Machine Learning and Integrative Analysis Reveals Stemness Index-associated Signature in Acute Myeloid Leukemia

Qiang Xu

Huazhong University of Science and Technology Tongji Medical College First Clinical College: Wuhan Union Hospital

Hao Zhou

Huazhong University of Science and Technology Tongji Medical College First Clinical College: Wuhan Union Hospital

Dedong Cao

Wuhan University Renmin Hospital

Bin Fang

Huazhong University of Science and Technology Tongji Medical College First Clinical College: Wuhan Union Hospital

Siqi Yan

Huazhong University of Science and Technology Tongji Medical College First Clinical College: Wuhan Union Hospital

Ziping Li

Huazhong University of Science and Technology Tongji Medical College First Clinical College: Wuhan Union Hospital

Yu Hu

Huazhong University of Science and Technology Tongji Medical College First Clinical College: Wuhan Union Hospital

Tao Guo (✉ guotao1968@163.com)

Wuhan Union Hospital <https://orcid.org/0000-0003-0298-5571>

Primary research

Keywords: leukemia stem cells, machine learning, WGCNA analysis, DNA methylation, epigenetic regulation

Posted Date: October 8th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-86574/v1>

Abstract

Background: The rapidly growing research in leukemia has provided unequivocal proof that cancer stem cells exist in acute myeloid leukemia. However, the mechanisms of the growth and self-renewal of acute myeloid leukemia stem cells (LSCs), which play an important role in the development of AML, were still elusive.

Methods: We first used a trained stemness index model based on a recently described machine learning method one-class logistic regression (OCLR) to score AML samples. We then obtained the mRNA expression based-index (mRNAsi), DNA methylation based-index (mDNAsi), and epigenetic regulation based-index (EREG-mRNAsi). With the acquired three stemness indices, we investigated the relationship between the stemness index and the tumor immune microenvironment/somatic mutation, clinical characteristics of AML patients. Furthermore, a weighted gene co-expression network analysis (WGCNA) and protein-protein interaction (PPI) network were performed to obtain key genes associated with the stemness index. Besides, GO and KEGG enrichment analysis was performed to explore the biological functions. Finally, we validated the key genes by applying it to the Oncomine and GEO datasets.

Result: The results suggested that EREG-mRNAsi might accurately predict overall survival (OS) in AML patients from TCGA, while the mRNAsi and mDNAsi was not significantly statistically associated with the OS. Furthermore, the mRNAsi was significantly correlated with the clinical characteristics of the FAB category, age, gender, immune score, WT1, and CEBPA mutation. Additionally, the EREG-mRNAsi was significantly associated with the cytogenetics risk category, TP53 mutation. The mDNAsi was significantly associated with the FAB-category, NPM1, IDH1, IDH2, and DNMT3A mutation. Several key genes related to the stemness index of AML were indicated to be possible and promising therapeutic targets for inhibiting LSCs in AML.

Conclusions: Our research based on machine learning and WGCNA analysis revealed a set of key genes contributed to the stemness index of AML. These genes may become therapeutic targets to inhibit the LSCs of AML in the future.

Background

Although most patients initially respond to chemotherapy, half of the patients' relapse and some patients succumb to the disease within 5 years of diagnosis. Relapse after the achievement of remission from standard induction chemotherapy is the key barrier to cure acute myeloid leukemia (AML), and some patients have no response to induction therapy [1]. Many AML patients show evidence of a hierarchical cellular organization, with a minor fraction of self-renewing leukemia stem cells (LSCs) at the apex of this hierarchy which supports the disease to outgrow and maintain [2–5]. Stemness, defined as the potential for self-renewal and differentiation from the cell of origin, was originally attributed to normal stem cells that possess the ability to give rise to all cell types in the adult organism [6]. LSCs were leukemia cells that possess the ability to give rise to all tumor cell types [7, 8]. AML is a heterogeneous disease that

resides within a complex microenvironment, which composed a diverse, integrated ecosystem of undifferentiated cells, myeloid cells, erythroid cells, lymphoid cells, and LSCs [4]. LSCs were considered to be responsible for AML growth, maintenance, and clinical outcomes, were often resistant to conventional chemotherapy [9–13]. Therefore, it is necessary to explore the molecular mechanism of LSCs and find the potential biomarkers for the diagnosis and treatment of AML.

In the present research, we integratively analyzed cancer stemness in a cohort of AML samples ($n = 151$) from The Cancer Genome Atlas (TCGA) database. First of all, we applied a trained stemness index model based on the previously described one-class logistic regression (OCLR) machine-learning method [6], which contains mRNA expression-based signature, DNA methylation-based signature, and epigenetic-based signature, to quantify AML stemness to acquire three independent stemness indexes. mRNA expression based-index was reflective of gene expression; DNA methylation based-index was reflective of DNA methylation, and the epigenetic regulation based-index was reflected in the epigenetic features. Then, we evaluated correlations between the three stemness index and clinical and molecular features, and we identified a stemness signature that might be useful for guiding the prognostic status of AML. Besides, we found the genes related to stemness indexes by weighted gene co-expression network analysis (WGCNA) [14–16]. Furthermore, we assessed the key genes by Gene Ontology (GO) enrichment analysis [17] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [18]. Finally, we validated the key genes in GEO and Oncomine database.

Materials And Methods

Data download and processing

In this research, we download the RNA sequencing data and the corresponding clinical data of AML from the TCGA database (<https://portal.gdc.cancer.gov/>). Due to the TCGA do not contains normal samples of AML, we download all data of normal whole blood samples from GTEx Analysis Release V8 (<https://www.gtexportal.org/home/>). A complete description of the donor genders, multiple ethnicity groups, wide age range, the biospecimen procurement methods, and sample fixation was described in GTEx official annotation. By using data from Ensembl (<https://uswest.ensembl.org/index.html>) [19], we transferred the ensemble ids to gene symbols. The expression levels of 337 normal whole blood samples and 151 AML samples were presented by fragments per kilobase of transcript per million mapped reads (FPKM) normalized estimation and log₂-based transformation. Then, “limma” package version 3.45.0 (<http://www.bioconductor.org/>) [20] was used in R language, version 4.0.0 (<https://cran.r-project.org/>) to merge the two datasets into one with normalization and the differentially expressed genes (DEGs) were calculated, and the genes with $FDR < 0.05$, $\log_2|\text{Fold change}| > 1.4$ were selected as the significant DEGs. To remove the batch effect, we used the Combat function in R package “sva” [21].

Machine learning based-identification of stemness index for AML

The stemness index was calculated by building a predictive model using an OCLR algorithm on pluripotent stem cell samples from the Progenitor Cell Biology Consortium dataset (PCBC) [22]. AML

samples were stratified by the mRNAsi, mDNAsi, and EREG-mDNAsi, which is based on the stemness index model, were utilized for the integrative analyses. The detailed methodologic information can be referred to as a previous report by Malta et al.[6].

Relationship of stemness index and clinical characteristic/somatic mutation/immune microenvironment

The prognostic roles of the stemness index were identified in the R package survival, and the correlation between the stemness index and clinical characteristic was also analyzed in R. To investigate the relationship of the somatic mutation and stemness index, we used the R package “maftools” [23]. By using CIBERSORT (a gene expression-based deconvolution algorithm) [24], we scored 22 immune cell types for their relative abundance in the AML samples. For the TCGA AML sample, we computed the associations between the stemness index and the estimated fractions of individual immune cell types. By applying ESTIMATE [25], we calculated the individual immune score to predict the level of infiltrating immune cells in the AML sample. We calculated the correlations between the stemness index and immune score or PD-L1 expression.

Construction of weighted gene co-expression network analysis

The co-expression network was constructed based on the DEGs by the “WGCNA” package (version 1.69) [14] in R. First, the normal samples were deleted and the cases were checked for missing data. Second, the cases were clustered according to the gene expression levels, reducing the outlier. Merging the data with the mRNAsi, mDNAsi, and EREG-mRNAsi data, the prepared input was used in the following analysis. For a scale-free gene co-expression network, an appropriate parameter β value was selected as a soft threshold to construct the adjacency matrix. A topological overlap matrix (TOM) was then created from the adjacency matrix to measure the network connection between genes, defined as the sum of adjacent genes, with a minimum of 50 genes per gene module, and module dissimilarity was calculated to construct a dendrogram. We selected two modules with the highest correlation of mRNAsi, and performed correlation analysis for each merged module, filtering out the key genes correlating with both the module and mRNAsi index.

Pathway enrichment and functional annotation of key genes

We selected the ClusterProfiler package [26] in R to construct Gene Ontology (GO) function annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis to visualize the biological function of key genes. A p-value < 0.05 was considered statistically significant.

Construction of protein-protein interaction (PPI) network analysis and Co-expression analysis of key genes

We applied an online tool, Search Tool for the Retrieval of Interacting Genes (STRING), to assess the protein-protein interaction (PPI) among key genes. Then, by using the Corrplot package in R to calculate the Pearson correlations at the transcription level, we identified the co-expression relationship between key genes.

Validated key genes in GEO and Oncomine database

We selected the GEO datasets (www.ncbi.nlm.nih.gov/geo/), GSE76008, which was used to verify the mRNA expression of key genes. The Oncomine database (<http://www.oncomine.org>) was selected to validate the key genes in the pan-cancer. GSE12417 and GSE37642 were used to identify the overall survival of key genes.

Statistical analysis

All statistical analysis was carried out in R version 4.0.0 and corresponding packages. The Wilcox test function in R was applied to evaluate the difference in the stemness index between the two groups. The Kruskal test in R was used to test the correlation between the stemness index and two or more independent samples. For all data, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

Results

mRNAsi, EREG-mRNAsi, and mDNAsi in AML

To obtain a stemness index, we first employed a previously existing stemness index model based on the OCLR machine learning method. We then obtained three stemness indices, including mRNAsi, EREG-mRNAsi, and mDNAsi. The mRNA expression based-index was reflective of gene expression, the DNA methylation based-index mDNAsi was reflective of DNA methylation, and the epigenetic regulation based-index EREG-mRNAsi was reflected in the epigenetic features. To explore the role of these three stemness indices in AML overall survival, by the Kaplan-Meier survival analysis, we observed the mRNAsi and mDNAsi were not associated with the overall survival of AML (Fig. 1a, b), but in patients with lower EREG-mRNAsi had better overall survival than that higher EREG-mRNAsi (Fig. 1c). Additionally, we evaluated the correlation between mRNAsi, mDNAsi, EREG-mRNAsi, and the clinical characteristics of AML. In terms of FAB-category, the results showed that there was a statistical correlation between mRNAsi, mDNAsi, and FAB-category (Fig. 1d, g). Moreover, mRNAsi was associated statistically with gender (Fig. 1e), age (Fig. 1f). According to the risk-category of AML, risk-category was statistically significant (Fig. 1h). These results reflected that the stemness index was tightly associated with LSCs in AML.

Relationship of stemness index and somatic mutation in AML/immune microenvironment

Analyzing the somatic mutation data in AML, we found a statistical association with mRNA and somatic mutation in AML, including WT1 mutation and CEBPA mutation (Fig. 2a, b). Moreover, mDNAsi was positively correlated to NPM1 and DNMT3A mutation (Fig. 2c, f); but IDH-1 and IDH2 mutation were negatively associated with the mDNAsi (Fig. 2d, e). In Fig. 2g, the EREG-mRNAsi of TP53 mutation was higher than those of the TP53 wild type. To identification of the relationship of stemness index and immune microenvironment, we calculated the correlation between the stemness index and individual types of immune cells, PD-L1 expression, and immune score (Fig. 3).

Key modules related to stemness index in AML

We identified 5276 differential expression genes by R package limma, including downregulated 2648 genes and upregulated 2628 genes (Figure S1a, 1b). Then, all DEGs were used to construct a weighted gene co-expression network to identify key modules that contained key genes related to the stemness index. In our research, we choose the soft threshold power of β was 7 (scale-free $R^2 = 0.88$) to establish the scale-free network and the gene cluster tree was acquired to identify the modules related to the stemness index of AML according to the topological overlap value (Fig. 4a). The Module-trait relationships were obtained by the correlation between modules and clinical phenotypes, which made it easier to identify highly correlated modules and stemness index (Fig. 4b). The pink modules had a significantly positive correlation with mRNAsi (MEpink: $r = 0.79$, $p = 7e-28$), then was the tan modules (MEtan: $r = 0.69$, $p = 3e-19$); EREG-mRNAsi had the highest association with the pink modules (MEpink: $r = 0.55$, $p = 2e-11$). Additionally, the red module was the highest negative correlation with the pink modules (MEred: $r = -0.45$, $p = 9e-08$), and the scatter diagrams showed the significant genes in modules related to the stemness index (Fig. 4c-f).

Analysis of key genes in key modules

To explore the genes in key modules, the standard of screening key genes in modules was defined as $\text{cor.MM} > 0.8$ and $\text{cor.GS} > 0.5$. Thus, the pink and tan modules were selected for subsequent analyses to find key genes. Finally, we identified 25 key genes (DKC1, TRMT10C, HSP90AB1, CCT2, GART, NPM1, UBA2, NOP58, CCT4, NIFK, and so on) in the pink modules and 29 key genes (CCNA2, RAD51AP1, CLSPN, TOP2A, ORC1, XRCC2, UBE2T, TUBB, and so forth) in the tan modules. We further extracted the key genes expression matrix to plot boxplot (Fig. 5a, b) and heatmap (Figure S2, S3), which showed the key genes were significantly upregulated in AML. We also identified the correlation of the key genes, which demonstrated the relevance of the genes in the pink and tan module (Fig. 5c, d). The protein-protein interaction relationships between key genes were analyzed using STRING, and a wide-ranging and strong relationship between key genes was shown in figure (S4a, b).

Functional annotation of key genes

The Gene Ontology (GO) and KEGG analyses were applied to complete the functional enrichment analysis of key genes. As figure(6a) showed that the key genes in the pink module mainly enriched in ribonucleoprotein complex biogenesis, RNA localization, ribosome biogenesis, regulation of protein localization to nucleus, positive regulation of telomerase activity, and so forth, while in the tan module the key genes enriched in DNA replication, DNA conformation change, nuclear division, organelle fission, and so on (Fig. 6c). In figure(6b), the key genes of the pink module enriched in Ribosome biogenesis in eukaryotes, Purine metabolism, One carbon pool by folate, and so on, while in the tan module the key genes were associated with cell cycle, DNA replication, Mismatch repair, Nucleotide excision repair and so forth (Fig. 6d).

Prognostic values of key genes in AML

To gain an incisive view of key genes in leukemia stem cells, we downloaded a dataset GSE76008 [13] from GEO to analyze their expression levels in LSCs- and LSCs + samples. In figure S5, the majority of key genes were differentially expressed in LSCs- and LSCs + AML samples, which showed that these genes

were enriched in the LSCs + group. And we further explored the key genes from the pink (Figure S6a) and tan (Figure S6b) module in the Oncomine database, which revealed that expression levels of the key genes in tumor and normal samples differed greatly in many types of tumor. To investigate the role of these key genes in AML, we conducted a survival analysis of these genes by using the Kaplan–Meier curve in two GEO datasets GSE12417 and GSE37642. In Fig. 8, we can see the genes (DSCC1, HSPD1, LRPPRC, MRPL3, NCAPG, NOCL1, NPM1, RCF3, UBA2) were significantly associated with the overall survival; in figure S7, the key genes (CEBPZ, CLSPN, TYMS, MCM4, DSCC1, GART, HSPD1, TUBB, UBA2, CCT2) were statistically significant with the overall survival.

Discussion

In this study, by applying a newly reported stemness index model-based OCLR machine-learning method to the AML samples, we obtained three distinct stemness indexes and then employed these indexes to assess the stemness features of AML based on their molecular and clinical information. Regarding the relevance of the stemness index in prognosis, there was a statistical correlation between EREG-mRNasi and the overall survival of AML patients. However, mRNasi and mDNasi were no statistical correlation with the overall survival of AML patients. Using CIBERSORT and ESTIMATE analysis, we obtained insight into the interaction of the stemness index and the immune microenvironment. Moreover, we identified 16 prognostic genes that could effectively predict the survival of AML patients and revealed the significant correlations between the stemness index and the somatic mutation in AML, including TP53, IDH1, IDH2, DNMT3A, NPM1, CEBPA, WT1 mutation. These results imply that AML patients might have a high degree of heterogeneity concerning the stemness phenotype.

In colon and breast cancer, stem cell signature expression correlates inversely with patient survival [27, 28]. Similarly, stem cell signatures shared by leukemia and hematopoietic stem cells predict clinical outcomes in AML patients [10]. These studies revealed that for multiple tumors, including AML, patients whose cancer exhibits higher expression levels of stem cell genes experience significantly worse clinical outcomes. In this study, we identified 5276 DEGs and used WGCNA to classify them into 14 gene modules based on their correlation with the stemness index. GO and KEGG functional enrichment analysis showed that 54 key genes were mainly involved in DNA replication, DNA conformation change, nuclear division, positive regulation of the establishment of protein localization to the telomere, and positive regulation of telomerase activity, which indicated the genes mainly involved in cell proliferation, cell cycle. The result was consistent with the infinite proliferation characteristic of stem cells. Many human tumor cancers, including AML, were characterized by robust telomerase activity relative to the normal cellular counterpart [29, 30], and telomerase had been recorded for the development of chronic myeloid leukemia induced by BCR-ABL [31]. These findings had identified telomerase as a potential therapeutic target in AML [32].

To verify the accuracy of our analysis results, stemness-related key genes in multiple tumor tissues were overexpressed in the Oncomine and GEO database. Among all key genes, 9 genes (UBA2, RFC3, NPM1, NOLC1, NCAPG, MRLP3, DSCC1, HSPD1, LRPPRC) from the GEO dataset GSE12417 and 10 genes (CCT2,

UAB2, TUBB, HSPD1, GART, DSCC1, MCM4, TYMS, CLSPN, CEPBZ) from GSE37642 were significantly correlated with the overall survival of AML patients. After a comprehensive literature review in Pubmed, we found RCF3 may have a role AML by promoting the G1/S progression and was required for chromatin loading of proliferating cell nuclear antigen (PCNA) [33], which was the key genes in the tan module. MCM4 is part of the complex of MCM2-7 that is important for licensing of DNA origins before the S phase and also serves as the core of the replicative helicase that unwinds DNA at replication forks [34]. Thymidylate synthase (TYMS) played an important role in the G1/S transition; knockdown of TYMS reduced proliferation, promoted apoptosis, facilitated cell cycle progression from G1 to S phase [35]. These key genes were mainly involved in DNA replication, mitotic nuclear division and chromosome segregation.

Furthermore, we identified the association between somatic mutation, immune microenvironment, and the stemness index. We found mRNasi had a negative association with the immune score for AML patients, suggesting that LSCs in AML patients may repress immune cells by affecting the activation of immune cells, which was similar to Van Galen's results [29]. The absence of PD-L1 expression in AML might explain partly why the stemness index had no significant associations with PD-L1 expression, indicating that the potential of immunotherapy with PD-L1 inhibitors seems limited in AML patients, which is consistent with the clinical data. Analyzing somatic mutation data, we identified high mRNasi was associated positively with mutation in WT1 and CEBPA. High mDNasi and low mDNasi were statistically correlated NPM1, DNMT3A mutation, and IDH1 and IDH2 mutation respectively. Besides, high EREG-mRNasi was associated with TP53 mutation. This result is similar to the previous report that TP53 mutation clones isolated from the AML patients cells also show increased chemoresistance potential and have a higher population of LSCs maker positive cells[36]. However, the mechanisms underlying our findings have not been elucidated here, further improvements were expected to result in a profound understanding of the mechanism details of the key genes that need various basic experiments to further explore in AML patients.

Conclusions

In summary, a total of 54 genes were identified to be closely correlated to the AML stemness index and 16 genes showed significant prognostic values of the overall survival. Besides, the main signaling pathway related to stemness in AML was DNA replication, cell cycle, and telomerase activity, which suggests that the key genes in these pathways might be therapeutic targets to eradicate the LSCs in AML. Since the conclusions were based on bioinformatic analysis, further biological and basic researches were needed to validate our results in the future.

Abbreviations

AML: acute myeloid leukemia

LSCs: leukemia stem cells

OCLR: one-class logistic regression

mRNA_{si}: mRNA expression based-index

mDNA_{si}: DNA methylation based-index

EREg-mRNA_{si}: epigenetic regulation based-index

WGCNA: weighted gene co-expression network analysis

DEGs: differentially expressed genes

PPI: protein-protein interaction

TCGA: the Cancer Genome Atlas

STRING: Search Tool for the Retrieval of Interacting Genes

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data used to support the findings of this study are included in the article. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding:

This work was supported by the National Natural Science Foundation of China (No. 81974008) and the Clinical Innovation Funds of Union Hospital (No. 2019-125).

Authors' contributions:

QX and HZ, drafting and revision of the submitted article.

DC and BF, constructive suggestions.

SY and ZL, critically reviewing.

YH and TG design of the study and revision of the submitted article.

Acknowledgements

Not applicable

References

1. Roellig C, Bornhaeuser M, Thiede C, Taube F, Kramer M, Mohr B, Aulitzky W, Bodenstein H, Tischler H-J, Stuhlmann R *et al*: **Long-Term Prognosis of Acute Myeloid Leukemia According to the New Genetic Risk Classification of the European LeukemiaNet Recommendations: Evaluation of the Proposed Reporting System.** *Journal of Clinical Oncology* 2011, **29**(20):2758-2765.
2. Pollyea DA, Jordan CT: **Therapeutic targeting of acute myeloid leukemia stem cells.** *Blood* 2017, **129**(12):1627-1635.
3. Ye HB, Adane B, Khan N, Sullivan T, Minhajuddin M, Gasparetto M, Stevens B, Pei S, Balys M, Ashton JM *et al*: **Leukemic Stem Cells Evade Chemotherapy by Metabolic Adaptation to an Adipose Tissue Niche.** *Cell Stem Cell* 2016, **19**(1):23-37.
4. van Galen P, Hovestadt V, Wadsworth MH, II, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Story JL *et al*: **Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity.** *Cell* 2019, **176**(6):1265-+.
5. Lagadinou ED, Sach A, Callahan K, Rossi RM, Neering SJ, Minhajuddin M, Ashton JM, Pei SS, Grose V, O'Dwyer KM *et al*: **BCL-2 Inhibition Targets Oxidative Phosphorylation and Selectively Eradicates Quiescent Human Leukemia Stem Cells.** *Cell Stem Cell* 2013, **12**(3):329-341.
6. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kaminska B, Huelsken J, Omberg L, Gevaert O *et al*: **Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation.** *Cell* 2018, **173**(2):338-+.
7. Bjerkvig R, Tysnes BB, Aboody KS, Najbauer J, Terzis AJA: **Opinion - The origin of the cancer stem cell: current controversies and new insights.** *Nature Reviews Cancer* 2005, **5**(11):899-904.
8. Kreso A, Dick JE: **Evolution of the Cancer Stem Cell Model.** *Cell Stem Cell* 2014, **14**(3):275-291.
9. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA: **Association of a Leukemic Stem Cell Gene Expression Signature With Clinical Outcomes in Acute Myeloid Leukemia.** *Jama-Journal of the American Medical Association* 2010, **304**(24):2706-2715.
10. Eppert K, Takenaka K, Lechman ER, Waldron L, Nilsson B, van Galen P, Metzeler KH, Poepl A, Ling V, Beyene J *et al*: **Stem cell gene expression programs influence clinical outcome in human leukemia.** *Nature Medicine* 2011, **17**(9):1086-U1091.
11. Li Z, Herold T, He C, Valk PJM, Chen P, Jurinovic V, Mansmann U, Radmacher MD, Maharry KS, Sun M *et al*: **Identification of a 24-Gene Prognostic Signature That Improves the European LeukemiaNet Risk**

- Classification of Acute Myeloid Leukemia: An International Collaborative Study.** *Journal of Clinical Oncology* 2013, **31**(9):1172-1181.
12. Boyd AL, Aslostovar L, Reid J, Ye W, Tanasijevic B, Porras DP, Shapovalova Z, Almakadi M, Foley R, Leber B *et al*: **Identification of Chemotherapy-Induced Leukemic-Regenerating Cells Reveals a Transient Vulnerability of Human AML Recurrence.** *Cancer Cell* 2018, **34**(3):483-+.
 13. Ng SWK, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, Arruda A, Popescu A, Gupta V, Schimmer AD *et al*: **A 17-gene stemness score for rapid determination of risk in acute leukaemia.** *Nature* 2016, **540**(7633):433-+.
 14. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *Bmc Bioinformatics* 2008, **9**.
 15. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: **Functional organization of the transcriptome in human brain.** *Nature Neuroscience* 2008, **11**(11):1271-1282.
 16. Langfelder P, Horvath S: **Fast R Functions for Robust Correlations and Hierarchical Clustering.** *Journal of Statistical Software* 2012, **46**(11):1-17.
 17. Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov D, Burgess S, Buza T, Gresham C *et al*: **Gene Ontology Consortium: going forward.** *Nucleic Acids Research* 2015, **43**(D1):D1049-D1056.
 18. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28**(1):27-30.
 19. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al*: **Ensembl 2015.** *Nucleic Acids Research* 2015, **43**(D1):D662-D669.
 20. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Research* 2015, **43**(7).
 21. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28**(6):882-883.
 22. Salomonis N, Dexheimer PJ, Omberg L, Schroll R, Bush S, Huo J, Schriml L, Sui SH, Keddache M, Mayhew C *et al*: **Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium.** *Stem Cell Reports* 2016, **7**(1):110-125.
 23. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP: **Maftools: efficient and comprehensive analysis of somatic variants in cancer.** *Genome Research* 2018, **28**(11):1747-1756.
 24. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA: **Robust enumeration of cell subsets from tissue expression profiles.** *Nature Methods* 2015, **12**(5):453-+.
 25. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA *et al*: **Inferring tumour purity and stromal and immune cell admixture from expression data.** *Nature Communications* 2013, **4**.
 26. Yu G, Wang L-G, Han Y, He Q-Y: **clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.** *Omics-a Journal of Integrative Biology* 2012, **16**(5):284-287.

27. Merlos-Suarez A, Barriga FM, Jung P, Iglesias M, Virtudes Cespedes M, Rossell D, Sevillano M, Hernando-Momblona X, da Silva-Diz V, Munoz P *et al*: **The Intestinal Stem Cell Signature Identifies Colorectal Cancer Stem Cells and Predicts Disease Relapse.** *Cell Stem Cell* 2011, **8**(5):511-524.
28. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF: **The prognostic role of a gene signature from tumorigenic breast-cancer cells.** *New England Journal of Medicine* 2007, **356**(3):217-226.
29. Aalbers AM, Calado RT, Young NS, Zwaan CM, Wu C, Kajigaya S, Coenen EA, Baruchel A, Geleijns K, de Haas V *et al*: **Telomere length and telomerase complex mutations in pediatric acute myeloid leukemia.** *Leukemia* 2013, **27**(8):1786-1789.
30. Bernard L, Belisle C, Mollica L, Provost S, Roy DC, Gilliland DG, Levine RL, Busque L: **Telomere length is severely and similarly reduced in JAK2V617F-positive and -negative myeloproliferative neoplasms.** *Leukemia* 2009, **23**(2):287-291.
31. Vicente-Duenas C, Barajas-Diego M, Romero-Camarero I, Gonzalez-Herrero I, Flores T, Sanchez-Garcia I: **Essential role for telomerase in chronic myeloid leukemia induced by BCR-ABL in mice.** *Oncotarget* 2012, **3**(3):261-266.
32. Bruedigam C, Bagger FO, Heidel FH, Kuhn CP, Guignes S, Song A, Austin R, Vu T, Lee E, Riyat S *et al*: **Telomerase Inhibition Effectively Targets Mouse and Human AML Stem Cells and Delays Relapse following Chemotherapy.** *Cell Stem Cell* 2014, **15**(6):775-790.
33. Chae HD, Mitton B, Lacayo NJ, Sakamoto KM: **Replication factor C3 is a CREB target gene that regulates cell cycle progression through the modulation of chromatin loading of PCNA.** *Leukemia* 2015, **29**(6):1379-1389.
34. Bagley BN, Keane TM, Maklakova VI, Marshall JG, Lester RA, Cancel MM, Paulsen AR, Bendzick LE, Been RA, Kogan SC *et al*: **A Dominantly Acting Murine Allele of Mcm4 Causes Chromosomal Abnormalities and Promotes Tumorigenesis.** *Plos Genetics* 2012, **8**(11).
35. Zhang S, Yan L, Cui C, Wang Z, Wu J, Zhao M, Dong B, Guan X, Tian X, Hao C: **Identification of TYMS as a promoting factor of retroperitoneal liposarcoma progression: Bioinformatics analysis and biological evidence.** *Oncology reports* 2020, **44**(2):565-576.
36. Yan B, Chen Q, Xu J, Li W, Xu B, Qiu Y: **Low-frequency TP53 hotspot mutation contributes to chemoresistance through clonal expansion in acute myeloid leukemia.** *Leukemia* 2020, **34**(7):1816-1827.

Figures

Figure 1

Correlation between the stemness index and the clinical characteristics. a Kaplan-Meier curve was drawn to analyze the relationship between mRNAasi and overall survival of AML. b Kaplan-Meier curve was drawn to analyze the relationship between mDNAasi and overall survival of AML. c Kaplan-Meier curves show that the low EREG-mRNAasi group had lower mortality than does the high EREG-mRNAasi group. d The correlation between mRNAasi and FAB-category. e The correlation between mRNAasi and gender. f The correlation between mRNAasi and age. g The correlation between mDNAasi and FAB-category. h The correlation between EREG-mRNAasi and Risk-category.

Figure 1

Correlation between the stemness index and the clinical characteristics. a Kaplan-Meier curve was drawn to analyze the relationship between mRNAasi and overall survival of AML. b Kaplan-Meier curve was drawn to analyze the relationship between mDNAasi and overall survival of AML. c Kaplan-Meier curves show that the low EREG-mRNAasi group had lower mortality than does the high EREG-mRNAasi group. d The correlation between mRNAasi and FAB-category. e The correlation between mRNAasi and gender. f The correlation between mRNAasi and age. g The correlation between mDNAasi and FAB-category. h The correlation between EREG-mRNAasi and Risk-category.

Figure 2

Relationship between the stemness index and somatic mutation in AML. a The correlation between mRNA_{si} and WT1-mutation. b The correlation between mRNA_{si} and CEBPA-mutation. c The correlation between mDNAsi and NPM1-mutation. d The correlation between mDNAsi and IDH2-mutation. e The correlation between mDNAsi and IDH1-mutation. f The correlation between mDNAsi and DNMT3A-mutation. g The correlation between EREG-mRNA_{si} and TP53-mutation.

Figure 2

Relationship between the stemness index and somatic mutation in AML. a The correlation between mRNA_{Si} and WT1-mutation. b The correlation between mRNA_{Si} and CEBPA-mutation. c The correlation between mDNA_{Si} and NPM1-mutation. d The correlation between mDNA_{Si} and IDH2-mutation. e The correlation between mDNA_{Si} and IDH1-mutation. f The correlation between mDNA_{Si} and DNMT3A-mutation. g The correlation between EREG-mRNA_{Si} and TP53-mutation.

Figure 3

Tumor immune microenvironments in AML The correlation between stemness index and proportion of immune cells, PD-L1 expression in AML.

Figure 3

Tumor immune microenvironments in AML The correlation between stemness index and proportion of immune cells, PD-L1 expression in AML.

Figure 4

The most significant module was screened by WGCNA. a Dendrogram of all 5276 differential expression genes clustered based on a dissimilarity measure. The dynamic tree cut corresponds to the original modules and the merged dynamic corresponds to the final modules. b Heatmap of the correlation between modular significant (MS) and stemness index of AML. c Scatter plot for correlation between gene module membership in the pink module and gene significance. d Scatter plot for correlation between gene module membership in the red module and gene significance. e Scatter plot for correlation between gene module membership in the tan module and gene significance. f Scatter plot for correlation between gene module membership in the pink module and gene significance.

Figure 4

The most significant module was screened by WGCNA. a Dendrogram of all 5276 differential expression genes clustered based on a dissimilarity measure. The dynamic tree cut corresponds to the original modules and the merged dynamic corresponds to the final modules. b Heatmap of the correlation between modular significant (MS) and stemness index of AML. c Scatter plot for correlation between gene module membership in the pink module and gene significance. d Scatter plot for correlation between gene module membership in the red module and gene significance. e Scatter plot for correlation between gene module membership in the tan module and gene significance. f Scatter plot for correlation between gene module membership in the pink module and gene significance.

Figure 5

Key genes in the respective module a The differential expression value of key genes in the pink module. b The differential expression value of key genes in the tan module. Samples were sorted into normal (N) and tumor (T) groups. c Correlation of key genes at the transcriptional level in the pink module. d Correlation of key genes at the transcriptional level in the tan module.

Figure 5

Key genes in the respective module a The differential expression value of key genes in the pink module. b The differential expression value of key genes in the tan module. Samples were sorted into normal (N) and tumor (T) groups. c Correlation of key genes at the transcriptional level in the pink module. d Correlation of key genes at the transcriptional level in the tan module.

Figure 6

Functional and pathway enrichment analysis. a GO functional enrichment analysis of the genes in the pink module. b KEGG pathway functional enrichment analysis of the genes in the pink module. c GO functional enrichment analysis of the genes in the tan module. d KEGG pathway functional enrichment analysis of the genes in the tan module.

Figure 6

Functional and pathway enrichment analysis. a GO functional enrichment analysis of the genes in the pink module. b KEGG pathway functional enrichment analysis of the genes in the pink module. c GO functional enrichment analysis of the genes in the tan module. d KEGG pathway functional enrichment analysis of the genes in the tan module.

GSE12417

Figure 7

Survival analysis of key genes in the pink module.

GSE12417

Figure 7

Survival analysis of key genes in the pink module.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.pdf](#)
- [Additionalfile1.pdf](#)