Construction of protein interaction network involved in lung adenocarcinomas using a novel algorithm

JUAN CHEN[1,2], HAI-TAO YANG[2], ZHU LI[3], NING XU[4], BO YU[2], JUN-PING XU[2], PEI-GE ZHAO[2], YAN WANG[2], XIU-JUAN ZHANG[2] and DIAN-JIE LIN[1]

[1]Department of Respiratory Medicine, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, Shandong 250014; Departments of [2]Respiratory Medicine and [3]Hepatobiliary Surgery, People's Hospital of Liaocheng, Liaocheng, Shandong 252000; [4]Department of Respiratory Medicine, Weihai Municipal Hospital, Weihai, Shandong 264200, P.R. China

Received February 25, 2015; Accepted December 1, 2015

DOI: 10.3892/ol.2016.4822

Abstract. Studies that only assess differentially-expressed (DE) genes do not contain the information required to investigate the mechanisms of diseases. A complete knowledge of all the direct and indirect interactions between proteins may act as a significant benchmark in the process of forming a comprehensive description of cellular mechanisms and functions. The results of protein interaction network studies are often inconsistent and are based on various methods. In the present study, a combined network was constructed using selected gene pairs, following the conversion and combination of the scores of gene pairs that were obtained across multiple approaches by a novel algorithm. Samples from patients with and without lung adenocarcinoma were compared, and the RankProd package was used to identify DE genes. The empirical Bayesian (EB) meta-analysis approach, the search tool for the retrieval of interacting genes/proteins database (STRING), the weighted gene coexpression network analysis (WGCNA) package and the differentially-coexpressed genes and links package (DCGL) were used for network construction. A combined network was also constructed with a novel rank-based algorithm using a combined score. The topological features of the 5 networks were analyzed and compared. A total of 941 DE genes were screened. The topological analysis indicated that the gene interaction network constructed using the WGCNA method was more likely to produce a small-world property, which has a small average shortest path length and a large clustering coefficient, whereas the combined network was confirmed to be a scale-free network. Gene pairs that were identified using the novel combined method were mostly enriched in the cell cycle and p53 signaling pathway. The present study provided a novel perspective to the network-based analysis. Each method has advantages and disadvantages. Compared with single methods, the combined algorithm used in the present study may provide a novel method to analyze gene interactions, with increased credibility.

Correspondence to: Dr Dian-Jie Lin, Department of Respiratory Medicine, Shandong Provincial Hospital Affiliated to Shandong University, 9677 Olympic Sports Middle Road, Jinan, Shandong 250014, P.R. China
E-mail: dianjielinsd@yeah.net

Key words: protein interaction network, lung adenocarcinomas, topological analysis, empirical Bayesian, weighted gene coexpression network analysis

Introduction

Lung cancer is the main cause of cancer-associated mortality, and annually results in >1 million mortalities globally (1). Lung adenocarcinomas (ADCs) constitute a biologically heterogeneous group of lung tumors, and are, at present, the most common type of lung cancer (2). Previous studies have reported that gene expression profiling can be used to divide lung ADC into several subgroups and to distinguish primary cancers from metastases of extrapulmonary origin. Lung ADCs show striking variation in expression patterns compared with squamous cell lung carcinomas or small cell lung carcinomas (3).

A method that is often used to investigate the histopathology of a disease is the study of microarray data to identify genetic signatures. The identification of genes that are differentially expressed (DE) across two types of tissue samples or samples obtained under two experimental conditions is a typical task in the analysis of microarray data (4). RankProd is a method often used for detecting DE genes in replicated microarray experiments (5). RankProd is a non-parametric statistical method derived from biological reasoning that detects items that are consistently highly ranked in a number of lists (6). The method confers a number of advantages over linear modeling, including the biological intuition of fold-change (FC) criterion, fewer assumptions under the model, and increased performance with noisy data or low numbers of replicates (7). However, the method does not accommodate for other types of differential regulation, including differential coexpression (DC). Therefore, the empirical Bayesian (EB) approach was introduced. The EB method provides a false discovery rate (FDR) controlled list of significant pairs and pair-specific posterior probabilities that

may be used in the identification of particular DC types (8). EB may also be used for the model-based inference of cellular signaling networks (9).

A necessary requirement for any systems-level understanding of cellular functions is the correct identification and annotation of all functional interactions among cell proteins (10). Functional links between proteins may often be inferred from genomic associations between their encoding genes (11). The search tool for the retrieval of interacting genes/proteins (STRING) database is a precomputed global resource for the investigation and analysis of protein associations (12). The database provides uniquely comprehensive coverage and ease of access to experimental and predicted interaction information. Interactions in STRING are provided with a confidence score and accessory information, including protein domains and 3 dimensional structures, is made available within a stable and consistent identifier space (10). In addition, correlation networks are increasingly being used in bioinformatics applications. The weighted gene coexpression network analysis (WGCNA) package is a comprehensive collection of R functions designed to perform various aspects of weighted correlation network analysis. The package includes functions for network construction, module detection, gene selection, calculations of topological properties, data simulation, visualization and interfacing with external software (13). WGCNA has been used to identify the endometrial cancer prognosis markers (14). In addition, from the perspective of systems biology, gene coexpression analysis is useful for investigating gene interconnection at the expression level. The differentially-coexpressed genes and links (DCGL) R package may be used to identify DCGs and links from gene expression microarray data (15).

A comparison between cellular networks may provide insight into biological understanding and therapeutics. However, the comparison between large networks is infeasible; therefore, heuristic methods, including the degree distribution, clustering coefficient, diameter and relative graphlet frequency distribution, were used (16). The analysis of network topological features may elucidate the complex cellular mechanisms and processes and provide insight into the evolutionary aspects of the proteins involved in (17). Previously, the topological analysis on mass-balanced signaling networks has been performed and used as a framework to obtain network properties, including crosstalk (18). Similar to numerous other biological and real-world networks, protein interaction networks also exhibit the established small-world phenomenon (19) and scale-free property (20). The small-world network, which has a small average shortest path length and a large clustering coefficient, may enable a rapid integration of information (21). The scale-free network, of which the node degree distribution follows a power law, is characterized by a small number of highly connected nodes, whereas the majority of nodes interact with only a few neighbors. The network also demonstrates an increased robustness to endure random failure.

In the present study, samples from patients with and without lung ADC were compared in order to find novel molecular targets for lung ADC treatment. First, the RankProd package was used to identify DE genes. Next, the EB coexpression meta-analysis, STRING approach, WGCNA package and DCGL package were used for gene interaction network construction. Each method has various advantages and weaknesses. In order to take the non-uniform outcomes from various approaches into consideration, a novel algorithm was applied to combine 4 existing methods to identify gene pairs and networks in the present study. The topological features of the 5 networks, including clustering coefficient, average shortest path length and degree distribution, were compared and analyzed. The present study may increase the future understanding of gene interactions, increase the credibility of current methods and be important for the understanding of the molecular mechanisms of lung ADC.

Materials and methods

Data collection and preprocessing. The microarray expression profiles of patients with and without lung ADC were downloaded from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under the E-GEOD-10072 (22), E-GEOD-19188 (23), E-GEOD-31210 (24,25) and E-MEXP-231 (26) access numbers. In all datasets, only lung ADC and non-lung ADC control patient data were retained. The sample characteristics, platform and gene expression data were also extracted from each dataset and the associated study. The characteristics of the studies are shown in Table I.

Prior to analysis, the original expression data from all conditions were subjected to data preprocessing. The probe-level data in the CEL files were converted into expression measures. Gene probes from each dataset were acquired and read by the affy package (bioconductor.org/packages/affy). A background correction was performed using the robust multiarray average algorithm (27) to eliminate non-specific hybridization. Data normalization was conducted using quantiles (28). The modification of perfect match/mismatch values was performed using the Micro Array Suite 5.0 algorithm (29). The expression value was aggregated by the median polish summarization method (30). The featureFilter function in the GeneFilter package (bioconductor.org/packages/genefilter) was used to filter data and for probe annotation. The GetSYMBOL function of the annotate package (bioconductor.org/packages/annotate) was used to map the association between the probes and gene symbols (31). All preprocessing was performed using the espresso function of the Limma package (bioconductor.org/packages/limma) (32,33). The average values of the gene symbols with multiple probes were obtained.

Identification of DE genes. The RankProd approach (6,34) was employed to identify the DE genes associated with lung ADC. The software RankProd is implemented in the statistical programming language R as a package of the open-resource Bioconductor project (35). The microarray expression data were combined to detect DE genes using the RPadvance function in the RankProd package. The P-values for all genes were converted into the form $-\log_2$. Only the genes with a percentage of false-positives (PFP) value of ≤ 0.05 were considered to be DE between treatments and controls.

Construction of protein interaction networks for DE genes

Identification of DC using the EB approach. At present, numerous approaches have been used to identify DC gene pairs; however, the gene pairs are often prone to false identification under the conditions of large cardinality of the space

Table I. Characteristics of the individual studies included in the present study.

First author	Year	Access no.	Sample size, total (cases/controls)	Platform	Gene size, bases	Ref.
Shiraishi *et al*	2010	E-GEOD-10072	107 (58/49)	Affymetrix HG-U133A	12,493	(22)
Hou *et al*	2010	E-GEOD-19188	110 (45/65)	Affymetrix HG-U133Plus2	20,109	(23)
Okayama *et al* and Yamauchi *et al*	2012	E-GEOD-31210	246 (226/20)	Affymetrix HG-U133Plus2	20,109	(24) (25)
Yap *et al*	2005	E-MEXP-231	58 (49/9)	Affymetrix HG-U133A	12,493	(26)

to be interrogated (36). Therefore, the EB method was applied, which provided an FDR controlled list of significant pairs and pair-specific posterior probabilities (8). To achieve this, the EBcoexpress package in R was employed to conduct the differential co-expression analysis (37). The EB approach is applicable within a single study and across multiple studies. In the single study analysis, 3 inputs were required, including *X*, the array conditions and the pattern object (37). For *X*, an *m*-by-*n* matrix of expression values was used, where *m* is the number of genes or probes under consideration and *n* is the total number of microarrays over all conditions. The values were normalized using background normalization and median correction methods to give all the arrays equal median expression. Generally, gene expression levels are transformed on a $\log_2$ scale. For the array conditions, the members of an array with length *n* were provided values '1, ..., *K*', where *K* is the total number of conditions. All microarrays and assays were placed in the same order as the *n* columns of *X*. An object EBarrays Pattern was used to define the equivalent coexpression/DC classes. Next, the function makeMyD() of biweight midcorrelation was used to the calculate intra-group associations for all $p = m(m - 1) / 2$ gene pairs.

The initializeHP() function of the Mclust algorithm was used to identify the component normal mixture model that best fits the correlations of D. The Mclust algorithm may identify the normal mixture that best fits the empirical distribution of correlations, including component means, standard deviations and weights. These values played a role in initializing the expectation-maximization (EM) algorithm. In total, 3 functions accounted for the various versions of the modified EM approach, including the zero-step, one-step and full versions. The full version runs a complete two-cycle alternating expectation-conditional maximization. The zero-step version uses the initial estimates of the hyperparameters to generate posterior probabilities of DC. Subsequent to using the aforementioned algorithms, the priorDiagnostic() function was used to check the prior distribution selected by the EM. Finally, the crit.fun() function was used to provide a soft threshold and simulations to identify the DC gene pairs. DC genes were distinguished from gene pairs with invariant expression by controlling the posterior expected FDR at 0.05, and the coexpression network was constructed to account for the correlation between each pair of genes in the study. The curve was fit to the node degree distribution of the network.

Protein interactions obtained from STRING database. At present, protein or gene interactions and associations are annotated at various levels of detail that range between raw data repositories and highly formalized pathway databases in online resources. STRING aims to simplify access to information by providing a comprehensive, yet quality-controlled collection of protein-protein associations for a large number of organisms with a global perspective. The majority of the available information on protein or gene associations may be aggregated, scored and weighted with known and predicted interactions. Therefore, protein interactions across diverse experimental conditions may be measured and used as a predictor of functional associations in STRING, as in the present study. STRING employs 2 strategies to transfer known and predicted associations between organisms (11). Subsequent to the assignment of association scores and transfer between species, a combined score between any pair of proteins was computed, which increased confidence levels with an increased score compared with the individual sub-scores. The combined score accounted for the predicted and known scores obtained for each protein interaction from the STRING database, and was calculated according to the following formula:

$$S_{AB} = 1 - \prod_i (1 - S_i)$$

where S_{AB} is the score for the interaction between proteins *A* and *B*, and S_i is the score normalized by the biggest value calculated for the method *i*.

A graphical protein-protein interaction (PPI) network was then constructed and the topological features of the network were analyzed.

Identification of weighted correlation network. Correlation networks are increasingly being used in bioinformatics applications, and WGCNA has been used to describe the correlation patterns among genes across microarray samples (38). WGCNA may be used to identify clusters or modules of highly associated genes, to summarize clusters using the module eigengene or an intramodular hub gene, to associate modules with one other and with external sample traits using eigengene network methodology, and to calculate module membership measures (13). The WGCNA R package may be used to compute a gene selection score, termed 'p.weighted', based on the significance of the gene and module membership. The smaller the p.weighted value, the stronger the proof that the gene is a disease-associated hub gene. In the present study, the threshold value p.weighted score was set at ≤0.55. The weighted coexpression was determined by calculating a correlation matrix that contained all pairwise Pearson correlations between all probe sets spanning all subjects. The network nodes corresponded to gene expression

and the edges between genes were determined by the pairwise Pearson correlation between gene expression. Subsequent to raising the absolute value of the Pearson correlation to a power β≥1 (soft thresholding), the weighted gene coexpression network construction emphasized the stronger correlations. The adjacency of an unsigned weighted gene coexpression network was calculated by $a_{ij} = |\text{cor}(x_i,x_j)|\beta$. The soft threshold β=6 was chosen using the scale-free topology criterion (39). The positive and negative correlations of the network were treated equally and provided a value between 0-1. Following the selection of the weighted correlation networks, the topological features of the network were analyzed.

Identification of DC network. In order to identify DCGs and differentially-coexpressed links (DCLs) from gene expression microarray data (40), the DCGL 2.0 package in R program was introduced (15,41). In the process, the DCp and DCe functions were used to extract DCGs and DCLs. DCp and DCe are involved in the DC analysis module of the DCGL package (40). DCp plays a role in filtering sets of gene coexpression value pairs. Each pair is composed of 2 coexpression values that are calculated under 2 varying conditions, separately. The subset of the pairs was written as 2 vectors, *X* and *Y*, where *n* is coexpression neighbors for a gene.

$$X = (x_{i1}, x_{i2}, \ldots x_{in})$$

$$Y = (y_{i1}, y_{i2}, \ldots y_{in})$$

The DC of the gene was defined with the following equation:

$$DC_n(i) = \sqrt{\frac{(x_{i1} - y_{i1})^2 + (x_{i2} - y_{i2})^2 + \cdots + (x_{in} - y_{in})^2}{n}}$$

The novel Pearson correlation coefficient (PCC) was calculated and gene pairs were filtered based on the novel PCC with a q-value of 0.05.

The DCe function may also be used to identify DCGs and DCLs, which are based on the limit fold-change (LFC) model. The correlation pairs were divided into 3 parts, according to the pairing of signs of coexpression values and the multitude of coexpression values, as follows: Pairs with same signs (N_1); pairs with differing signs (N_2); and pairs with differently-signed high coexpression values (N_3). N_1 and N_2 were processed with the LFC model separately to produce 2 subsets of DCLs (K_1,K_2). N_3 was added to the set of DCLs directly. For a gene (g_i), the total number of links (n_i) and DCLs in particular (k_i) associated with it were counted. The DC of gene *i* measured using the DCe method was expressed by the following equation:

$$p(g_i) = \sum_{x=k_i}^{n_i} c_{n_i}^{x} \left(\frac{K}{N}\right)^{x} \left(1 - \frac{K}{N}\right)^{n_i - x}$$

In the process, gene pairs with a correlation value of ≥0.65 were considered to be significantly co-expressed (15,42). Finally, DCGs were mapped into Cytoscape software (www.cytoscape.org) for construction of the coexpression network, and topological features of the network were analyzed.

Conversion and combination of the gene association scores of the 4 methods. The score of each pair was obtained following the analysis of gene interactions using the aforementioned methods. Considering that variation in the results was obtained by the varying approaches, all the scores were analyzed in order to maintain a uniform standard. Therefore, a novel algorithm was applied to convert the scores of all the gene pairs in the present study. The conversion equation was as follows:

$$S_{com} = \frac{1}{n} \sum_{1<N<M} (-2 \log N/M)$$

where S_{com} was the combined score of each gene pair with integrated multiple results, *n* was the number of methods (n=4 in the present study), *M* was the number of gene pairs of the DE genes and *N* was the rank of a pair of genes.

A novel score of each gene pair was obtained by calculating the mean. The mean was obtained by dividing the combined score by the number of methods. Next, gene pairs were ranked based on the novel scores, and the pairs that satisfied the criteria $N/M \leq 10\%$ or $-2\log N/M \geq 6.643856$ were selected. The combined gene interaction network of the selected gene pairs was then constructed and the topological features of the network were analyzed.

Topological analysis. The clustering coefficient and short average path length of the aforementioned 5 networks were obtained and compared to investigate whether the networks constructed from the 5 methods exhibited the small-world network properties. In addition, the fit of the R^2 coefficient of the power-law $y = ax^b$ of the 5 networks was also compared, as PPI networks in general are modular and scale-free, which meant that the networks had power-law (or scale-free) degree distributions (28,43). Network Analyzer 2.7 plugin in Cytoscape 3.1.0 was used for the evaluation of topological parameters.

Functional enrichment analysis. Highly connected gene pairs generally participate in similar biological processes and pathways. In order to investigate the biological functional enrichment of the identified gene pairs, a pathway enrichment analysis was performed, based on the Kyoto encyclopedia of genes and genomes (KEGG; www.genome.jp/kegg/). The DE genes identified by RankProd were first imported to the online database for annotation, visualization and integrated discovery (http://david.abcc.ncifcrf.gov/tools.jsp), and all the pathways that the DE genes enriched were obtained. Next, with the DE genes in each pathway as a background, the number of enriched gene pairs identified by the 4 existing methods and the combined approach were calculated and compared. The terms with P<0.01 were considered to indicate a significant difference.

Results

Identification of DE genes. Following the normalization and preprocessing of the expression profile datasets, a total of 12,493 genes in E-GEOD-10072, 20,109 genes in E-GEOD-19188, 20,109 genes in E-GEOD-31210 and

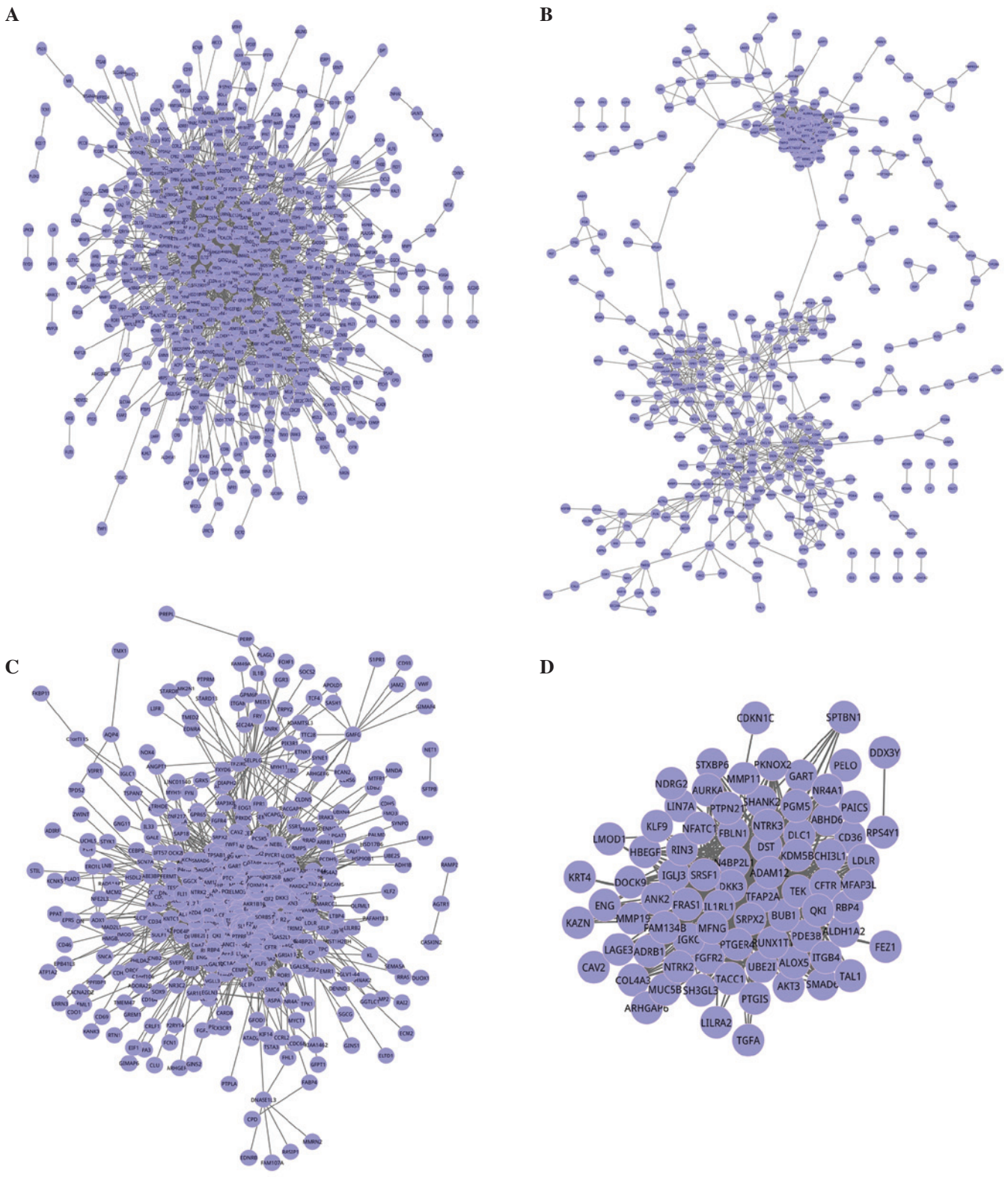

Figure 1. Graphical representation of the topological structures of the gene interaction networks constructed by 4 existing methods. Genes were denoted as nodes, and interactions between gene pairs were presented as edges (lines) in the images. (A) Network identified by empirical Bayesian method. (B) Network based on search tool for the retrieval of interacting genes/proteins database. (C) Coeexpression network constructed using the differentially-coexpressed genes and links approach. (D) Network based on weighted gene co-expression network analysis.

12,493 genes in E-MEXP-231 were obtained. Of those genes, 12,493 were common. By applying the RankProd package for meta-analysis, a total of 941 DE genes, 386 upregulated and 555 downregulated, were considered to be DE, with a PFP value of ≤0.05 and FC value of >2.

Topological analysis of 5 protein interaction networks. The protein interaction networks of DE genes were constructed using EB, STRING, DCGL and WGCNA (Fig. 1), and the association between gene pairs was determined. Subsequently, a novel algorithm was implemented to combine the score values

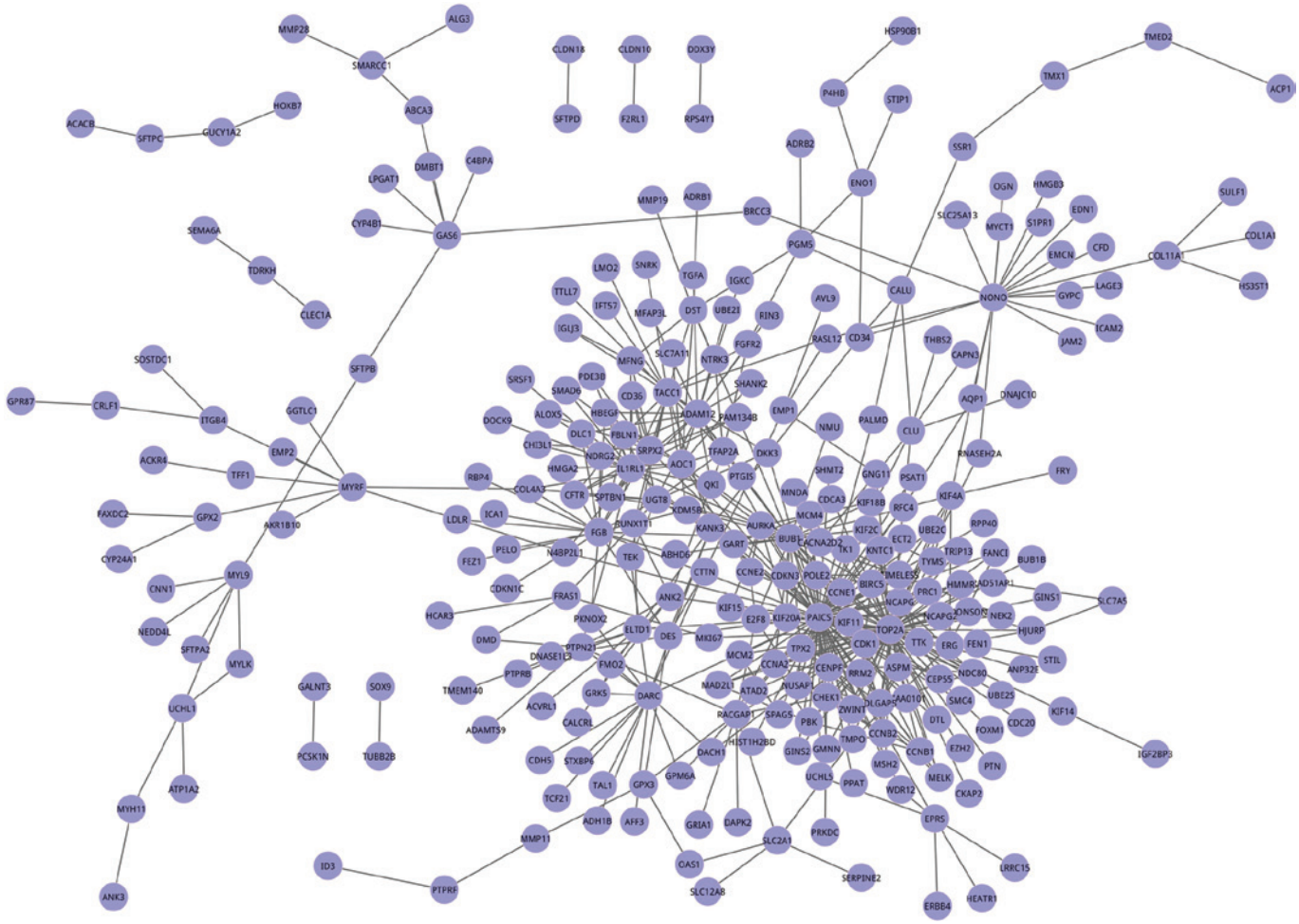

Figure 2. Combined gene interaction network based on the novel scores of each gene pairs across 4 methods. Genes were denoted as nodes and interactions between gene pairs were presented as edges (lines) in the image. A total of 280 nodes and 515 edges composed the combined network.

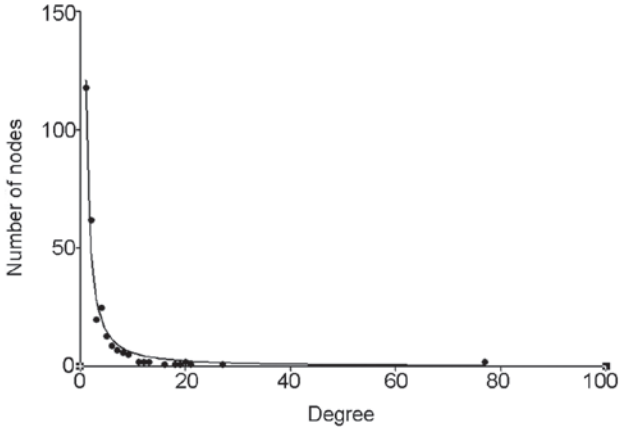

Figure 3. Scatter-gram of gene degree in the combined network. The combined network is a scale-free network of which the degree distribution followed a power law ($y = ax^b$, where a=121.0, b=-1.315) with the highest fitting coefficient (R^2=0.977).

of all gene pairs obtained from the 4 existing approaches. A novel matrix with a combined score of each gene pair was produced and a simple rank-based permutation procedure was used. Next, the combined gene interaction network was also constructed, consisting of 280 nodes and 515 edges (Fig. 2).

Network analysis showed that 4/5 networks exhibited the scale-free property, with a degree distribution that follows the power law with high fitting coefficients R^2, with the exception of the network constructed using the WGCNA method (R^2=0.264). The combined network showed the highest fitting coefficient (R^2=0.977) compared with the other 4 networks (Fig. 3), which indicates the evident scale-free property and increased robustness against the random failure of the network, compared with the other networks. However, the network constructed by the WGCNA method was more likely to be a small-world network, with the smallest mean shortest path length (1.783) and the largest clustering coefficient (0.813). The detailed parameters of the 5 networks are shown in Table II.

Functional enrichment analysis. All the KEGG pathways that the DE genes enriched were obtained as background, and 7 significant terms were identified, including extracellular matrix-receptor interaction (P=0.0000977), cell adhesion molecules (P=0.000991), p53 signaling pathway (P=0.00147), focal adhesion (P=0.00151), vascular smooth muscle contraction (P=0.00265), cell cycle (P=0.00335), and complement and coagulation cascades (P=0.00519). In order to investigate the enriched pathways of the gene pairs identified by various methods, the number of gene pairs enriched in each pathway was calculated and compared (Table III). Following the

Table II. Parameters of 5 networks constructed using 4 existing approaches and a novel algorithm.

Characteristic	EB	STRING	DCGL	WGCNA	Combination
Nodes	703.000	419.000	537.000	79.000	280.000
Edges	2,064.000	3,734.000	6379.000	649.000	515.000
R^2	0.963	0.931	0.938	0.264	0.977
Clustering coefficient	0.024	0.453	0.118	0.813	0.211
Mean shortest path length	3.673	5.337	2.715	1.783	4.195

EB, empirical Bayesian; STRING, search tool for the retrieval of interacting genes/proteins; DCGL, differentially-coexpressed genes and links; WGCNA, weighted gene coexpression network analysis.

Table III. Enriched Kyoto encyclopedia of genes and genomes pathways of gene pairs identified by 4 existing methods and a novel algorithm.

			Number of gene pairs				
Pathway	Category	P-value	EB	STRING	DCGL	WGCNA	Combination
ECM-receptor interaction	hsa04512	0.000098	0	36	3	0	1
Cell adhesion molecules	hsa04514	0.000991	1	5	1	0	0
p53 signaling pathway	hsa04115	0.001466	1	21	1	0	4
Focal adhesion	hsa04510	0.001510	1	38	3	0	2
Vascular smooth muscle contraction	hsa04270	0.002649	0	7	1	0	1
Cell cycle	hsa04110	0.003350	1	95	8	2	10
Complement and coagulation cascades	hsa04610	0.005190	0	3	0	0	0

ECM, extracellular matrix; EB, empirical Bayesian; STRING, search tool for the retrieval of interacting genes/proteins; DCGL, differentially-coexpressed genes and links; WGCNA, weighted gene coexpression network analysis.

combination of the 4 existing methods, the gene pairs mostly enriched the cell cycle and p53 signaling pathway. The common pathway that gene pairs enriched across the 5 methods was the cell cycle.

Discussion

In the present study, a novel algorithm that combined multiple existing approaches was applied in order to better understand the molecular mechanisms of lung ADC. First, samples from patients with and without lung ADC were compared. Next, the RankProd package was used to identify DE genes, and a total of 941 DE genes were screened across 4 datasets. Based on these DE genes, gene interaction networks were constructed, and the score value of each gene pair was obtained using the EB coexpression approach, STRING database, DCGL method and WGCNA package. A novel algorithm was applied to convert and combine the score values that were obtained from the aforementioned methods; a novel matrix with a combined score of each gene pair was then produced and sorted using a rank-based method. Finally, the combined gene interaction network was constructed via linking gene pairs.

A map of PPIs may provide useful revelations with regard to the cellular function and machinery of a proteome (44). A variety of methods have been proposed for the analysis of gene expression microarray data; however, few methods exist that use microarray data to quantify the interassociated behavior of genes within a gene interaction network (45). The incidence of cancer is considered to be closely associated with the abnormal expression of numerous genes; however, the previous methods used to study DE genes are inadequate, as there is a large difference between identifying DE genes and understanding the complex mechanisms of cancer. Therefore, the study of gene interactions is essential, as gene interactions are important for biological processes (46). Network-based approaches utilizing interaction information between gene pairs have emerged as powerful tools for the systematic understanding of the molecular mechanisms underlying biological processes, and a number of algorithms have been created to study these biological networks. Barter *et al* (47) performed a comparative analysis and indicated that the network-based method was more stable compared with single-gene and gene-set methods. Wu *et al* (48) also developed a network-based differential gene expression (nDGE) analysis, and demonstrated that nDGE outperformed existing methods for the prioritization of deregulated genes and the identification of deregulated gene modules using simulated data sets. Furthermore, a study conducted by Li *et al* (49) identified several key genes that were closely associated with survival in patients with lung ADC using a network-based approach.

The topological properties of gene interaction networks have been studied widely. Gene interaction networks have

been indicated to exhibit small-world and scale-free properties (50,51), which are typical of biological networks. Featherstone and Broadie (52) demonstrated that the scale-free property of the gene interaction network aided organisms by conferring the ability of resistance to the deleterious effects of mutation. Similar architecture was also indicated in the gene coexpression network of gastric cancer (53). The small-world property of biological networks was also confirmed in multiple data sources (43). In particular, Arita (54) indicated that the metabolic world of *Escherichia coli* was not a small biological network, but a network with a mean shortest path length that was much longer than previously hypothesized. In the present study, 5 gene interaction networks of lung ADC were constructed using 4 existing approaches and a novel combined algorithm. The network built using the WGCNA method was the most likely to be a small-world network, with the smallest mean shortest path length and the largest clustering coefficient. However, the combined network was revealed to be a scale-free network that possessed a node degree distribution that followed a power law with the highest fitting coefficient.

Generally, gene pairs that are connected closely participate in the same pathway. Li *et al* (49) suggested that alterations in cell cycle genes and pathways were associated with tumor grade and contributed to the survival of lung ADC patients, regardless of smoking status, using a systems biology-based network approach. The study conducted by Wu *et al* (48) also identified that cell cycle-associated genes played a role in the molecular variations between smoker and non-smoker lung ADC. A study of cisplatin in lung ADC demonstrated that cisplatin exerted a cytotoxic effect through the blockage of the cell cycle pathway, and may be partly regulated by the p53 signaling pathway. Consistent with previous studies, the findings in the present study suggested that the gene pairs mainly enriched the cell cycle and p53 signaling pathway subsequent to combination, and that the cell cycle pathway was the common pathway that gene pairs enriched across 5 methods.

In the present study, 4 existing network-based approaches were presented. Evidently, varying methods often possess varying abilities. Therefore, a novel merged approach was created to enhance stability and reliability. The combined gene interaction network was constructed by reassembling the scores of gene pairs from 4 existing methods. Network analysis showed that the network constructed by the WGCNA method was more inclined to be a small-world property and that the combined network was revealed to demonstrate scale-free network features. In addition, pathway analysis demonstrated that the cell cycle pathway was involved in the pathogenesis of lung ADC. When considering the applications and limitations of each of the methods, the novel merged algorithm outlined in the present study may provide a more credible and robust outcome for genetic network analyses, and is recommended for future application.

References

1. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, *et al*: Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455: 1069-1075, 2008.
2. Chen G, Gharib TG, Huang C-C, Thomas DG, Shedden KA, Taylor JM, Kardia SL, Misek DE, Giordano TJ, Iannettoni MD, *et al*: Proteomic analysis of lung adenocarcinoma: Identification of a highly expressed set of proteins in tumors. Clin Cancer Res 8: 2298-2305, 2002.
3. Miura K, Bowman ED, Simon R, Peng AC, Robles AI, Jones RT, Katagiri T, He P, Mizukami H, Charboneau L, *et al*: Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. Cancer Res 62: 3244-3250, 2002.
4. Pan W: A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18: 546-554, 2002.
5. Koziol JA: Comments on the rank product method for analyzing replicated experiments. FEBS Lett 584: 941-944, 2010.
6. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL and Chory J: RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22: 2825-2827, 2006.
7. Breitling R and Herzyk P: Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J Bioinform Comput Biol 3: 1171-1189, 2005.
8. Dawson JA and Kendziorski C: An empirical Bayesian approach for identifying differential coexpression in high-throughput experiments. Biometrics 68: 455-465, 2012.
9. Klinke DJ II: An empirical Bayesian approach for model-based inference of cellular signaling networks. BMC Bioinformatics 10: 371, 2009.
10. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, *et al*: The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561-D568, 2011.
11. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA and Bork P: STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433-D437, 2005.
12. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B: STRING: A database of predicted functional associations between proteins. Nucleic Acids Res 31: 258-261, 2003.
13. Langfelder P and Horvath S: WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9: 559, 2008.
14. Zhu X-L, Ai Z-H, Wang J, Xu Y-L and Teng Y-C: Weighted gene coexpression network analysis in identification of endometrial cancer prognosis markers. Asian Pac J Cancer Prev 13: 4607-4611, 2012.
15. Liu B-H, Yu H, Tu K, Li C, Li Y-X and Li Y-Y: DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics 26: 2637-2638, 2010.
16. Pržulj N: Biological network comparison using graphlet degree distribution. Bioinformatics 23: e177-e183, 2007.
17. Jonsson PF and Bates PA: Global topological features of cancer proteins in the human interactome. Bioinformatics 22: 2291-2297, 2006.
18. Papin JA and Palsson BO: Topological analysis of mass-balanced signaling networks: A framework to obtain network properties including crosstalk. J Theor Biol 227: 283-297, 2004.
19. Watts DJ and Strogatz SH: Collective dynamics of 'small-world' networks. Nature 393: 440-442, 1998.
20. Albert R: Scale-free networks in cell biology. J Cell Sci 118: 4947-4957, 2005.
21. Sporns O and Zwi JD: The small-world of the cerebral cortex. Neuroinformatics 2: 145-162, 2004.
22. Shiraishi T, Matsuyama S and Kitano H: Large-scale analysis of network bistability for human cancers. PLoS Comput Biol 6: e1000851, 2010.
23. Hou J, Aerts J, den Hamer B, van Ijcken W, den Bakker M, Riegman P, van der Leest C, van der Spek P, Foekens JA, Hoogsteden HC, *et al*: Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 5: e10312, 2010.
24. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, Furuta K, Tsuta K, Shibata T, Yamamoto S, *et al*: Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 72: 100-111, 2012.

25. Yamauchi M, Yamaguchi R, Nakata A, Kohno T, Nagasaki M, Shimamura T, Imoto S, Saito A, Ueno K, Hatanaka Y, *et al*: Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. PLoS One 7: e43923, 2012.
26. Yap YL, Lam DC, Luc G, Zhang XW, Hernandez D, Gras R, Wang E, Chiu SW, Chung LP, Lam WK, *et al*: Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. Nucleic Acids Res 33: 409-421, 2005.
27. Kim Y, Doan BQ, Duggal P and Bailey-Wilson JE: Normalization of microarray expression data using within-pedigree pool and its effect on linkage analysis. BMC Proc 1 (Suppl 1): S152, 2007.
28. Rifai N and Ridker PM: Proposed cardiovascular risk assessment algorithm using high-sensitivity C-reactive protein and lipid screening. Clin Chem 47: 28-30, 2001.
29. Zhang L, Miles MF and Aldape KD: A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol 21: 818-821, 2003.
30. Durinck S: Pre-processing of microarray data and analysis of differential expression. Methods Mol Biol 452: 89-110, 2008.
31. Gentleman R, Maintainer MBP and Biobase I: AnnotationDbi D, Biobase S and Amat CARAR: Package 'annotate'. 2013.
32. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V and Notredame C: Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34: W604-W608, 2006.
33. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: e3, 2004.
34. Breitling R, Armengaud P, Amtmann A and Herzyk P: Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573: 83-92, 2004.
35. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5: R80, 2004.
36. Cho SB, Kim J and Kim JH: Identifying set-wise differential coexpression in gene expression microarray data. BMC Bioinformatics 10: 109, 2009.
37. Dawson JA, Ye S and Kendziorski C: R/EBcoexpress: An empirical Bayesian framework for discovering differential coexpression. Bioinformatics 28: 1939-1940, 2012.
38. Presson AP, Sobel EM, Papp JC, Suarez CJ, Whistler T, Rajeevan MS, Vernon SD and Horvath S: Integrated weighted gene coexpression network analysis with an application to chronic fatigue syndrome. BMC Syst Biol 2: 95, 2008.
39. Saris CG, Horvath S, van Vught PW, van Es MA, Blauw HM, Fuller TF, Langfelder P, DeYoung J, Wokke JH, Veldink JH, *et al*: Weighted gene coexpression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. BMC Genomics 10: 405, 2009.
40. Yang J, Yu H and Liu B-H: Using the DCGL 2.0 Package. 2013. http://lifecenter.sgst.cn/main/en/dcgl/DCGL.pdf. Accessed November 23, 2014.
41. Yu H, Liu B-H, Ye Z-Q, Li C, Li Y-X and Li Y-Y: Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinformatics 12: 315, 2011.
42. Li C, Shen W, Shen S and Ai Z: Gene expression patterns combined with bioinformatics analysis identify genes associated with cholangiocarcinoma. Comput Biol Chem 47: 192-197, 2013.
43. Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabási A-L: Hierarchical organization of modularity in metabolic networks. Science 297: 1551-1555, 2002.
44. Wu X, Zhu L, Guo J, Zhang D-Y and Lin K: Prediction of yeast protein-protein interaction network: Insights from the Gene Ontology and annotations. Nucleic Acids Res 34: 2137-2150, 2006.
45. Toyoshiba H, Yamanaka T, Sone H, Parham FM, Walker NJ, Martinez J and Portier CJ: Gene interaction network suggests dioxin induces a significant linkage between aryl hydrocarbon receptor and retinoic acid receptor beta. Environ Health Perspect 112: 1217-1224, 2004.
46. Toyoshiba H, Sone H, Yamanaka T, Parham FM, Irwin RD, Boorman GA and Portier CJ: Gene interaction network analysis suggests differences between high and low doses of acetaminophen. Toxicol Appl Pharmacol 215: 306-316, 2006.
47. Barter RL, Schramm SJ, Mann GJ and Yang YH: Network-based biomarkers enhance classical approaches to prognostic gene expression signatures. BMC Syst Biol 8 (Suppl 4): S5, 2014.
48. Wu C, Zhu J and Zhang X: Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma. BMC Bioinformatics 14: 365, 2013.
49. Li Y, Tang H, Sun Z, Bungum AO, Edell ES, Lingle WL, Stoddard SM, Zhang M, Jen J, Yang P, *et al*: Network-based approach identified cell cycle genes as predictor of overall survival in lung adenocarcinoma patients. Lung Cancer 80: 91-98, 2013.
50. Jordan IK, Mariño-Ramírez L, Wolf YI and Koonin EV: Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol 21: 2058-2070, 2004.
51. van Noort V, Snel B and Huynen MA: The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5: 280-284, 2004.
52. Featherstone DE and Broadie K: Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. BioEssays 24: 267-274, 2002.
53. Aggarwal A, Guo DL, Hoshida Y, Yuen ST, Chu KM, So S, Boussioutas A, Chen X, Bowtell D, Aburatani H, *et al*: Topological and functional discovery in a gene coexpression meta-network of gastric cancer. Cancer Res 66: 232-241, 2006.
54. Arita M: The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101: 1543-1547, 2004.