© Kamla-Raj 2018

Int J Hum Genet, 18(2): 109-115 (2018)
DOI: 10.31901/24566330.2018/18.2.694

Prediction of Key Gene Functions Associated with Chronic Hepatitis B Utilizing Network-Based Gene Function Inference Method

Wei-Li He[1], Wei Gou[2], Yan-Ling Wang[3], Bing Qiao[2] and Hai-Feng Zhang[2]

[1]Infections Department, Dongzhimen Hospital Beijing University of Chinese Medicine, Beijing, 100700, China
[2]Department of Hepatology (No.6), QingDao No.6 People's Hospital, QingDao, 266033, Shandong Province, China
[3]Outpatient Department, QingDao No.6 People's Hospital, QingDao, 266033, Shandong Province, China

KEYWORDS Chronic Hepatitis B. Co-Expression Network. Differentially Expressed Genes. Gene Ontology. Guilt by Association. Prediction

ABSTRACT The objective of this paper was to reveal key gene functions associated with chronic hepatitis B (CHB) patients utilizing network-based guilt by association (GBA) method. Firstly, differentially expressed genes (DEGs) and gene ontology (GO) annotation data were prepared. Subsequently, a co-expression matrix for DEGs was constructed based on Spearman correlation coefficient (SCC) method. And then prediction of optimal gene functions were conducted through extended GBA algorithm. Finally, key gene functions were selected according to the area under the receiver operating characteristics curve (AUC) index. Results of predictions showed that 241 GO terms had a good classification performance with AUC > 0.5. In particular, AUC for 8 ones was more than 0.7 and defined as key gene functions. The results might provide potential biomarkers for early detection and treatment of CHB patients, and give great insights to revealing molecular mechanism underlying this progression.

INTRODUCTION

Approximately one third of the world's population has serological evidence of past or present infection with hepatitis B virus (HBV) and 350 - 400 million people are chronic HBV surface antigen carriers (Desai et al. 2017). The spectrum of disease and natural history of chronic HBV infection are diverse and variable, ranging from an inactive carrier state to progressive chronic hepatitis B (CHB), which may evolve to cirrhosis hepatic decompensation, and hepatocellular carcinoma (HCC) (Brown et al. 2016; Atiase et al. 2018). Morbidity and mortality in CHB are linked to persistence of viral replication and evolution to cirrhosis HCC (Organization 2015). Longitudinal studies of untreated patients with CHB indicate that, after diagnosis, the 5-year cumulative incidence of developing cirrhosis ranges from 8 percent to 20 percent (Carlin et al. 2017). Hence effective prevention and treatment for CHB patients are vital and urgent at present.

Generally, network-based approach is capable of extracting informative and significant genes dependent on bio-molecular networks rather than individual genes (Liu et al. 2012; Blum et al. 2018). Utlizing networks for co-expression analysis has been an attractive proposition for a number of reasons, such as statistical confidence of individual connections, overlap with protein interaction, and mathematical convenience (Ma et al. 2013). Meanwhile, previous studies have shown that by using variants of guilt-by-association (GBA), gene function predictions can be made with very high statistical confidence assuming that the associations in the data of a gene are necessary in establishing guilt (Gillis and Pavlidis 2011a). The principle of GBA forms the basis for most gene function prediction algorithms, which typically apply relational information in order to predict new gene membership in gene function categories. The integration of network and GBA method, network-based GBA method, may provide a new

Address for correspondence:
Wei Gou
Department of Hepatology (No.6),
QingDao No.6 People's Hospital,
No.9 on Fushun Road, Shibei District,
QingDao, 266033,
Shandong Province, China
Telephone & Fax: 86-0532-81636808
E-mail: weigou112@126.com

way to identifying biomarkers and revealing molecular mechanisms for many diseases.

Objectives

In the present study, the researchers took Gene Ontology (GO) annotations and gene expression data as study objectives, employed the network-based GBA method to predict gene functions for CHB patients. Of which, a co-expression network (CEN) was constructed for differentially expressed genes (DEGs) based on Spearman correlation coefficient (SCC) method. In addition, key gene functions were selected according to the area under the receiver operating characteristics curve (AUC) index. These gene functions might be potential biomarkers for early detection and target treatment of this disease.

METHODOLOGY

Gene Lists

Gene expression data for CHB with accessing number GSE58208 were collected from the public online National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (https://www.ncbi. nlm. nih. gov/geo/). In details, the GSE58208 was deposited on Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2] Platform. Besides, it was comprised of two kinds of samples, normal controls and CHB samples, dependent on peripheral blood mononuclear cells (PBMCs) from healthy individuals and patients with CHB carriers, respectively. A total of 5 normal controls and 12 CHB samples were detected in this data. Prior to facilitate the subsequent analysis, standard normalizations and pretreatments were conducted on GSE58208 (Ding et al. 2018). Ultimately, 20,514 genes were obtained in the gene expression data to identify DEGs for CHB patients.

Specifically, Linear Models for Microarray Data (Limma) is an R/Bioconductor software that provides an integrated solution for analyzing data from gene expression experiments (Ritchie et al. 2015). Hence in this paper, the researchers employed the Limma package to identify DEGs between CHB and normal controls. The lmFit function implemented in Limma was utilized to perform empirical Bayes statistics and false discovery rate (FDR) calibration of the P values on the data (Kim et al. 2018; Liang et al. 2018). Only genes which satisfied with the thresholds of $P < 0.05$ and $|\log_2 \text{FoldChange}| > 2$ were defined as DEGs across CHB patients and normal controls.

Gene Sets

The Gene Ontology Consortium (http:// geneontology.org/) is a community-based bioinformatics resource that supplies information about gene product function using ontologies to represent biological knowledge. All human GO annotations were recruited from the Gene Ontology Consortium, which was consisted of 19,003 sets and 18,402 genes. Subsequently, the researchers propagated over the GO structure and filtered for GO terms on size such that each remaining term had between 20 and 1000 associated genes while excluding inferred from electronic annotation, a range generally gives stable performance (Gillis and Pavlidis 2011a,b). Furthermore, to make these GO terms correlated to CHB closely, DEGs between CHB patients and normal controls were mapped to them. If the number of DEGs for a GO term was smaller than 20, it would be removed. In other words, only GO terms including equal or more than 20 DEGs were reserved.

Gene Networks

CEN Construction

Using DEGs of CHB, a CEN or association matrix was constructed based on the SCC algorithm. Here, SCC is a measure of the correlation between two variables, giving a value between -1 and +1 inclusive. If SCC had a positive value, there was a positive linear correlation between two genes, otherwise, a negative relationship. Besides, for an interaction between gene *i* and *j*, the absolute SCC value was denoted as its weight value (*W*). The SCC was computed as following:

$$SCC = \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{g(i,k)-\bar{g}(i)}{\sigma(i)}\right)\cdot\left(\frac{g(j,k)-\bar{g}(j)}{\sigma(j)}\right)$$

Where *n* was the number of samples in the gene expression data; $g(i, k)$ or $g(j, k)$ was the expression level of gene *i* or *j* in the sample *k* under a specific condition; $g(i)$ or $g(j)$ represented the mean expression level of gene *i* or *j*, respectively. As a result, a co-expression matrix was gained, of which node was DEG, and the weight indicated the strength for the interaction or DEG-DEG pair.

Network Topological Analysis

To the best of the researchers' knowledge, topological centrality is shown to be effective for identifying essential molecules in well-characterized interaction networks (Prifti et al. 2010). Thus the researchers conducted the topological centrality analysis to understand the functionality of complex systems of DEGs in the CEN. Degree quantifies the local topology of each node, by summing up the number of its adjacent nodes (Gao et al. 2018). It gives a simple count of the number of interactions of a given node. Further, an assortativity coefficient was computed for node degrees in the CEN, which is the Pearson correlation coefficient of degree between pairs of linked nodes (Drozd-Dabrowska et al. 2017). Its positive value indicated a correlation between nodes of similar degree, while negative value meant relationships between nodes of different degree. Particularly, when assortativity coefficient = 1, the network is said to have perfect assortative mixing patterns, while at -1, the network is completely disassortative. An assortativity coefficient of 0 indicates the characteristic has no influence on partnering (random mixing).

Gene Function Predictions

During this step, the GBA method was implemented to predict significant gene functions in the progression of CHB patients. For GBA analysis, its principle forms the basis for most gene function prediction algorithms, which typically uses relational information to predict new gene membership in gene function categories (Mostafavi and Morris 2010). Taking GO as functional annotations, a multi-functionality score (MFS) was assigned to each gene *i* in the given co-expression matrix (Gillis and Pavlidis 2011a),

$$MFS(i) = \sum_{x|i\varepsilon \mathrm{GO}x} \frac{1}{Num_{in_x} * Num_{out_x}}$$

Of which Num_{inx} stood for the number of genes within GO group *x*, whose weighting had the effect of giving contribution to a GO group; and Num_{inx} was the number of genes outside GO group *x* in the CEN, whose weighting provided a corresponding weighting to genes not within the GO group. In other words, being the only gene outside a large GO group subtracted as much from that one gene's score as being the only gene within a GO group would add to another gene's score. Weighting referred to effect of counting membership in group by how much the gene contributes to that GO group.

Evaluation of Prediction Performance

The AUC was utilized as the main measure of performance in prediction. An AUC of 0.5 represents classification at chance levels, while an AUC of 1.0 represents a perfect classification. In the gene function prediction literature, AUC > 0.7 are considered good (Gillis and Pavlidis 2011b). Hence GO terms were defined with AUC > 0.7 as optimal gene functions for CHB patients. Briefly, the researchers applied 3-fold cross-validation to determine a MFS ranked list scoring genes as to how well they belonged within the known gene set and compute the AUC for assessing the classification performances between CHB samples and normal controls. To the best of the researchers' knowledge, AUC has been introduced as a better measure for evaluating the predictive ability of machine learners in support vector machines (SVM) model than accuracy to assess the clinical classification performance (Hansen et al. 2018). In consequence, the researchers could obtain the AUC values for predicted GO terms, and select these terms of AUC > 0.7 as key gene functions for CHB.

RESULTS

Gene Lists and Sets

In the present study, 263 DEGs were obtained between CHB patients and normal controls based on Limma package when setting the thresholds as$|\log_2 \mathrm{FoldChange}| > 2$ and $P < 0.05$. Meanwhile, total 19,003 GO terms involved in 18,402 genes were downloaded from the Gene Ontology Consortium. By removing terms with gene size out of 20 ~ 1000 and intersected DGEs amounts < 20, total 275 terms were reserved. As described above, the 263 DEGs were denoted as the gene lists, whereas the 275 GO terms were defined as the gene sets.

CEN

In order to explore the biological activities and correlations among DEGs, the researchers constructed a CEN with 263 nodes and 34,716 interactions for CHB patients based on the SCC method. By accessing the topological degree centrality analysis on the CEN, the degrees for all nodes were obtained. The node degree distribution was illustrated in Figure 1A. As the figure shown, degree for a large number of DEGs ranged from 130 to 160, which suggested that these DEGs could interact with each other felicitously. What's more, an assortativity coefficient was computed to evaluate degree assortative mixing pattern extent, which lies between -1 and +1 in general. The result showed that the assortativity coefficient for the CEN was 0.875, meaning that the network had perfect assortative mixing patterns. Besides, in the Figure 1B, the weight distribution for all interactions in the CEN was also displayer. In the figure, a square represented an edge in the CEN. The darker of the square, the larger of its weight value was. Interestingly, a good liner correlation was uncovered among interactions, which suggested that the CEN had a good network scale property.

Moreover, an interaction with high weight might be more important than the low ones, and thus the researchers selected interactions with weight > 0.8 to construct a sub-network. The sub-network was visualized by Cytoscape software which is an open project for integrating bio-molecular interactions with high-throughput expression data and other molecular states into a unified conceptual network (Fatemipour et al. 2017). As clarified in Figure 2, there were 50 nodes and 988 interactions. Interestingly, *WBP2*, *NFE2* and *ANKRD22* were the top three DEGs with highest degree either in CEN or its sub-network.

Key Gene Functions

Given a co-expression matrix, GBA method was used to predict gene functions and AUC index to determine key gene functions for CHB progression. In brief, a MFS was calculated for each gene in a GO term, which affected on counting membership in a GO by how much the gene contributed to that GO. If a gene had a high MFS, it would be a candidate for having any function. Hence a single ranked list of genes which best captured candidacy across all functions was equivalent to a list of genes ranked by MFS. Intuitively, if one was forced to choose a

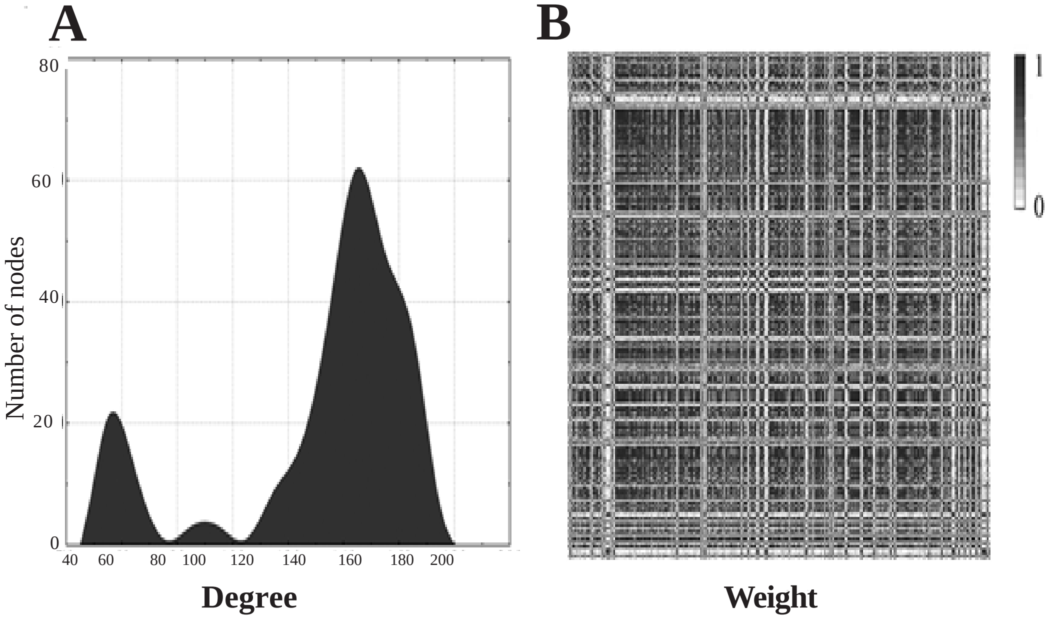

Fig. 1. Specific topological properties for co-expression network (CEN).
A: Node degree distribution; B: Weight distribution
Source: Author

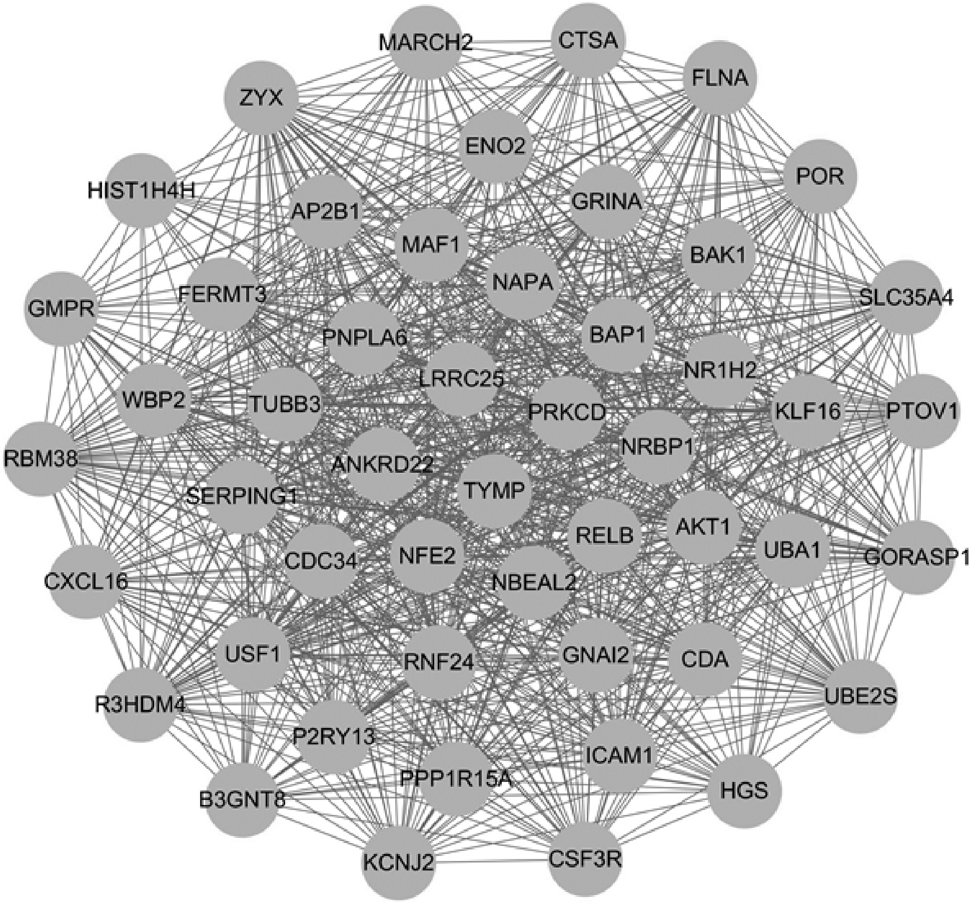

Fig. 2. Sub-network extracted from the co-expression network (CEN). Nodes were differentially expressed genes (DEGs), whereas an edge stood for the interaction between two DEGs with weight > 0.8
Source:Author

single ranking, the gene with the most GO annotations could be predicted as being in all GO categories. This was because if one gene was in 100 GO categories (high MFS), and another was in only one (low MFS), by placing the former gene ahead of the latter gene in a fixed ranking, the researchers made a correct prediction more often across all GO categories. Therefore, 3-fold cross-validation on MFS was carried out to calculate AUC for GO terms which aimed to classify CHB patients and normal controls.

Consequently, AUC values were obtained for 275 GO terms. The AUC distribution was shown in Figure 3. The AUC for most amount of GO terms contributed to the section of 0.5 ~ 0.7, especially 0.55 ~ 0.65. When used it as a predictor of GO category membership, the researchers should get values of the AUC of over 0.5 for many GO terms, and obtained 241 of 275 GO terms of AUC > 0.5 (about 87.64 %). Importantly, 8 GO terms had the AUC > 0.7 and were denoted as key gene functions in Table 1. As could be seen, the top five terms were regulation of protein modification process (GO:0031399, AUC = 0.745), organic substance transport (GO:0071702, AUC = 0.739), cellular protein modification process (GO:0006464, AUC = 0.725), positive regulation of cellular biosynthetic process (GO:0031328, AUC = 0.726), and nucleoplasm (GO:0005654, AUC = 0.719).

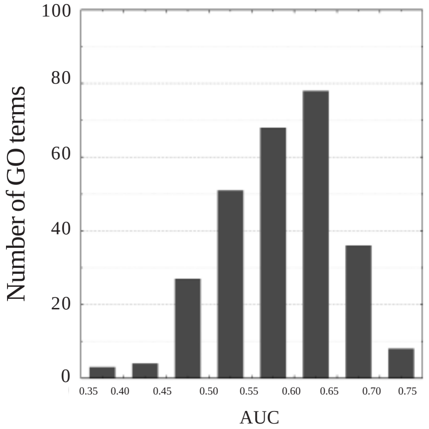

Fig. 3. Gene function prediction performance by guilt by association (GBA)
Source: Author

Table 1: 8 GO terms

GO ID	*Term*	*AUC*
GO:0031399	regulation of protein modification process	0.745
GO:0071702	organic substance transport	0.739
GO:0006464	cellular protein modification process	0.725
GO:0031328	positive regulation of cellular biosynthetic process	0.726
GO:0005654	nucleoplasm	0.719
GO:0007599	hemostasis	0.711
GO:0050817	coagulation	0.704
GO:0042592	homeostatic process	0.702

DISCUSSION

From a systemic view, hard thresholding in a network might increase the possibility of less robust results (Zhang and Horvath 2011). Whereas a soft-thresholding approach such as in weighted co-expressions has been shown to work well in the analysis of functional modules within the network (An et al. 2018), combining both greater sparsity with similarity to the original correlation matrix (Hansen et al. 2017). In consequence, the SCC method was applied to weight the CEN and its weight distribution suggested that the CEN had good scale network property. The CEN was comprised of 263 DEGs and 34,716 interactions, of which a sub-network with high weight ($W > 0.8$) was extracted to illustrate these important interactions in more detail. There were 50 nodes and 988 interactions in the sub-network of CEN.

Currently, numbers of researches have focused on extending GBA method to indirect connections, including voting within a fixed radius, clustering into function classes, specialized support vector machine, prediction by path length, weighting indirect connections by local topology, network propagation, topological overlap as well as others (Hu et al. 2018; Jiang et al. 2018). Most of them report improvement over GBA between direct connections, although where methods are compared they tend to perform comparably and only slightly better than direct GBA (Jing et al. 2018). In a word, these outcomes don't meet to the expectations perfectly. Therefore in this study, the researchers proposed a network-based GBA method to predict both direct and indirect key gene functions for CHB patients based on the integration of GO annotations, CEN and GBA method. Broadly speaking, the extended-GBA approach could make exhaustively examining issues faster and easier (less subject to fine-tuning) than the simple GBA.

Particularly, a MFS was computed for each gene enriched in the GO term, which played a role in the prediction of gene function from genomics data. The researchers wished to examine whether MFS was reflected in other properties of genes which were used in data interpretation. This is potentially important because it was usually assumed that when genes were assigned a function, it was due to either a valid prediction or a false positive due to promiscuity or other issues with the data. Furthermore, the AUC was utilized to assess the prediction performance for each GO term based on the MFS. As a result, ranking terms by AUC would be a good way to get good performance from a gene function prediction algorithm, which validated feasibility and confidence of the network-based GBA method. In details, about 87.64 percent of all GO terms had a good classification performance with AUC > 0.5. Especially 8 ones were more than 0.7 and defined as key gene functions, such as regulation of protein modification process, organic substance transport and cellular protein modification process.

CONCLUSION

In conclusion, the researchers have predicted 8 key gene functions in the progression of CHB utilizing the network-based GBA method.

The findings might give great insights to revealing molecular mechanism underlying CHB, and provide potential biomarkers for its prevention and target treatment.

RECOMMENDATIONS

Nevertheless, several limitations must be taken into consideration. To begin with, there were limited samples. Besides, how these key gene functions work together or regulate the disease is still unknown, and future study should focus on it.

REFERENCES

An J, Kim JW, Shim JH et al. 2018. Chronic hepatitis B infection and non-hepatocellular cancers: A hospital registry-based, case-control study. *PloS One,* 13: e0193232.

Atiase Y, Yorke E, Akpalu J, Opoku-Asare B, Adjei P, Amissah-Arthur MB, Akpalu A 2018. Pachydermoperiostosis in a patient with chronic hepatitis B virus infection referred as acromegaly: A case report. *Journal of Medical Case Reports,* 12: 59.

Blum CF, Heramvand N, Khonsari AS, Kollmann M 2018. Experimental noise cutoff boosts inferability of transcriptional networks in large-scale gene-deletion studies. *Nature Communications,* 9: 133.

Brown RS, Mcmahon BJ, Lok ASF, Wong JB, Ahmed AT, Mouchli MA, Wang Z, Prokop LJ, Murad MH, Mohammed K 2016. Antiviral therapy in chronic hepatitis B viral infection during pregnancy: A systematic review and meta-analysis. *Hepatology,* 63(1): 319-333.

Carlin DE, Paull EO, Graim K, Wong CK, Bivol A, Ryabinin P, Ellrott K, Sokolov A, Stuart JM 2017. Prophetic granger causality to infer gene regulatory networks. *Scientific Reports,* 12: e0170340.

Desai JS, Sartor RC, Lawas LM, Jagadish SVK, Doherty CJ 2017. Improving gene regulatory network inference by incorporating rates of transcriptional changes. *Sci Rep,* 7(1): 17244.

Ding P, Luo J, Liang C, Xiao Q, Cao B 2018. Human disease MiRNA inference by combining target information based on heterogeneous manifolds. *J Biomed Inform,* 80: 26-36.

Drozd-Dabrowska M, Ganczak M, Karpinska E 2017. Concerns related to CCR5 gene delta 32 mutation role in hepatitis B virus infection. *Przeglad Epidemiologiczny,* 71: 571-581.

Fatemipour M, Arabzadeh SAM, Molaei H, Geramizadeh B, Dabiri S, Fatemipour B, Vahedi SM, Malekpour-Afshar R 2017. Evaluation of STAT3 rs1053004 single nucleotide polymorphism in patients with chronic hepatitis B and hepatocellular carcinoma. *Cellular and Molecular Biology,* 63: 45-50.

Gao H, Yang M, Zhang X 2018. Investigating a multigene prognostic assay based on significant pathways for luminal A breast cancer through gene expression profile analysis. *International Journal of Genomics,* 15: 5027-5033.

Gillis J, Pavlidis P 2011a. The impact of multifunctional genes on “Guilt by association” analysis. *PLoS One,* 6: e17258.

Gillis J, Pavlidis P 2011b. The role of indirect connections in gene networks in predicting function. *Bioinformatics,* 27: 1860-1866.

Hansen BO, Meyer EH, Ferrari C, Vaid N, MovahedI S, Vandepoele K, Nikoloski Z, Mutwil M 2018. Ensemble gene function prediction database reveals genes important for complex I formation in Arabidopsis thaliana. *PloS One,* 217: 1521-1534.

Hansen J, Meretzky D, Woldesenbet S, Stolovitzky G, Iyengar R 2017. A flexible ontology for inference of emergent whole cell function from relationships between subcellular processes. *PloS One,* 7: 17689.

Hu Y, Fang Z, Yang Y, Fan T, Wang J 2018. Analyzing the pathways enriched in genes associated with nicotine dependence in the context of human protein-protein interaction network. *Journal of Biomolecular Structure and Dynamics,* 24: 1-26.

Jiang Q, Liu Y, Xu B, Zheng W et al. 2018. Analysis of T cell receptor repertoire in monozygotic twins concordant and discordant for chronic hepatitis B infection. *Biochemical and Biophysical Research Communications,* 497: 153-159.

Jing R, Liang Y, Ran Y, Feng S, Wei Y 2018. Ensemble methods with voting protocols exhibit superior performance for predicting cancer clinical endpoints and providing more complete coverage of disease-related genes. *Int J Genomics,* 2018(2018): 14.

Kim MY, Kim JS, Son SH, Lim CS et al. 2018. Mbd2-CP2c loop drives adult-type globin gene expression and definitive erythropoiesis. *Nucleic Acids Research,* 14: gky193.

Liang Y, Yano Y, Putri WA, Mardian Y, Okada R, Tanahashi T, Murakami Y, Hayashi Y 2018. Early changes in quasispecies variant after antiviral therapy for chronic hepatitis B. *Molecular Medicine Reports,* 17: 5528-5537.

Liu Z-P, Wang Y, Zhang X-S, Chen L 2012. Network-based analysis of complex diseases. *IET Systems Biology,* 6(1): 22-33.

Ma S, Shah S, Bohnert HJ, Snyder M, Dineshkumar SP 2013. Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. *Plos Genetics,* 9: 2184-2196.

Mostafavi S, Morris Q 2010. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics,* 26: 1759-1765.

Prifti E, Zucker J-D, Clement K, Henegar C 2010. Interactional and functional centrality in transcriptional co-expression networks. *Bioinformatics,* 26: 3083-3089.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research,* 43(7): e47.

WHO 2015. *Guidelines for the Prevention, Care and Treatment of Persons with Chronic Hepatitis B Infection.* Geneva: World Health Organization.

Zhang B, Horvath S 2011. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology,* 4: 1-45.

Paper received for publication on March 2018
Paper accepted for publication on May 2018