Gene coexpression network analysis revealed biomarkers correlated with blast cells and survival in acute myeloid leukemia

YUMING PAN1,2*, QIAOXIA ZHANG2*, XIAOPENG DENG2, NA AN2, XIN DU2 and JIAJUN LIU1

1Department of Hematology, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou, Guangdong 510630; 2Shenzhen Bone Marrow Transplantation Public Service Platform, Department of Hematology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen, Guangdong 518035, P.R. China

Received September 3, 2019; Accepted January 30, 2020

Abstract. Acute myeloid leukemia (AML) is a hematological malignancy with a poorly understood pathogenesis, especially among patients with no known cytogenetic abnormalities. Furthermore, there is a lack of therapeutic gene targets and diagnostic biomarkers for the effective treatment of AML. The present study aimed to identify candidate biomarkers correlated with the clinical prognosis of patients with AML. Leukemic cells from 5 patients with AML exhibiting a normal karyotype, and hematopoietic cells from 5 healthy donors were processed for RNA sequencing (RNA-seq), and the obtained RNA expression profiles were subjected to weighted gene correlation network analysis. A novel group of genes (the red module) were identified to be significantly associated with AML, and this module contained a closely connected network with 114 mRNAs. Analysis of the correlation between these mRNAs and blast cell percentage, overall survival (OS) and disease-free survival (DFS) using cases from The Cancer Genome Atlas (TCGA) database revealed that CSF3R, ALPL, and LMTK2 were negatively associated with the percentage of blast cells, while high expression of these genes was associated with longer OS and DFS in patients with AML. The differential expression of these three genes between patients with AML and healthy control subjects was supported using the Genotype-Tissue Expression and TCGA databases and was further confirmed using reverse transcription-quantitative (RT-qPCR). These genes exhibited significantly lower expression in patients with AML compared with control subjects. The results indicated that CSF3R, ALPL, and LMTK2 exhibit the potential to be prognostic biomarkers. However, the biological functions of these three candidate genes need to be assessed in further studies.

Introduction

Acute myeloid leukemia (AML) is a hematologic malignancy characterized by aberrant clonal amplification of undifferentiated myeloid progenitors in bone marrow (BM) and results in dysregulated hematopoiesis (1). In the USA, the morbidity and mortality of AML are 13 and 7.1 per one hundred thousand persons, respectively (2), and these values in China are increasing (3). AML outcomes are classified as favorable, intermediate and adverse. Young patients with AML in the favorable and adverse groups had three-year overall survival (OS) rates of 66 and 12% and three-year disease-free survival (DFS) rates of only 55 and 10%, respectively (4,5). Prognosis is related not only to age, sex, karyotype, white blood cell (WBCs) count and blast cell count but also to the expression and mutation of some critical genes (6). Several biomarkers have been proven to be useful in the diagnosis and prognosis of AML according to recent studies (7,8) and some reports have shown that the expression levels of some genes, such as SETBP1, VEGFC and EVII, are associated with the risk level and the survival of patients (9-11). However, OS and DFS of patients with AML remain poor (12,13). Therefore, identifying additional AML-related genes is urgently needed.

A typical feature of cancer is altered transcriptional networks originating from genetic aberrances, which drive disease occurrence and development (14,15). These genetic abnormalities can act in conjunction with suitable upstream and downstream molecules to exert procarcinogenic activities. Therefore, the identification of novel transcriptional networks and key nodes should help combat abnormal transcription. Weighted gene correlation network analysis (WGCNA) is a statistical technique based on functions in the R software...
package and is used to identify groups of genes among microarray or transcriptome sequencing data that are highly correlated with each other and with biological traits (16). This biological analysis method has been used in pancreas, colon and bladder studies to identify important prognostic and therapeutic targets in data from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) (17-19). In this study, we used WGCNA to identify genes in the coexpressed network associated with AML.

Materials and methods

Clinical specimens. The clinical specimens used for RNA-seq were obtained from five patients with de novo AML (three females and two males) with normal karyotypes. The five AML patients were adults aged 17, 24, 26, 31 and 44 years. The French-American and British (FAB) types were M1 (2), M4 (2) and M5 (1). The WBC count per liter of blood ranged from 106x10⁹ to 642x10⁹. Hematopoietic stem cells from mobilized peripheral blood (PB) of five healthy male donors with a mean age of 33.8 years were used as controls. The clinical specimens used to verify the differences in gene expression included samples from patients with AML (30) and healthy individuals (43). The thirty AML patients included 17 males and 13 females, and the average age was 41.1 years, ranging from 24 to 64 years old. Conventional cytogenetic analysis showed that 15 AML patients had a normal karyotype, while the karyotype of the other patients showed abnormalities. The FAB classification of the patients was stratified as follows: 1 M0, 4 M1s, 7 M2s, 7 M4s and 11 M5s. The healthy donors included 32 males and 11 females, and they ranged in age from 11 to 58 years, with a mean age of 33.6. The collection products from BM and PB were processes with red blood cell lysis buffer (Beijing Solarbio Science & Technology Co., Ltd.) and the leukocytes remaining after centrifugation were used in subsequent RNA_seq or RT-qPCR.

This study protocol was approved by the institutional medical Ethics Committee of Shenzhen Second People’s Hospital (Shenzhen, China). All patients and donors provided informed consent for the molecular analysis of their samples.

Online data resources. Clinicopathological data, including blast cells, OS, DFS and cytogenetic karyotype, of 200 AML specimens and the corresponding mRNA expression data were downloaded from the TCGA (https://gdc-portal.nci.nih.gov/). Because some data were invalid, the clinical data of 170 PB blast cells, 173 BM blast cells, 160 OS, 171 DFS and 156 karyotype samples were finally included to investigate the prognostic potential of candidate genes for AML. The mRNA expression data of whole blood from 337 healthy donors were downloaded from the Genotype-Tissue Expression (GTEX) database (https://gtexportal.org/home/index.html).

RNA sequencing. mRNA and IncRNA were isolated by removing the ribosomal RNA from total RNA. Then, the remaining RNA was fragmented (200-500 bp) and reverse transcribed into cDNA using random primers. A cDNA template with an adapter was used for fragment amplification and library construction. The libraries were sequenced using an Illumina HiSeq 2000 system (Total Genomics Solution Pte. Ltd.). Clean reads were retrieved after filtering out sequences with poor quality and adaptor sequences from the raw reads and were aligned with the reference genome (UCSC hg 19) by HISAT (20).

Analysis of mRNA and IncRNA expression. The transcripts of the samples were reconstructed by StringTie (21), and redundant transcripts were eliminated using Cuffcompare software (22). The IncRNAs were collected through four filtering steps as follows: The short transcripts (<200 bp) were removed, the background transcripts were removed, the known transcripts and pre-mRNAs were removed, and the transcripts with protein-coding potential were removed. The number of reads mapped to the exon regions was calculated using HTseq software, and the expression levels of IncRNAs and mRNAs were calculated as the RPKM. CircRNA was selected as the intersection of the results, which were predicted by find_circ and CIRI software. The expression levels of the circRNAs were calculated with the pseudo RPKM method.

Weighted correlation network analysis. The mRNAs, IncRNAs and circRNAs were screened from the transcriptome profiles according to the following criteria: The expression levels of mRNAs and IncRNAs must be ≥ one in all specimens; the coefficient of variation must be at least 0.5; and circRNA must be expressed in 80% of the specimens. The resulting RNAs were used to construct the weighted gene coexpression network by WGCNA (https://labs.genetics.ucla.edu/horvath/htdocs/CoeXpressionNetwork/Rpackages/WGCNA/). First, Pearson's method was used to calculate the pairwise correlation coefficients of the genes and to construct the gene expression correlation matrix. Next, the appropriate value of the soft-thresholding power (β) was selected to build a weighted adjacency matrix, which was further transformed into a topological overlap matrix (TOM) and dissimilarity matrix. The latter was used for hierarchical clustering and dynamic cutting. The main modules were identified after an appropriate cutHeight point for cutting the tree was chosen and modules with similar eigengenes were merged.

Identifying the module associated with AML and functional enrichment analysis. The module eigengene (ME) represents a distinctive gene expression pattern of a module in a sample. The module-trait relationships were calculated using the correlation between modules’ MEs and traits of AML. The gene significance (GS) was used to combine the clinical traits with the coexpression network. The higher the absolute value of GS, the more biologically meaningful the gene in a module is. Module significance (MS) is defined as the average absolute GS measured for all genes in a given module. The genes in the module of interest were subjected to Gene Ontology (GO) analyses. A P-value < 0.05 was considered to be the cut-off criterion for significance.

Candidate prognostic target selection. The nodes (genes) in an undirected, weighted gene network corresponded to gene expression profiles. The edges between genes were determined by pairwise correlations between the expression levels of the genes. The genes in the module that were highly associated
expression levels of genes were statistically analyzed by blast cells in the BM or PB of AML patients and the Statistical analysis. Assessment of the biological processes showed that each module were subjected to GO functional enrichment analysis. Although six modules were obtained through the dynamic branch cutting method (cutHeight=0.18; Fig. 1C). The red module had the greatest MS value (slightly higher than the firebrick4 module, Fig. 2B). This finding suggested that the red module was closely related to AML. Furthermore, the protein-coding genes presented in each module were subjected to GO functional enrichment analysis. Assessment of the biological processes showed that the genes within the red module were enriched in processes associated with the biological characteristics of hematopoietic cells such as neutrophil activation (P-value=2.38x10^-43), neutrophil degranulation (P-value=7.47x10^-45) and leukocyte migration (P-value=2.09x10^-18; Fig. 3B), while the genes in firebrick4 module were enriched in genes related to ribosome biogenesis (P-value=2.51x10^-15) and ncRNA metabolic process (P-value=3.14x10^-11; Fig. 3B); genes in mediumpurple3 were mainly enriched in T cell activation (P-value=5.31x10^-27; Fig. 3C); and genes in the darkseagreen4 were enriched in the regulation of the smoothened signaling pathway (P-value=1.71x10^-5; Fig. 3D). The functional annotation of the genes in the red module revealed that they are

with AML were selected as candidate genes with the criterion of a weighted value (edge width) that was not smaller than 0.4 between any two genes in the module. The weighted value between the genes was derived from the TOM matrix. We graphed the candidate gene coexpression network using Cytoscape software.

RNA extraction and RT-qPCR. Clinical specimens were washed with RBC lysis buffer (Beijing Solarbio Science & Technology Co., Ltd.) to remove the red blood cells (RBCs) and washed at least once with PBS buffer. Then, for the extraction of total RNA, the remaining white cells were suspended in RNAiso Plus reagent (Takara Bio, Inc.) and placed in a -80°C refrigerator according to the manufacturer’s instructions. A Prime Script II cDNA synthesis kit (Takara Bio, Inc.) was used to perform reverse transcription. In total, 2 µg of RNA was converted into cDNA with random primers. The RT-qPCR system was prepared with TB Green Premix Ex Taq II (Takara Bio, Inc.) and the reaction was performed on a QuantStudio DX (Applied Biosystems; Thermo Fisher Scientific, Inc.). The primers for the candidate and reference genes are listed in Table SI.

OS, overall survival; DFS, disease free survival.

Table I. The optimal cut-off points of the candidate genes for OS and DFS.

<table>
<thead>
<tr>
<th>Genes</th>
<th>OS Cut-off point</th>
<th>OS Statistic</th>
<th>DFS Cut-off point</th>
<th>DFS Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTPRI</td>
<td>10.41</td>
<td>1.33</td>
<td>10.05</td>
<td>1.49</td>
</tr>
<tr>
<td>WLS</td>
<td>3.92</td>
<td>1.21</td>
<td>1.51</td>
<td>2.50</td>
</tr>
<tr>
<td>EXT1</td>
<td>4.20</td>
<td>3.04</td>
<td>7.00</td>
<td>1.85</td>
</tr>
<tr>
<td>KRMEN1</td>
<td>5.52</td>
<td>1.90</td>
<td>5.52</td>
<td>2.15</td>
</tr>
<tr>
<td>ALPL</td>
<td>0.94</td>
<td>1.89</td>
<td>3.30</td>
<td>3.57</td>
</tr>
<tr>
<td>QPCT</td>
<td>2.22</td>
<td>2.46</td>
<td>6.31</td>
<td>1.17</td>
</tr>
<tr>
<td>CR1</td>
<td>7.99</td>
<td>1.51</td>
<td>11.60</td>
<td>2.34</td>
</tr>
<tr>
<td>RASSF5</td>
<td>11.31</td>
<td>2.70</td>
<td>10.98</td>
<td>2.22</td>
</tr>
<tr>
<td>RAB43</td>
<td>7.44</td>
<td>1.94</td>
<td>5.25</td>
<td>2.23</td>
</tr>
<tr>
<td>SEMA4B</td>
<td>8.68</td>
<td>1.99</td>
<td>8.14</td>
<td>2.06</td>
</tr>
<tr>
<td>GLT1D1</td>
<td>8.34</td>
<td>1.36</td>
<td>8.34</td>
<td>1.07</td>
</tr>
<tr>
<td>SLC25A37</td>
<td>10.94</td>
<td>1.71</td>
<td>11.99</td>
<td>1.83</td>
</tr>
<tr>
<td>PIK3CD</td>
<td>12.46</td>
<td>1.73</td>
<td>12.46</td>
<td>1.99</td>
</tr>
<tr>
<td>LMTK2</td>
<td>9.34</td>
<td>3.29</td>
<td>9.34</td>
<td>2.99</td>
</tr>
<tr>
<td>IG5F6</td>
<td>5.69</td>
<td>1.80</td>
<td>5.69</td>
<td>1.52</td>
</tr>
<tr>
<td>ECE1</td>
<td>9.52</td>
<td>4.78</td>
<td>9.22</td>
<td>3.65</td>
</tr>
<tr>
<td>STEAP4</td>
<td>5.81</td>
<td>1.58</td>
<td>3.09</td>
<td>1.91</td>
</tr>
<tr>
<td>SLC44A2</td>
<td>11.09</td>
<td>3.12</td>
<td>10.95</td>
<td>2.49</td>
</tr>
<tr>
<td>CSF3R</td>
<td>14.50</td>
<td>4.14</td>
<td>13.43</td>
<td>4.31</td>
</tr>
<tr>
<td>DOK3</td>
<td>11.80</td>
<td>1.24</td>
<td>10.52</td>
<td>2.51</td>
</tr>
</tbody>
</table>
PAN et al.: THE ASSOCIATION OF CSF3R, LMTK2 AND ALPL WITH THE PROGNOSIS OF AML

Figure 1. Modules were identified from the gene coexpression network using WGCNA. (A) The scale-free fitted curve shows the value of the degree of fitting according to $R^2$ based on the various soft threshold powers ($\beta$). (B) The mean connectivity curve assisted in determining the value of $\beta$. (C) The modules were identified from the cluster dendrogram after dynamic branch cutting of the clustering tree to formed the original modules (cut Height=0.18) and merging of some modules based on the eigengene similarity of the original modules. WGCNA, Weighted gene correlation network analysis.

Figure 2. The genes in the red module were revealed to be negatively correlated with AML. (A) The heatmap displays the relationship between the ME of the modules and the traits; the values without parentheses are the correlation coefficients; red corresponds to a high correlation while blue indicates a low correlation; the values in parentheses are the P-values; P<0.05 indicated a statistically significant difference. (B) The mean absolute value of GS (MS) of the eighteen modules; larger MS values correspond to the stronger relationships between the module and AML. AML, Acute myeloid leukemia; ME, module eigengene; GS, gene significance.
Figure 3. The genes from the red module were enriched in biological processes associated with the properties of hematopoietic cells. The biological processes for genes in the (A) red module, (B) firebrick4 module, (C) mediumpurple3 module and (D) darkseagreen4, (top 8). P<0.05 indicated a statistically significant difference.

Figure 4. The gene coexpression network of the red module (edge ≥0.4). Yellow dots indicate circRNAs, pink triangles indicate lncRNAs and green squares indicate mRNAs. The lines denote a coexpression relationship between the two points. The size of the point denotes the total connectivity value.
intimately associated with the physiological development of hematopoietic cells and significantly correlated with AML. Based on the high pairwise-weighted values of genes in the red module, we drew the central nodes and their pattern of connectivity (Fig. 4). The network was composed of 147 nodes and 482 edges, corresponding to 123 mRNAs, 20 IncRNAs and 4 circRNAs (Table SIV).

**Genes associated with the blast cells of AML patients.** The circRNAs (4), IncRNAs (20), and some mRNAs (9) were removed from the 147 nodes in the network because of the lack of expression data on these RNAs in the TCGA database. The correlation between the expression levels of the remaining 114 genes and the blast cell percentage in BM (173) and PB (170) of patients diagnosed with AML were investigated. The results demonstrated that 23 genes had negative correlations with the percentage of blast cells in BM, although the relationships were weak (Pearson’s r value ranged from -0.33 to -0.15; Fig. S1). In addition, 20 of the 23 genes had a significant negative relationship with the percentage of blast cells in PB (Pearson’s value is from -0.22 to -0.54; Fig. 5). These 20 genes were correlated with blast cells in both BM and PB and may play important roles in regulating the growth of leukemic cells.

**Genes associated with OS and DFS of AML patients.** Next, we used the endpoints of OS and DFS to analyze the association of the 20 genes with the survival of AML patients. The
AML patients were categorized into low (≤ cut-off point) and high (> cut-off point) groups based on the expression levels of the 20 genes. Log-rank analysis showed that the OS of AML patients between the two groups was significantly different when the patients were stratified by the expression levels of the EXT1, SLC44A2, ALPL, CSF3R, ECE1, LMTK2, QPCT and RASSF5 genes. Patients with low expression levels of EXT1, SLC44A2, ECE1 and RASSF5 exhibited higher survival rates than those with high expression levels of these genes, while patients with high expression levels of ALPL, CSF3R, LMTK2 and QPCT had better survival rates than those with low expression levels of these genes (Fig. 6A). In addition, all these genes (with the exception of QPCT) were correlated with DFS (Fig. 6B).

Validation of the differences in expression levels of candidate genes between the AML and healthy groups. Our RNA-seq data showed that the expression levels of EXT1, SLC44A2, ALPL, CSF3R, ECE1, LMTK2, QPCT and RASSF5 were significantly lower in AML patients than in healthy controls (Fig. S2). More gene expression data of normal donors (337) and AML patients (156) were downloaded from the GTEx and TCGA databases to verify the differences in the expression levels of the genes of interest initially observed between the healthy and AML groups. This larger dataset also showed that the expression levels of these eight genes were significantly decreased in AML patients. In addition, CSF3R and SLC44A2 had remarkably reduced expression in the normal karyotypic group compared with the abnormal karyotypic group, while RASSF5 tended to have lower expression in AML patients with an abnormal karyotype (Fig. 7).

Further confirmation that the expression levels of CSF3R, LMTK2 and ALPL were downregulated in AML patients compared with healthy controls was established using clinical samples (Fig. 8A-C). This result is logically consistent with the positive correlation of the three genes with OS and DFS in AML patients. Therefore, CSF3R, LMTK2 and ALPL show great potential as new prognostic markers of AML.

Figure 6. A total of eight genes were related to survival in patients with AML. (A) The genes associated with OS in patients with AML, n=160. (B) The genes related to DFS in patients with AML, n=171. The cut-off point was determined in R with the tidyverse and survminer packages. The log-rank test was used to analyze the survival difference between the two groups and to determine the P-value (<0.05 as significant) and the hazard ratio. AML, Acute myeloid leukemia; OS, overall survival; DFS, disease-free survival.
Discussion

WGCNA is an efficient bioinformatics method used to reduce complicated transcriptomes into several gene modules with high interconnectivity and to determine the associations of these modules with clinical traits (16). In this way, we identified that the constructed red module was strongly negatively associated with AML, and the functional annotations revealed that the genes in the red modules are enriched in processes relating to neutrophils activation, neutrophil degranulation and leukocyte migration. Through a series of correlation analyses, we found that three genes with anomalous low expression levels were significantly inversely correlated with the percentage of blast cells but positively correlated with the survival of AML patients.

The three genes CSF3R, ALPL, and LMTK2 had extremely downregulated expression levels in AML patients compared with healthy controls, and the downregulation of these genes was associated with worse OS and DFS in AML patients.

Receptor for colony stimulating factor 3 (CSF3R) is well
known to regulate the production, differentiation, and function of granulocytes (23). Mutations in this gene are frequently present in patients with chronic neutrophilic leukemia (CNL) and can be used as accurate diagnostic markers for CNL (24). Mutations in CSF3R are rare in AML and have been reported to highly overlap with CEBPα mutations in AML patients, which predicts a poor outcome (25,26). Our data show that CSF3R is tended to underexpressed in AML patients with a normal karyotype and may serve as a special genetic biomarker for the prognosis and treatment of AML patients with a normal karyotype. Tissue-nonspecific alkaline phosphatase (ALPL) plays a role in bone mineralization, and mutations in this gene are used to diagnose hypophosphatasia (27). Further studies are needed to reveal the functions of ALPL in AML. Lemur tyrosine kinase 2 (LMTK2) is a tumor suppressor that is downregulated in some neurodegenerative diseases (28) and can inhibit the activity of PP1C by controlling GSK3β phosphorylation (29). The effect of LMTK2 on the pathogenesis of AML has not been studied, but LMTK2 is predicted to enhance the cytotoxic activity of natural killer cells to kill leukemic blast cells via inhibition of GSK3β (30).

In the present study, we using RNA_seq combined with WGCNA statistical method finding CSF3R, ALPL and LMTK2 are potential prognostic markers for AML but need to be studied more thoroughly to confirm their biological functions in this disease. However, the limitation of RNA_seq is the result simply represents the mean expression of genes in white blood cells which contain diverse cell populations (31). The newly developed single-cell RNA sequencing can compensate for the defect and provide more huge and accurate data. The latest method would help in finding exceptional subpopulations and genes of interest in the future.

Acknowledgements

Not applicable.

Funding

The present study work was supported by grants from the National Natural Science Foundation of China (grant no. 81350027), the Major Projects of the Guangdong Natural Team (grant no. 2014A030312012) and the Clinic Research Program of Shenzhen Second People's Hospital (grant no. 20173357201802).

Availability of data and materials

The RNA_seq data from clinical samples were analyzed in the current study and are available in the public repository in the NCBI database (SRA accession: PRJNA576718). The data used to confirm this research are available from TCGA (https://gdc-portal.nci.nih.gov/) and GTEx (https://gtexportal.org/home/index.html).

Authors' contributions

JL, QZ and XD contributed to design and supervision of the project. YP and QZ were responsible for writing the manuscript. YP and XD analyzed the RNA data using WGCNA and NA contributed to the collection of the final specimens. All authors have read and approved the final version of this manuscript.

Ethics approval and consent to participate

This study protocol was approved by the institutional medical Ethics Committee of Shenzhen Second People's Hospital (Shenzhen, China). All patients and donors provided informed consent for the molecular analysis of their samples.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


