Identification of hub genes and its correlation with the prognosis of acute myeloid leukemia based on high-throughput data analysis

Wei Fu1 | Guo-bin Cheng1 | Yao Ding1 | Ya-jie Deng1 | Peng-xiang Guo2

1Department of Hematology, 925th Hospital of PLA, Guiyang, China
2Department of Hematology, Guizhou Provincial People’s Hospital, Guiyang, China

Correspondence
Fu Wei, Department of Hematology, 925th Hospital of PLA, Guiyang 550009, P.R. China.
Email: fmmufw@foxmail.com

Funding information
Science Foundation of National Health Commission, Grant/Award Number: gzwjkj-2017-1-017

Abstract
Objective: Acute myeloid leukemia (AML) is one of the most common forms of leukemia in the world, but its molecular mechanism is still not well understood. The aim of this study was, by using multiple AML datasets, to obtain genes with differential expression, and to identify key genes in the development and progression of AML.

Methods: The AML microarray dataset (GSE24395, GSE30029, GSE38865, GSE90062) was downloaded from the Gene Expression Omnibus database. Differentially expressed genes (DEGs) were identified and functionally enriched, and a protein–protein interaction network was constructed. Cytoscape was used for module analysis to obtain fundamental genes (hub genes), and combined with The Cancer Genome Atlas clinical data for survival and expression analysis.

Results: A total of 134 DEGs were identified (fold change > 1, P < 0.01). Gene enrichment analysis included positive regulation of RNA polymerase II promoter transcriptions (c = 17.967, P = 0.011), neutrophil degranulation (c = 18.625, P = 0.017), integrin-mediated cells (c = 17.862, P = 0.017), adhesion regulation, neutrophil-mediated immunity (c = 17.624, P = 0.017), transcriptional disorders in cancer, (c = 14.786, P = 0.031). A total of 16 hub genes were identified. Based on The Cancer Genome Atlas clinical data expression analysis, CYBB (t = 0.368, P = 0.012) and CYFIP2 (t = 2.097, P = 0.038) were abnormally expressed in different living conditions of AML. Survival analysis showed that SERPINE1 (P = 0.031) and ITGAM (P = 0.049) might be involved in invasion or recurrence of AML.

Conclusion: The analysis of DEGs and hub analysis of genes can contribute to understanding the molecular mechanism of AML occurrence and progression, and provides a candidate for the treatment and prognosis of AML.

Keywords
acute myeloid leukemia, high-throughput data, hub genes, prognosis

Acute myeloid leukemia (AML) is a genetically heterogeneous disease characterized by the accumulation of acquired genetic changes in hematopoietic progenitors, altering the mechanisms of self-renewal, proliferation, and differentiation.1,2 Although some progress has been made in understanding the biology of AML, it has been slow to translate this knowledge into a viable treatment. Therefore, an in-depth understanding of the precise molecular mechanisms involved in carcinogenesis, proliferation, and recurrence of AML is essential to provide a basis for developing effective therapeutic strategies and prognostic judgments. Over the past few decades, microarray technology and bioinformatics analysis have been widely used to screen for genetic changes at the genomic level, which helps identify...
differentially expressed genes (DEGs), and those involved in AML carcinogenesis and progression. Functional pathway analysis. However, the false positive rate in independent microarray analysis makes it difficult to obtain reliable results. Therefore, in this study, the Gene Expression Omnibus (GEO) was downloaded and analyzed from the gene expression database. The four mRNA microarray data sets of differential genes between AML and normal tissues take the intersection, and then perform gene ontology (GO), Kyoto Gene and Genome Encyclopedia (KEGG) pathway enrichment analysis and protein-protein interaction (PPI) Network analysis. The analysis results provide an understanding of the molecular mechanism of cancer development and progression.

1 | METHODS

1.1 | Microarray data

Four gene expression datasets – GSE24395, GSE30029, GSE38865, and GSE90062 – were downloaded from GEO, and the probes were converted into corresponding gene symbols based on platform annotation information. The inclusion criteria for the gene set were first studied in human blood samples. The study design was a comparison between AML tumor stem cells and normal hematopoietic stem cells (HSCs), and the relevant literature impact factors cited in the gene set were >10. The sample quality is reliable. The GSE24395 dataset contains 12 AML samples and five normal HSC samples. The gse30029 contains 46 AML samples and 31 normal HSC samples. Gse38865 contains 25 AML samples and five normal HSC samples. Gse90062 contains three AML samples and three normal HSC samples.

1.2 | DEG recognition

DEGs between AML and non-cancer patients can be obtained using GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r). GEO2R is an interactive web tool that allows users to compare two or more datasets in the GEO series to identify DEGs under experimental conditions. First, the differential genes of each dataset were obtained, and then the intersections were obtained, and the differentially expressed genes were found in different gene sets.

1.3 | Hub key core gene screening

The PPI network of DEGs in the present study was constructed using the STRING database, and interacted with the combined interaction of Cytoscape for molecular complex detection (MCODE; version 1.4.2) and CytoHubba, which are applications used to explore important nodes/hubs and fragile motifs in an interactome network. Cytoscape was used to draw a PPI network, and MCODE was used to identify the most important modules in the PPI network. The selection criteria were as follows: MCODE degree cut-off = 2, node score cut-off = 0.2, maximum depth = 100, k score = 2, and CytoHubba selects the top 10.

1.4 | DEGs KEGG and GO enrichment analysis annotation and visualization

Enrichr (https://icahn.mssm.edu/research/labs/maayan-laboratory the Ma’ayan Lab.New York, US) is an online bioinformatics database that combines biological data and analytical tools to aggregate data from 35 databases, providing comprehensive gene and protein functional annotation information. KEGG is a database resource for understanding high-level functional and biological systems from large-scale molecular data generated by high-throughput experimental techniques. GO is a major bioinformatics tool for annotating genes and analyzing the biological processes of these genes. The KEGG and GO analyses of the differential genes in the analysis module were carried out, and the combined scores (c) were ranked in the top 20 (c = log(p · z)). The PPI network was predicted using an online database search tool to retrieve the interacting genes (STRING 10.0; http://string-db.org).

1.5 | Hub gene bioinformatics analysis

The cBioPortal (http://www.cbioportal.org) online gene network and its co-expressed gene analysis platform were used. The biological process analysis of the hub gene was visualized using Cytoscape’s BioNetwork Gene Oncology Tool (BINGO 3.0.3) plug-in. Hierarchical clustering of the hub gene was constructed using the UCSC Cancer Genomics Browser (http://gray-cancer.ucsc.edu). The Kaplan–Meier curve was used in cBioPortal to show the overall survival and disease-free survival analysis of the hub gene. The online database, Oncomine (http://www.oncomine.com), was used to analyze the association of gene expression patterns with AML grades and survival status.

1.6 | Statistical analysis

Differential genetic analysis used the R language package. The adjusted P-value and the Benjamini and Hochberg false discovery rates were used to strike a balance between the discovery of genes and limitations of statistically significant false positives. Probe sets without corresponding gene symbols, or genes with more than one probe set were removed, respectively. LogFC > 1 and adjusted P < 0.01 were considered statistically significant. Enrichr provides an enrichment analysis using Fisher’s exact test or hypergeometric test. The grade score or Z-score was calculated using a modification to Fisher’s exact test. The survival analysis was carried out by cBioPortal Kaplan–Meier, and a log-rank P-value was obtained to compare the overall survival rates of the two groups. The Oncomine TCGA-AML dataset hub gene expression and tumor grade were analyzed using one-way ANOVA. Test level α = 0.05 (both sides).
FIGURE 1 Analysis of differential expression genes Venn diagram, protein interaction network, and cytoscape module analysis. (a) In the mRNA expression profiling set, select differentially expressed genes with multiples >1 and $P < 0.01$, and four datasets contain three sets with overlapping 134 genes. (b) Protein–protein interaction networks of differentially expressed genes were constructed using Cytoscape. (c) The most important module is obtained from the protein–protein interaction network using Molecular Complex Detection and CytoHubba plug-ins. It shows the top 10 genes of important nodes and the red marker node multi-gene.

2 RESULTS

2.1 Identification of DEGs in AML

Figure 1a shows that after the microarray results were normalized, the data was overlaid in four datasets, and the sum of the genes containing the three datasets contained 134 genes.

2.2 DEG, KEG and GO enrichment analysis

The DEG was analyzed by Enrichr, and its function and pathway were analyzed by biological classification and enrichment. The results show that the biological process has changed. DEG significantly enriched the positive regulation of RNA polymerase II promoter transcription ($P = 0.017$), neutrophil degranulation ($P = 0.017$), integrin-mediated cell adhesion negative regulation ($P = 0.017$), integrin-mediated cell adhesion regulation, neutrophil-mediated immunity ($P < 0.01$), cell adhesion negative regulation, membrane lipid catabolism ($P = 0.02$), and so on. (Figure 2). Changes in molecular functions were shown, mainly in RNA polymerase II regulatory region sequence-specific DNA binding ($P = 0.011$), phospholipid anti-enzyme activity, transcription factor activity ($P = 0.011$), Toll-like receptor binding, protein homology source dimerization activity ($P = 0.011$), acyl-CoA oxidase activity, folate receptor binding, and CD4 receptor binding (Figure 2). KEGG pathway analysis showed that DEGs were mainly enriched in transcriptional disorders in cancer ($P = 0.031$), leukocyte transendothelial migration, complement and coagulation, HIF-1 signaling pathway, and p53 signaling pathway.

2.3 PPI network construction and module analysis

The PPI built the DEG network (Figure 1b) using Cytoscape to obtain a significant module (Figure 1c). GO enrichment was used to analyze the functional analysis of the genes involved in the module (Figure 3). The results showed that the gene in this module mainly involved abnormal regulation of tumors, migration of leukocytes across endothelial cells, regulation of RNA polymerase II, activation of transcription factors, binding, specific DNA sequences, and so on. (Figure 3).

2.4 Hub gene selection and analysis

A total of 16 genes were identified as the hub gene A network of hub genes, and their co-expressed genes were analyzed using the cBioPortal online platform (Figure 4a). The biological process analysis of
the hub gene is shown in Figure 4b. Hierarchical clustering showed that the Hub gene can substantially distinguish between AML samples and healthy groups (Figure 4c). Subsequently, overall survival was carried out using Kaplan–Meier for survival analysis of the hub gene (Figure 5). Using data from cBioPortal, it was noted that patients with a genomic alteration association in CYFIP2, HMOX1, CDC42, and CDKN1A showed a reduction in overall and disease-free survival; however, the differences in these observations were not statistically
FIGURE 4  Hub gene interaction network, Gene Ontology biological process association analysis, and hierarchical cluster analysis. (a) Using the cBioPortal to analyze the hub gene and its co-expressed genes, the nodes with bold black outlines represent the hub gene. Nodes with a thin black outline represent co-expressed genes. (b) Bioprocess analysis of building hub genes using Cytoscape analysis of building expressions. (c) Hierarchical clustering of hub genes using UCSC, upregulation of genes is marked in red; downregulation of genes is marked in blue.

FIGURE 5  Hub gene cBioPortal survival analysis. (a) Overall survival of the hub gene using the cBioPortal online platform. (b) Disease-free survival analysis.

significant. AML patients with ITGAM changes showed worse overall survival, suggesting a role in the development or progression of AML, $P = 0.049$. AML patients with changed SERPINE1 showed an improvement in disease-free survival, $P = 0.031$ (Figure 5).

Oncomine analysis of cancer and normal tissues showed that ATF3, KLF6, and FHL2 were overexpressed in AML in different datasets (Figure 6). In the TCGA-AML dataset, ATF3 ($f = 6.936$, $P = 0.001$), CYBB ($f = 5.797$, $P = 0.004$), CDKN1A ($f = 11.69$, $P < 0.001$), and
SERPINE1 ($t = 4.628, P = 0.011$) showed higher mRNA levels, which were associated with tumor grade, and CYBB ($t = 0.368, P = 0.012$) and CYFIP2 ($t = 2.097, P = 0.038$) were correlated with disease survival (Figure 7).

3 | DISCUSSION

In the USA, there are 19,000 new cases per year of AML, which is one of the common hematological diseases. There are currently stratified
treatments based on karyotype and mutation, but most of the newly diagnosed AML subtypes have no major changes in the standard treatment regimen.21,22 Until recently, most AML patients had no revolutionary chemotherapy regimen.23,24 Given that the long-term prognosis of most adult AML patients is poor, especially in older AML patients, there are many new drugs being developed, combined with new molecular targeted therapies, which are superior to traditional chemotherapy alone.25–27

In the past 10 years, with the development of sequencing and chip technology, people have better understood the genetic drivers of AML and prognostic factors.29,30 Although the new research has promoted the treatment and prognosis of AML, it still has not reached expectations. Therefore, there is an urgent need to find potential markers of treatment and prognosis.

The present study analyzed four mRNA microarray datasets to obtain DEGs between tissue and non-cancerous tissues. A total of 134 DEGs were identified. GO and KEGG enrichment analysis was carried out to explore the interaction between DEGs. Differential genes are mainly enriched for abnormal regulation of tumors, regulation of RNA polymerase II, activation of transcription factors, binding to specific DNA sequences, and the like. The obtained tumor abnormality regulating genes (CDKN1A, ITGAM, PPARG, TCF3, PTK2, PBX1, and ETV5) suggest an important role in the development of AML disease. GO enrichment analysis showed that the most important module changes focused on the positive regulation of RNA polymerase II promoter transcription, integrin-mediated cell adhesion negative regulation, integrin-mediated cell adhesion regulation, negative regulation of cell adhesion, membrane lipid catabolism, and so on. In combination with recent studies, the NR4A chemosensitizer activates the elongation of the promoter-suppressed RNA polymerase II by recruiting transcriptional elongation complex components to achieve tumor suppressive effects of AML.31 Cell adhesion and migration are closely related to the spread of tumors. In short, all these theories are consistent with the results of the present study.

This study screened 16 genes as hub genes. Among these Hub genes, JUN and TLR2 showed the highest degree of node, indicating that they might play an important role. Among them, studies have suggested that reactive oxygen species activate the differentiation gene transcription of acute myeloid leukemia cells through the JNK/c–JUN signaling pathway. Eriksson et al. also found that TLR2 is abnormally highly expressed in AML, and targeting TLR2 can produce a tumor suppressor effect on AML.23 This study evaluated the overall and disease-free survival of ITGAM and SERPINE1 expression. Changes in serpine1 showed an increase in disease-free survival, whereas changes in ITGAM showed a significant correlation with a poor overall survival rate, but not with a disease-free survival rate. It is speculated that the results of the survival analysis are inconsistent – the overexpression of the relevant gene is usually caused by a mutation or amplification, and the survival analysis of cBioPortal is based on the relationship between gene mutation and prognosis. Oncomine analysis showed that ATF3, CYBB, SERPINE1, and CDKN1A mRNA levels were associated with tumor grade, and CYBB and CYFIP2 were associated with disease survival, and both analyses showed the important role of CYBB in AML. However, studies have shown that high levels of CYBB expression are associated with long-term survival of AML, whereas CYFIP2 (lowly expressed in both analyses, and statistically significant) has not been reported. In addition, the hub gene was hierarchically clustered.34 The results show that these hub genes distinguish AML samples from non-cancer samples, suggesting that they might be candidate biomarkers.

In summary, the present study aimed to identify DEGs that might be involved in AML carcinogenesis or progression. A total of 134 DEG and 16 hub genes were identified, which can be used as a potential biomarker for the diagnosis and prognosis of AML. The innovation of this study was using different gene sets to obtain the hub gene, to eliminate the bias of individual analysis, and to analyze the effects of these genes in multiple dimensions (including enrichment analysis, cluster analysis, survival analysis, expression analysis, etc.). However, the in-depth biological functions of these genes in AML require clarification through further research.

ACKNOWLEDGMENTS

The author Supported by Science Foundation of National Health Commission in Guizhou province (gzwjkj-2017-1-017).

CONFLICT OF INTEREST

The authors declare that they have read the article and there are no competing interests.

ORCID

Wei Fu https://orcid.org/0000-0001-8155-431X

REFERENCES

1. Patel JP, Gonen M, Figueroa ME, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia[J]. N Engl J Med. 2012;366(12):1079-1089.
2. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia[J]. N Engl J Med. 2016;374(23):2209-2221.
3. Handschuh L, Kaczmierczak M, Milewski MC, et al. Gene expression profiling of acute myeloid leukemia samples from adult patients with AML-M1 and -M2 through boutique microarrays, real-time PCR and droplet digital PCR[J]. Int J Oncol. 2018;52(3):656-678.
4. Wouters BJ, Lowenberg B, Delwel R. A decade of genome-wide gene expression profiling in acute myeloid leukaemia: flashbacks and prospects[J]. Blood. 2009;113(2):291-298.
5. Wilson CS, Davidson GS, Martin SB, et al. Gene expression profiling of acute myeloid leukemia identifies novel biologic clusters for risk classification and outcome prediction[J]. Blood. 2006;108(2):685-696.
6. Grimwade D, Haferlach T. Gene-expression profiling in acute myeloid leukemia[J]. N Engl J Med. 2004;350(16):1676-1678.
7. Clough E, Barrett T. The gene expression omnibus database[J]. Methods Mol Biol. 2016;1418:93-110.
8. Huntley RP, Sawford T, Mutowo-Meullenet P, et al. The GOA database: gene ontology annotation updates for 2015[J]. Nucleic Acids Res. 2015;43(Database issue):D1057-1063.
9. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D361.
10. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update[J]. Nucleic Acids Res. 2013;41(Database issue):D991-995.
11. Demchak B, Otasek D, Pico AR, et al. The cytoscape automation app article collection[J]. F1000Res. 2018;7:800.
12. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC Bioinformatics. 2003;4:2.
13. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update[J]. Nucleic Acids Res. 2016;44(W1):W90-97.
14. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Res. 2017;45(D1):D353-D361.
15. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life[J]. Nucleic Acids Res. 2015;43(Database issue):D447-452.
16. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal[J]. Sci Signal. 2013;6(26):pl1.
17. Maere S, Heymans K, Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks[J]. Bioinformatics. 2005;21(16):3448-3449.
18. Goldman M, Craft B, Swatloski T, et al. The UCSC cancer genomics browser: update 2015[J]. Nucleic Acids Res. 2015;43(Database issue):D812-817.
19. Cline MS, Craft B, Swatloski T, et al. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser[J]. Sci Rep. 2013;3:2652.
20. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles[J]. Neoplasia. 2007;9(2):166-180.
21. Kadia T, Kantarjian H, Garcia-Manero G, et al. Prognostic significance of the medical research council cytogenetic classification compared with the European LeukaemiaNet risk classification system in acute myeloid leukaemia[J]. Br J Haematol. 2015;170(4):590-593.
22. Grimwade D, Hills RK, Moorman AV, et al. Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the united kingdom medical research council trials[J]. Blood. 2010;116(3):354-365.
23. Kobayashi Y. Current diagnosis and treatment for acute myeloid leukemia[J]. Rinsho Ketsueki. 2017;58(4):364-372.
24. Roboz GJ. Current treatment of acute myeloid leukemia[J]. Curr Opin Oncol. 2012;24(6):711-719.
25. O’Donnell MR, Tallman MS, Abboud CN, et al. Acute myeloid leukemia, version 3.2017. nccn clinical practice guidelines in oncology[J]. J Natl Compr Canc Netw. 2017;15(7):926-957.
26. O’Donnell MR, Abboud CN, Altman J, et al. NCCN clinical practice guidelines acute myeloid leukemia[J]. J Natl Compr Canc Netw. 2012;10(8):984-1021.
27. Duer MBA, Hong L, Wuer GZLNW, et al. Therapeutic effect of chemotherapy combined with micro-transplantation on elderly patients with acute myeloid leukemia [j]. Chinese Journal of Cancer Prevention and Treatment. 2018;25(51):160-161.
28. McKerrell T, Park N, Moreno T, et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis[J]. Cell Rep. 2015;10(8):1239-1245.
29. Jianping J, Fang Y, Liang Z, et al. Combining multi-omics and human protein interaction networks to study the interplay network of mutant genes in acute myeloid leukemia[j]. Journal of Fudan University(Natural Science). 2015;54(4):459-468. 541.
30. Lindley RC, Mar BG, Mazzola E, et al. Acute myeloid leukemia ontogeny is defined by distinct somatic mutations[J]. Blood. 2015;125(9):1367-1376.
31. Boudreaux SP, Duren RP, Call SG, et al. Drug targeting of NR4A nuclear receptors for treatment of acute myeloid leukemia[J]. Leukemia. 2019;33(1):52-63.
32. Lam CF, Yeung HT, Lam YM, et al. Reactive oxygen species activate differentiation gene transcription of acute myeloid leukemia cells via the JNK/c-JUN signaling pathway[J]. Leuk Res. 2018;68:112-119.
33. Eriksson M, Pena-Martinez P, Ramakrishnan R, et al. Agonistic targeting of TLR1/TLR2 induces p38 MAPK-dependent apoptosis and NFkappaB-dependent differentiation of AML cells[J]. Blood Adv. 2017;1(23):2046-2057.
34. Garcia-Manero G, Tambaro FP, Bekele NB, et al. Phase II trial of vorinostat with idarubicin and cytarabine for patients with newly diagnosed acute myelogenous leukemia or myelodysplastic syndrome[J]. J Clin Oncol. 2012;30(18):2204-2210.

How to cite this article: Fu W, Cheng, G-b, Ding Y, Deng Y-j, Guo P-x. Identification of hub genes and its correlation with the prognosis of acute myeloid leukemia based on high-throughput data analysis. Prec Radiat Oncol. 2020;4:49–56. https://doi.org/10.1002/pro6.1089