesian network structures from experimental data. *J Am Stat Assoc*. 2008;103(482):778–89. doi:10.1198/016214508000000193.
- Liang FM, Zhang J. Learning bayesian networks for discrete data. *Comput Stat Data Anal*. 2009;53(4):865–76. doi:10.1016/j.csda.2008.10.007.

11. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. Inference in bayesian networks. *Nat Biotechnol.* 2006;24(1):51–3. doi:10.1038/nbt0106-51.
12. Sachs K, Gifford D, Jaakkola T, Sorger P, Lauffenburger DA. Bayesian network approach to cell signaling pathway modeling. *Sci STKE.* 2002;2002(148):38. doi:10.1126/stke.2002.148.pe38.
13. Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis. *Bmc Bioinforma.* 2008;9:1. doi:10.1186/1471-2105-9-559.
14. Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. *Biometrika.* 2007;94(1):19–35. doi:10.1093/biomet/asm018.
15. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics.* 2008;9(3):432–41. doi:10.1093/biostatistics/kxm045.
16. Witten DM, Friedman JH, Simon N. New insights and faster computations for the graphical lasso. *J Comput Graph Stat.* 2011;20(4):892–900. doi:10.1198/jcgs.2011.11051a.
17. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat.* 2006;34(3):1436–62. doi:10.1214/009053606000000281.
18. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *J Am Stat Assoc.* 2009;104(486):735–46. doi:10.1198/jasa.2009.0126.
19. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human b cells. *Nat Genet.* 2005;37(4):382–90. doi:10.1038/ng1532.
20. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla-Favera R, Califano A. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *Bmc Bioinforma.* 2006;7:1. doi:10.1186/1471-2105-7-51-57.
21. Zhang X, Zhao XM, He K, Lu L, Cao Y, Liu J, Hao JK, Liu ZP, Chen L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics.* 2012;28(1):98–104. doi:10.1093/bioinformatics/btr626.
22. Zhang X, Zhao J, Hao JK, Zhao XM, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Res.* 2015;43(5):31. doi:10.1093/nar/gku1315.
23. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol.* 2007;3:1. doi:10.1038/msb4100120.
24. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. *PLoS ONE.* 2012;7(1):29348. doi:10.1371/journal.pone.0029348.
25. Pan W. Network-based multiple locus linkage analysis of expression traits. *Bioinformatics.* 2009;25(11):1390–6. doi:10.1093/bioinformatics/btp177.
26. Pan W, Xie BH, Shen XT. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics.* 2010;66(2):474–84. doi:10.1111/j.1541-0420.2009.01296.x.
27. Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics.* 2008;24(3):404–11. doi:10.1093/bioinformatics/btm612.
28. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol.* 2004;14(3):283–91. doi:10.1016/j.sbi.2004.05.004.
29. Li JJ, Xie D. Rack1, a versatile hub in cancer. *Oncogene.* 2015;34(15):1890–8. doi:10.1038/onc.2014.127.
30. Selvanathan SP, Graham GT, Erkizan HV, Dirksen U, Natarajan TG, Dakic A, Yu S, Liu X, Paulsen MT, Ljungman ME, Wu CH, Lawlor ER, Uren A, Toretsky JA. Oncogenic fusion protein *ews-fl1* is a network hub that regulates alternative splicing. *Proc Natl Acad Sci USA.* 2015;112(11):1307–16. doi:10.1073/pnas.1500536112.
31. Liu Q, Ihler A. Learning scale free networks by reweighted L1 regularization. In: *AISTATS*; 2011. p. 40–48.
32. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD, Tyers M. Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.* 2006;4(10):317. doi:10.1371/journal.pbio.0040317.
33. Ekman D, Light S, Bjorklund AK, Elofsson A. What properties characterize the hub proteins of the protein-protein interaction network of *saccharomyces cerevisiae*? *Genome Biol.* 2006;7(6):45. doi:10.1186/gb-2006-7-6-r45.
34. Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol.* 2005;4:32. doi:10.2202/1544-6115.1175.
35. Efron B. Local false discovery rates. available at. 2005. <http://statweb.stanford.edu/~ckirby/brad/papers/2005LocalFDR.pdf>. Accessed 9 Mar.
36. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B-Methodological.* 1996;58(1):267–88.
37. Zhao P, Yu B. On model selection consistency of lasso. *J Mach Learn Res.* 2006;7:2541–63.
38. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Stat.* 2008;2(1):224–44. doi:10.1214/07-Aoas147.
39. Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J Mach Learn Res.* 2008;9:485–516.
40. Mazumder R, Hastie T. The graphical lasso: New insights and alternatives. *Electron J Stat.* 2012;6(0):2125–149. doi:10.1214/12-ejs740.
41. Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions. *J Comput Graph Stat.* 2000;9(1):1–20.
42. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of *escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 2007;5(1):8. doi:10.1371/journal.pbio.0050008.
43. Meyer PE, Lafitte F, Bontempi G. minet: A R/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinforma.* 2008;9:461.
44. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*, 2nd ed. Boston: The MIT Press; 2000.
45. Lauritzen SL. *Graphical Models*. New York: Oxford University Press Inc.; 1996. <http://books.google.com/books?id=mQWkx4guhAC>.
46. Yu D, Son W, Lim J, Xiao G. Statistical completion of a partially identified graph with applications for the estimation of gene regulatory networks. *Biostatistics.* 2015. doi:10.1093/biostatistics/kxv013.
47. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database 2009 update. *Nucleic Acids Res.* 2009;37(suppl 1):767–72. doi:10.1093/nar/gkn892.
48. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* 2008;36(Database issue):866–70. doi:10.1093/nar/gkm815.
49. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernandez S, Alquicira-Hernandez K, Lopez-Fuentes A, Porcion-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martinez YI, Pannier L, Olvera M, Labastida A, Jimenez-Jacinto V, Vega-Alvarado L, Del Moral-Chavez V, Hernandez-Alvarez A, Morett E, Collado-Vides J. Regulondb v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* 2013;41(Database issue):203–13. doi:10.1093/nar/gks1201.
50. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. *CA Cancer J Clin.* 2010;60(5):277–300. doi:10.3322/caac.20073.
51. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med.* 2008;14(8):822–7. doi:10.1038/nm.1790.
52. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics.* 2003;19(10):1236–42. doi:10.1093/bioinformatics/btg148.

53. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available at <https://www.R-project.org/>.
54. Zang X, Chen M, Zhou Y, Xiao G, Xie Y, Wang X. Identifying cdkn3 gene expression as a prognostic biomarker in lung adenocarcinoma via meta-analysis. *Cancer Inform*. 2015;14(Suppl 2):183–91. doi:10.4137/CIN.S17287.
55. Fan C, Chen L, Huang Q, Shen T, Welsh EA, Teer JK, Cai J, Cress WD, Wu J. Overexpression of major cdkn3 transcripts is associated with poor survival in lung adenocarcinoma. *Br J Cancer*. 2015;113(12):1735–43. doi:10.1038/bjc.2015.378.
56. Tomida S, Takeuchi T, Shimada Y, Arima C, Matsuo K, Mitsudomi T, Yatabe Y, Takahashi T. Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. *J Clin Oncol*. 2009;27(17):2793–9. doi:10.1200/JCO.2008.19.7053.
57. Jones MH, Virtanen C, Honjoh D, Miyoshi T, Satoh Y, Okumura S, Nakagawa K, Nomura H, Ishikawa Y. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *The Lancet*. 2004;363(9411):775–81. doi:10.1016/S0140-6736(04)15693-6.
58. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haippek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*. 2008;455(7216):1069–75.
59. Kollareddy M, Zheleva D, Dzubak P, Brahmikshatriya PS, Lepsik M, Hajduch M. Aurora kinase inhibitors: progress towards the clinic. *Invest New Drugs*. 2012;30(6):2411–32. doi:10.1007/s10637-012-9798-6.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

