Construction of gene/protein interaction networks for primary myelofibrosis and KEGG pathway-enrichment analysis of molecular compounds

C.G. Sun[1]*, X.J. Cao[2]*, C. Zhou[1], L.J. Liu[1], F.B. Feng[1], R.J. Liu[1], J. Zhuang[1] and Y.J. Li[2]

[1]Department of Cancer Center,
Weifang Traditional Chinese Medicine Hospital, Weifang, China
[2]Department of Clinical Institute,
Shandong University of Traditional Chinese Medicine, Jinan, Shandong, China

*These authors contributed equally to this study.
Corresponding author: C.G. Sun
E-mail: zhongliuyike@163.com

Genet. Mol. Res. 14 (4): 16126-16132 (2015)
Received June 30, 2015
Accepted September 25, 2015
Published December 8, 2015
DOI http://dx.doi.org/10.4238/2015.December.8.1

ABSTRACT. The objective of this study was the development of a gene/protein interaction network for primary myelofibrosis based on gene expression, and the enrichment analysis of KEGG pathways underlying the molecular complexes in this network. To achieve this, genes involved in primary myelofibrosis were selected from the OMIM database. A gene/protein interaction network for primary myelofibrosis was obtained through Cytoscape with the literature mining performed using the Agilent Literature Search plugin. The molecular complexes in the network were detected by ClusterViz plugin and KEGG pathway enrichment of molecular complexes was performed using DAVID online. We found 75 genes associated with primary myelofibrosis in the OMIM database. The gene/protein interaction network of primary myelofibrosis contained 608 nodes, 2086 edges, and

4 molecular complexes with a correlation integral value greater than 4. Molecular complexes involved in KEGG pathways are related to cytokine regulation, immune function regulation, ECM-receptor interaction, focal adhesion, actin cytoskeleton regulation, cell adhesion molecules, and other biological behavior of tumors, which can provide a reliable direction for the treatment of primary myelofibrosis and the bioinformatic foundation for further understanding the molecular mechanisms of this disease.

Key words*:* Gene/protein interaction networks; Primary myelofibrosis; Molecular complexes; KEGG pathway

INTRODUCTION

Primary myelofibrosis (PMF) is a disease that originates from an abnormal clone of steam cells that leads to the bone marrow developing a hyperplasia of fibrous tissue (Rajasekaran et al., 2015). The major symptoms of PMF are the swelling of the liver and spleen, due to extramedullary hematopoiesis, abnormal blood count, constitutional symptoms, cachexia, etc., with the disease having the potential to develop into leukemia. The pathogenesis of PMF is still unclear, but the development of bioinformatics, especially protein-protein interaction networks and the growing emergence of analysis methods, provides new tools for understanding the molecular mechanisms of PMF.

In this study, we screened the confirmed disease-related genes using the Online Mendelian Inheritance in Man (OMIM) database, created gene/protein interaction networks based on biological function using the Cytoscape software, detected molecule complexes that may be included in the network, and predicted relevant KEGG pathways that may be involved in pathogenesis. We provide a basis for further explaining the mechanism of PMF development.

MATERIAL AND METHODS

Data acquisition

OMIM is a comprehensive, authoritative, daily-updated human phenotype database, containing more than 12,000 genes of all human genetic diseases, mainly focusing on hereditary diseases. Text notes, related reference information, sequence records, maps, and related databases are available for each gene. On February 1st, 2015, after inputting “primary myelofibrosis” in the OMIM database home page (Hamosh et al., 2005; Amberger et al., 2009), we acquired the gene information related to PMF by a series of screenings.

Construction of gene/protein interaction networks

The genes associated with PMF were submitted to the Agilent Literature Search plugin (Vailaya et al., 2005) (version 2.7.7; Agilent Technologies, USA) of the Cytoscape software (version 2.8.2; Shannon et al., 2003) and the PubMed literature related to the submitted genes was mined (Vailaya et al., 2005). False positive interaction information was removed from the results. Then, gene/protein interaction networks were read and visualized in Cytoscape (version 2.8.2; Shannon et al., 2003).

Network analysis

ClusterViz (version 1.2), a plugin of Cytoscape (version 2.8.2), was used to make the correlation analysis for the area of the construction of biological networks, through the MCODE algorithm (Li et al., 2008). By analyzing the network structure, proteins were clustered and shown in Cytoscape according to their correlation integral value. Clusters with an integral value greater than 4 were regarded as molecular complexes. The gene/protein names in the molecular compounds were submitted to the Database for Annotation, Visualization and Integrated Discovery (DAVID; Dennis et al., 2003; Huang et al., 2009). Using the Kyoko Encyclopedia of Genes and Genomes (KEGG) database, biological pathways involved in PMF heredity were identified.

Main outcome measurements

Protein networks were constructed based on PMF-related genes, nodes (proteins) and edges (interactions between proteins), molecular complexes in the network and its associated interaction points and nodes (protein) and the edge (interaction between proteins), analyze the biological pathways involved in the molecular complexes.

RESULTS

PMF-related genes in OMIM

Through the OMIM database, seventy-five genes associated with PMF were identifed: ASXL1, BMP1, BMP6, BMP7, CALR, CD14, CD177, CD34, CEACAM6, CEACAM8, CRP, CXCL12, CXCL8, CXCR1, CXCR2, CXCR4, DNMT3A, ENG, ERCC2, ETV6, EZH2, FKBP5, GATA1, HDAC11, HDAC9, HMGA2, IDH1, IDH2, IL15, IL17A, JAK2, KIT, KRT7, LDHA, LDHB, LOX, LOXL1, LOXL2, LTBP1, LYN, MCAM, MMP13, MMP14, MMP8, MPL, MTHFR, MTOR, NFE2, NOG, NPM1, NR3C1, PHF20, PML, PRG2, PTH, RARB, RUNX1, SH2B3, SLCO2A1, SOCS3, SRSF2, STAT3, STAT5A, TAC1, TEK, TET2, TGFB1, THBS1, TNF, TNFRSF11B, TP53, TPO, U2AF1, VEGFA, and ZEPM1.

Gene/protein interaction networks

The text mining of the seventy-five disease-related genes resulted in a network diagram with 608 nodes (gene/proteins) and 2086 edges (interactions). In Figure 1, the red concave quadrilaterals represent the OMIM disease-related genes, while the green octagons represent the genes/proteins obtained from the mining.

Detection of molecular complexes

Cluster analysis using the MCODE algorithm resulted in four molecular complexes with correlation integral values greater than 4 (Figure 2).

KEGG pathway enrichment of the molecular complexes

The five protein molecular complexes were submitted to the KEGG database, which allowed the identification of the KEGG pathways associated with it (Table 1).

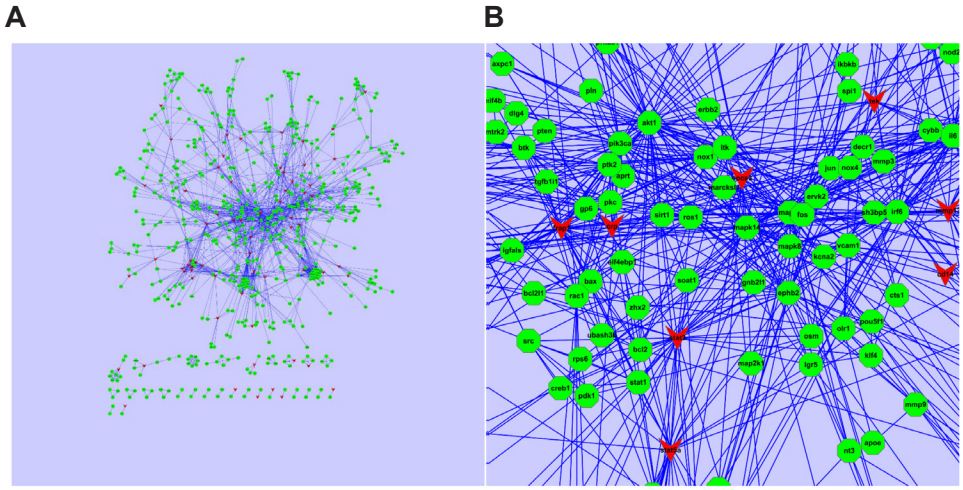

Figure 1. Network map of primary myelofibrosis gene/protein interactions. **A.** Overall view; **B.** Zoomed view.

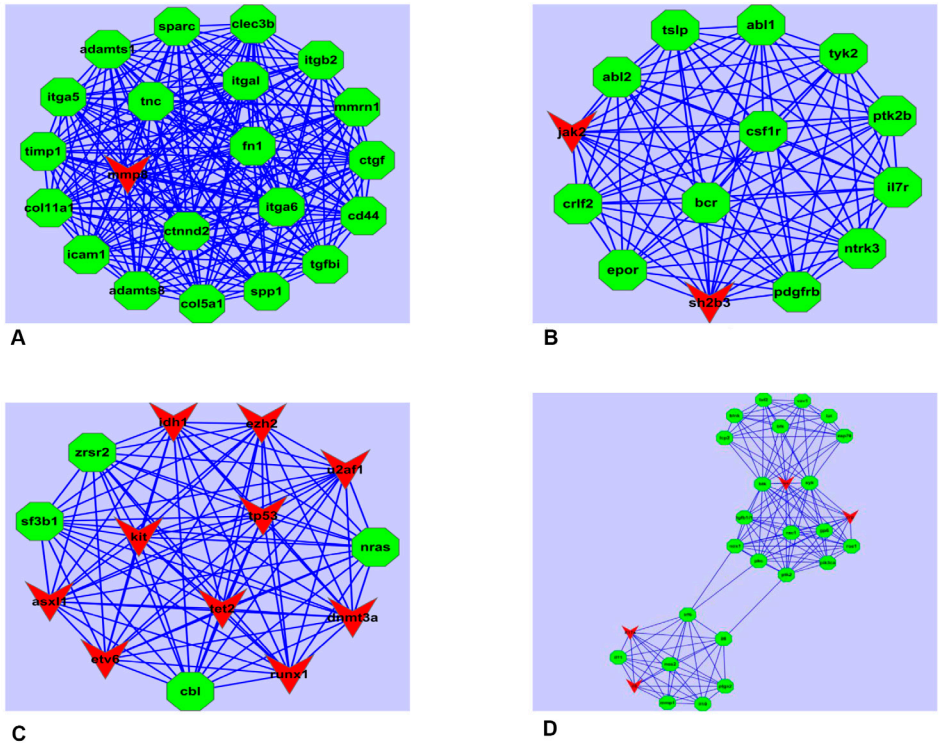

Figure 2. Molecular complexes obtained through the analysis using the MCODE algorithm. **A.** Complex 1 (relation score = 10, 21 nodes and 210 edges); **B.** complex 2 (relation score = 6.5, 14 nodes and 91 edges); **C.** complex 3 (relation score = 6.5, 14 nodes and 91 edges); **D.** complex 4 (relation score = 5.25, 28 nodes and 147 edges).

Table 1. Enrichment of KEGG pathways related to the molecular complexes.

Complex	KEGG pathway	P	Genes	FDR
Complex 1	hsa04512:ECM-receptor interaction	3,01E-11	CD44, ITGA6, ITGA5, TNC, COL11A1, COL5A1, SPP1, FN1	2,04E-08
	hsa04510:Focal adhesion	6,52E-07	ITGA6, ITGA5, TNC, COL11A1, COL5A1, SPP1, FN1	4,43E-04
	hsa04810:Regulation of actin cytoskeleton	5,34E-04	ITGAL, ITGA6, ITGA5, ITGB2, FN1	3,62E-01
	hsa04514:Cell adhesion molecules (CAMs)	1,80E-03	ICAM1, ITGAL, ITGA6, ITGB2	1,21E+00
	hsa05416:Viral myocarditis	8,05E-03	ICAM1, ITGAL, ITGB2	5,34E+00
Complex 2	hsa04630:Jak-STAT signaling pathway	2,61E-05	TYK2, TSLP, CRLF2, EPOR, JAK2, IL7R	2,16E-02
	hsa04060:Cytokine-cytokine receptor interaction	3,20E-04	TSLP, CRLF2, PDGFRB, EPOR, IL7R, CSF1R	2,65E-01
Complex 3	hsa05220:Chronic myeloid leukemia	2,43E-04	NRAS, CBL, TP53, RUNX1	2,33E-01
	hsa05200:Pathways in cancer	1,65E-03	NRAS, CBL, TP53, KIT, RUNX1	1,57E+00
	hsa05221:Acute myeloid leukemia	4,37E-03	NRAS, KIT, RUNX1	4,11E+00
Complex 4	hsa04664:Fc epsilon RI signaling pathway	1,00E-08	LAT, LYN, RAC1, PIK3CA, VAV1, LCP2, BTK, SYK	1,01E-05
	hsa04662:B cell receptor signaling pathway	2,77E-07	LYN, RAC1, PIK3CA, VAV1, BLNK, BTK, SYK	2,79E-04
	hsa04650:Natural killer cell mediated cytotoxicity	8,20E-06	LAT, RAC1, ZAP70, PIK3CA, VAV1, LCP2, SYK	8,28E-03
	hsa04666:Fc gamma R-mediated phagocytosis	2,54E-05	LAT, LYN, RAC1, PIK3CA, VAV1, SYK	2,57E-02
	hsa04660:T cell receptor signaling pathway	4,74E-05	LAT, ZAP70, PIK3CA, VAV1, IL10, LCP2	4,78E-02
	hsa05200:Pathways in cancer	1,62E-04	PTK2, IL6, PTGS2, IL8, RAC1, PIK3CA, NOS2, MMP1	1,64E-01
	hsa04062:Chemokine signaling pathway	6,30E-04	PTK2, IL8, LYN, RAC1, PIK3CA, VAV1	6,34E-01

P < 0.01. FDR = false-discovery rate.

DISCUSSION

There are three main categories to construct biomolecular regulatory networks: gene regulatory networks through a mathematical model; networks through literature mining, and integrating multiple data (Friedman et al., 2000; Hwang et al., 2005; Mohamed-Hussein and Harun, 2009). Building a network through literature mining means using bioinformatics, computational biology, and other tools of computer science to analyze the data in the literature, and build biomolecular regulatory networks using the relationships between gene/protein interactions of the existing literature. The advantage of this method is the establishment of an accurate and stable relationship between regulation and the network. Therefore, in this study, we use the OMIM database and literature mining methods to construct the protein interaction networks associated with PMF. Although the data of this experiment are genes, we need to elaborate and prove their functionality and relationship with other molecules at the protein level using the literature. For this reason, the constructed network should be a protein interaction network.

PMF belongs to the classical bone marrow hyperplastic diseases in which the Philadelphia chromosome abnormality is not present. Due to the limited knowledge on this disease, its pathogenesis remains controversial. In recent years, with the study of PMF genetic features through cell and molecular biology, we have gained a new understanding of the pathogenesis of this disease. With the application of a new generation of sequencing technology, it was found that patients with PMF have a series of genetic mutations, such as JAK2, ASXL1, EZH2, IDH1, IDH2, SRSF2, CALR, etc., which are closely related to the prognosis (Tefferi et al., 2009; Tefferi et al., 2010; Abdel-Wahab et al., 2011; Guglielmelli et al., 2011; Nangalia et al., 2013). Using the seventy-five genes provided by the OMIM database, we have built a protein interaction network for PMF containing 608 nodes (gene/proteins) and 2086 edges (interaction). The network covers the genes identified by the other studies to be involved in PMF, which indicates that the network can be used to describe gene/protein interactions in the development process of PMF.

Due to the size of the network, we used the MCODE clustering algorithm to evaluate the network's regional integration through the correlation integral. The correlation integral describes the degree of association of proteins within the region. Proteins of the same molecular complex

generally have the same biological function; therefore, we can discover unknown gene functions or new molecular functional groups (Bader and Hogue, 2003). Our results show four molecular complexes with a correlation integral greater than four. DAVID is not only an extensive database possessing gene annotations for different species but is also enriched to include the biological information of a single gene. There are several biological pathways involving these four molecular complexes; therefore, this study focused on the enrichment analysis of only the KEGG pathways, resulting in 17 KEGG biological pathways related to these molecular complexes, whose $P < 0.01$.

The KEGG database connects genome and function information. KEGG pathways bring together the molecular interactions and the reaction networks through an artificial pathway diagram. DAVID is a functional annotation system based on network access that, by taking advantage of DAVID analysis tool KEGG pathway, we can directly observe gene enrichment pathways, the protein that targets the corresponding genes, and their mutual relations. When analyzing a disease-related gene through its ID, we can focus on the important pathway determined according to the P ($P < 0.01$), which contributes to the understanding of the molecular pathology of the disease and provides effective targets for the research and development of new drugs.

According to the KEGG pathway enrichment, molecular complex 2 was predicted to be related to the cytokine-cytokine receptor interaction and JAK-STAT signaling pathway. Existing literature indicates that various cytokines, especially those discharged from the megakaryocyte progenitor cell during maturation arrest, increase the formation of fibroblasts, increase the synthesis, and decrease the split of collagen in the bone marrow (Schmitt et al., 2000). Such an imbalance between formation and degradation of collagen leads to its excess accumulation in the marrow stromal cells, forming the bone marrow fibrosis. The JAK2 gene mutation and the study of the JAK-STAT signal pathway have advanced studies of the pathogenesis of PMF, providing an excellent molecular marker for diagnosis and therapeutic targets. The KEGG pathways in which the molecular complex 4 is involved include Fc epsilon RI signaling pathway, B cell receptor signaling pathway, natural killer cell mediated cytotoxicity, Fc gamma R-mediated phagocytosis, T cell receptor signaling pathway, and chemokine signaling pathway, which are closely related to the immune function. Some studies have found that the PMF patients can have all sorts of immune function defects, such as circulating immune complex, lupus anticoagulant, antinuclear antibodies, rheumatoid factor, positive Coombs' test, complement activation, matrix protein antibody, and immune globulin (in which the main one is IgG), whose roles in the pathogenesis of PMF are still not clear (Vardiman et al). The pathogenesis of PMF can perhaps be studied from the point of view of these signaling pathways related to the immune system. The KEGG pathways in which the molecular complex 1 is involved include ECM-receptor interaction, focal adhesion, regulation of actin cytoskeleton, cell adhesion molecules (CAMs), while the molecular complex 3 shows extensive correlation with other cancers. These signal pathways and genes can provide a reliable direction for the study of the molecular mechanisms of PMF and treatments for it, which require further study and validation.

The molecular mechanisms of PMF development need more in-depth research. The construction of gene/protein interaction networks for PMF and KEGG pathway enrichment analysis can provide a reliable direction for clinical research, providing a variety treatment targets in the form of signal pathways that can inhibit the occurrence and development of PMF.

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES

Abdel-Wahab O, Pardanani A, Rampal R, Lasho TL, et al. (2011). DNMT3A mutational analysis in primary myelofibrosis, chronic myelomonocytic leukemia and advanced phases of myeloproliferative neoplasms. *Leukemia* 25: 1219-1220.

Amberger J, Bocchini CA, Scott AF, Hamosh A, et al. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* 37 (Database issue): D793-D796.

Bader GD and Hogue CW (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* 4: 2.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4: P3.

Friedman N, Linial M, Nachman I and Pe'er D (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7: 601-620.

Guglielmelli P, Biamonte F, Score J, Hidalgo-Curtis C, et al. (2011). EZH2 mutational status predicts poor survival in myelofibrosis. *Blood* 118: 5227-5234.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, et al. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33 (Database issue): D514-D517.

Huang DW, Sherman BT and Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4: 44-57.

Hwang D, Smith JJ, Leslie DM, Weston AD, et al. (2005). A data integration methodology for systems biology: experimental verification. *Proc. Natl. Acad. Sci. U. S. A.* 102: 17302-17307.

Li M, Wang J and Chen J (2008). A fast agglomerate algorithm for mining functional modules in protein interaction networks. Proceedings of 1st International Conference on Biomedical Engineering and Informatics (BMEI). Washington, 3-7.

Mohamed-Hussein ZA and Harun S (2009). Construction of a polycystic ovarian syndrome (PCOS) pathway based on the interactions of PCOS-related proteins retrieved from bibliomic data. *Theor. Biol. Med. Model.* 6: 18.

Nangalia J, Massie CE, Baxter EJ, Nice FL, et al. (2013). Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* 369: 2391-2405.

Rajasekaran A, Ngo TT and Abdelrahim M (2015). Primary myelofibrosis associated glomerulopathy: significant improvement after therapy with ruxolitinib. *BMC Nephrol.* 16: 121.

Schmitt A, Jouault H, Guichard J, Wendling F, et al. (2000). Pathologic interaction between megakaryocytes and polymorphonuclear leukocytes in myelofibrosis. *Blood* 96: 1342-1347.

Shannon P, Markiel A, Ozier O, Baliga NS, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13: 498-2504.

Tefferi A, Pardanani A, Lim KH, Abdel-Wahab O, et al. (2009). TET2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. *Leukemia* 23: 905-911.

Tefferi A, Lasho TL, Abdel-Wahab O, Guglielmelli P, et al. (2010). IDHl and IDH2 mutation studies in 1473 patients with chronic-, fibrotic- or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis. *Leukemia* 24: 1302-1309.

Vailaya A, Bluvas P, Kincaid R, Kuchinsky A, et al. (2005). An architecture for biological information extraction and representation. *Bioinformatics* 21: 430-438.

Vardiman JW, Thiele J, Arber DA, Brunning RD, et al. (2009). The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* 114: 937-951.