SPANDIDOS PUBLICATIONS EXPERIMENTAL AND THERAPEUTIC MEDICINE 11: 1707-1715, 2016

Co-expression network-based analysis of hippocampal expression data associated with Alzheimer's disease using a novel algorithm

HONG YUE, BO YANG, FANG YANG, XIAO-LI HU and FAN-BIN KONG

Department of Neurology (No. 2), Rizhao People's Hospital, Rizhao, Shandong 276826, P.R. China

Received December 29, 2014; Accepted January 7, 2016

DOI: 10.3892/etm.2016.3131

Abstract. Recent progress in bioinformatics has facilitated the clarification of biological processes associated with complex diseases. Numerous methods of co-expression analysis have been proposed for use in the study of pairwise relationships among genes. In the present study, a combined network based on gene pairs was constructed following the conversion and combination of gene pair score values using a novel algorithm across multiple approaches. Three hippocampal expression profiles of patients with Alzheimer's disease (AD) and normal controls were extracted from the ArrayExpress database, and a total of 144 differentially expressed (DE) genes across multiple studies were identified by a rank product (RP) method. Five groups of co-expression gene pairs and five networks were identified and constructed using four existing methods [weighted gene co-expression network analysis (WGCNA), empirical Bayesian (EB), differentially co-expressed genes and links (DCGL), search tool for the retrieval of interacting genes/proteins database (STRING)] and a novel rank-based algorithm with combined score, respectively. Topological analysis indicated that the co-expression network constructed by the WGCNA method had the tendency to exhibit small-world characteristics, and the combined co-expression network was confirmed to be a scale-free network. Functional analysis of the co-expression gene pairs was conducted by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. The co-expression gene pairs were mostly enriched in five pathways, namely proteasome, oxidative phosphorylation, Parkinson's disease, Huntington's disease and AD. This study provides a new perspective to co-expression analysis. Since different methods of analysis often present varying abilities, the novel combination algorithm may provide a more credible and robust outcome, and could be used to complement to traditional co-expression analysis.

Correspondence to: Dr Hong Yue, Department of Neurology (No. 2), Rizhao People's Hospital, 126 Taian Road, Rizhao, Shandong 276826, P.R. China
E-mail: yuehongrz@yeah.net

Key words: Alzheimer's disease, gene co-expression analysis, weighted gene co-expression network analysis, empirical Bayesian, differentially co-expressed genes and links, search tool for the retrieval of interacting genes/proteins database

Introduction

Generally, complex diseases result from a combination of genetic perturbations and their interactions (1). During the past few decades, a considerable number of gene biomarkers have been successfully identified to be associated with complex diseases through genome-wide analysis of gene expression profiles (2,3). However, biomolecules in living organisms rarely act individually but interact to achieve biological functions (4). Network-based approaches have been developed as powerful and informative tools to identify candidate biomarkers or therapeutic targets based on transcript data (5-7). These methods generally utilize the knowledge of physical or functional interactions between molecules, and have been successfully applied in various diseases, such as cancer.

Various types of intermolecular interactions have been disclosed, including protein-protein interactions, protein phosphorylation networks, DNA methylation networks and gene co-expression. These interactions can be represented as networks with nodes that denote molecules, and edges that denote interactions between them. Genes in the same pathways or functional complex often exhibit similar expression patterns across multiple experiments and various organisms (8). Thus, the creation of a co-expression network from high-throughput data has become a popular alternative to the conventional methods of analysis, as it allows researchers to study the whole spectrum of pairwise associations of genes (9). By constructing a co-expression network, the regulatory relationships underlying different conditions can be estimated (10).

Co-expression networks can have small-world (11) and scale-free properties (12). A scale-free network is a network in which the node degree distribution follows a power law, and is characterized by a small number of highly connected nodes, the majority of which interact with only a few neighbors, and a high robustness to withstand random failure. A small-world network is considered to be efficient, in that it enables the rapid integration of information (13). It has two independent structural features, comprising a low average shortest path length and a high clustering coefficient.

With the development of bioinformatics analysis, a variety of algorithms have been developed to evaluate these biological networks (14,15), both in terms of experimental measurements

and computational prediction techniques. Correlation-based methods are perceived as being the most straightforward for exploration of gene co-expression networks (15). Weighted gene co-expression network analysis (WGCNA), as a statistical approach based on correlations, has been widely used to analyze transcriptional profiles, and has been demonstrated to be an informative approach for the functional annotation of uncharacterized genes (16). In a recent study conducted by Allen *et al* (15), WGCNA was one of the best-performing methods for the construction of global co-expression networks. Moreover, an empirical Bayesian (EB) approach aims to identify differential co-expression by examining correlations among gene pairs (17). It effectively avoids the problem of inconsistent co-expression between different studies by producing a false discovery rate (FDR)-controlled list of differential co-expression pairs without sacrificing power. This approach is applicable within a single study and across multiple studies. Differentially Co-expressed Genes and Links (DCGL) is an R-package for the identification of differentially co-expressed genes and links from gene expression microarray data (18). It examines gene expression correlation using exact co-expression changes of gene pairs between two conditions, and thus can differentiate significant co-expression changes from relatively trivial ones (19). In addition, the database Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) provides a comprehensive, quality-controlled collection of protein-protein associations for a large number of organisms (20). It integrates and ranks associations derived from high-throughput experimental data, database and literature mining, and predictions based on genomic context analysis, respectively. STRING has an integrated scoring scheme for the interactions, and provides a high level of confidence.

The aforementioned co-expression-based methods have been used in a number of studies and have shown their usefulness in the interpretation of biological results and identification of important gene modules (17,21,22). Each method has certain advantages. However, different approaches often produce different co-expression data for the same experiment (15). Thus, in the present study, a novel algorithm was applied to combine four existing methods to identify co-expression gene pairs and networks. Topological features, including clustering coefficient, average shortest path length and degree distribution were investigated and compared to evaluate whether each network tended to be a scale-free or small-world network. The study initially focused on identifying differentially expressed (DE) genes between Alzheimer's disease (AD) patients and normal controls on the basis of hippocampal transcript profiles. To compare the approaches, the related scores of gene pairs were obtained using the STRING database, DCGL package, EB analysis and WGCNA algorithm, respectively. Considering the non-uniform outcomes from different approaches, all scores from the four methods were converted and united using a rank-based model and a combined score of each gene pairs was obtained. Then, gene co-expression networks obtained from the four approaches respectively and a combined network were constructed, and topological properties were further analyzed. The aim was to provide a novel tool for the analysis of gene interactions with a higher credibility and rapid transmission of information, concentrating on the scores of each gene pair across multiple approaches.

Materials and methods

Data recruitment and preprocessing. In the present study, three hippocampal transcript profiles of AD patients and normal controls deposited in ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) were examined: E-GEOD-1297 (23), E-GEOD-28146 (24), and E-GEOD-5281 (3,25). These datasets contained data for 54 patients with AD and 30 normal controls. The characteristics of the studies are shown in Table I.

Prior to analysis, the original expression information from all conditions was subjected to data preprocessing. For each dataset, in order to eliminate the influence of nonspecific hybridization, background correction and normalization were carried out by the robust multichip average (RMA) method (26) and quantile-based algorithm (27), respectively. Perfect match and mismatch values were revised using the Micro Array Suite 5.0 algorithm (28), the value of which was selected via the median method. The gene expression values of all data were transformed to a comparable level. The data were then screened using the feature filter method of the genefilter package (version 1.52.0; bioconductor.org/packages/genefilter). Each probe is mapped to one gene, where the probe is discarded if it does not match any genes.

Detection of DE genes. Since the three sets of AD data had different origins, a rank product (RP) algorithm was implemented to integrate the array datasets (RankProd; Version 2.42.0; bioconductor.org/packages/RankProd/). This method can determine how significant changes are and how many of the selected genes are likely to be truly differentially expressed. It also allows for the flexible control of the FDR and family-wise error rate in the multiple testing situation of a microarray experiment (29). Considering a situation of the microarray experiment with two replicates (A and B), RP for a certain gene g will be as follows:

$$RP_g = \left(rank_g^{replicateA} / n \right) \times \left(rank_g^{replicateB} / n \right)$$

where rank is the position of gene g in the list of genes in the replicate A. RPg can be taken as a P-value when all ranks are equally likely, but cannot be used directly to assess the significance of an observed change in. Therefore, a simple permutation-based estimation procedure is used to determine how likely it is to observe a given RP value or better in a random experiment, thus converting the RP value to an E value (30). Subsequently, for each gene g, a conservative estimate of the percentage of false-positive (pfp) is calculated if this gene is considered as significantly differentially expressed:

$$q_g = E(RPg) / rank(g)$$

Rank (g) denotes the position of gene g in a list of all genes sorted by increasing RP value. This method can decide how large a pfp will be accepted and extend the list of accepted genes up to the gene with this q_g value. In the present study, a pfp cut-off value of <0.01 was used.

Construction of gene co-expression network for DE genes

Scoring of gene co-expression using STRING database. Gene and protein interactions have been annotated at various levels of detail ranging from raw data repositories to highly formalized

Table I. Characteristics of the individual datasets included in this study.

Accession number	Year	Sample size (cases/controls)	Platform
E-GEOD-1297	2004	31 (22/9)	Affymetrix HG-U133A
E-GEOD-28146	2011	30 (22/8)	Affymetrix HG-U133Plus2
E-GEOD-5281	2007	23 (10/13)	Affymetrix HG-U133Plus2

pathway databases in online resources. In the present study, the possible functional associations of DE genes were investigated using STRING (http://string-db.org/), which provides a comprehensive, quality-controlled collection of protein-protein associations for a large number of organisms with a global perspective (31). In the STRING database, most of the available information on genes (proteins) can be aggregated, scored and weighted with known and predicted associations. A scored association between two proteins could be transferred between organisms. Following assignment of association scores and transfer between species, a final combined score between any pair of proteins was computed, which increased confidence with a higher score than the individual sub-scores. The combined score took into account the prediction and known scores obtained from each protein interaction. Subsequently, a graphical protein-protein network was constructed and the topological features of the network were further analyzed.

Identifying differential co-expression by DCGL. Biological functions result from numerous gene products acting together, and highly co-expressed genes take part in similar biological processes and pathways. The DCGL 2.0 package was applied to identify differentially co-expressed (DC) genes and links. DCGL (version 2.0; lifecenter.sgst.cn/main/en/dcgl.jsp) is a R Package for revealing differential regulation from differential co-expression. It contains four modules: Gene filtration, link filtration, differential co-expression analysis (DCEA) and differential regulation analysis (DRA) modules. Differential co-expression profile (DCp) and differential co-expression enrichment (DCe) are involved in the DCEA module for extracting DC genes and DC links. DCp operated on the filtered set of gene co-expression value pairs, where each pair comprised two co-expression values determined under two different conditions separately. The subset of co-expression value pairs associated with a particular gene, in two groups for the two conditions separately, was written as vectors X and Y (n is number of co-expression neighbors).

$$X=\left(x_{i1},x_{i2},\cdots,x_{in}\right)$$

$$Y=\left(y_{i1},y_{i2},\cdots,y_{in}\right)$$

A length-normalized Euclidean distance was used to measure the differential co-expression (dC) of this gene.

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

A permutation test was performed to assess the significance of dC. In this test, the disease samples and normal controls were randomly permuted, new Pearson correlation co-efficient (PCC) was calculated, gene pairs were filtered based on the new PCC, and new dC statistics were calculated. The sample permutation was repeated N times, and a large number of permutation dC statistics formed an empirical null distribution. The P-value for each gene could then be estimated.

DCe was also used to identify DC genes and DC links, which based on the 'Limit Fold Change' (LFC) model. First, correlation pairs were divided into three sets according to the pairing of signs of co-expression values and the multitude of co-expression values: Pairs with same signs (N1), pairs with different signs (N2) and pairs with differently-signed high co-expression values (N3). The first two sets were processed with the 'LFC' model separately to produce two subsets of DC links (K1, K2), while the third set (N3) was added to the set of DC links directly. Therefore, K = N3 + K1 + K2 DC links were determined from N gene links. For a gene (g_i), the total number of links (n_i) and DC links in particular (k_i) associated with it were counted. Binomial probability model was used to estimate the significance of the gene being a DC gene.

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

Differentially co-expression summarization (DCsum) was implemented to combine the results from the DCp and DCe methods.

Identification of differential co-expression by EB. Several approaches have been developed for differential regulation analysis by the identification of DC gene pairs. However, these methods are frequently underpowered, prone to false discoveries or computationally intractable under the conditions of large cardinality of the space to be interrogated and influential outliers (32). To address this limitation, Dawson and Kendziorski (17) presented an effective EB approach that provided a FDR controlled list of notable pairs along with pair-specific posterior probabilities to identify DC gene pairs without sacrificing power; the EB approach is suitable for use within and across experiments, has exhibited improved runtimes and may be a useful complement to existing DE methods by simulations and case studies respectively. In the present study, the identification of DC gene pairs was conducted using the following steps: Three inputs of matrix X, the conditions array and the pattern object were required. The expression values in an m-by-n matrix of X (where m is the number of genes/probes under consideration, and n is the total number of microarrays over all conditions) were normalized with background normalization and median correction and were represented on the log2 scale. The members of the

conditions array with length n took values 1-K (where K indicated the total number of conditions). It was used to define the equal co-expression/differential co-expression classes with an ebarraysPatterns object based on the unique values in the conditions array. Intra-group correlations for all $p=m^*(m-1)/2$ gene pairs from X and the conditions array were calculated using bi-weight mid-correlation. A p-by-K of D matrix with correlations was obtained. The mclust algorithm (33) was used to initialize the hyper-parameters to find the component normal mixture model that could best fit the empirical distribution of correlations. The values of the component in the normal mixture model with component means, standard deviations and weights would be used to initialize the Expectation-maximization (EM) algorithm. In this step, the initial estimates of the hyper-parameters were used to generate posterior probabilities of differential co-expression. Finally, a soft threshold was provided by controlling the posterior probabilities of differential co-expression to identify particular types of DC gene pairs. DC genes were distinguished from gene pairs having invariant expression by controlling the posterior expected FDR at 0.05 and a co-expression network was constructed to represent the correlation between each pair of genes.

Identifying differential co-expression by WGCNA. Gene co-expression networks, which represent a major application of correlation network methodology, are instrumental for describing the pairwise relationships among gene transcripts (34,35) and facilitate the understanding of their function and identification of their key players. In the present study, WGCNA (36), a systems biology method for performing a correlation network analysis of large and high-dimensional data sets, was used to describe correlation patterns among gene expression profiles. Also, co-expression network construction as a function in the WGCNA package was demonstrated. Genes were denoted as nodes of a gene co-expression network which were labeled by indices i, j=1,2,……n, and correlations between gene pairs were presented as edges. The network can be illustrated with its adjacency matrix A, a symmetric n x n matrix with entries a_{ij} in (0,1) which encodes the strength of the network link between genes i and j. An intermediate quantity of co-expression similarity is first defined to calculate the adjacency A of an unsigned network (value between 0 and 1), in which positive and negative correlations are treated equally. However, the use of an absolute value for the correlation may obscure biologically relevant information of the distinction between gene activation and repression. A signed co-expression measure between x_i and x_j is used to preserve the sign of the correlation, which is defined with a simple transformation of the correlation:

$$S_{ij}=\frac{1+cor(x_i,x_j)}{2}$$

The difference between signed and unsigned similarities lies in how they treated negatively correlated genes. There will be a high similarity in an unsigned network of genes with a high negative correlation compared with a low similarity in a signed network (37).

Then, $A=[a_{ij}]$ is defined using a thresholding procedure of the co-expression similarity. For an unweighted network, the adjacency is defined to be 1 ($a_{ij}=1$) and 0 otherwise if the absolute correlation between expression profiles is above a pre-defined threshold τ and deemed separated otherwise, as described in the following formula:

$$a_{ij}=signum(s_{ij},\tau)=\begin{cases}1 & if\ S_{ij}\geq\tau\\ 0 & otherwise\end{cases}$$

The hard thresholding of unweighted networks may lose the continuous nature of the underlying co-expression information (36). By contrast, a weighted network adjacency can be defined by raising the co-expression similarity s_{ij} to a power $\beta\geq1$, which is referred to as soft thresholding. It can allow the adjacency to take on values in succession between 0 and 1 to preserve the continuous nature of the co-expression information. The continuous measure for the assessment of gene connection strength is as follows:

$$a_{ij}=s_{ij}^{\beta}$$

This formula implies that the weighted adjacency a_{ij} between two genes is proportional to their similarity in the form of $\log(a_{ij}) = \beta \times \log(s_{ij})$.

Conversion and combination of gene association scores of the four methods. Following analysis of the gene interactions using the above four methods, the score of each gene pair was obtained. Since the results differed because of the various approaches taken, all score values of gene pairs were processed further to make them uniform at the same standard and converted to the form of rank/(total number of gene pairs). A novel algorithm was implemented to convert the scores of all gene pairs in this study. Four matrices were presented in three columns comprising gene pairs and the new score of each pair. By multiplication of the four matrices, a new matrix with a combined score of each gene pair was produced and sorting was conducted using a rank-based method similar to the application used in DE gene detection. Gene pairs were obtained ultimately following the management of all scores with a q-value package of FDR<0.1. A combined gene interaction network was then constructed by linking gene pairs.

Topological analysis. Following the calculation of scores using the four existing methods and the novel algorithm, and the construction of five networks, the clustering coefficient and short average path length of each were obtained and compared to investigate whether or not the networks had the classic small-world network property. Furthermore, considering that protein/gene interaction networks in general are scale-free (38), which means that they have power-law (or scale-free) degree distributions, the fitting coefficient R^2 of the power-law $y=ax^b$ of the five networks was also compared. The evaluation of topological parameters was conducted using the Network Analyzer Version 2.7 (39) plugin in Cytoscape Version 3.1.0 (40).

Functional enrichment analysis. Highly co-expressed genes generally participate in similar biological processes and pathways. To further investigate the biological functional enrichment of the co-expression gene pairs that were

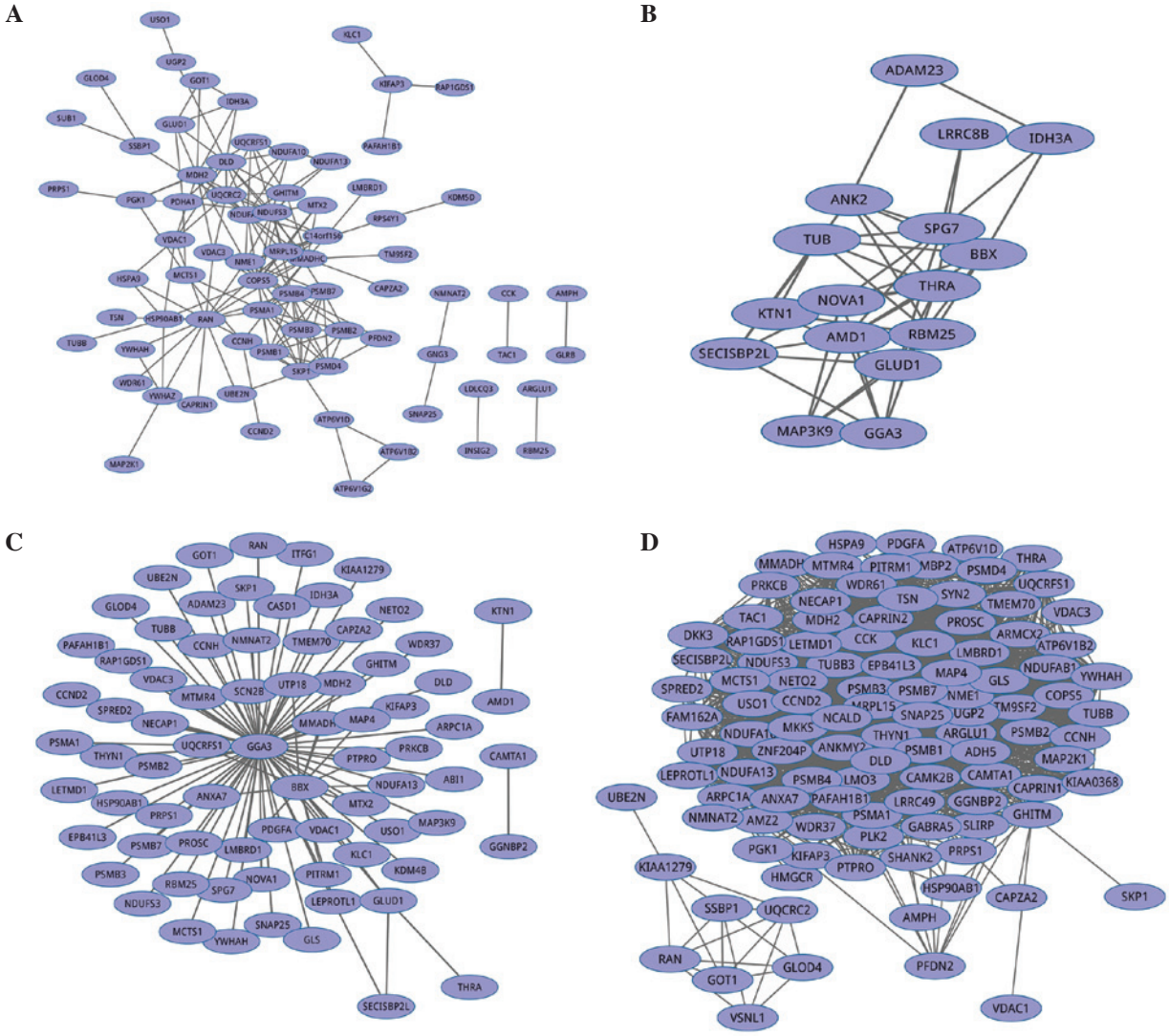

Figure 1. Graphical representation of co-expression networks identified by four existing method. Genes are denoted as nodes and interactions between gene pairs are presented as edges. (A) Search tool for the retrieval of interacting genes/proteins database, (B) differentially co-expressed genes and links, (C) empirical Bayesian and (D) weighted gene co-expression network analysis.

identified, a signaling pathway analysis was performed to assess the functional relevance of selected genes based on Kyoto Encyclopedia of Genes and Genomes (KEGG) database (www.genome.jp/kegg/), a widely used comprehensive resource for the pathway mapping of genes. DE genes identified by RP were first imported to the online tool Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov/tools.jsp), and all pathways these genes were enriched in was obtained. Then, on the basis of the DE genes in each pathway, the number of enriched co-expression gene pairs identified by the four existing methods and the new combined approach, respectively, were calculated and compared.

Results

Integrated analysis of DE genes in multiple studies. In the present study, three sets of hippocampal expression data associated with AD were integrated to identify DE genes using the RP method. After data preprocessing of three different datasets, the number of genes in E-GEOD-1297, E-GEOD-28146 and E-GEOD-5281 were 12,493, 20,109 and 20,109, respectively. Finally, a total of 144 DE genes were detected, including 8 upregulated genes and 136 downregulated genes, under an estimated pfp<0.01.

Co-expression analysis of four existing methods. Co-expression networks of DE genes were constructed using STRING, DCGL, EB and WGCNA analysis, respectively, and the co-expression relationships between gene and gene or co-expressed gene pairs were determined.

Scoring of gene associations based on STRING. A combined score was computed using the known and predicted associations, considering that various sources of association data are benchmarked independently in the STRING database. The combined score indicates a higher confidence level when more than one type of information supports a given association. A graphical protein-protein interaction network was constructed with 74 nodes and 166 edges (Fig. 1A). Also, all scores of gene pairs were obtained in the context of inputting 144 DE genes. A clustering coefficient of 0.300 and mean shortest path of 2.925 were computed. After conducting

Table II. Topological parameters of co-expression networks constructed using four existing approaches and the new algorithm.

Measure	STRING	DCGL	EB	WGCNA	Combined
R^2	0.786	0.037	0.477	0.071	0.810
Clustering coefficient	0.300	0.178	0.0	0.820	0.172
Mean shortest path length	2.925	1.783	2.038	1.578	3.618

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

degree distribution by nonlinear curve fitting according to the power law ($y=ax^b$), a fitting coefficient ($R^2=0.786$) was produced.

Construction of a gene co-expression network using DCGL. The DCGL 2.0 package in R was applied to identify DC genes and DC links, in which DCp and DCe methods involved in the DCEA module were employed. A total of 43 co-expression gene pairs were identified, and the two genes in each gene pair were DC genes. Finally, a co-expression network with 16 nodes and 43 edges was built using Cytoscape (Fig. 1B). A clustering coefficient of 0.178 and mean shortest path of 1.783 were computed. Likewise, the degrees of all nodes were determined and a fitting coefficient ($R^2=0.037$) of their degree distribution was obtained, which indicated that this network was not a scale-free network.

Construction of a gene co-expression network using EB methods. The EB approach was used to identify DC gene pairs based on 144 DE genes. A total of 88 protein pairs with FDR≤0.05 were produced and the relational values of all pairs were yielded following the analysis of gene expression relationships using meta-analysis. A gene interaction network containing 76 nodes and 88 edges was constructed using the 88 protein pairs in this analysis (Fig. 1C). The network was binary, with all interactions being unweighted and undirected. In addition, a clustering coefficient of 0.0 and mean shortest path of 2.038 were obtained. The degrees of all proteins were determined and a fitting coefficient of $R^2=0.477$ for their degree distribution was obtained following nonlinear regression according to the power law.

Construction of gene co-expression network using WGCNA. Using the WGCNA package, a total of 2,271 protein pairs were produced, and a co-expression network with 107 nodes and 2,271 edges was built using Cytoscape (Fig. 1D). The degrees of all nodes were determined and a fitting coefficient ($R^2=0.071$) of their degree distribution was obtained following nonlinear regression, which also presented a non scale-free property.

Combination of all gene pairs and construction of a co-expression network. In the present study, a novel algorithm was implemented to convert the score values of all gene pairs obtained from the four existing approaches in the form of rank/(total number of gene pairs). Multiplication of the four matrices produced a new matrix containing a combined score for each gene pair, and a simple rank-based permutation procedure was conducted. Then, a combined gene co-expression network was constructed that comprised 37 nodes linked by a total of 57 connections (Fig. 2A). The distribution of the number of links per node was scale free with $R^2=0.881$. Thus, the results conformed to a scale-free network whose degree distribution followed the power law ($y=ax^b$, a=12.464, b=-0.840; Fig. 2B).

Topological analysis of the five networks. Topological parameters of the five networks were compared, including the clustering coefficient, mean shortest path length and the fitting coefficient R^2 (Table II). The results showed that the network constructed by the WGCNA method had the greatest tendency to display small-world characteristics, as it had the smallest mean shortest path length and the largest clustering coefficient. However, the combined network showed a higher fitting coefficient R^2 than the other four networks, indicating its scale-free property.

Functional enrichment analysis. Firstly, all pathways that DE genes enriched were identified as background. To investigate the biological functional enrichment of the co-expression gene pairs identified by the different methods, the number of gene pairs enriched in each pathway was calculated and compared. The top five pathways were shown in Fig. 3. Co-expression gene pairs obtained using the EB and DCGL methods could not be enriched in any of the pathways that were identified, while co-expression gene pairs identified by STRING, WGCNA and the novel method were enriched in similar pathways. Following combination of the four existing methods, the co-expression gene pairs were found to be mostly enriched in proteasome, oxidative phosphorylation, Parkinson's disease, Huntington's disease, and AD pathways.

Discussion

Co-expression network-based approaches are powerful tools for the systematic identification of molecular mechanisms underlying biological processes, and a variety of algorithms have been developed to study these biological networks. Co-expression networks present binary relationships between individual genes, and also encode obscure higher level forms of cellular communication. In the present study, a co-expression network was constructed using a list of gene pairs with combined scores across multiple approaches. Three sets of hippocampal data associated with AD were employed and a total of 144 DE genes were identified using the RP package. From these DE genes, co-expression gene pairs were extracted by STRING, DCGL, EB and WGCNA approaches respectively,

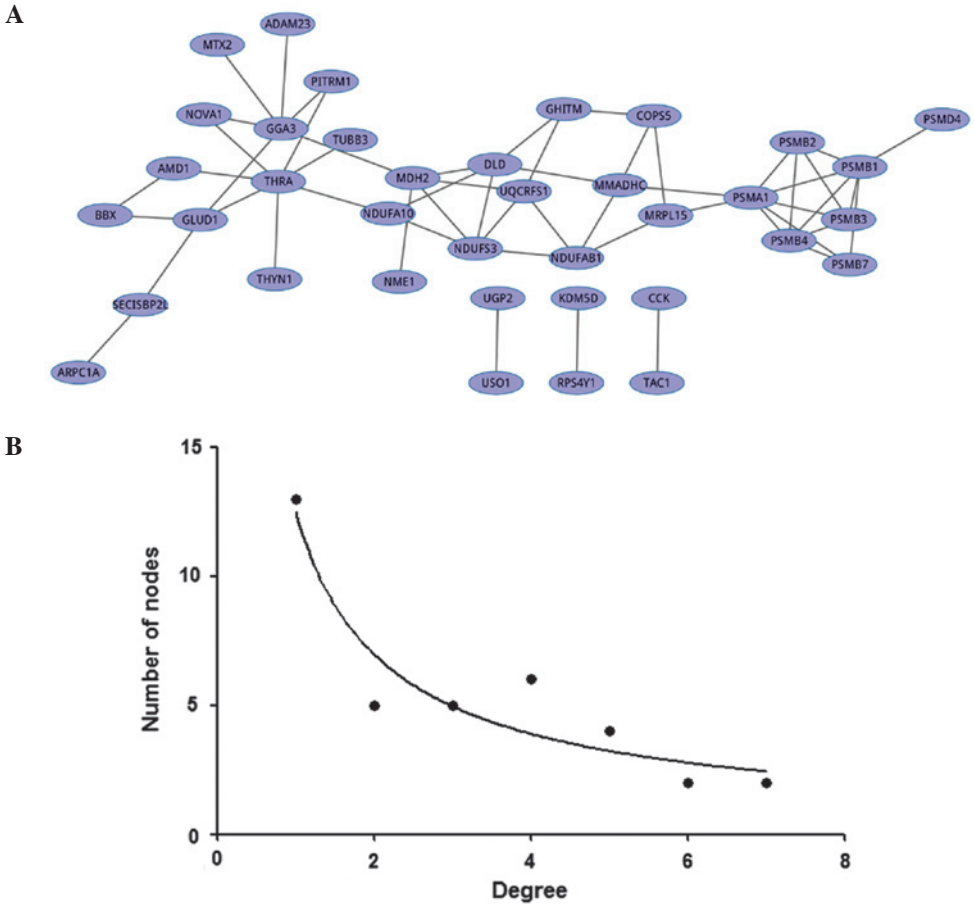

Figure 2. Combined co-expression network using the novel algorithm and its degree distribution. (A) Combined co-expression network based on the novel scores of each gene pair across four methods. A total of 37 nodes and 57 edges composed this combined network. (B) Scatter-gram of gene degree in this co-expression network. The combined co-expression network was a scale-free network whose degree distribution followed a power law ($y=ax^b$, where a=12.464, b=-0.840, R^2=0.881).

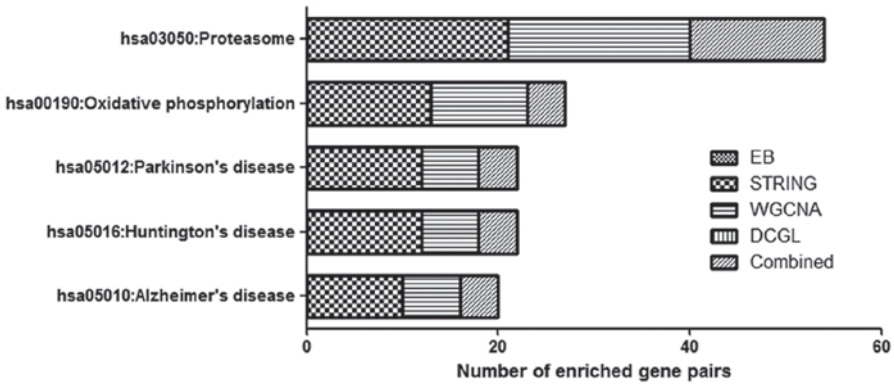

Figure 3. Five most enriched pathways of co-expression gene pairs identified by four existing methods and the novel algorithm. Co-expression gene pairs identified by EB and DCGL methods could not be enriched in any of the identified pathways. The five pathways were proteasome, oxidative phosphorylation, Parkinson's disease, Huntington's disease and Alzheimer's disease. EB, empirical Bayesian; STRING, search tool for the retrieval of interacting genes/proteins database; WGCNA, weighted gene co-expression network analysis; DCGL, differentially co-expressed genes and links.

and the score value of each gene pair was computed. Different approaches often give different results. To achieve a more reliable result, a novel algorithm was presented to produce a new score for each gene pair by combining the above four methods. Then, five networks were constructed, and their degree distribution and network topological properties (clustering coefficient and mean shortest path length) were compared.

Previous studies have analyzed the topological properties of gene co-expression networks, and have indicated that co-expression networks have small-world and scale-free properties (41,42). Such properties are typical of biological networks in which the nodes are connected when they are involved in the same biological process. Featherstone and Broadie (43) demonstrated that the uneven distribution of gene degrees in

a network, that is, a scale-free organization, helped organisms to resist the deleterious effects of mutation. A similar architecture was also found in the gene co-expression network of gastric cancer, which exhibited a hierarchical scale-free architecture (44). Furthermore, previous studies have confirmed the small-world property of biological networks with multiple data sources (45,46). However, a study conducted by Arita (47) indicated that the mean shortest path length of the biological network of *Escherichia coli* was much longer than previously thought, and the topology of this organism was not small. In the present study, co-expression networks for AD were built using four existing approaches and a novel algorithm, respectively. The results showed that the co-expression network constructed using the WGCNA method exhibited greater small-world network properties than the other four networks did, as it had the smallest mean shortest path length and the largest clustering coefficient. When analyzing the degree distributions of these co-expression networks, the combined gene interaction networks whose node degree distributions followed a power law with a high fitting coefficient clearly exhibited scale-free network characteristics.

Gene interactions are considered to be highly effective for use in the determination of gene functions and the identification of groups of genes that encode proteins in the same pathway. Previous studies have investigated the pathway enrichments associated with AD. Karim *et al* (48) demonstrated using an Ingenuity Pathway Analysis tool that synapse-associated pathways in neurons were tightly associated with the development and progression of AD. A more recent study highlighted cell adhesion molecules and purine metabolism pathways in AD by integrating genome-wide association study and brain expression data (49). In the present study, the co-expression gene pairs identified by the novel algorithm were mostly enriched in proteasome, oxidative phosphorylation, Parkinson's disease, Huntington's disease and AD. Consistent with this, Zabel *et al* (50) confirmed that proteasome and oxidative phosphorylation changes were closely associated with neurodegenerative disorders, such as AD, Parkinson's disease and Huntington's disease. Furthermore, in the present study, it was found that co-expression gene pairs identified by the EB and DCGL methods could not be enriched in any pathways that were identified, which was in contrast to the STRING and WGCNA analysis, and the novel method of the present study. Different methods for conducting co-expression network-based analysis often present varying abilities; thus, careful consideration is required when selecting synthetic methods, dependent on the nature of the research being undertaken.

In this study, a novel merged approach was used to identify co-expression gene pairs and enriched pathways, and this approach was compared with various network construction methods. Network analysis showed that the co-expression network constructed by the WGCNA method was most inclined to exhibit small-world properties, and the combined co-expression network exhibited scale-free network features. Moreover, the co-expression gene pairs were mostly enriched in proteasome, oxidative phosphorylation, Parkinson's disease, Huntington's disease and AD pathways. Each method of analysis has certain advantages and disadvantages. Considering the applications and limitations of each co-expression method, the novel algorithm developed in the present study may provide a new method for the analysis of gene interactions with a greater credibility and strength.

References

1. Schadt EE: Molecular networks as sensors and drivers of common human diseases. Nature 461: 218-223, 2009.
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511, 2000.
3. Liang WS, Reiman EM, Valla J, Dunckley T, Beach TG, Grover A, Niedzielko TL, Schneider LE, Mastroeni D, Caselli R, *et al*: Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. Proc Natl Acad Sci USA 105: 4441-4446, 2008.
4. Barabási AL and Oltvai ZN: Network biology: Understanding the cell's functional organization. Nat Rev Genet 5: 101-113, 2004.
5. He D, Liu ZP, Honda M, Kaneko S and Chen L: Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. J Mol Cell Biol 4: 140-152, 2012.
6. Van Leene J, Hollunder J, Eeckhout D, Persiau G, Van De Slijke E, Stals H, Van Isterdael G, Verkest A, Neirynck S, Buffel Y, *et al*: Targeted interactomics reveals a complex core cell cycle machinery in *Arabidopsis thaliana*. Mol Syst Biol 6: 397, 2010.
7. Wen Z, Liu ZP, Liu Z, Zhang Y and Chen L: An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc 20: 659-667, 2013.
8. Stuart JM, Segal E, Koller D and Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249-255, 2003.
9. Elo LL, Järvenpää H, Oresic M, Lahesmaa R and Aittokallio T: Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics 23: 2096-2103, 2007.
10. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R and Califano A: Reverse engineering of regulatory networks in human B cells. Nat Genet 37: 382-390, 2005.
11. Watts DJ and Strogatz SH: Collective dynamics of 'small-world' networks. Nature 393: 440-442, 1998.
12. Albert R: Scale-free networks in cell biology. J Cell Sci 118: 4947-4957, 2005.
13. Sporns O and Zwi JD: The small world of the cerebral cortex. Neuroinformatics 2: 145-162, 2004.
14. Zhang W, Zang Z, Song Y, Yang H and Yin Q: Co-expression network analysis of differentially expressed genes associated with metastasis in prolactin pituitary tumors. Mol Med Rep 10: 113-118, 2014.
15. Allen JD, Xie Y, Chen M, Girard L and Xiao G: Comparing statistical methods for constructing large scale gene networks. PLoS One 7: e29348, 2012.
16. Childs KL, Davidson RM and Buell CR: Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One 6: e22196, 2011.
17. Dawson JA and Kendziorski C: An empirical Bayesian approach for identifying differential coexpression in high-throughput experiments. Biometrics 68: 455-465, 2012.
18. Liu BH, Yu H, Tu K, Li C, Li YX and Li YY: DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics 26: 2637-2638, 2010.
19. Yu H, Liu BH, Ye ZQ, Li C, Li YX and Li YY: Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinformatics 12: 315, 2011.
20. 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 (Database Issue), 2005.
21. Mason MJ, Fan G, Plath K, Zhou Q and Horvath S: Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics 10: 327, 2009.
22. 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.

23. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR and Landfield PW: Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci USA 101: 2173-2178, 2004.
24. Blalock EM, Buechel HM, Popovic J, Geddes JW and Landfield PW: Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. J Chem Neuroanat 42: 118-126, 2011.
25. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, Caselli RJ, Kukull WA, McKeel D, Morris JC, *et al*: Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genomics 28: 311-322, 2007.
26. Ma L, Robinson LN and Towle HC: ChREBP*Mlx is the principal mediator of glucose-induced gene expression in the liver. J Biol Chem 281: 28721-28730, 2006.
27. 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.
28. Pepper SD, Saunders EK, Edwards LE, Wilson CL and Miller CJ: The utility of MAS5 expression summary and detection call algorithms. BMC Bioinformatics 8: 273, 2007.
29. 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.
30. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: Basic local alignment search tool. J Mol Biol 215: 403-410, 1990.
31. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C and Jensen LJ: STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808-D815 (Database Issue), 2013.
32. Cho SB, Kim J and Kim JH: Identifying set-wise differential co-expression in gene expression microarray data. BMC Bioinformatics 10: 109, 2009.
33. Raftery CFAE: Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97: 611-631, 2002.
34. Carey VJ, Gentry J, Whalen E and Gentleman R: Network structures and algorithms in Bioconductor. Bioinformatics 21: 135-136, 2005.
35. Cokus S, Rose S, Haynor D, Grønbech-Jensen N and Pellegrini M: Modelling the network of cell cycle transcription factors in the yeast *Saccharomyces cerevisiae*. BMC Bioinformatics 7: 381, 2006.
36. Zhang B and Horvath S: A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4: Article17, 2005.
37. Langfelder P and Horvath S: WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9: 559, 2008.
38. Barabasi AL and Albert R: Emergence of scaling in random networks. Science 286: 509-512, 1999.
39. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T and Albrecht M: Computing topological parameters of biological networks. Bioinformatics 24: 282-284, 2008.
40. Morris JH, Lotia S, Wu A, Doncheva NT, Albrecht M, Pico AR and Ferrin TE: SetsApp for Cytoscape: Set operations for Cytoscape Nodes and Edges. F1000Res 3: 149, 2014.
41. 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.
42. 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.
43. Featherstone DE and Broadie K: Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. Bioessays 24: 267-274, 2002.
44. 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.
45. Fell DA and Wagner A: The small world of metabolism. Nat Biotechnol 18: 1121-1122, 2000.
46. Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabási AL: Hierarchical organization of modularity in metabolic networks. Science 297: 1551-1555, 2002.
47. Arita M: The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101: 1543-1547, 2004.
48. Karim S, Mirza Z, Ansari SA, Rasool M, Iqbal Z, Sohrab SS, Kamal MA, Abuzenadah AM and Al-Qahtani MH: Transcriptomics study of neurodegenerative disease: Emphasis on synaptic dysfunction mechanism in Alzheimer's disease. CNS Neurol Disord Drug Targets 13: 1202-1212, 2014.
49. Xiang Z, Xu M, Liao M, Jiang Y, Jiang Q, Feng R, Zhang L, Ma G, Wang G, Chen Z, *et al*: Integrating genome-wide association study and brain expression data highlights cell adhesion molecules and purine metabolism in alzheimer's disease. Mol Neurobiol 52: 514-521, 2015.
50. Zabel C, Nguyen HP, Hin SC, Hartl D, Mao L and Klose J: Proteasome and oxidative phoshorylation changes may explain why aging is a risk factor for neurodegenerative disorders. J Proteomics 73: 2230-2238, 2010.