Transcriptome Research of Key Modules and Hub Genes Associated with Primary Myelofibrosis by Weighted Co-expression Network Analysis.

Weihang Li

Department of Orthopaedics, Xijing Hospital, The Fourth Military Medical University. Xi'an, China

Bin Yuan

Department of Orthopaedics, Xijing Hospital, The Fourth Military Medical University, Xi'an, China.

Shilei Zhang

Department of Orthopaedics, Xijing Hospital, The Fourth Military Medical University, Xi'an, China.

Ziyi Ding

Department of Orthopaedics, Xijing Hospital, The Fourth Military Medical University, Xi'an, China.

Yingjing Zhao

College of Clinical Medicine, Jilin University.

Sheng Zhong

Department of Bioinformatics, Harvard Medical school, Boston, MA 02115, USA.

Guangxun Gao

Department of Hematology, Xijing Hospital, The Fourth Military Medical University, Xi'an, China.

Ming Yan (✉ yanming_spine@163.com)

Xijing Hospital, the Fourth Military Medical University.

Research article

Keywords: Bioinformatics, WGCNA, Myeloproliferative Neoplasm, Primary Myelofibrosis, Prognostic markers

Posted Date: July 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-41776/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: This study aimed to identify novel targets of diagnosis, therapy as well as prognosis for primary myelofibrosis (PMF).

Methods: The gene expression profiles of GSE26049 was obtained from GEO dataset, weighted gene co-expression network analysis (WGCNA) was then performed to identify the most related modules with PMF. Subsequently, GO (Gene Ontology), KEGG (Kyoto Encyclopedia Genes and Genomes), GSEA (Gene Set Enrichment Analysis) and PPI (Protein-Protein Interaction) network were conducted to fully understand the detailed information of the green module.

Results: Green module was strongly correlated with PMF disease after WGCNA analysis. 20 genes in green module were identified as hub genes responsible for the progression of PMF. Functional annotation and pathway analysis revealed that these hub genes were primarily enriched in erythrocyte differentiation, transcription factor binding, hemoglobin complex, transcription factor complex and cell cycle et al. Of which, EPB42, CALR, SLC4A1 and MPL had the most correlations with PMF.

Conclusions: This study elucidated that genes EPB42, CALR, SLC4A1 and MPL were significantly more highly expressed in PMF samples than in normal samples. These four genes may be considered candidate prognostic biomarkers and potential therapeutic targets for early stage of PMF. Meanwhile, EPB42 and SLC4A1 were firstly found to be highly correlated with the progression of PMF.

Background

The World Health Organization (WHO) revised and updated the hematopoietic tumor classification system in 2016, which confirmed seven major categories of myeloid malignancies including acute myeloid leukemia (AML) and related neoplasms, myeloproliferative neoplasms (MPN), myelodysplastic syndromes (MDS), mastocytosis, eosinophilia-associated myeloid/lymphoid neoplasms with specific mutations, MDS/MPN overlap and myeloid neoplasms with germline predisposition[1]. Of which MPNs are clonal hematopoietic disorders characterized by excessive production of differentiated hematopoietic cells in chronic phase[2]. The Philadelphia-negative MPNs contained 3 major diseases including polycythemia vera (PV), essential thrombocythemia (ET) and primary myelofibrosis (PMF). Among MPNs, PMF is the most essential neoplasms which could evolve from the other MPNs such as PV and ET. Later in their courses, both PV and ET disorders may evolve into myelofibrotic phases termed “post polycythemia vera myelofibrosis” or “post-essential thrombocythemia myelofibrosis”, respectively[3, 4]. In this situation, PMF is referred to as post-PV MF or post-ET MF.

Primary myelofibrosis (PMF) is a nascent, myeloproliferative, neoplastic disorder characterized by clonal proliferation of myeloid cells in the bone marrow resulting in fibrosis and leading to the devastation of healthy marrow[3, 5, 6]. Clinical manifestations often include marked splenomegaly, anemia as well as constitutional symptoms such as fatigue, fever, weight loss and night sweats. Patients with severe symptoms of PMF may show upper abdomen flatulence feeling, bone pain, bleeding and cachexia et

al[7–9]. Besides, the overall treatment and prognosis of PMF is generally poor. Epidemiology findings showed that the estimated prevalence of PMF is between 4 to 6 per 100000 people, with a median survival time of 15 years for patients younger than 60 and 6 years for patients older than 60, respectively[10, 11]. Meanwhile, chemotherapy, including JAK2 inhibitors does not provide a promising prospective view[12, 13], which may result from inadequate exploration of genes responsible for the occurrence and development of PMF. Therefore, there is an urgent need to discover novel chemotherapy as well as to explore original biomarkers to assess their malignant potential and prognosis for a better understanding in genetic and molecular biology significance behind the PMF.

Recent decades, bioinformatics combined with microarray technology have displayed prospective view for analyzing molecular and genetic mechanisms of malignant neoplasms[14], which prompted study on the initiation, development and metastasis of tumor at the molecular level, making it possible to analyze the genetic alteration and molecular mechanisms in the development of neoplasm. Weighted Gene Co-expression Network Analysis (WGCNA) is an effective and systematic biological analysis method among bioinformatics, which is used to perform weighted correlation network analysis for exploring the correlation between genes and a given feature[15, 16]. The mathematical principle of WGCNA was firstly proposed in 2005[17], and the algorithm was implemented in R environment in 2008, which was named as WGCNA package[16]. Recent 5 years, WGCNA has been witnessed for extensive application in bioinformatics to identify the candidate biomarkers through a variety of tumors such as uveal melanoma, breast cancer and adrenocortical carcinoma et al[18–22]. With the assistance of WGCNA, researchers could explore potential mechanisms among highly co-related genes and discover novel diagnostic biomarkers or therapeutic targets from gene cluster associated with disease. To the best knowledge, the application of WGCNA in the identification of biomarkers of PMF has not been reported up to now.

In the current study, microarray dataset GSE26049 has been analyzed to construct a weighted co-expression network as well as to find a module with specific interest to the development of PMF to identify potential therapeutic targets and key signaling pathways. Subsequently, functional enrichment analysis including GO (Gene Ontology) and KEGG (Kyoto encyclopedia genes and genomes) were carried out to discover molecular function and aberrant regulated pathways through displaying their cellular component, biological process, and molecular functions. Then protein-protein interaction (PPI) network analysis of interested modules responsible for the progression of PMF was performed to visualize the connections between highly co-related genes. The results in this study provided a comprehensive understanding in the oncogenesis of PMF under transcriptome level, which may contribute to assess the prognosis of PMF and ultimately offer new ideas into the treatment of PMF.

Methods

Microarray Data

Gene expression profile of GSE26049 was downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>), which is a functional

public genomics repository including high throughout gene expression data, chips and microarray. The profile totally contained 90 samples, of which 9 PMF (primary myelofibrosis) samples, 19 ET (essential thrombocythemia) samples, 41 PV (polycythemia vera) samples and 21 normal samples.

Gene Expression Profiles' Preprocessing

The method of quality control was performed by R ("affy", "affyPLM" package) and standard samples were identified through calculating their normalized unscaled standard error. Data preprocessing procedure was used to handle the raw data (".CEL" file format) using robust multi-array average (RMA) background correction and normalization (RMA function, "affy" and "affyPLM" package in R environment). The Affymetrix annotation files from GPL570 platform were applied to annotate corresponding probes, and probes without annotation were removed.

Weighted Gene Co-expression Network Analysis (WGCNA)

WGCNA analysis was performed to identify corresponding expression modules. In this part, the steps of WGCNA analysis as well as the way to identify significant modules associated with clinical features were clearly explained, together with how to screen the most representative genes in the interested module.

The whole WGCNA analysis procedures were conducted under R environment ("WGCNA" package[16]). Firstly, hierarchical clustering analysis was applied to check the heterogeneity of samples to detect and eliminate the outliers. After excluding outliers, clustering was performed according to gene expression levels in each sample to construct the network connection to get access to the real biological network state (scale-free network). The soft threshold power was determined to select a value by computing a correlation raised to a power between every pair of genes, which could amplify disparity among strong and weak correlations so that the network state could approach scale-free network topology. The proper soft threshold power was selected when the scale-free topology index (R^2) reached 0.90 and mean connectivity decreased to 0.

Next, weighted gene co-expression network was constructed based on the correlations of inter-gene expression levels, and co-expression modules were identified by dynamic pruning method. The minimum number of genes in module was set as 30. Eigengenes adjacency was calculated to evaluate interaction of various gene modules. The expression pattern of the gene modules in different features was shown by the heat map (the X axis was the features and Y axis was the module), module-trait relationships were estimated to identify the correlation between modules and PMF.

The degree of correlation between genes was calculated by WGCNA using the Topological overlap measure (TOM)[23, 24], and Pearson's correlation coefficient was also calculated to evaluate correlation between module eigengene and phenotype. $P < 0.01$ was considered to be statistically significant. Then relationships between Gene Significance (GS) and Module Membership (MM) was plotted. Ultimately, the

module with the highest weighted correlation coefficient among all clustered modules was chosen as the interested module and was pooled for further analysis.

Functional and Pathway Enrichment Analysis on Interested Module

Functional annotations and interpretations were applied to discover their biological meaning behind many genes. This study used DAVID database (Database for Annotation, Visualization and Integrated Discovery, <http://david.abcc.ncifcrf.gov/>), to get comprehensive information, which is a crucial repository that provides a set of functional annotation tools for researchers to explore the real biological significance underlying different genes. GO (Gene Ontology) is a powerful tool to analyze the molecular function, cellular component, and biological process of genes. KEGG (Kyoto Encyclopedia Genes and Genomes) pathway enrichment analysis was performed to understand links among different genes and signaling pathways. This study performed GO and KEGG in DAVID database and $P < 0.05$ was set as the cutoff to find out statistically significant information. Additionally, GSEA (Gene set enrichment analysis) was further conducted to determine which set of genes showed statistical significance. Based on the size of set, a normal enrichment score (NES) was assessed for each gene set, FDR and p value were calculated as the cutoff criteria.

PPI (Protein-Protein Interaction) Network Construction and Hub Genes Selection

PPI analysis was then conducted by STRING online tool, (Search tool for retrieval of interacting genes, <https://string-db.org/>), an online database that provided PPI analysis for bioinformatics. Subsequently, we used Cytoscape software (version 3.7.0) to screen hub genes and modules through Molecular Complex Detection (MCODE) plug-in, which was used for clustering given network links based on topology to identify densely connected regions. Each node in the network represented a gene, and edges represented regulatory relationships between genes. MCODE recognized the most significant modules through mathematical algorithm that number of nodes > 7 and MCODE scores > 8 . $P < 0.05$ was considered to be significant difference and hub driver genes were selected as degree ≥ 36 .

Results

Profile's Quality Control

The microarray dataset GSE26049 were applied into this study, gene expression profiles of normal samples as well as different tumor samples ($n = 90$) were normalized and generated. Firstly, we filter and identify the samples which we could use according to the quality control of these gene chips. Standard samples used in this study were selected by calculating their normalized unscaled standard errors (NUSE)

and NUSE plot was also generated for visualization (Fig. 1A). Results illustrated that these 90 samples all had a good chip quality and could be used in our further research. Subsequent to data preprocessing using RMA algorithm, the boxplot of normalized and background corrected gene expression level was plotted (Fig. 1B), results showed that the median amount of gene expression in each sample was on a straight line, indicating that the preprocessed data was served as standard data and could be analyzed for the following study.

Construction of Weighted Gene Co-expression Networks

Firstly, hierarchical clustering analysis for each sample was applied to check the heterogeneity of samples to detect and remove the outliers (Fig. 2A), cutoff threshold height was set as 85 to filter outliers. Finally, one sample GSM639674 was excluded and the gene expression matrix contained 20482 genes in the rest of 89 samples were used for WGCNA analysis. Next, the optimal soft threshold power for WGCNA was calculated. As shown in Fig. 2B, when the soft threshold power β was equal to 16, the scale free topology model fit index (R^2) was higher than 0.90 and mean connectivity was infinitely approaching zero. As a result, power 16 was selected as the optimal soft threshold power value.

Based on the co-expression relationships, hierarchical clustering analysis was then performed to obtain the weighted co-expression network (Fig. 3A), results displayed that after generating the clustered dendrogram, 14 distinct gene co-expression modules, characterized by their unique module color, were identified and clustered. The module colors were salmon, tan, blue, purple, yellow, brown, pink, magenta, black, red, turquoise, green, green-yellow, grey, respectively. Interaction relationships of the 14 co-expression modules in all of the genes was shown in Fig. 3B, topological overlap heatmap revealed that each co-expression module could independently validate each other in the network, and dendrogram branches indicated that genes in each module were highly heterogeneous. Consequently, it was necessary to further identify the interested modules in each subgroup.

Key Module Identification of PMF

After obtaining WGCNA network data, the module-trait relationships heat map was also performed. The interaction relationships between each module and each different feature including control, PV, ET, and PMF subgroups were fully assessed (Fig. 4A, 4B). From this heat map and histogram, we observed that the modules significantly associated with PV was black module, no co-expression module was significantly related to ET. As for PMF, results visualized that correlations and p values of PMF samples in the modules green and red had strong positive correlation compared to normal samples, while purple and yellow modules had a strong negative correlation compared to normal samples, indicating that modules green, red, purple and yellow could significantly differentiate the PMF samples from the normal samples and genes in these modules could be up-regulated or down-regulated in the progression of PMF. Of which, green module had the strongest correlation ($r = 0.46$) and the lowest p value ($p = 8e-5$), elucidating that the green module was highly correlated with PMF, the genes in green module played a crucial roles in the

pathogenesis and oncogenesis of PMF. Besides, Module Membership (MM) in PMF versus Gene Significance (GS) scatter plot of green module was also generated (Fig. 4C), results illustrated that MM was highly correlated with GS in green module ($\text{cor} = 0.48$, $p = 3.6\text{e-}5$).

Module eigengene adjacency was subsequently calculated to cluster and assess the identified modules with feature PMF, and heatmap was plotted to depict their interactions within modules and PMF subgroup (Fig. 5A, 5B). Based on the dendrogram, result illustrated that the PMF subgroup clustered with green module, implying that these two eigenvalues had a highly correlations with each other. As for heat map, each module exhibited their independent validation to the other. Blue colors and red colors represented the different strength of co-expression interactions, and results displayed that the PMF subgroup had the highly correlations with green module.

Then we performed the hierarchical clustering of the gene expression profiles in green module between four subgroups: control, PV, ET and PMF in green module. As shown in Fig. 5C, hierarchical clustering analysis displayed that the genes in green module could significantly differentiate PMF group from other three groups, and PMF samples clustered tightly with each other compared to other groups. From the results above, green module was identified as key module accounting for the progression of PMF and subsequent research was pooled based on the green module genes.

Functional and Pathway Enrichment Analysis of Genes in Green Module

Totally the blue module contained 217 genes (**as shown in supplementary table 1**), to obtain a comprehensive understanding of the biological functions and signaling pathway relevance of the genes in green module, GO, KEGG and GSEA researches were conducted. These genes were uploaded into DAVID database for further investigation, the detailed results of GO, KEGG was shown in Table 1 **and** Fig. 6. GO results indicated that these genes were mainly associated with several biological process including erythrocyte differentiation, hemopoiesis, positive regulation of transcription from RNA polymerase II promoter and negative regulation of apoptosis process et al; cellular component including cytosol, hemoglobin complex, transcription factor complex et al; and several molecular functions such as protein binding, transcription factor binding, heme binding et al. KEGG analysis revealed that some signaling pathways were identified to be significantly changed, such as cell cycle, protein processing in endoplasmic reticulum, bile secretion and porphyrin metabolism. Then GSEA analysis was also conducted to screen the potential biological function and aberrant regulated pathways in PMF patients. As shown in Fig. 7, results indicated that the mutual genes in green module were commonly evolved in critical signaling pathways that were correlated with carcinogenesis of tumor, including cell cycle, hematopoietic cell lineage, JAK-STAT signaling pathway, oocyte meiosis, P53 signaling pathway and toll-like receptor signaling pathway et al.

Table 1
Functional and pathway (GO, KEGG) enrichment analysis of the genes in green module.

Category	Term	Count	%	P Value
GOTERM_BP_DIRECT	GO:0030218 ~ erythrocyte differentiation	5	2.304147	9.37E-04
GOTERM_BP_DIRECT	GO:0030097 ~ hemopoiesis	5	2.304147	0.003971
GOTERM_BP_DIRECT	GO:0043066 ~ negative regulation of apoptotic process	13	5.990783	0.004454
GOTERM_BP_DIRECT	GO:0008285 ~ negative regulation of cell proliferation	11	5.069124	0.012111
GOTERM_BP_DIRECT	GO:0000122 ~ negative regulation of transcription from RNA polymerase II promoter	16	7.373272	0.012984
GOTERM_BP_DIRECT	GO:0007050 ~ cell cycle arrest	6	2.764977	0.019582
GOTERM_BP_DIRECT	GO:0008584 ~ male gonad development	5	2.304147	0.019829
GOTERM_BP_DIRECT	GO:0045944 ~ positive regulation of transcription from RNA polymerase II promoter	19	8.75576	0.022256
GOTERM_BP_DIRECT	GO:0043161 ~ proteasome-mediated ubiquitin-dependent protein catabolic process	7	3.225806	0.024531
GOTERM_BP_DIRECT	GO:0006810 ~ transport	9	4.147465	0.03818
GOTERM_BP_DIRECT	GO:0030308 ~ negative regulation of cell growth	5	2.304147	0.044148
GOTERM_BP_DIRECT	GO:0043065 ~ positive regulation of apoptotic process	8	3.686636	0.047575
GOTERM_CC_DIRECT	GO:0005829 ~ cytosol	63	29.03226	2.93E-06
GOTERM_CC_DIRECT	GO:0030863 ~ cortical cytoskeleton	6	2.764977	3.03E-06
GOTERM_CC_DIRECT	GO:0005833 ~ hemoglobin complex	4	1.843318	2.47E-04
GOTERM_CC_DIRECT	GO:0016020 ~ membrane	39	17.97235	0.001861
GOTERM_CC_DIRECT	GO:0005737 ~ cytoplasm	76	35.02304	0.001901
GOTERM_CC_DIRECT	GO:0080008 ~ Cul4-RING E3 ubiquitin ligase complex	3	1.382488	0.013939
GOTERM_CC_DIRECT	GO:0005667 ~ transcription factor complex	7	3.225806	0.017789
GOTERM_CC_DIRECT	GO:0016323 ~ basolateral plasma membrane	6	2.764977	0.044445

Category	Term	Count	%	P Value
GOTERM_CC_DIRECT	GO:0005856 ~ cytoskeleton	9	4.147465	0.046641
GOTERM_MF_DIRECT	GO:0005515 ~ protein binding	124	57.14286	9.16E-05
GOTERM_MF_DIRECT	GO:0051537 ~ 2 iron, 2 sulfur cluster binding	4	1.843318	0.002571
GOTERM_MF_DIRECT	GO:0008134 ~ transcription factor binding	9	4.147465	0.014121
GOTERM_MF_DIRECT	GO:0042803 ~ protein homodimerization activity	16	7.373272	0.015948
GOTERM_MF_DIRECT	GO:0003714 ~ transcription corepressor activity	7	3.225806	0.025659
GOTERM_MF_DIRECT	GO:0004842 ~ ubiquitin-protein transferase activity	9	4.147465	0.030192
KEGG_PATHWAY	hsa05144: Malaria	5	2.304147	0.003013
KEGG_PATHWAY	hsa04976: Bile secretion	5	2.304147	0.010206
KEGG_PATHWAY	hsa00860: Porphyrin and chlorophyll metabolism	4	1.843318	0.014778
KEGG_PATHWAY	hsa04141: Protein processing in endoplasmic reticulum	7	3.225806	0.017796
KEGG_PATHWAY	hsa00910: Cell Cycle	3	1.382488	0.018195
KEGG_PATHWAY	hsa05211: Renal cell carcinoma	4	1.843318	0.047755

Identification of Hub Genes and PPI (Protein-Protein Network) Construction

After screening the interested module, the PPI network was then analyzed into STRING database, and the weighted co-expression network was constructed and uploaded into Cytoscape software.

Co-related modules in these genes were also identified by the Molecular Complex Detection (MCODE) plug-in in Cytoscape, totally 155 nodes and 863 edges were acquired, together with the top 2 significant modules, as shown in Fig. 8A. Module 1 contained 10 nodes and 55 edges, module 2 contained 7 nodes and 30 edges. The functional annotation and pathway enrichment of screened modules were also conducted using DAVID database, as shown in Table 2. Results showed that genes in module 1 mainly played roles in hemoglobin metabolic process, erythrocyte differentiation, iron ion binding and hemoglobin complex. Genes in module 2 were primarily enriched in protein polyubiquitination, protein binding and cytoplasm et al.

Table 2.
Functional and pathway enrichment analysis of MCODE identified genes.

Module	Term	Count	%	P Value
1	GO:0020027~hemoglobin metabolic process	5	50	0.002856
	GO:0030218~erythrocyte differentiation	4	40	0.018902
	GO:0005833~hemoglobin complex	4	40	0.005912
	GO:0005887~integral component of plasma membrane	3	30	0.027505
	GO:0005506~iron ion binding	2	20	0.078693
2	GO:0000209~protein polyubiquitination	5	71.42857	2.06E-07
	GO:0016567~protein ubiquitination	4	57.14286	1.85E-04
	GO:0005737~cytoplasm	5	71.42857	0.025964
	GO:0005654~nucleoplasm	4	57.14286	0.027962
	GO:0004842~ubiquitin-protein transferase activity	7	100	5.24E-11
	GO:0005515~protein binding	7	100	0.019848

Subsequently, hub genes in this weighted network were filtered with degrees ≥ 36 , degrees meant the level of correlation between two genes. Altogether, 20 genes were identified as hub genes including EPB42, SLC4A1, CALR, MPL, FECH, GYPB, KLF1, DMTN, RBX1, HBD, GYPA, GLRX5, UBE2H, KAT2B, RHAG, SELENBP1, CDC34, TAL1, NEDD4L, SNCA, as listed in Table 3. Meanwhile, the gene expression levels of the hub genes between PMF and normal samples were tested. The results of gene expression levels were shown in heat map (Fig. 8B). Illustration showed that these hub genes expression levels in PMF samples were much higher compared to normal samples, indicating that these genes was responsible for the development of PMF. Additionally, EPB42, CALR, SLC4A1 and MPL associated genes including HBM, KLF1, GYPB, MYL4, AHSP, RHAG were also expressed abnormally, which demonstrated that EPB42, CALR, SLC4A1 and MPL associated pathways were upregulated aberrantly.

Table 3
Detailed information of the hub genes screened in green module.

Gene Symbol	Degree	Betweenness	Gene Symbol	Degree	Betweenness
CALR	68	0.178851	GYPA	40	0.013157
EPB42	63	0.171656	GLRX5	38	0.048597
MPL	62	0.090448	UBE2H	38	0.104637
SLC4A1	53	0.026492	KAT2B	38	0.106015
FECH	52	0.073644	RHAG	37	0.00664
GYPB	52	0.01537	SELENBP1	37	0.050267
KLF1	46	0.013714	CDC34	36	0.025117
DMTN	43	0.077598	TAL1	36	0.080359
RBX1	43	0.050277	NEDD4L	36	0.048756
HBD	41	0.004011	SNCA	36	0.093156

Among those genes, the node degree of genes EPB42, CALR, SLC4A1 and MPL ranked highest, showing that EPB42, CALR, SLC4A1 and MPL had the most links with other proteins, which indicated that these four genes could play a vital role in the pathogenesis and oncogenesis of PMF.

Discussion

Primary myelofibrosis (PMF) is one of the most complex and common malignant tumors in myeloproliferative neoplasm (MPN), which is frequently but not always accompanied by JAK2, CALR mutation, aberrant cytokine expression, bone marrow fibrosis and anemia et al[25–27]. Despite the fact that great progress has been made in the diagnosis of PMF during the last decades, the overall treatment and prognosis of PMF is still unsatisfactory, followed by a short median survival time as well as an inadequate exploration of chemotherapy[10–13]. Consequently, to better discover novel and precise molecular biomarkers that could effectively and accurately predict the progression and prognosis of PMF patients, this study extracted gene expression profile of GSE26049 from GEO database to investigate and validate potential key modules and hub genes responsible for the pathogenesis and oncogenesis of PMF through bioinformatics analysis with WGCNA.

WGCNA is a powerful global research tool for data mining to analyze multiple genes in large-scale datasets, which is developed in recent years characterized by filtering significant modules and gene signatures associated with phenotype features or clinical traits. WGCNA has already been used in the past five years in some kinds of neoplasms and has made a great progress whether in the identification of biomarkers in disease or treatment of targeted genes[28–32]. To our knowledge, the application of

WGCNA methods in the identification of potential targeted genes in PMF has never been reported so far. Therefore, a comprehensive understanding in the pathogenesis of PMF under transcriptome level is imperative for its treatment and prognosis in order that researchers could provide effective therapeutic strategies as well as make an overall prognostic assessment at the early stage of PMF.

In the present study, WGCNA was conducted and 14 co-expression modules were generated based on 20482 genes from the 89 samples in GEO database. The aim of this study was to elucidate the molecular pattern involved in PMF and the association of key modules with PMF features and excavate the real biological meaning behind hub genes. Firstly, hierarchical clustering analysis was generated to detect outliers and cutoff threshold was set as 85 to remove the outliers. Next, the optimal soft threshold power was picked by “pickSoftThreshold” function in R to construct a scale-free topology network in order to get access to the real biological network state. Soft threshold power was a key value applied to amplify the disparity among strong and weak correlation genes, which could affect the mean connectivity and the scale independence of co-expression modules. Ultimately, soft threshold power β was determined as 16.

The weighted co-expression network was constructed after soft threshold power was selected, then hierarchical clustering was performed to visualize the weighted network. Altogether, 14 distinct co-expression modules were generated, and each module contained the genes that had the most correlations with each other. After acquiring WGCNA network data, the module-trait heat map was displayed, WGCNA tool was conducted because of its merit in revealing co-expression modules and their correlation with different groups. from the heat map, we observed that green module had the strongest positive correlation and the lowest p value to PMF groups, illustrating that genes in green module could significantly differentiate the PMF groups from the normal groups. Meanwhile, purple and yellow modules had a highly negative correlation and low p value in PMF groups whereas these modules had a highly positive correlation and low p value in normal groups, genes in purple and yellow modules were highly expressed in normal samples while poorly expressed in PMF samples, indicating that genes in these modules may benefit for the normal organism development. Additionally, the correlation and p value in PV and ET groups were not as significant as PMF group (Fig. 4A). Meanwhile, hierarchical clustering analysis in green module visualized that genes expression level could significantly differentiate PMF groups from other three groups (Fig. 5C). These findings implied, consistent with previous studies, that PMF is the most major neoplasms in MPNs, later in their courses, both PV and ET disorders may be evolved into post-PV MF and post-ET MF due to gene mutation or over-expression[3, 4].

To further study the function and pathway regulation mechanisms of tumorigenesis, GO, KEGG and GSEA researches were carried out to identify significant biological functions and abnormal signaling pathways related to PMF. Functional annotation revealed that genes in green module were mainly enriched in erythrocyte differentiation, hemoglobin complex, transcription factor complex and protein binding, which may explain why the fast multiplication of cancer cells as well as the generation of tumor cells. Our results implied that abnormal changes such as erythrocyte differentiation may occur in hemoglobin complex, which agreed with previous researches that myeloproliferative differentiation disorders may lead into aplastic anemia in the development of PMF[33]. This analysis also found that molecular

functions were primarily enriched in transcription factor binding in transcription factor complex, which was consistent with recent studies that cytokines mediated corresponding receptors into the nucleus to participate in transcriptional regulation, and finally lead into tumor growth[5]. Furthermore, the analysis of KEGG and GSEA revealed that these genes were primarily enriched in cell cycle, hematopoietic cell lineage, JAK-STAT signaling pathway, oocyte meiosis, P53 signaling pathway. The oocyte meiosis was firstly contacted to PMF, and the mechanism was presumed to be related to progesterone-mediated oocyte maturation, which was reported to be related with many neoplasms[34]. Next, toll-like receptor signal pathway was found to be highly activated in normal samples, indicating that toll-like receptor behaved as role in resisting tumor in the normal individual. Besides, protein processing in endoplasmic reticulum, bile secretion and porphyrin metabolism were also aberrantly activated in PMF, the reason caused bile secretion may due to hepatomegaly and hyperfunction caused by gene over expression. Advanced research reported that porphyria was a group of porphyrin metabolic disorders caused by the deficiency of special enzyme or the decrease of enzyme in the pathway of heme synthesis. Based on the findings in this study, genes responsible for the progression of PMF in green module may play roles in heme binding, porphyrin metabolism et al, we have reason to deduce that the development of PMF may also lead to the occurrence of porphyria, the links between these two diseases had not been reported so far. Further research needs to be conducted to find out the correlations between these two disorders. Additionally, studies also reported that dysregulation of cell cycle played a crucial role in the oncogenesis of tumor[35, 36], which was agreed with our findings that cell cycle was aberrantly activated in PMF.

With the aim of screening hub genes among green module identified in our previous work, the 217 mutual genes were analyzed with construction of PPI network based on the STRING database. Altogether, 20 genes were selected as hub genes with high degrees. Heat map of these gene expression between PMF samples and normal samples showed that these hub genes could significantly distinguish these two groups (Fig. 8B). Particularly, EPB42, CALR, SLC4A1 and MPL ranked highest in their degree value, indicating that these genes had the most correlations with the progression of PMF.

CALR, located on chromosome 19, whose full name was calreticulin, was a multifunctional protein that acted as a major Ca^{2+} binding protein in the lumen of the endoplasmic reticulum. Advanced research proved that it also existed in nucleus, suggesting that it may play a role in transcription regulation. Many of existed studies had reported their highly correlation with the development of PMF[5, 37, 38], driver mutations involving CALR in 90% of patients mediated constitutive JAK-STAT signaling, thereby leading to changes in the cytokine and growth factor milieu and accordingly potentiated fibrosis and finally played a fundamental role in disease pathogenesis. MPL, located on chromosome 1, mediated the expression of thrombopoietin and its receptor, which played an important role in cellular signaling pathway, they were essential to the expansion and regulation of megakaryocytes as well as self-renewal of hematopoietic stem cells. Previous researches showed that mutations in both MPL and CALR could increase the JAK-STAT activation, of which CALR may induce JAK-STAT signaling pathway by increasing recruitment of proteins to the MPL promoter site[39]. EPB42, erythrocyte membrane protein band 4.2, was an ATP-binding protein which may regulate the correlation of protein 3 with ankyrin. It was reported that

EPB42 could manipulate the shape and mechanical property of erythrocyte and have a high association with hereditary spherocytosis[40, 41]. SLC4A1, the protein encoded by this gene was part of the anion exchanger family and was expressed in the erythrocyte plasma membrane. The protein comprised two domains which were structurally and functionally distinct, of which the N-terminal 40 kDa domain was in the cytoplasm and acted as an attachment site for erythrocyte skeleton by binding with ankyrin. Many mutations in this gene had been known in human, the over expression of SLC4A1 could lead to hereditary spherocytosis caused by destabilization of cell membrane[42]. It was well known that the clinical manifestation of primary myelofibrosis was obvious enlargement of spleen (mostly splenomegaly) and extramedullary hematopoiesis of each organ. Hematological features were the presence of large amount of erythrocyte in the peripheral blood cell smear. Based on the findings above, this study firstly found that EPB42 and SLC4A1 had a strong correlation with PMF, consequently, we hypothesize that patients with PMF may have a complication of hereditary spherocytosis due to the over expression of EPB42 and SLC4A1 in the progression of the disorder. Due to the fact that there were no reports of the association between these two disorders, further research needs to be conducted to validate our hypothesis.

Overall, this study used WGCNA analysis for a whole situation to conduct an elaborate, precise, and systematic view to analyze the differentially expressed genes from PMF patients. However, considering all the results above, there were still some limitations, on the one hand, small size of samples may limit the statistical power to identify the hub genes; on the other hand, molecular biological experiments needs to be applied to further validate these hub genes to determine whether they may be beneficial in the diagnosis or treatment of PMF.

Conclusions

In conclusion, this study attempted to explore the potential molecular regulatory mechanism of PMF based on WGCNA analysis. 14 distinct co-expression modules were identified from GEO dataset GSE26049. Of which, green module was most significantly correlated with PMF. EPB42, CALR, SLC4A1 and MPL in green module were recognized as key therapeutic targets for PMF, EPB42 and SLC4A1 were firstly found to be highly correlated with PMF.

Abbreviations

DAVID: Database for Annotation, visualization, and integrated discovery

ET: Essential Thrombocythemia

GO: Gene Ontology

GSEA: Gene Set Enrichment Analysis

KEGG: Kyoto Encyclopedia Genes and Genomes

PPI: Protein-Protein Interaction

PMF: Primary Myelofibrosis

PV: Polycythemia Vera

STRING: Search tool for retrieval of interacting genes

WGCNA: Weighted Co-expression Network Analysis

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

All contributing authors agree to the publication of this article.

Availability of data and materials

The data used and analyzed during the current study are available upon reasonable request.

Competing Interests

All authors declare no conflicts of interest related to this manuscript, and all authors have approved the publication of this work.

Funding

Not applicable.

Authors' Contribution

This study was completed with teamwork. Every author had made corresponding contribution to the study:

Conceived the idea: WL, MY, GG.

Wrote the main manuscript: WL, BY.

Used the software: WL, SZ.

Downloaded and collected data: ZD, YZ, WL.

Analyzed the data: WL, BY.

Prepared figures: WL, ZD, YZ.

Checked and redressed the manuscript: MY, GG.

Reviewed the manuscript: MY, GG.

Statement: all authors have read and approved the manuscript, and all authors have approved the publication of this work.

Acknowledgements

Not applicable.

References

1. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405. *Blood*. 2016; 128(3):462-463.
2. Nangalia J and Green AR. Myeloproliferative neoplasms: from origins to outcomes. *Blood*. 2017; 130(23):2475-2483.
3. Mesa RA, Verstovsek S, Cervantes F, Barosi G, Reilly JT, Dupriez B, Levine R, Le Bousse-Kerdiles M-C, Wadleigh M, Campbell PJ, Silver RT, Vannucchi AM, Deeg HJ, et al. Primary myelofibrosis (PMF), post polycythemia vera myelofibrosis (post-PV MF), post essential thrombocythemia myelofibrosis (post-ET MF), blast phase PMF (PMF-BP): Consensus on terminology by the international working group for myelofibrosis research and treatment (IWG-MRT). *Leuk Res*. 2007; 31(6):737-740.
4. Tefferi A, Thiele J, Orazi A, Kvasnicka HM, Barbui T, Hanson CA, Barosi G, Verstovsek S, Birgegard G, Mesa R, Reilly JT, Gisslinger H, Vannucchi AM, et al. Proposals and rationale for revision of the World Health Organization diagnostic criteria for polycythemia vera, essential thrombocythemia, and primary myelofibrosis: recommendations from an ad hoc international expert panel. *Blood*. 2007; 110(4):1092-1097.
5. Garmezay B, Schaefer JK, Mercer J and Talpaz M. A provider's guide to primary myelofibrosis: pathophysiology, diagnosis, and management. *Blood Rev*. 2020:100691.
6. Tefferi A and Gilliland G. Classification of chronic myeloid disorders: from Dameshek towards a semi-molecular system. *Best Pract Res Clin Haematol*. 2006; 19(3):365-385.

7. Finazzi G, Vannucchi AM and Barbui T. Prefibrotic myelofibrosis: treatment algorithm 2018. *Blood Cancer J.* 2018; 8(11):104.
8. Tefferi A and Barbui T. Polycythemia vera and essential thrombocythemia: 2017 update on diagnosis, risk-stratification, and management. *Am J Hematol.* 2017; 92(1).
9. Tefferi A, Nicolosi M, Mudireddy M, Szuber N, Finke CM, Lasho TL, Hanson CA, Ketterling RP, Pardanani A, Gangat N, Mannarelli C, Fanelli T, Guglielmelli P, et al. Driver mutations and prognosis in primary myelofibrosis: Mayo-Careggi MPN alliance study of 1,095 patients. *Am J Hematol.* 2018; 93(3):348-355.
10. Leiva O, Ng SK, Chitalia S, Balduini A, Matsuura S and Ravid K. The role of the extracellular matrix in primary myelofibrosis. *Blood Cancer J.* 2017; 7(2):e525.
11. Mehta J, Wang H, Iqbal SU and Mesa R. Epidemiology of myeloproliferative neoplasms in the United States. *Leuk Lymphoma.* 2014; 55(3):595-600.
12. Tefferi A, Partain DK, Palmer JM, Slack JL, Roy V, Hogan WJ, Litzow ML, Ketterling RP and Patnaik MM. Allogeneic hematopoietic stem cell transplant overcomes the adverse survival effect of very high risk and unfavorable karyotype in myelofibrosis. *Am J Hematol.* 2018; 93(5):649-654.
13. Samuelson Bannow BT, Salit RB, Storer BE, Stevens EA, Wu D, Yeung C, Fang M, Petersdorf EW, Linenberger ML, Woo J, Sorrow ML, Doney K, Sandmaier BM, et al. Hematopoietic Cell Transplantation for Myelofibrosis: the Dynamic International Prognostic Scoring System Plus Risk Predicts Post-Transplant Outcomes. *Biol Blood Marrow Transplant.* 2018; 24(2):386-392.
14. Tsimberidou A-M. Targeted therapy in cancer. *Cancer Chemother Pharmacol.* 2015; 76(6):1113-1132.
15. Liu X, Hu A-X, Zhao J-L and Chen F-L. Identification of Key Gene Modules in Human Osteosarcoma by Co-Expression Analysis Weighted Gene Co-Expression Network Analysis (WGCNA). *J Cell Biochem.* 2017; 118(11):3953-3959.
16. Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008; 9:559.
17. Zhang B and Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol.* 2005; 4:Article17.
18. Xia W-X, Yu Q, Li G-H, Liu Y-W, Xiao F-H, Yang L-Q, Rahman ZU, Wang H-T and Kong Q-P. Identification of four hub genes associated with adrenocortical carcinoma progression by WGCNA. *PeerJ.* 2019; 7:e6555.
19. Zhao Q, Song W, He DY and Li Y. Identification of key gene modules and pathways of human breast cancer by co-expression analysis. *Breast Cancer.* 2018; 25(2):213-223.
20. Shi K, Bing Z-T, Cao G-Q, Guo L, Cao Y-N, Jiang H-O and Zhang M-X. Identify the signature genes for diagnose of uveal melanoma by weight gene co-expression network analysis. *Int J Ophthalmol.* 2015; 8(2):269-274.
21. Wan Q, Tang J, Han Y and Wang D. Co-expression modules construction by WGCNA and identify potential prognostic markers of uveal melanoma. *Exp Eye Res.* 2018; 166:13-20.

22. Shi Z, Derow CK and Zhang B. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst Biol.* 2010; 4:74.
23. Yip AM and Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics.* 2007; 8:22.
24. Li A and Horvath S. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics.* 2007; 23(2):222-231.
25. Barbui T, Thiele J, Gisslinger H, Finazzi G, Vannucchi AM and Tefferi A. The 2016 revision of WHO classification of myeloproliferative neoplasms: Clinical and molecular advances. *Blood Rev.* 2016; 30(6):453-459.
26. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, Bloomfield CD, Cazzola M and Vardiman JW. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood.* 2016; 127(20):2391-2405.
27. Tefferi A. Primary myelofibrosis: 2017 update on diagnosis, risk-stratification, and management. *Am J Hematol.* 2016; 91(12):1262-1271.
28. Wang Y, Chen L, Wang G, Cheng S, Qian K, Liu X, Wu C-L, Xiao Y and Wang X. Fifteen hub genes associated with progression and prognosis of clear cell renal cell carcinoma identified by coexpression analysis. *J Cell Physiol.* 2019; 234(7):10225-10237.
29. Liu B, Huang G, Zhu H, Ma Z, Tian X, Yin L, Gao X and He X. Analysis of gene co-expression network reveals prognostic significance of CNFN in patients with head and neck cancer. *Oncol Rep.* 2019; 41(4):2168-2180.
30. Giulietti M, Occhipinti G, Principato G and Piva F. Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development. *Cell Oncol (Dordr).* 2016; 39(4):379-388.
31. Chen J, Wang X, Hu B, He Y, Qian X and Wang W. Candidate genes in gastric cancer identified by constructing a weighted gene co-expression network. *PeerJ.* 2018; 6:e4692.
32. Foroughi K, Amini M, Atashi A, Mahmoodzadeh H, Hamann U and Manoochchri M. Tissue-Specific Down-Regulation of the Long Non-Coding RNAs PCAT18 and LINC01133 in Gastric Cancer Development. *Int J Mol Sci.* 2018; 19(12).
33. Tefferi A, Finke CM, Lasho TL, Wassie EA, Knudson R, Ketterling RP, Hanson CA and Pardanani A. U2AF1 mutations in primary myelofibrosis are strongly associated with anemia and thrombocytopenia despite clustering with JAK2V617F and normal karyotype. *Leukemia.* 2014; 28(2):431-433.
34. Liu M, Li X, Fan R, Liu X and Wang J. A Systematic Analysis of Candidate Genes Associated with Nicotine Addiction. *Biomed Res Int.* 2015; 2015:313709.
35. Wang Y, Cheng J, Xu C, Liu S, Jiang S, Xu Q, Chen X, Zhuang H and Lu F. Quantitative methylation analysis reveals gender and age differences in p16INK4a hypermethylation in hepatitis B virus-related hepatocellular carcinoma. *Liver Int.* 2012; 32(3):420-428.

36. Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, Zhang Y, Gorospe M, Prasanth SG, Lal A and Prasanth KV. Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet.* 2013; 9(3):e1003368.
37. Pronier E, Cifani P, Merlinsky TR, Berman KB, Somasundara AVH, Rampal RK, LaCava J, Wei KE, Pastore F, Maag JL, Park J, Koche R, Kentsis A, et al. Targeting the CALR interactome in myeloproliferative neoplasms. *JCI Insight.* 2018; 3(22).
38. Gangat N and Tefferi A. Myelofibrosis biology and contemporary management. *Br J Haematol.* 2020.
39. Kaushansky K, Lok S, Holly RD, Broudy VC, Lin N, Bailey MC, Forstrom JW, Buddle MM, Oort PJ and Hagen FS. Promotion of megakaryocyte progenitor expansion and differentiation by the c-Mpl ligand thrombopoietin. *Nature.* 1994; 369(6481):568-571.
40. Chonat S, Risinger M, Sakthivel H, Niss O, Rothman JA, Hsieh L, Chou ST, Kwiatkowski JL, Khandros E, Gorman MF, Wells DT, Maghathe T, Dagaonkar N, et al. The Spectrum of -Associated Hereditary Spherocytosis. *Front Physiol.* 2019; 10:815.
41. Qin L, Nie Y, Zhang H, Chen L, Zhang D, Lin Y and Ru K. Identification of new mutations in patients with hereditary spherocytosis by next-generation sequencing. *J Hum Genet.* 2020; 65(4):427-434.
42. Lux SE, John KM, Kopito RR and Lodish HF. Cloning and characterization of band 3, the human erythrocyte anion-exchange protein (AE1). *Proc Natl Acad Sci USA.* 1989; 86(23):9089-9093.

Figures

Figure 1

(A), Normalized unscaled standard error (NUSE) plot of GSE26049 for quality control. (B), box plot of gene expression level in GSE26049 after RMA background correction and normalization.

Figure 2

(A), Cluster dendrogram of samples in GSE26049 to detect outliers. The dendrogram branches represented the clustered samples. (B), Selection of soft threshold power value through WGCNA analysis. The left panel showed the scale-free model fit index (y axis) as function of soft thresholding power value (x axis). Red line represented the y-axis value. Right panel displayed the mean connectivity (y axis) as a function of soft thresholding powers (x axis).

Figure 3

(A), Cluster dendrogram of all genes enriched based on the dissimilarity measure and the assignment modules. Highly interconnected groups of genes were clustered. Every color below represented one co-expression module, totally 14 distinct modules were identified with the hierarchical clustering tree analysis. (B), Topological overlap heatmap of the gene co-expression network. Each row and column represented a gene. Light color indicated high topological overlap and dark color represented low

topological overlap. Different colors on x and y axis indicated different modules. The dendrogram suggested that the clustering of these genes was based on the similarity of their gene expression levels.

Figure 4

(A), The module-trait relationship heatmap between MEs and features. Each row corresponded to a module eigengene and column to a feature. The first column was normal samples; the second column was essential thrombocythemia; the third column was primary myelofibrosis; and the fourth column was polycythemia vera. Each square contained corresponding correlation (first line) and p value (second line).

The left side of heat map indicated the module name, right side of heat map indicated the colors of correlation (blue represented negatively correlated, red represented positively correlated). (B), histogram of gene significance across modules in PMF group. (C), The correlation between gene significance and module membership in green module. The y-axis indicated gene significance in green module, and x-axis represented the module membership in green module.

Figure 5

(A), the cluster dendrogram of adjacencies in the eigengene network. (B), the heat map of adjacencies in the eigengene network. Blue represented a negative correlation, while red represented a positive correlation. (C), Green module genes expression heat map of different groups (Control, ET, PV and PMF).

Figure 6

Functional and pathway enrichment analysis of genes in green module.

Figure 7

(A-F), Gene set enrichment analysis of KEGG pathways in green module.

Figure 8

(A), Visualization of protein-protein interaction (PPI) network of genes in green module as well as the top 2 modules from PPI network. Every edge represented the interaction between two genes. (B), hub genes expression heat map identified in green module.