e-ISSN:2320-3528
p-ISSN:2347-2286

Research & Reviews: Journal of Microbiology and Biotechnology

Genome-wide Screening of Cell Wall-related Modules in *Mycobacterium tuberculosis* based on Large-scale Co-Expression Analysis

Huaidong Wang[1], Bingqiang Liu[2], Zhuoyuan Xin[1], Zipeng Duan[1], Xiaoyu Sun[1], Yan Lin[1], Zhifeng Yang[2], Guoqing Wang[1]*, Fan Li[1]*

[1]Department of Pathogenobiology, The Key Laboratory of Zoonosis, Chinese Ministry of Education, College of Basic Medicine, Jilin University, Changchun, Jilin,130021, P. R. China
[2]School of Mathematics, Shandong University, Jinan 250100, P. R. China

Research Article

Received date: 30/11/2015
Accepted date: 05/08/2016
Published date: 15/08/2016

***For Correspondence**

Guoqing Wang, Department of Pathogenobiology, The Key Laboratory of Zoonosis, Chinese Ministry of Education, College of Basic Medicine, Jilin University, Changchun, Jilin,130021, P. R. China

Fan Li, Department of Pathogenobiology, The Key Laboratory of Zoonosis, Chinese Ministry of Education, College of Basic Medicine, Jilin University, Changchun, Jilin,130021, P. R. China

E-mail: qing@jlu.edu.cn

Keywords: *Mycobacterium tuberculosis,* Cell wall-related modules, Motif analysis, Cluster analysis

ABSTRACT

The cell wall of *Mycobacterium tuberculosis* plays an important role in pathogenesis. It is impossible to analyze the cell wall-associated genes one by one since lacking of functional annotation. Here, we performed clustering analysis of gene microarray expression data and acquired 33 co-expression modules by the construction of three nodes co-regulated networks. A total of 555 cell wall-related genes were predicted in the modules using a multi-factor logistic regression model and motif predicting analysis. The module analysis identified 15 genes without annotation that were also associated with the cell wall. Twenty-five modules contained significant motifs, and genes in 10 of these 25 modules shared a common motif. The methodological approach utilized herein may be applied to identification and description of other function-associated genes in the *M. tuberculosis* genome. The results of this study might improve understanding of the *M. tuberculosis* cell wall, and in finding new targets for anti-TB drugs.

INTRODUCTION

Tuberculosis (TB) is one of the most serious global health issues, the pathogenesis of which is still not clearly understood [1]. Emergence of multidrug resistant (MDR) and, more recently, extremely drug-resistant (XDR) *Mycobacterium tuberculosis* strains, along with TB-HIV co-infection, have become the new major challenge for TB therapy and control [2]. Anti-TB drugs, such as rifampin and isoniazid, were discovered in 1963, and since then, there have not been any discovery of novel, efficient anti-TB drugs [3]. The number of effective therapeutic targets for TB treatment is insufficient, especially for treatment of MDR-TB. Thus, high-throughput screening for therapeutic targets is the first and most important step for the development of novel anti-TB drugs and improved control of TB [4].

The cell wall of *M. tuberculosis* is mainly composed of capsule, mycolic acid, peptidoglycan, arabinose and intima [5]. These components play an important role in the processes that maintain the integrity of *M. tuberculosis* cell morphology, act against erosion by chemicals, escape host immune response, and lead to development of drug resistance and on a whole, increase the pathogenicity of *M. tuberculosis*. In this view, the *M. tuberculosis* cell wall-related components and biosynthesis pathways could be used as targets of anti-TB drugs. Traditional anti-TB drugs, such as Isonicotinic acid hydrazide (INH), target the mycolic acid synthesis pathway [6]. However, inadequate or interrupted treatment with INH results in INH resistance through acquisition

of mutations in *inhA, ahpC, nadh, katG, or KasA* in *M. tuberculosis* clinical isolates [7-9]. Hence, it is prudent to screen new *M. tuberculosis* cell wall-relevant genes. In addition, the components and functions of *M. tuberculosis* cell walls are quite complex and various. Traditional methods for screening *M. tuberculosis* cell wall genes are expensive and inefficient since they use the technologies of gene knockout and RNA interference to screen the potential cell wall synthesis genes or the target molecules in metabolism. Along with the limitations of experimental methodology, there is still no efficient technology to systematically screen molecules relevant to cell wall synthesis. Thus, the development of a novel approach for panoramic scanning and screening of the *M. tuberculosis* genes related to cell wall synthesis is necessary. For the above reasons, in our study, we retrieved all published gene microarray data for *M. tuberculosis* H37Rv, and established the co-regulatory networks of *M. tuberculosis* genes associated with cell wall synthesis, unknown genes and other genes by means of integration and clustering [10]. The module analysis and the high-throughput annotation of *M. tuberculosis* genes associated with cell walls provide a molecular basis for the research and development of novel efficient and sensitive anti-TB drugs with less harmful effects.

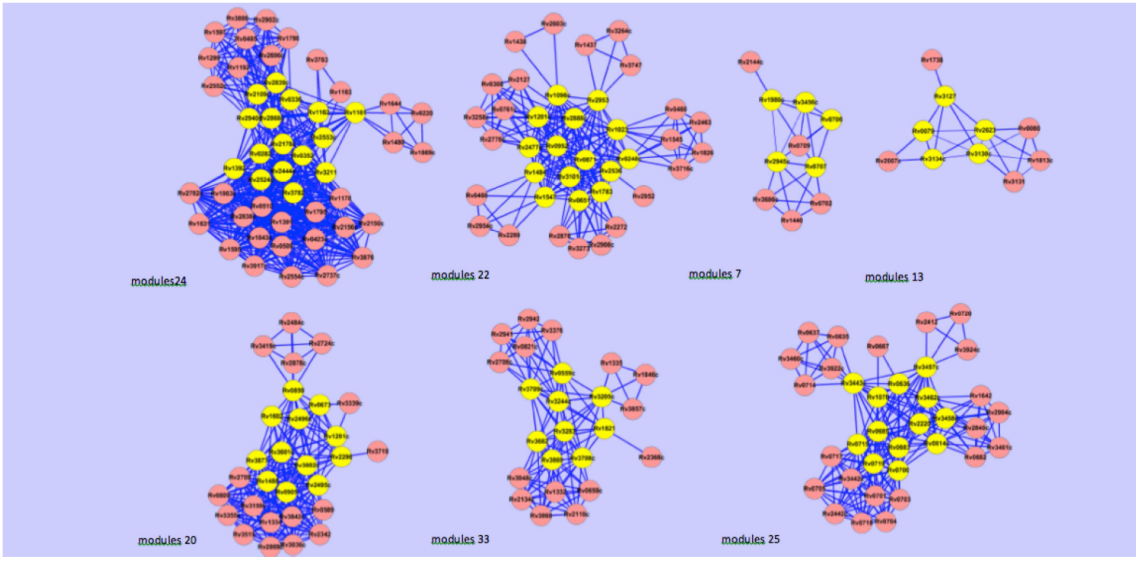
RESULTS AND DISCUSSION

Cell wall-related modules in *M. tuberculosis*

Microarray data for *M. tuberculosis* H37Rv were downloaded from the NCBI database (as of May 2013), which totaled 2863 microarrays of 43 series. These microarrays were related to DNA methylation (n=15), Drug action on tuberculosis (n=1076), Growth and Growth condition to *M. tuberculosis* (n=758), infection (n=910), gene mutation (n=78) and regulating factors (n=21) **(Table S1).** A total of 727 genes that were annotated in the Gene Ontology database (www.geneontology.org) were utilized as “seed” genes, and processed for cluster analysis by means of hierarchical clustering, K-means clustering and integrated clustering **(Table 1).** Based on the results of hierarchical clustering, every module contained a large number of genes, and the largest module contained 342 genes. This method decreased the discrimination of gene functions, although hierarchical clustering revealed the interaction between genes [11]. On the contrary, based on K-means clustering, the genes associated with cell walls were scattered in 201 modules. This method had higher discrimination, but was less clear regarding the interaction between genes [12]. Considering the outcomes of these two methods, integrated clustering was performed based on both K-means clustering and hierarchical clustering (detailed in Materials and Methods). Using this integrated cluster analysis, we identified 163 modules, to which all the known cell wall-associated genes were allocated. Statistical analysis demonstrated that 33 of the 163 modules were closely associated with cell wall synthesis, which contained a total of 555 genes. The correlation of these modules was calculated by Pearson correlation coefficient, and the integrated clustering results were visualized with Cytoscape software version 3.0.2. The results of GO and Pfam analysis illustrated that these genes correlated significantly with several biological processes, such as pathogenesis and response to stimulation **(Figure 1)**, which were consistent with the function of the cell wall in *M. tuberculosis* pathogenesis. These results showed a better discrimination of the integrated cluster analysis since at most 49 genes were contained in one module and at most 16 genes were associated with the cell wall.

Table 1. Cluster analysis of *Mycobacterium tuberculosis* H37Rv gene expression

	Total number of modules	No. of modules containing cell wall-associated genes	Maximum No. of genes in one module	Significantly related modules	Maximum No. of cell wall-associated genes in one module
K-means cluster	308	201	54	24	19
Hierarchical cluster	308	131	342	29	70
Integrated cluster	163	163	49	33	16



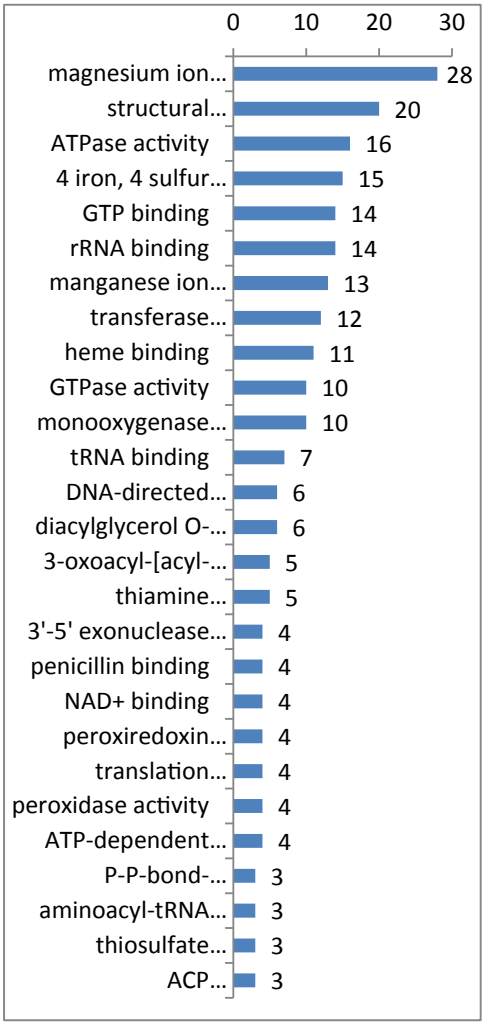
The yellow spots represent the known *Mycobacterium tuberculosis* genes associated with cell walls. The red spots represent candidate genes. The lines represent the relationships between genes. The thickness of the lines represents the strength of correlation.

Figure 1. Cell wall-related modules in *Mycobacterium tuberculosis*.

e-ISSN:2320-3528
p-ISSN:2347-2286

Multi-factor logistic regression analysis of cell wall related modules

A multi-factor logistic regression was used to analyze the correlation between observations and several factors due to polygenic co-effect of *M. tuberculosis* cell wall functions [13]. Each factor was evaluated by means of multi-factor linear regression equations [14]. After that, we analyzed the significance of GO-results using multi-factor logistic regression analysis **(Figure 2)**, and constructed a predictive model (detailed in Materials and Methods).



A. Items in biological processes; B. Items in molecular functions.

Figure 2. Significant items (number of genes) associated with cell wall.

We randomly chose 126 genes from the 727 known cell wall-associated genes, and 128 genes from the 1703 known genes not associated with the cell wall to perform verification. Our predictive model identified 125 of the 126 cell wall-associated genes with an accuracy 99.2%, while 122 of the 128 non-cell-wall-associated genes were correctly determined, the accuracy of which was about 95.3%. So we utilized this predictive model to identify the genes in the 33 cell wall-associated modules described above; 120 of the 197 annotated genes were found associated with cell wall **(Table S2).**

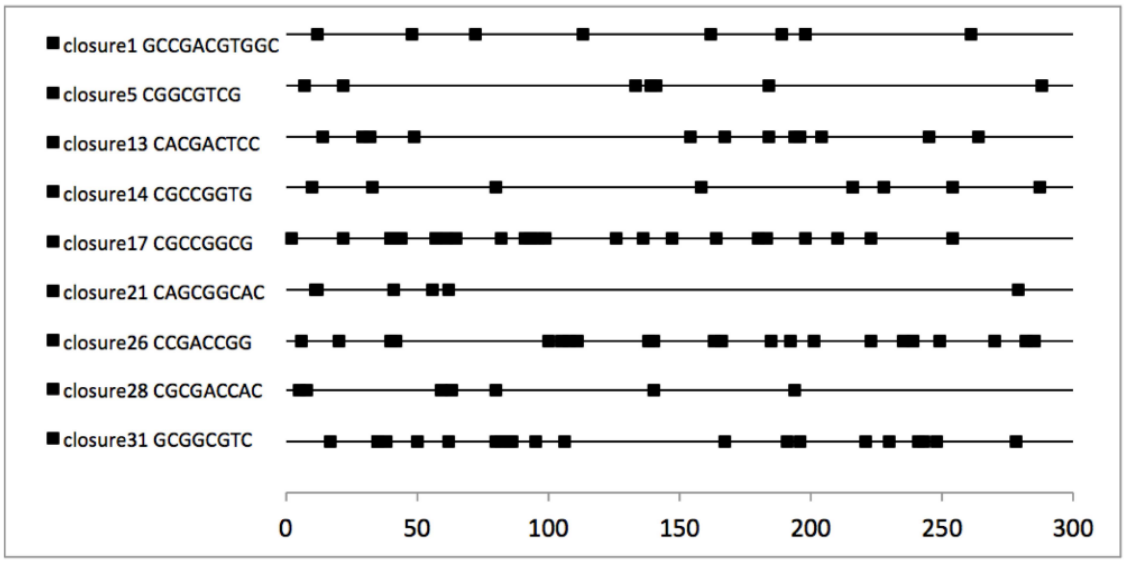
Motif analysis of cell wall-related modules

The 33 modules associated with the cell wall were subjected to motif analysis using BOBRO software [15]. As a result, 25 modules were found to contain significant motifs whereas genes in 10 of these 25 modules shared a common motif **(Figure 3).** In addition, these motifs were located at several sites, and the number of motifs before operons was different, which was in accord with the general law of the distribution of motifs.

According to the results of motif prediction, 15 genes without annotation in the 10 modules were associated with cell wall. In particular, module 5 included three cell wall-associated genes (Rv2005c, Rv3132c, Rv3133c) and one unknown gene (Rv0082). Using DOOR2 (http://csbl.bmb.uga.edu/DOOR/), we found that the four genes in module 5 were regulated by three operons (NO.7810, NO.8253, NO.8521), which had a common motif sequence of CGGCGTCG. This revealed the relationship between Rc0082 and the function of the cell wall. In addition, CAAT-box located at 70~100 bps, demonstrated the accuracy of the upstream motif prediction. The cell wall-related genes, Rv1440 and Rv0702, in module 13 were also identified as cell wall-associated genes through the model analysis. These two genes shared common transcription factors with other cell wall-associated genes in module 13. Therefore, we assumed that Rv1440 and Rv0702 were closely related to cell wall function. It is

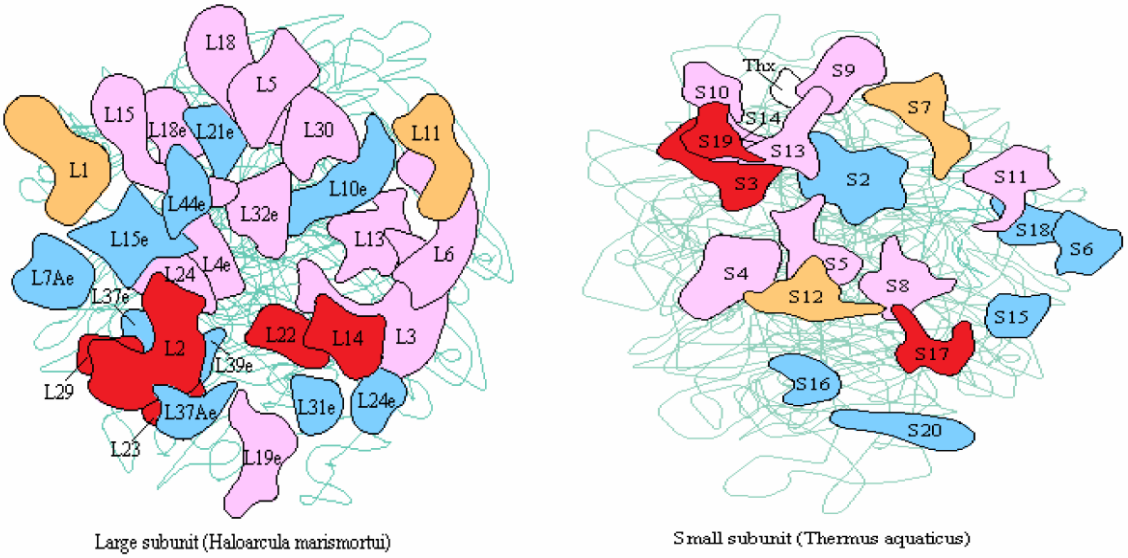
e-ISSN:2320-3528
p-ISSN:2347-2286

remarkable that Rv0702 is a part of the Rv0702~Rv0710 gene cluster, whose expression was associated with ribosome proteins as we showed by using the KEGG pathway analysis tool (http://www.genome.jp/kegg/) **(Figure 4).**



The rectangles represent motif locations. The base-pair sequences represent the motifs. The cluster numbers represent the modules in which the motifs exist.

Figure 3. Landscape of motifs in ten cell wall-related modules.



The red subunits represent the proteins expressed by the *Rv0702~Rv0710* gene cluster. The pink subunits represent the ribosomal proteins in *Mycobacterium tuberculosis*. The blue subunits represent the proteins unassociated with ribosomes.

Figure 4. Schematic view of the ribosome subunit proteins.

Previous studies demonstrated that genes involved in the synthesis of ribosome proteins also participate in the synthesis of cell wall [16], and the restriction of the function of genes associated with cell wall synthesis can change the structure of ribosomes [17]. Hence, genes associated with cell wall synthesis also play an important role in ribosome synthesis. In addition to the above findings, the genes, Rv0702, Rv0706, Rv0707, and Rv0709, in module 13 not only were close in proximity, but also shared the same operon, No. 7948. Other genes regulated by operon No. 7948 contained 33 modules in all, including Rv0703, Rv0704, Rv0705, Rv0708, and Rv0710. Furthermore, it was verified that Rv0706 and Rv0707 were associated with *M. tuberculosis* cell wall. Together with the characteristics of prokaryotic genic expression, we believe that genes in the modules regulated by operon No. 7948 might be associated with *M. tuberculosis* cell wall.

MATERIALS AND METHODS

Collection and mining of data

Gene microarray data for strain H37Rv of *M. tuberculosis* were downloaded from NCBI (www.ncbi.nlm.nih.gov/gds). After selection, 2710 H37Rv microarrays of 43 series were retained. Owing to the different probe number of several microarrays from different companies, we unified the genes in different microarrays according to the *M. tuberculosis* gene number and names

published by KEGG database in 2013; missing items were represented as NA. We further standardized the microarray data by means of min-max procedure as follows: the 5% largest and 5% smallest data in every microarray were given the same maximum value or minimum value, respectively, to remove the effect of extreme values. The values ranged from 1 (min) to 100 (max). A total of 727 *M. tuberculosis* genes were selected through GO analysis.

Cluster analysis

A 3994 × 2710 matrix was built using the standardized microarray data. The rows of the matrix represented the expression of each gene under 2710 conditions, and the columns represented the expression of each condition for every gene. The cluster analysis was performed using the bioinformatics toolbox of Matlab software. During hierarchical clustering, we used the Spearman function to calculate correlation. The complete function was best for calculating the linkages within clusters. Finally, the correlation function was used to calculate correlation in K-means clustering.

Construction of co-expression networks and predicitive model

Every gene in the same module had a certain correlation. To display the correlation, we calculated Pearson correlation coefficients for each pair of genes in every module. A positive co-expression was one with an R-value greater than 0.90, while negative co-expression was indicated by an R-value less than -0.90 (Usadel et al.). A co-expression network was constructed using Cytoscape software version 3.0.2.

Multi-factor logistic regression analysis

Highly significant cell wall-associated items ($P<0.01$) were selected in the GO database, and a multi-factor logistic regression analysis model was constructed using those items as dependent variables. To verify the sensitivity and specificity of the multi-factor logistic regression model, 126 genes were randomly chosen among 727 cell wall-associated genes and 128 genes of 1703 genes not associated with cell wall. We marked the items, which had the dependent variables in the annotations, with 1. Then, by the theory of logic regression, the gene was not associated with *M. tuberculosis* cell wall if all of the results were 0. This model was used to predict cell wall-associated the genes in the 33 modules.

Motif analysis

A total of 33 modules associated with cell walls were analyzed using BOBRO software [18]. BOBRO is accurate motif prediction software focused on the features of prokaryotic genomes, which uses the algorithms of motif closures and graph theory. It is based on the hypothesis that the internal genes might be regulated by the same transcription factors and prokaryotic regulatory regions [19]. Upstream 300-bp DNA sequences were selected as regulatory regions, and the conservative sites were searched upon the regulatory regions, used as candidate regulatory motifs. The significant motifs were selected utilizing the characteristics of transcriptional regulation motif sequences.

Statistical analysis

The significances were analysed using hypergeometric distribution, and P-values less than 0.01 were considered statistically significant. Correlation was calculated using the Pearson correlation coefficient, and Chi-square test was used to analyze the discrepancy between the three clusters.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (61303084 and 81271897), Specialized Research Fund for the Doctoral Program of Higher Education of China (20110061120093), China Postdoctoral Science Foundation (20110491311 and 2012T50285, Foundation of Jilin Provincial Health Department (2011Z049), Foundation of Jilin Province Science and Technology Department (20130522013JH and 20140414048GH) and the Norman Bethune Program of Jilin University (No. 2012219). We also thank the Medjaden Bioscience Limited for editing and proofreading this manuscript.

REFERENCES

1. Dutta NK, et al. Genetic requirements for the survival of tubercle bacilli in primates. J Infect Dis 2010; 201: 1743-1752.
2. Cohn DL, et al. Drug-resistant tuberculosis: review of the worldwide situation and the WHO/IUATLD Global Surveillance Project. International Union against Tuberculosis and Lung Disease. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America 1997; 24 1: S121-130.
3. Brigden G, et al. Principles for designing future regimens for multidrug-resistant tuberculosis. Bulletin of the World Health Organization 2014; 92: 68-74.
4. Brennan PJ. Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. Tuberculosis (Edinb) 2003; 83: 91-97.
5. Vilcheze C, et al. Inactivation of the inhA-encoded fatty acid synthase II (FASII) enoyl-acyl carrier protein reductase induces accumulation of the FASI end products and cell lysis of Mycobacterium smegmatis. Journal of bacteriology 2000; 182: 4059-4067.

6. Bergler H, et al. Protein EnvM is the NADH-dependent enoyl-ACP reductase (FabI) of Escherichia coli. J Biol Chem 1994; 269: 5493-5496.
7. Wilson TM and Collins DM. ahpC, a gene involved in isoniazid resistance of the Mycobacterium tuberculosis complex. Molecular microbiology 1996; 19: 1025-1034.
8. Wade MM, et al. Accurate mapping of mutations of pyrazinamide-resistant Mycobacterium tuberculosis strains with a scanning-frame oligonucleotide microarray. Diagnostic microbiology and infectious disease 2004; 49: 89-97.
9. Yu SX and Shi J. Segmentation given partial grouping constraints. IEEE Trans Pattern Anal Mach Intell 2004; 26: 173-183.
10. Vermunt JK. K-means may perform as well as mixture model clustering but may also be much worse: comment on Steinley and Brusco (2011). Psychol Meth 2011; 16: 82-88; discussion 89-92.
11. Baltar VT, et al. A structural equation modelling approach to explore the role of B vitamins and immune markers in lung cancer risk. Eur J Epidemiol 2013; 28: 677-688.
12. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic acid Res 2013; 41: 991-995.
13. Galagan JE, et al. The Mycobacterium tuberculosis regulatory network and hypoxia. Nature 2013; 499: 178-183.
14. Fu LM. Machine learning and tubercular drug target recognition. Curr Pharm Des 2014; 20: 4307-4318.
15. Usadel B, et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ 2009; 32: 1633-1651.
16. Shannon PT, et al. RCytoscape: tools for exploratory network analysis. BMC Bioinformatics 2013; 14: 217.
17. Do JH and Choi DK. Clustering approaches to identifying gene expression patterns from DNA microarray data. Mol Cells 2008; 25: 279-288.
18. Li G, et al. A new framework for identifying cis-regulatory motifs in prokaryotes. Nucleic Acids Res 2011; 39: e42.
19. Ma Q, et al. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. Bioinformatics 2013; 29: 2261-2268.