The core genome M^5C plays an important role in methylation modification and immune infiltration of acute myelocytic leukemia samples

Qiang Wen

Cancer Hospital of University of Chinese Academy of Sciences, Zhejiang Province, China

ShouJun Wang

HangZhou FuYang Hospital of Traditional Chinese Medicine, Zhejiang Province, China

Lili Hong

The First Affiliated Hospital of Zhejiang Chinese Medical University, Hang Zhou, China

Xianfu Sheng

The First Affiliated Hospital of Zhejiang Chinese Medical University, Hang Zhou, China

Xiaofen Zhuang

HangZhou FuYang Hospital of Traditional Chinese Medicine, Zhejiang Province, China

Yu Chen

Hangzhou Medical College, Hang Zhou, China

haifeng zhuang (✉ zhuanghaifeng5@163.com)

The First Affiliated Hospital of Zhejiang Chinese Medical University, Hang Zhou, China

Research Article

Keywords: Bioinformatics, Acute myelocytic leukemia, CNV, M^5C , TCGA

Posted Date: January 4th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-135560/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Currently, the pathogenesis of acute myelocytic leukemia(AML) is still unclear. We found the core genome m5C plays a vital role in methylation modification and immune infiltration of AML. At the same time, we created a new m5C score model to define the high and low-risk groups of AML. Our research showed the expression levels of the three molecular subtypes of m5C (C1, C2 and C3); as well as different clinical features, the results showed significant differences in age\RUNX1-RUNX1T1 fusion, but not in RUNX1 mutation group. We constructed a prognostic risk model based on m5C phenotype from 417 samples in the GSE37642 data set and found 5 differential genes using lasso regression method. And the prognostic KM curve of the 5-gene signature was obtained, from which it can be seen that: all the five genes could significantly reduce the high and low risk of GSE37642 training set samples ($P < 0.05$). Finally, the robustness of m5C related 5-gene signature for AML prediction was verified by internal and external data using single factor and multifactor COX regression analysis. 5-gene signature has strong robustness and can play a stable prediction performance in external validation data sets (GSE12417, TCGA-LAML).

Introduction

AML is one of the most common hematological malignancies in adults, which is highly heterogeneous, and the biological characteristics of AML originated from different stages of progenitor cells are not the same(1). Physical factors (such as X-ray, gamma ray and other ionizing radiation), chemical factors (such as occupational benzene exposure, long-term use of alkylating agent, etc.), and genetic factors are all the high-risk factors of AML(2). However, how these factors lead to the occurrence of AML, and its specific pathogenesis are not fully understood.

There are more than 100 chemical modifications carried by RNA in eukaryotes, about 60% of which are RNA methylation modification(3). Among them, m5C is the most common methylation modification, and highly abundant and stable in tRNA and rRNA. Methylation of m5C RNA plays an important role in regulating total protein synthesis and cell fate(4). Therefore, activation of RNA methylation or inhibition of tRNA cleavage is essential for survival of tumor-initiating cells in response to cytotoxic stress(5). Although m5C has been found to be associated with the development of different types of tumors, its relationship with AML is poorly understood. In the process of this study, we hope to investigate the role of m5C methylation-related genes in acute leukemia, by creating a new m5C-score model to define high and low risk groups for AML, which shed more lights on the AML prognostic mechanism at the molecular level.

Materials And Methods

Data extraction and processing

All the data we used in our study are publicly accessible at TCGA and NCBI GEO (accession number: GSE37642, GSE12417) database (search terms: TCGA-LAML). The UCSC cancer browser ([https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Acute%20Myeloid%20Leukemia%20\(LAML\)&removeHub=https%3A%2F%2Fxcna.treehouse.gi.ucsc.edu%3A443](https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Acute%20Myeloid%20Leukemia%20(LAML)&removeHub=https%3A%2F%2Fxcna.treehouse.gi.ucsc.edu%3A443)) was also utilized to download CNV, clinical follow-up, TCGA RNA-Seq and SNP 6.0 chip data. The download date is November 3, 2020. Additionally, mutation annotation file (MAF) was collected based on the GDC client, GSE37642(6) and GSE12417(7) expression patterns, and clinical follow-up data were obtained from GEO database.

417 AML samples who had sufficient follow-up data were screened for TCGA RNA-Seq data, and randomized as two groups, namely, test set (n=209) and training set (n=208); in addition, both GSE37642 (n=417) and GSE12417 (n=163) data sets were adopted to be external verification sets(Figure1). Table 1 shows the details of samples of every group. For GEO data processing, we first download the mini format files from the GEO platform. According to the background file, the probe ID is converted to the gene symbol. The average value of multiple genes corresponding to a single probe was calculated, and the probes correspond to multiple genes were eliminated. Then, the expression spectrum matrix was further normalized.

Univariate Cox proportional hazard regression analysis

Univariate Cox proportional hazard regression analysis was conducted in order to identify genes whose expression levels were markedly correlated with patient overall survival (OS) in training set, at the threshold of $P < 0.01$.

Analysis of CNV data

GISTIC has been extensively adopted for detecting focal and broad (probably overlapping) recurrent events(8). Consequently, the GISTIC 2.0 was adopted for CNV data from TCGA, so as to found the significantly deleted or amplified genes, at the thresholds of $p < 0.05$ and

fragments that had >0.1 deletion or amplification length.

Analysis of gene mutation

For identifying genes with significant mutation, Mutsig 2.0(9) software was used for recognizing those genes with significant mutation based on the MAF of TCGA mutation data, at the threshold of $P < 0.05$.

Prognosis-related gene signature construction

First of all, the lasso cox regression was adopted for refining those above-identified prognostic genes using the glmnet function of R package(10). Secondly, the MASS function of R package was utilized to carry out stepwise regression analysis in accordance with the Akaike information criterion for obtaining the eventual 5-gene risk model. Typically, the formula was shown below:

$$\text{RiskScore} = -0.2283059 \times \text{ITGA4} - 0.1575680 \times \text{IGLL1} - 0.2686156 \times \text{LAPTM4B} - 0.1220958 \times \text{HIST1H2AE} - 0.1472148 \times \text{HOPX}$$

Then, the risk score values were z-score normalized, and samples with the processed z-score value of >0 were classified as the high-risk group, while those of <0 were the low-risk group.

Functional enrichment analyses

The cluster Profiler (v3.8.1)(11) was adopted to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) pathway enrichment analyses on genes, so as to recognize those enriched KEGG pathways and GO terms among the three categories, including cellular component (CC), molecular function (MF), and biological processes (BP). Notably, the false discovery rate (FDR) value of <0.05 indicated statistical significance. The expression matrix of genes between different samples was first converted into the expression matrix of gene sets, to evaluate which metabolic pathways were enriched. The correlations between the risk score and pathways were further calculated using Pearson correlation analysis. Signaling pathways with correlation coefficient of >0.35 were considered to be related to the risk score.

Statistical methods

The median risk score of every data set was adopted to be the threshold to plot the Kaplan-Meier (KM) curves, and then the survival risks were compared in high-risk group with those in low-risk group. On the other hand, the feasibility of using the gene markers as the factors to independently predict prognosis was examined through multivariate Cox regression analysis. $P < 0.05$ indicated statistical significance. The R version 3.6.0 was adopted for all statistical analyses.

Results

Identification and functional analysis of m5C modified isoforms

Determination of m5C modified subtypes

Firstly, we extracted the expression levels of m5C regulatory factors from the geo expression profile matrix, where nsun2, nsun4, tet2 and alyref genes did not exist, so we finally extracted the expression profiles of 9 m5C genes for subsequent analysis. Then, consensusclusterplus (v1.48.0; parameters: reps = 100, pitem = 0.8, pfeature = 1, distance = Euclidean ") was used for consistent clustering. Km and Euclidean distance were used as clustering algorithm and distance measure respectively; k = 3 was selected as the optimal clustering according to CDF value and delta area (Figure 2A-C). We further analyzed the prognostic relationship among the three groups, and the results showed that C1, C2, and C3 were significantly different (Figure 2D, log rank, $P < 0.0001$).

Relationship between m5C modified subtypes and clinical features

We compared the distribution of different clinical features in the three molecular subtypes to see whether clinical features differed in the different subtypes. The results showed that: 1) C1, C2, C3 had significant difference in survival rate; 2) C2 and C3 had significant difference in age group and runx1-runx1t1 fusion group; 3) C1, C2, C3 had no significant difference in Runx1 mutation group.(Figure 3A-D)

Correlation analysis of immune infiltration of m5C modified subtypes

To identify the relationship between the immune cell scores of molecular subtypes, we used mcpcounter to evaluate the scores of 10 kinds of immune cells and the ssgsea method of gsva package to evaluate the scores of 28 kinds of immune cells (cell markers from references)

pmid:28052254 [1]. And then we compared the difference of immune score in molecular subtypes. The results are shown in Figure 4A-B. Meanwhile, we analyzed the expression of 9 m5c modification-related genes in three subtypes (Figure 4C). It can be seen from the figure that there are significant differences in the expression of m5C gene in different subtypes.

Analysis of differentially expressed genes among m5C modified subtypes

Identification of differentially expressed genes

The DEGs between C1~C3, C2~C1 and C2~C3 were calculated by limma package, and filtered according to the threshold $FDR < 0.05$ and $|\log_2 fc| > \log_2(2)$. The volcano diagram of C1~C3 up-regulated and down-regulated differentially expressed genes was shown in Figure 5A, including 3 up-regulated genes and 45 down-regulated genes; There were 133 up-regulated and 28 down-regulated differentially expressed genes in Figure 5B from C2~C1. The results showed that the main up-regulated differential expression genes were between C1~C2; the volcano diagram of C2~C3 up-regulated and down-regulated differentially expressed genes was shown in Figure 5C, of which 21 were up-regulated and 52 were down-regulated. The main differential expression genes were down-regulated. The detailed differential expression genes were found in S1.xlsx.

Functional analysis of differentially expressed genes among m5C molecular subtypes

Furthermore, we used webgestalt (v0.4.2) to perform the KEGG pathway and GO functional enrichment analysis of 230 differentially expressed genes among C1~C3, C2~C1 and C2~C3 molecular subtypes. For the GO functional annotation of AML differential genes, 248 ($P < 0.01$) with significant difference in BP were annotated. The annotation results of the first 15 genes were shown in Figure 6A; 29 items ($P < 0.01$) with significant difference from CC, and the results of the first 15 items were shown in Figure 6B; 26 items ($P < 0.01$) with significant difference from MF were noted. The results of the first 15 items were shown in Figure 6C; See S2,CSV for more details. For AML differential gene KEGG pathway enrichment, annotated to a significant pathway ($P < 0.05$).

Construction of prognostic risk model based on m5C phenotype correlation

Random grouping of training set samples

Firstly, 417 samples in GSE37642 data set were divided into training set and validation set. The final training set data was 208 samples, and the validation set data was 209 samples. Finally, training set and validation set of GSE37642 data were shown in Table 2. Chi square test was used to test the training set and validation set samples. The results showed that our group had no preference and there was no significant difference between groups ($P > 0.05$).

Single factor risk analysis of training set

Using training set data, univariate Cox proportional risk regression model was performed for differentially expressed genes between subtypes (230 in total) as well as survival data using the R package survival coxph function. $P < 0.01$ was selected as the threshold value for filtering, with a total of 18 genes with differences at the end. The univariate Cox analysis results were shown in S3.txt.

Multi factor risk analysis of training set

At present, 18 differentially related genes have been identified. Lasso regression was used to further compress the 18 genes to reduce the number of genes in the risk model. We used the R software package glmnet to carry out lasso Cox regression analysis. First, we analyze the change track of each independent variable, as shown in Figure 7A. We can see that with the gradual increase of lambda, the number of independent variable coefficient tending to 0 also gradually increased. We used 10-fold cross validation to build the model, and analyzed the confidence interval under each lambda, as shown in Figure 7B, from which we can see that when $\lambda = 0.04845916$, the model is optimal. For this purpose, we selected 9 genes at $\lambda = 0.04845916$ as the target genes for the next step.

The step method in stats package starts from the most complex model and removes one variable in turn to reduce AIC. The smaller the value, the better the model. It shows that the model has enough fitting degree with fewer parameters. Using this algorithm, we finally reduced 9 genes to 5 genes, which are ITGA4, IGLL1, LPTM4B, HIST1H2AE and HOPX.

The prognostic KM curve of the five genes is shown in Figure 8. It can be seen that all the five genes can significantly reduce the high and low risk of GSE37642 training set samples ($P < 0.05$). The final 5-gene signature formula is as follows:

$$\text{RiskScore} = -0.2283059 \times \text{ITGA4} - 0.1575680 \times \text{IGLL1} - 0.2686156 \times \text{LPTM4B} - 0.1220958 \times \text{HIST1H2AE} - 0.1472148 \times \text{HOPX}$$

Construction and evaluation of risk model

We calculated the risk score of each sample according to the expression level of the samples, and drew the risk score distribution of the sample, as shown in Figure 9A. It can be seen from the figure that the death rate of the samples with high risk score is significantly greater than that of those with low score, which indicates that the high risk score samples have a worse prognosis. The changes in the expression of five different signature genes with increasing risk values identified that high expression of LAPTM4B, HIST1H2AE and HOPX were associated with high risk. They were risk factors, and ITGA4 and IGLL1 were protective factors.

Further, we used the R software package timeroc to analyze the prognosis classification of riskscore. We analyzed the classification efficiency of one year, three years and five years respectively, as shown in Figure 9B. Finally, we performed zscale on riskscore, and classified the samples with riskscore greater than zero as high-risk group and those with less than zero as low-risk group. The KM curve is drawn, as shown in Figure 9C. It can be seen that there is a significant difference between them ($P < 0.0001$). 106 samples are classified as high-risk group and 102 samples as low-risk group.

Verification of risk model

Internal data set to verify the robustness of 5-gene signature

In order to determine the robustness of the model, we used the same model and the same coefficient as the training set in GSE37642, calculate the riskscore of each sample according to the expression level of the samples, and draw the riskscore distribution of the samples.

The riskscore distribution of GSE37642 validation set is shown in Figure 10A. It can be seen from the figure that the proportion of death in samples with high RiskScore is significantly higher than that of samples with low RiskScore, which is consistent with the performance of GSE37642 training set. Further, we used the R software package timeroc to analyze the prognosis classification of RiskScore, and analyzed the classification efficiency of one year, three years and five years, as shown in Figure 10B. Finally, we performed zscale on riskscore, and classified the samples with riskscore greater than zero as high-risk group and those with less than zero as low-risk group. The KM curve is drawn, as shown in Figure 10C. It can be seen that there is a very significant difference between them ($P < 0.0001$). 105 samples are classified as high-risk group and 104 samples as low-risk group.

The RiskScore distribution of GSE37642 data set is shown in Figure 11A. It can be seen from the figure that the proportion of death in the samples with high RiskScore is significantly higher than that with low RiskScore, which is consistent with the performance of TCGA training set. Further, we used the R software package timeROC to analyze the prognosis classification of high and low risk groups of RiskScore. We analyzed the classification efficiency of prognosis prediction in one year, three years and five years, as shown in Figure 11B. Finally, we performed zscale on RiskScore, and classified the samples with RiskScore greater than zero as high-risk group. The KM curve is drawn for samples less than zero, as shown in Figure 11C. It can be seen that there is a very significant difference between them ($P < 0.0001$). 127 samples are classified as high-risk group and 130 samples as low-risk group.

External data sets to verify the robustness of 5-gene signature

We use the same model and the same coefficient as the training set in the external validation data sets GSE12417 andTCGA-LAML, and then calculate theRiskScore of each sample according to the expression level of the samples, and draw the RiskScore distribution of the samples.

The RiskScore distribution of independent external validation dataset of GSE12417 is shown in Figure 12A. It can be seen from the figure that the proportion of death in the samples with high RiskScore is significantly higher than those with low RiskScore, which is consistent with the performance of GSE37642 training set. Further, we used the R software package timeROC to analyze the prognosis classification of RiskScore. We analyzed the classification efficiency of one year, two years and three years, as shown in Figure 12B. Finally, we performed zscale on RiskScore, and classified the samples with riskscore greater than zero as high-risk group and those with less than zero as low-risk group. The KM curve is drawn, as shown in Figure 12C. It can be seen that there is a significant difference between them ($P < 0.001$). 81 samples are classified as high-risk group and 82 samples as low-risk group.

The RiskScore distribution of independent external validation data set of TCGA-laml is shown in Figure 13A. It can be seen from the figure that the proportion of death in the samples with high RiskScore is significantly higher than that with low RiskScore, which is consistent with the performance of GSE37642 training set. Furthermore, we used the R software package timeROC to analyze the prognosis classification of RiskScore. We analyzed the classification efficiency of one year, three years and five years, as shown in Figure 13B. Finally, we performed zscale on RiskScore, and classified the samples with RiskScore greater than zero as high-risk group and those with less than

zero as low-risk group . The KM curve is drawn, as shown in Figure 13C, from which we can see that there is a very significant difference between them ($P < 0.001$). Among them, 68 samples are classified as high-risk group and 72 samples as low-risk group.

Risk model and prognosis analysis of clinical features

Further, we conducted the RiskScore analysis of 5-gene markers, and found that 5-gene signature could significantly distinguish the high and low risk groups by age, RUNX1-RUNX1T1 fusion and RUNX1 mutation (Figure 14A-F, $P < 0.05$). This further indicates that our model still has good predictive ability in different clinical signs.

The expression of risk score in different clinical characteristics and different molecular subtypes

By comparing the distribution of RiskScore among clinical characteristics groups, we found that there were significant differences in age, RUNX1-RUNX1T1 fusion and RUNX1 mutation (Figure 15A-D, $P < 0.05$). At the same time, we compared the difference of risk scores in molecular subtypes. The results showed that the risk score of C2 subtype with a poorer prognosis was significantly higher than that of C3 subtype with a better prognosis.

The relationship between RiskScore and biological function

In order to observe the relationship between the risk score and biological function of different samples, we selected the gene expression profile corresponding to these samples, used R software package gsva to conduct single sample GSEA analysis, and calculated the ssgsea scores of each sample for each function. We further calculated the correlation between these functions and risk score, and selected the function with correlation greater than 0.3, as shown in Figure 16A-F, from which we can see the junction pathway of KEGG_ ABC_ TRANSPORTERS, KEGG_ TIGHT_ was positively related with the sample's risk score, while KEGG_ NON_ SMALL_ CELL_ LUNG_ CANCER, KEGG_ Glioma and KEGG_ PYRIMIDINE_ were negatively related.

Single factor and multi factor analysis of 5-gene signature

In order to identify the independence of 5-gene signature model in clinical application, we used univariate and multivariate Cox regression to analyze the relevant HR, 95% CI of HR and P value in the clinical information carried by GSE37642 data. We systematically analyzed the clinical information of GSE37642 patients, including age, RUNX-RUNX1T1 fusion, RUNX1 mutation and riskscore (Figure 17).

In the GSE37642 dataset, univariate Cox regression analysis found that riskscore was significantly associated with survival, and the corresponding multivariate Cox regression analysis found the same result (HR = 1.57, 95% CI = 1.38 – 1.79, $P < 1e-5$).

The above results show that our model 5-gene signature has good predictive performance in clinical application value.

Construction of nomogram by riskscore and clinical features

According to the results of univariate and multivariate analysis, we constructed nomographic model with clinical features age, Runx1 mutaion and riskscores. We used all data sets of GSE37642 to construct nomogram (Figure 18A). From the model results, the risk score feature has the greatest impact on the survival rate prediction, indicating that the 5-gene based risk model can better predict the prognosis. Furthermore, we use the calibration curve to evaluate the prediction accuracy of the model, such as Figure 18B. We can observe that the predicted calibration curves of 1, 3 and 5 years nearly overlapped with the standard curve, which indicates that the model has good predictive performance. In addition, we also used DCA (decision curve) to evaluate the reliability of the model, such as Figure 18C. It can be observed that the benefits of riskscore and nomogram are significantly higher than those of extreme curve, where nomogram is higher than riskscore, and age and Runx1 mutaion are close to the extreme curve. The result indicates that riskscore and nomogram have good reliability.

Discussion

AML is a malignant clonal and proliferative disease derived from myeloid hematopoietic stem / progenitor cells(12). It is characterized by abnormal proliferation of primitive and immature myeloid cells and ineffective hematopoiesis of bone marrow. Its clinical manifestations are anemia, hemorrhage, infection, fever, organ infiltration and metabolic abnormalities(13). Physical factors, chemical factors, and genetic factors are all the high-risk factors of AML(14). However, how these factors lead to the occurrence of AML and its specific pathogenesis are not fully understood.

m5C is highly abundant and stable in tRNA and rRNA. At present, there are three types of genes related to m5C, which are: 1) Writers: methylase, which mediates RNA methylation modification, including NSUN1, NSUN2, NSUN3, NSUN4, NSUN5, NSUN6, NSUN7, DNMT1, DNMT2, Dnmt3a and DNMT3b. 2) Erasers: demethylase, including TET2. 3) Readers: m5C Binding protein recognizes and binds to the m5C site of mRNA, including ALYREF(15). "Writer" formed a methyltransferase complex and enhanced the level of m5C, while "eraser" was an m5C demethylase, which was opposite to "writer". In addition, the "reader" is a decoding effector, which converts the m5C methylation information into a functional signal(16).

Post transcriptional methylated tRNA of nsun2 on cytosine-5 (m5C) inhibited total protein synthesis(17). Methylation of m5C RNA plays an important role in regulating the total protein synthesis and cell fate(18). Therefore, activation of RNA methylation or inhibition of tRNA cleavage is essential for survival of tumor initiating cells in response to cytotoxic stress(19). Although m5C has been found to be associated with the development of different types of tumors, its relationship with AML is poorly understood(20).

In the process of this study, we designed a research protocol. In the first step, the expression level of m5C regulatory factor was extracted from the GEO expression profile matrix, and the results showed that the expression level of m5C regulatory factor was higher than that of m5C regulatory gene. The results showed that: 1) C1, C2, C3 had significant difference in survival rate; 2) C2 and C3 had significant difference in age group and RUNX1-RUNX1T1 fusion group; 3) C1, C2 and C3 were significantly different in RUNX1-RUNX1T1 fusion group; 3) C1, C2 and C3 were significantly different in RUNX1. There was no significant difference in mutation group. It can be seen that the gene expression of C1, C2 and C3 subtypes in m5C molecule is different from each other, and it is related to the clinical characteristics of patients.

In the second step, we used mcpcounter to evaluate the scores of 10 immune cells and GSVA package of ssgsea method for 28 immune cells; From our research data, we can see that M5c related genes can infiltrate into AML, and then affect the function of immune cells. However, the specific mechanism needs to be further verified by cell experiments and animal experiments. At the same time, we also analyzed the expression of 9 m5c modification-related genes in three subtypes. DEGs after interaction between C1-C3, c2-c1 and C2-C3 molecular subtypes were calculated by limma package. The results showed that the major down-regulated differential expression was mainly between C1 / C3, the main upregulation differential expression between C1 / C2, and the main downregulation difference between C2 / C3. Furthermore, we analyzed the KEGG pathway and go function enrichment of 230 differentially expressed genes among C1-C3, c2-c1 and C2-C3 molecular subtypes by R software package webgestalt (v0.4.2). Among them, 248 genes with significant difference in BP ($P < 0.01$) and 29 genes with significant difference in CC ($P < 0.01$) were annotated < 26 items with significant difference in MF ($P < 0.01$). For AML differential gene KEGG pathway enrichment, annotated to a significant pathway ($P < 0.05$). The results showed that the changes of immune cells of three subtypes of m5C were different from each other, and the target cells of m5C were different.

Finally, 417 samples from GSE37642 dataset were used to construct a prognostic risk model based on m5C phenotype. A total of 18 genes with significant differences were selected and further compressed by lasso regression to reduce the number of genes in the risk model. Tibshirani (1996)) method is a kind of compressed estimation. By constructing a penalty function, a more refined model is obtained, which makes it compress some coefficients and set some coefficients to zero(21). Therefore, it retains the advantage of subset contraction, and it is a biased estimation to deal with multicollinearity data, which can achieve the selection of variables and parameter estimation at the same time, and better solve the multicollinearity problem in regression analysis. With this algorithm, we finally reduce the number of genes to five, which are ITGA4, IGLL1, LAPTM4B, HIST1H2AE and HOPX. According to the prognosis KM curve of the five genes, it can be seen that all of the five genes could significantly reduce the high and low risk of GSE37642 training set samples ($P < 0.05$); the high expression of LAPTM4B, HIST1H2AE and HOPX were identified as risk factors, and ITGA4 and IGLL1 were protective factors. Finally, through the internal and external data, the robustness of 5-gene signature was verified; further, we conducted the risk score analysis of 5-gene markers, and found that 5-gene signature could significantly distinguish the high-risk group from the high-risk group by grouping age, RUNX1-RUNX1T1 fusion and RUNX1 mutation. At the same time, we compared the difference of risk scores in molecular subtypes. The results showed that the risk score of C2 subtype with a poorer prognosis was significantly higher than that of C3 subtype with a better prognosis. We have developed a 5-genes signature prognostic stratification system, which has good AUC in both training set and independent validation set, and it is a model independent of clinical characteristics. Nonetheless, certain limitations must be noted. First of all, some samples did not have complete clinical follow-up data, so it was impossible to examine the feasibility of using these biomarkers in distinguishing patient prognosis according to their additional health status factors. Secondly, bioinformatic analysis results alone were not enough, and experiments must be performed for verification. Consequently, more large-scale experimental and genetic studies are warranted for verification.

Conclusions

Based on consistent clustering of genes related to m5c modification, AML was classified into three subtypes, which showed significant differences in prognosis between subtypes.

The limma package analysis was used to identify differential genes among subtypes, and finally 230 DEGs were obtained, based on which we constructed a 5-gene prognostic risk model.

The 5-gene signature has strong robustness and can play a stable predictive performance in external validation data sets (GSE12417, TCGA-LAML).

In conclusion, in this study, we developed a 5-genes signature prognostic stratification system with good AUC in both training set and independent verification set, and it is a model with independent clinical characteristics. Therefore, we suggest using this classifier as a molecular diagnostic test to assess the prognosis risk of patients with acute leukemia.

Declarations

Author Contribution

Haifeng Zhuang: project development, administration and supervision

Qiang Wen and Shou-Jun Wang: methodology development, manuscript writing

Lili Hong and Xianfu Sheng: data collection and analysis, manuscript review and editing

Yu Chen and Xiaofen Zhuang: data collection, figure organization

Compliance with ethical standards

Conflict of interest

The authors declare no conflicts of interest in this work.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

This work was supported by Zhejiang Provincial Natural Science Foundation of China (LY19H290003) (LQ20H280002) Zhejiang Provincial Medical and Health Science and Technology Project(2020KY196), Foundation of Zhejiang province Chinese medicine science and technology planes 2020ZA044, Key project of the 2017 school research fund of Zhejiang Chinese Medical University (2017ZZ02).

References

1. N. Duployez *et al.*, The stem cell-associated gene expression signature allows risk stratification in pediatric acute myeloid leukemia. 1-27 (2019).
2. M. G. J. M. v. Bergen, B. A. V. D. J. H. Reijden, Targeting the GFI1/1B—CoREST Complex in Acute Myeloid Leukemia. **1**, (2019).
3. J. Zhang, X. Zhang, X. Zhang, X. Zhang, <RNA m5C methylation.pdf> **22**, 4093-4097 (2019).
4. Y. L. a. D. V. Santi, m5C RNA and m5C DNA methyl transferases use different cysteine residues as catalysts. *PNAS*, (2020).
5. L. Trixl, A. Lusser, The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. *Wiley Interdiscip Rev RNA* **10**, e1510 (2019).
6. T. Herold *et al.*, A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *haematol*.178442 (2017).
7. K. H. Metzeler, M. Hummel, C. D. Bloomfield, K. Spiekermann, C. J. B. Buske, An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. **112**, 4193-4201 (2008).
8. N. Pećina-Šlaus *et al.*, Comparable Genomic Copy Number Aberrations Differ across Astrocytoma Malignancy Grades. **20**, (2019).
9. C. S. R. A, A. K. B, K. H. J. C. G. A, Establishing a human adrenocortical carcinoma (ACC)-specific gene mutation signature. **230**, 1-12 (2019).

10. S. Engebretsen, J. J. C. E. Bohlin, Statistical predictions with glmnet. **11**, (2019).
11. N. Zhu, J. J. C. B. Hou, Chemistry, Exploring the mechanism of action Xianlingubao Prescription in the treatment of osteoporosis by network pharmacology. **85**, 107240 (2020).
12. J. C. M. Abstracts, The usage of comprehensive geriatric assessment in elderly patients with acute myeloid leukemia:a multicenter,prospective study. **v.36**, 52-52 (2019).
13. J. Cheng, L. Qu, J. Wang, L. Cheng, Y. Wang, High expression of FLT3 is a risk factor in leukemia. *Mol Med Rep* **17**, 2885-2892 (2018).
14. H. Döhner *et al.*, Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. **129**, 424-447 (2017).
15. B. Ramakrishnan, M. Sundaralingam, Crystal structure of the A-DNA decamer d(CCIGGCCm5CGG) at 1.6 Å showing the unexpected wobble I.m5C base pair. *Biophys J* **69**, 553-558 (1995).
16. L. P. Orbons, G. A. van der Marel, J. H. van Boom, C. Altona, An NMR study of the polymorphous behavior of the mismatched DNA octamer d(m5C-G-m5C-G-T-G-m5C-G) in solution. The B, Z, and hairpin forms. *J Biomol Struct Dyn* **4**, 939-963 (1987).
17. Miaomiao Xue, Gene signatures of m5C regulators may predict prognoses of patients with head and neck squamous cell carcinoma. (2020).
18. Q. S. Miaomiao Xue , Lian Zheng, Qingbin Li, Liya Yang, Yuanyuan Zhang, Gene signatures of m5C regulators may predict prognoses of patients with head and neck squamous cell carcinoma. *Am J Transl Res*, (2020).
19. L. H. Larsen, A. Rasmussen, A. M. Giessing, G. Jogl, F. Kirpekar, Identification and characterization of the *Thermus thermophilus* 5-methylcytidine (m5C) methyltransferase modifying 23 S ribosomal RNA (rRNA) base C1942. *J Biol Chem* **287**, 27593-27600 (2012).
20. K. Baumann, m(5)C mRNAs on the move. *Nat Rev Mol Cell Biol* **20**, 512-513 (2019).
21. R. J. J. o. t. R. S. S. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective. **73**, 267-288 (2011).

Tables

Table 1:GSE37642 dataset sample clinical statistics

Clinical Features	GSE37642	GSE12417	TCGA-LAML
OS			
0	109	60	53
1	308	103	87
FAB			
M0	14		
M1	84		
M2	117		
M3	19		
M4	104		
M5	47		
M6	15		
M7	2		
RUNX1-RUNX1T1			
Yes	23		
No	394		
RUNX1 mutation			
Yes	59		
No	311		
Age			
≤60	238		
>60	179		

Table 2: Sample information of training set and verification set of GSE37642 data

Clinical Features	GSE37642-train	GSE37642-test	P value
OS			
0	57	52	1
1	151	157	
FAB			
M0	8	6	0.2424142
M1	37	47	
M2	55	62	
M3	9	10	
M4	58	46	
M5	23	24	
M6	7	8	
M7	2	0	
RUNX1-RUNX1T1			
Yes	11	12	1
No	197	197	
RUNX1 mutation			
Yes	33	26	1
No	152	159	
Age			
≤60	109	80	1
>60	99	129	

Figures

Figure 1

Technology roadmap.

Figure 2

A: The results showed that the heat map of samples with consistency clustering k = 3; B: CDF curve of consistent clustering; C: Delta area curve of consistent clustering; D: OS survival curve of molecular subtype.

Figure 3

Comparison of the distribution of different clinical features between the two molecular subtypes in the GSE37642 dataset: A. Comparison of survival proportion; B. Comparison of age difference; C. Comparison of runx1-runx1t1 fusion difference; D. Comparison of Runx1 mutation difference.

Figure 4

A: Comparison of ssgsea immune scores among molecular subtypes of GSE37642 data set; B: comparison of mcpcounter immune score between molecular subtypes of GSE37642 data set; C: distribution of nine genes related to m5C modification in m5C modification related subtypes.

Figure 5

A: The volcano map of C1 and C3 genes; B: volcano map of C2 and C1 genes; C: Volcano map of C2 and C3 genes.

Figure 6

A: BP annotation map of differentially expressed genes of molecular subtypes; B: CC annotation map of differentially expressed genes of molecular subtypes; C: MF annotation map of differentially expressed genes of molecular subtypes.

Figure 7

A: The horizontal axis represents the log value of the independent variable lambda, and the vertical axis represents the coefficient of the independent variable; B: the confidence interval under each lambda.

Figure 8
Km curve of 5 genes (GSE37642 training set).

Figure 9

A: GSE37642 training set riskscore, survival time and survival state and expression of 5-gene; B: ROC curve and AUC of 5-gene signature classification; C: km survival curve distribution of 5-gene signature in training set.

Figure 10

A:GSE37642 validation set riskscore, survival time and survival state and expression of 5-gene; B: ROC curve and AUC of 5-gene signature classification; C: km survival curve distribution of 5-gene signature in validation set.

Figure 11

A:RiskScore, survival time and survival state and 5-gene expression of TCGA samples; B:ROC curve and AUC of 5-gene signature classification; C:KM survival curve distribution of 5-gene signature in all data.

Figure 12

A:Independent verification of the data set GSE12417 RiskScore, survival time and survival state and the expression of 5-gene; B: ROC curve and AUC of 5-gene signature classification; C: km survival curve distribution of 5-gene signature in all data.

Figure 13

A:Independent verification of the data set TCGA-LAML RiskScore, survival time and survival state and the expression of 5-gene; B: ROC curve and AUC of 5-gene signature classification; C: KM survival curve distribution of 5-gene signature in all data.

Figure 14

The performance of our risk model in different clinical characteristics: A. The survival of patients aged > 60 years was compared; B. Comparison of survival of patients aged ≤ 60 years; C. The survival of patients with runx1-runx1t1 fusion (+) was compared. D. Comparison of the survival of patients with runx1-runx1t1 fusion (-); E. The survival of patients with Runx1 mutation (+) was compared; F. The survival of patients with Runx1 mutation (-) was compared.

Figure 15

A: Comparison of RiskScore between age group samples; B: Comparison of RiskScore between RUNX1-RUNX1T1 fusion group samples; C: Comparison of RiskScore between RUNX1 mutation group samples; D: Comparison of RiskScore between our molecular subtype samples.

Figure 16

The correlation coefficient between KEGG pathway and RiskScore was greater than 0.3: A. The pathway of KEGG_OTHER_GLYCAN_DEGRADATION was negatively correlated with the RiskScore of the samples; B. The pathway of KEGG_NON_SMALL_CELL_LUNG_CANCER was negatively correlated with the RiskScore of the samples; C. The pathway of KEGG_ABC_TRANSPORTERS was positively correlated with the RiskScore of the samples; D. The pathway of KEGG_TIGHT_JUNCTION was positively correlated with the RiskScore of the samples; E. The pathway of KEGG_GLIOMA was negatively correlated with the RiskScore of the samples; F. The pathway of KEGG_PYRIMIDINE_METABOLISM was negatively correlated with the RiskScore of the samples.

Figure 17

A: Univariate analysis results of clinical features and RiskScore; B: Multivariate analysis results of clinical features and RiskScore.

Figure 18

A: Construction and evaluation of nomogram model. A: The results show that the model is composed of age, RUNX1 mutaion and RiskScores; B:Calibration curve of nomogram in 1, 3 and 5 years;C is the decision curve of age, RUNX1 mutaion and RiskScores.