Int J Clin Exp Med 2019;12(7):8565-8571
www.ijcem.com /ISSN:1940-5901/IJCEM0088648

Original Article

Application of weighted gene co-expression network analysis to rheumatoid arthritis

Xing Song, Xiaofeng Zeng

Department of Rheumatology, Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical Sciences, Beijing, China

Received November 19, 2018; Accepted April 10, 2019; Epub July 15, 2019; Published July 30, 2019

Abstract: The goal of this study was to identify hub genes as potential targets in rheumatoid arthritis (RA) using weighted gene co-expression network analysis (WGCNA). Gene expression profiles of GSE17755 were downloaded from the GEO database and screened for differentially expressed genes (DEGs) with the limma package in R. Significant modules in the network were identified via WGCNA. Then, Gene Ontology (GO) functional enrichment of genes in the most significant module was analyzed using Database for Annotation, Visualization, and Integrated Discovery (DAVID). Finally, the disease-related gene co-expression network was visualized using Cytoscape, and hub genes were identified on CytoHubba. Overall, 3666 DEGs and 8 modules were identified. The turquoise module including 1044 genes was identified as the most relevant to RA. GO functional enrichment showed genes in the most relevant module were mainly related to the inflammatory response and the type I interferon signaling pathway. Ten hub genes, including PIGL, PRKAA1, and MRPS10, were identified. Genes related to the inflammatory response and the type I interferon signaling pathway possibly play critical roles in RA pathogenesis. PIGL, PRKAA1, and MRPS10 may be new targets for treating RA.

Keywords: WGCNA, rheumatoid arthritis, hub genes, PIGL, PRKAA1

Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune disease involving multiple systems and is characterized by persistent inflammatory synovitis of peripheral joints. Its incidence is approximately 1% worldwide and it mostly affects women [1]. It is known that failure to control inflammation over time causes cartilage damage, bone erosion, and joint ankylosis, which leads to joint deformities and functional loss [2], but the pathogenesis of RA remains unclear. Nevertheless, great progress has been made in RA treatment following the advent of targeted drugs, such as tumor necrosis factor (TNF)-α antagonists, interleukin (IL)-6 inhibitors, and JAK pathway inhibitors. Thus, studying the molecular mechanism of RA is beneficial for early diagnosis and prognosis improvement.

With the rapid development of genomics, transcriptional and sequencing technology in recent years, the application of gene co-expression networks is gradually expanding in biological research. These networks are widely used in high-throughput chip data, RNA sequencing data, DNA methylation data, and other data analysis. The most representative analysis is weighted gene co-expression network analysis (WGCNA) [3], which has been widely used and has provided meaningful findings in the gene analysis of many diseases [4]. However, literature review returned only two studies of applying WGCNA to RA. Their samples were small in size and came from the synovium and T cells, respectively [5, 6]. Since RA is a systemic disease, its lesions are not restricted to the synovium, and peripheral blood is easy to collect. Therefore, WGCNA was first used to analyze the gene expression profiles in peripheral blood, clarify the association between modules and RA as well as find potential biomarkers by exploring the genes in relevant modules.

Materials and methods

Data collection

The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database provides a large collection of microarray expression data [7]. The gene expression

profiles were searched using "rheumatoid arthritis" as the key word according to two criteria: 1) the dataset was generated from peripheral blood cells of patients with RA; 2) the expression profiles of peripheral blood cells of RA patients and healthy individuals were available in the dataset. Finally, GSE17755, GSE15573, GSE68689, and GSE100191 were selected. There were 112 patients with RA and 45 healthy individuals in the GSE17755 dataset, which had the largest sample size among the selected datasets.

Therefore, the GSE17755 series was selected as the research object, and the microarray data were obtained on the GPL1291 platform and submitted by Lee H and colleagues [8]. A total of 244 samples are included in this series, of which there were 112 RA and 45 healthy control samples for a total of 157 samples that were selected in this study for subsequent analysis. Ethics approval was not required because the expression profiles were downloaded from a public database and did not directly perform any experiments with patients or animals.

Data preprocessing

GSE17755 series matrix files were downloaded, and the probe names for platform GPL1291 were transformed into gene symbols based on the Hitachisoft AceGene Human Oligo Chip 30K Chip, Version 1. If multiple probes corresponded to the same gene symbol, the mean value was calculated using the aggregate function in R as the expression value of that particular gene. If the expression value of the probe was absent, the nearest neighbor average (KNN) algorithm in the impute package of R was used [9]. The obtained data were standardized.

Differentially expressed genes (DEGs) analysis

DEGs in RA samples compared to healthy controls were screened using the t test in the limma package of R [10]. Then, the *P* values were adjusted using the FDR method. Only genes with adjusted $P<0.05$ were recognized as significantly and differentially expressed genes.

Construction of weighted gene co-expression network

Analysis in the WGCNA package (version 1.63) of R was performed as described later [11]. The soft threshold of an adjacency matrix was selected to ensure closeness to the scale-free network (model index = 0.9), but the minimum threshold was chosen to make the curve smooth. Such settings allowed the network to contain sufficient information for module mining. The adjacency matrix was computed and transformed into a topological overlap matrix (TOM), through which a hierarchical clustering tree was generated. With the dynamic tree cut method, the minimum module size was set to 30, and several gene modules were determined in which module 0 preserved the genes outside all modules and was expressed in gray [12].

Detection of disease-related modules

Two methods were used to determine the association between each module and RA [3]. 1) The biological significance of each gene was evaluated using gene significance (GS). The differential expression between RA patients and controls was examined via the t test. Later, the correlation between the module and RA status was assessed using module significance (MS), which was defined as the average GS of all genes in a module. 2) The modular eigengene (ME) in each module was used as the characteristic expression of all genes in the module to correlate with the sample phenotypes, and the most disease-related module was identified.

GO functional enrichment of genes in the disease-related module

The genes in the disease-related module were submitted to DAVID (Database for Annotation, Visualization and Integrated Discovery) for functional enrichment analysis based on the Gene Ontology (GO) database [13]. Significant enrichment was set as $P<0.05$.

Identification of hub genes in the disease-related module

The collective weighted value was calculated among genes in selected modules and set the threshold at >0.6. The gene co-expression network was visualized on Cytoscape 3.6.0, and then hub genes in the network were identified on CytoHubba, which is a plugin of Cytoscape [14].

Results

Data preprocessing and DEGs screening

After data preprocessing, expression matrices of 13101 genes were obtained from the 157 samples. Overall, 3666 DEGs with adjusted

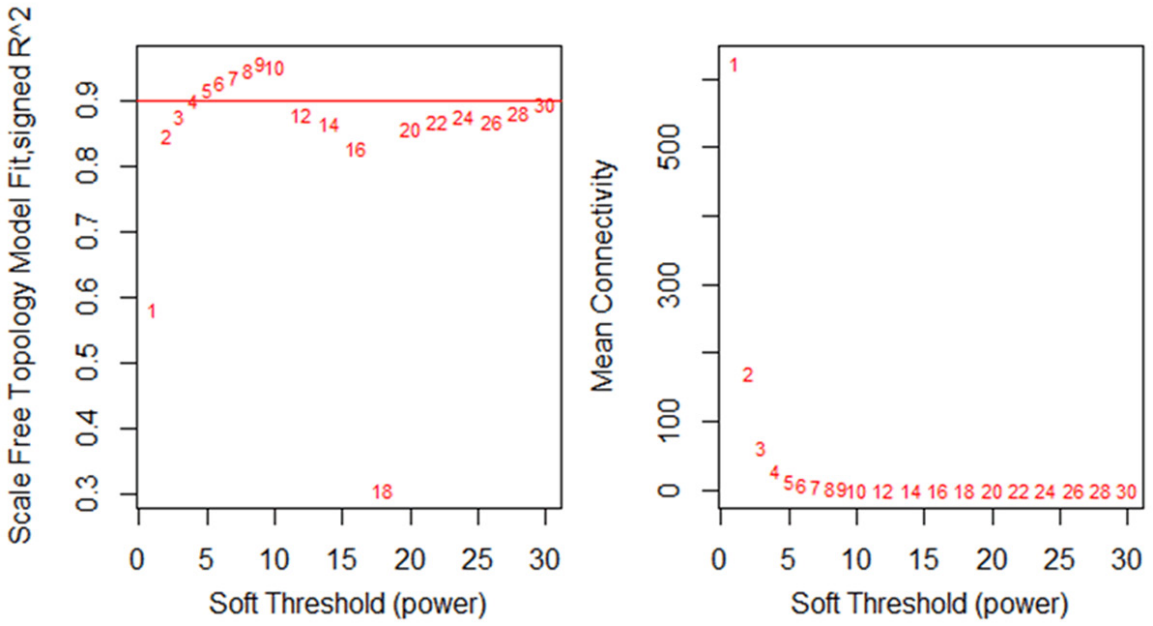

Figure 1. Threshold value analysis. The X-axis in the left figure represents the candidate thresholds and the Y-axis corresponds to the index of a scale-free network model; the right figure shows the mean connectivity of a network corresponding to the thresholds.

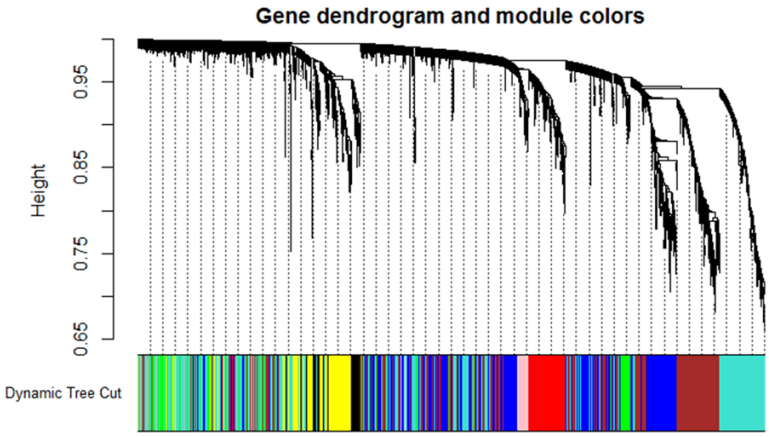

Figure 2. In the tree diagram, every leaf node represents a gene. Modules represented by different colors were identified using the dynamic tree cut algorithm. The distance between two genes was shown as height on the y-axis.

P<0.05 were screened out for subsequent analysis.

Detection of RA-related co-expression modules

Using the WGCNA package of R, a set of candidate soft thresholds were selected and returned to the detected model (**Figure 1**). Clearly, the appropriate soft threshold was five. In the hierarchical clustering tree, eight co-expression modules were identified with the dynamic tree cut method (**Figure 2**).

Two methods were used to test the relevance between each module and RA. First, the MS of each module was calculated, and larger MS suggested higher relevance (**Figure 3**). The results showed the turquoise module had the highest MS among all the selected modules. Second, the relevance between the ME of modules and phenotypes was calculated (**Figure 4**). Clearly, the turquoise module was still the most relevant (correlation coefficient = 0.92). The results of the two methods were consistent, and thus the turquoise module involving 1044 genes was identified as the most relevant module to RA.

GO functional enrichment of genes in the disease-related module

Genes in the turquoise module were tested via GO functional enrichment analysis (**Table 1**). It was found that the selected genes were significantly associated with inflammatory response, vasculogenesis, transforming growth factor

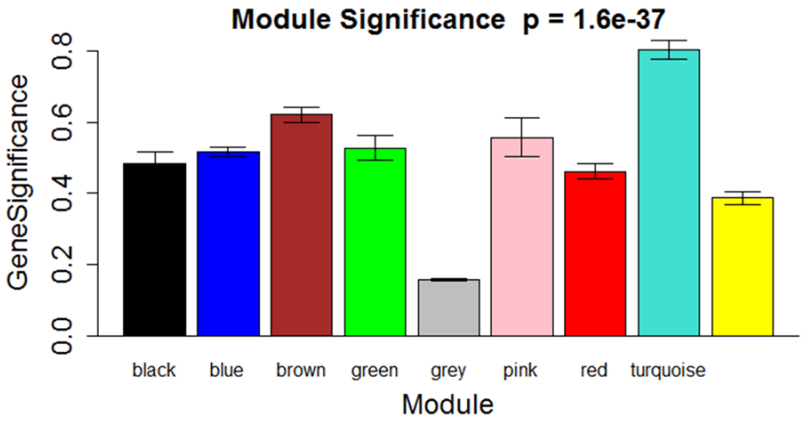

Figure 3. Module significance (MS) values of each module. Different colors indicate different modules.

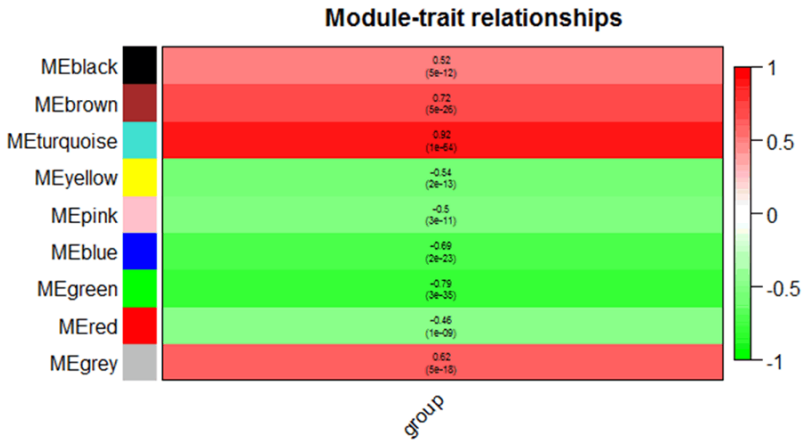

Figure 4. Relevance between each module and RA. Each row represents a module. In each module, the numbers above represent correlation coefficients, and the numbers in brackets below represent *P* values.

Table 1. GO functional enrichment analysis of the turquoise module

Term	Count	*P* Value
Inflammatory response	40	9.2E-06
Vasculogenesis	13	1.3E-05
Transforming growth factor beta receptor signaling pathway	16	3.66E-05
Cell proliferation	37	5.22E-05
Positive regulation of the protein kinase B signaling	14	2.05E-04
Type I interferon signaling pathway	12	2.44E-04

beta receptor signaling pathway, cell proliferation, positive regulation of protein kinase B signaling and type I interferon (IFN) signaling pathway in the category Biological Process (BP).

Identification of hub genes in the disease-related module

According to the collective weighted values among genes, the co-expression network in the turquoise module was visualized using Cytoscape software, and its hub genes were identified via cytoHubba. Among the 12 methods, MCC, which captures more essential proteins in the top of the ranked list for both high-degree and low-degree proteins, performs better than the others [14]. Therefore, the MCC method was used to identify 10 hub genes and generated a circular co-expression network (**Figure 5**). The hub genes include PIGL, MRPS10, PRKAA1, SLC22-A17, LEPRE1, C19orf61, ZNF646, ZNF84, C8orf17, and FOXM1 (**Table 2**).

Discussion

RA is a common autoimmune disease, which probably causes joint deformation, interstitial lung disease and vasculitis. RA also disturbs patients' normal lives and induces severe property losses. Currently, the pathogenesis of RA remains unclear, but the advent of targeted therapy highlights the great significance of investigating RA at the molecular level, as researchers have done for tumors.

Currently, bioinformatics analysis is becoming more common in basic research on rheumatoid arthritis. Most of these investigations are based on analysis of DEGs. WGCNA divides genes into multiple modules by analyzing the association between genes. Then, through the correlation analysis between these modules and sample phenotypes, the molecular characteristics of specific phenotypes could be found. WGCNA is obviously more scientific and reasonable than DEGs analysis. Before this study, there was no study that applied WGCNA to analyze the expression of genes from the RA

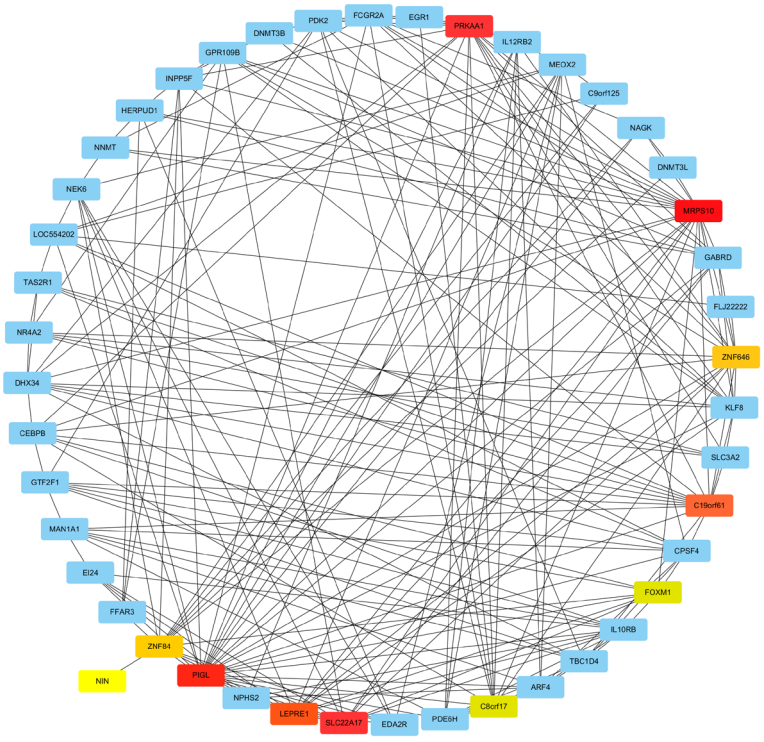

Figure 5. A circular gene co-expression network in the turquoise module. Ten hub genes were identified with the MCC method in cytoHubba. Different colors for the genes indicate different degrees of importance that decrease from red to orange to dark yellow.

Table 2. Ten hub genes and their MCC scores in the circular gene co-expression network

Name	Score
PIGL	170
MRPS10	141
PRKAA1	76
SLC22A17	68
LEPRE1	62
C19orf61	50
ZNF646	44
ZNF84	40
C8orf17	38
FOXM1	38

peripheral blood gene chip. Moreover, the selected series contains a large number of samples, which ensures the results are accurate and stable.

GSE17755 is currently the largest gene expression series in the GEO database that meets our study requirements. After screening 3666 DEGs, we utilized WGCNA to identify 8 modules, and the turquoise module involving 1044 genes was found to be the most relevant to RA.

GO enrichment analysis showed the selected genes were mainly enriched in GO terms related to the inflammatory response, vasculogenesis, and the type I IFN signaling pathway. First, the 40 genes enriched in the inflammatory response were partial chemokines and their receptors, such as CCL5, CCL7, CCL13, CCL17, CCL20, and CCR7. Chemokines, which induce migration of inflammatory cells, were abundantly expressed in the synovial tissues of RA, which suggests chemokines are promising targets for RA therapy [15]. Second, the type I IFN signaling pathway includes 12 genes, such as JAK1 and HLA-E. In recent years, JAK1 has become a focus for researchers in the field of rheumatoid arthritis. Tofacitinib, which is a JAK inhibitor, was approved as a novel drug for RA treatment. Type I IFN signaling, which can interact with Toll-like receptor and TNF-α, was reportedly involved in the pathogenesis of RA [16, 17].

Of the top 10 hub genes identified in the disease-related module, PIGL is a protein-coding gene and its related GO annotations include N-acetylglucosaminylphosphatidylinositol deacetylase activity. PIGL mutations are associated with CHIME syndrome, which is characterized by colobomas, heart defects, ichthyosiform dermatosis, mental retardation, and ear anomalies. Although there is no report concerning PIGL in RA, PIGL mutations can impair glycosylphosphatidylinositol (GPI) biosynthesis, which is associated with RA [18]. A study showed that soluble GPI was released from activated neutrophils and was present at high concentrations in synovial fluids but not the sera of RA patients [19]. Therefore, PIGL may affect RA by regulating GPI biosynthesis.

PRKAA1 encodes a catalytic alpha subunit of AMP-activated protein kinase (AMPK). As re-

ported, mice deficient in PRKAA1 mildly elicited an increase of clinical arthritis versus wild type controls [20]. Recent studies showed that activation of AMPK can limit JAK-STAT-dependent signaling pathways for RA treatment [21, 22]. Methotrexate protects the vascular endothelium of RA patients from inflammatory injury via activation of AMPK-CREB signaling [23].

MRPS10 encodes a subtype of mammalian mitochondrial ribosomal proteins that facilitate protein synthesis in the mitochondria. LEPRE1 is a gene that is highly associated with a rare genetic disease called osteogenesis imperfecta, and it may play a role in the biology of cellular senescence [24, 25]. FOXM1 encodes a protein that is a transcriptional activator in cell proliferation, and this hot gene has been exploited as a biomarker for diagnosis, prognosis and treatment of many different tumors [26, 27]. Although the roles of these genes in RA have not been reported, predictions for the aforementioned genes were made using the same methods from previous studies, and it was reasonable to believe these genes were probably involved in RA pathogenesis. Therefore, these newly identified genes reflect the innovative value of our research, and this study explores new directions for future basic research of RA.

Nevertheless, this study is still in the bioinformatics analysis stage, and the findings should be verified by several appropriate experiments. In addition, the samples from the GEO database provide very few clinical data, and there is especially a lack of data concerning epidemiology, disease activity and treatment, which prevented us from analyzing the relationships between the selected modules and additional clinical data for RA.

Conclusions

A total of 8 modules was detected using GSE17755 data and this study identified the most relevant one for RA. Genes in the most relevant module were related to the inflammatory response and type I IFN signaling pathway according to GO Biological Process. Ten hub genes, including PIGL, PRKAA1, and MRPS10, were identified in this module, which presumably play critical roles in RA pathogenesis.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Xiaofeng Zeng, Department of Rheumatology, Peking Union Medical College Hospital, No. 1 Shuaifuyuan, Dongcheng District, Beijing 100730, China. E-mail: zengxfpumc@foxmail.com

References

[1] Abhishek A, Doherty M, Kuo CF, Mallen CD, Zhang W and Grainge MJ. Rheumatoid arthritis is getting less frequent-results of a nationwide population-based cohort study. Rheumatology (Oxford) 2017; 56: 736-744.

[2] Gay S, Gay RE and Koopman WJ. Molecular and cellular mechanisms of joint destruction in rheumatoid arthritis: two cellular mechanisms explain joint destruction? Ann Rheum Dis 1993; 52: S39-47.

[3] Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.

[4] Guo S, Wang J, Li J, Xu F, Wei Q, Wang H, Huang H, Zheng S, Xie Y and Zhang C. Identification of gene expression profiles and key genes in subchondral bone of osteoarthritis using weighted gene co-expression network analysis. J Cell Biochem 2018; 119: 7687-7695.

[5] Ma C, Lv Q, Teng S, Yu Y, Niu K and Yi C. Identifying key genes in rheumatoid arthritis by weighted gene co-expression network analysis. Int J Rheum Dis 2017; 20: 971-979.

[6] Sumitomo S, Nagafuchi Y, Tsuchida Y, Tsuchiya H, Ota M, Ishigaki K, Nakachi S, Kato R, Sakurai K, Hanata N, Tateishi S, Kanda H, Suzuki A, Kochi Y, Fujio K and Yamamoto K. A gene module associated with dysregulated TCR signaling pathways in CD4+ T cell subsets in rheumatoid arthritis. J Autoimmun 2018; 89: 21-29.

[7] Edgar R, Domrachev M and Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002; 30: 207-210.

[8] Lee HM, Sugino H, Aoki C and Nishimoto N. Underexpression of mitochondrial-DNA encoded ATP synthesis-related genes and DNA repair genes in systemic lupus erythematosus. Arthritis Res Ther 2011; 13: R63.

[9] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. AM STAT 1992; 46: 175-185.

[10] Smyth GK. limma: Linear models for microarray data. Bioinformatics & Computational Biology Solutions Using R & Bioconductor 2005; 397-420.

[11] Horvath S and Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 2008; 4: e1000117.

[12] Langfelder P, Zhang B and Horvath S. Defining clusters from a hierarchical cluster tree: the

dynamic tree cut package for R. Bioinformatics 2008; 24: 719-720.
[13] Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC and Lempicki RA. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007; 8: R183.
[14] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT and Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014; 8 Suppl 4: S11.
[15] Nanki T. Treatment for rheumatoid arthritis by chemokine blockade. Nihon Rinsho Meneki Gakkai Kaishi 2016; 39: 172-80.
[16] López de Padilla CM, Niewold TB. The type I interferons: basic concepts and clinical relevance in immune-mediated inflammatory diseases. Gene 2016; 576: 14-21.
[17] Smiljanovic B, Grun JR, Biesen R, Schulte-Wrede U, Baumgrass R, Stuhlmuller B, Maslinski W, Hiepe F, Burmester GR, Radbruch A, Haupl T and Grutzkau A. The multifaceted balance of TNF-alpha and type I/II interferon responses in SLE and RA: how monocytes manage the impact of cytokines. J Mol Med (Berl) 2012; 90: 1295-1309.
[18] Brett SJ, Baxter G, Cooper H, Rowan W, Regan T, Tite J and Rapson N. Emergence of CD52-, glycosylphosphatidylinositol-anchor-deficient lymphocytes in rheumatoid arthritis patients following Campath-1H treatment. Int Immunol 1996; 8: 325-334.
[19] Huang J, Takeda Y, Watanabe T and Sendo F. A sandwich ELISA for detection of soluble GPI-80, a glycosylphosphatidyl-inositol (GPI)-anchored protein on human leukocytes involved in regulation of neutrophil adherence and migration-its release from activated neutrophils and presence in synovial fluid of rheumatoid arthritis patients. Microbiol Immunol 2001; 45: 467-471.
[20] Guma M, Wang Y, Viollet B and Liu-Bryan R. AMPK activation by A-769662 controls il-6 expression in inflammatory arthritis. PLoS One 2015; 10: e140452.
[21] Rutherford C, Speirs C, Williams JJ, Ewart MA, Mancini SJ, Hawley SA, Delles C, Viollet B, Costa-Pereira AP, Baillie GS, Salt IP and Palmer TM. Phosphorylation of Janus kinase 1 (JAK1) by AMP-activated protein kinase (AMPK) links energy sensing to anti-inflammatory signaling. Sci Signal 2016; 9: a109.
[22] Speirs C, Williams JJL, Riches K, Salt IP and Palmer TM. Linking energy sensing to suppression of JAK-STAT signalling: a potential route for repurposing AMPK activators? Pharmacol Res 2018; 128: 88-100.
[23] Thornton CC, Al-Rashed F, Calay D, Birdsey GM, Bauer A, Mylroie H, Morley BJ, Randi AM, Haskard DO, Boyle JJ and Mason JC. Methotrexate-mediated activation of an AMPK-CREB-dependent pathway: a novel mechanism for vascular protection in chronic systemic inflammation. Ann Rheum Dis 2016; 75: 439-448.
[24] Cabral WA, Chang W, Barnes AM, Weis M, Scott MA, Leikin S, Makareeva E, Kuznetsova NV, Rosenbaum KN, Tifft CJ, Bulas DI, Kozma C, Smith PA, Eyre DR and Marini JC. Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta. Nat Genet 2007; 39: 359-365.
[25] Succoio M, Comegna M, D'Ambrosio C, Scaloni A, Cimino F and Faraonio R. Proteomic analysis reveals novel common genes modulated in both replicative and stress-induced senescence. J Proteomics 2015; 128: 18-29.
[26] Teh MT, Hutchison IL, Costea DE, Neppelberg E, Liavaag PG, Purdie K, Harwood C, Wan H, Odell EW, Hackshaw A and Waseem A. Exploiting FOXM1-orchestrated molecular network for early squamous cell carcinoma diagnosis and prognosis. Int J Cancer 2013; 132: 2095-2106.
[27] Meng FD, Wei JC, Qu K, Wang ZX, Wu QF, Tai MH, Liu HC, Zhang RY and Liu C. FoxM1 overexpression promotes epithelial-mesenchymal transition and metastasis of hepatocellular carcinoma. World J Gastroenterol 2015; 21: 196-213.