Molecular characterization of the highest risk adult patients with acute myeloid leukemia (AML) through multi-omics clustering

trinh nguyen (tinh.nguyen@nih.gov)
NCI: National Cancer Institute https://orcid.org/0000-0002-6606-6948
John Pepper
NIH: National Institutes of Health
Cu Nguyen
NIH: National Institutes of Health
Yu Fan
NIH: National Institutes of Health
Ying Hu
NIH: National Institutes of Health
Qingrong Chen
NIH: National Institutes of Health
Chunhua Yan
NIH: National Institutes of Health
Daoud Meerzaman
NIH: National Institutes of Health

Research article

Keywords: multi-omics, unsupervised clustering, intrinsic subtypes, acute myeloid leukemia

Posted Date: January 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-138491/v1

Abstract

Background: Acute myeloid leukemia (AML) is a clinically heterogeneous group of diseases with poor outcomes that are partly due to its complex and poorly understood heterogeneity.

Methods: Here, we use a multi-omics approach to identify a molecular subgroup with the worst response to chemotherapy, and to identify promising drug targets specifically for this AML subgroup.

Results: Among the three primary clusters of 166 AML adult TCGA cancer cases, we found a High Risk Molecular Subgroup with the worst overall survival at two years – only about 10% survival. TP53 was mutated in most patients in this High Risk Molecular Subgroup , but not in any other patients, constituting a highly significant difference. The top 5 Genes over-expressed in the High Risk Molecular Subgroup included E2F4, CD34, CD109, MN1, and MMLT3. This High Risk Molecular Subgroup also showed higher activation than other patients of molecular pathways related to immune function, cell proliferation, and DNA damage.

Conclusion: Some AML patients are not successfully treated with the current standard of care chemotherapy, and urgently need targeted therapeutics. Potential drug targets include over-expressed genes E2F4 , and MN1 , as well as mutations in TP53, and several molecular pathways .

Introduction

AML not only represents one of the most fatal leukemias but also ranks among the deadliest of all cancers. It presents a myriad of chromosomal alterations and gene mutations, comprising a clinically heterogeneous group of diseases [1]. The poor outcomes of AML are at least in part due to its complex and poorly understood heterogeneity, which has impeded the rational targeting of defined molecular subgroups [1]. The most widely used criteria for clinical subgroups are based on cytogenetic abnormalities, which do not translate directly or easily into drug targets. Although over- and under-expression of genes is known to affect prognosis, such data are not included in the most recent revision of the WHO classification of subgroups [2]. AML treatment concepts have not drastically changed since the 1970s, and AML generally remains a mostly incurable disease [1]. For many other cancers, effective targeted therapeutics have become available, but treatment of AML continues to rely mostly on untargeted therapies. This is largely due to the difficulties of molecular targeting for an extremely heterogeneous collection of diseases.

The current standard of care for AML, using untargeted chemotherapy, is effective against those cytogenetic subgroups recognized as having good prognosis under the current treatment regimen, but chemotherapy offers very low survival rates to those cytogenetic subgroups recognized as having poor prognosis under this treatment regimen, with only about 20% survival beyond two years [3]. To address this problem, we examined multi-omics data to seek intrinsic molecular subgroups that could potentially guide the development of more effective targeted therapies specifically for those patients who would respond poorly to untargeted chemotherapy. The existing system of AML subtyping predates many

modern techniques for molecular analysis, and to our knowledge, this is the first study using molecular multi-omics analysis to contribute to an augmented system of molecular subgroups of AML.

Methods

We began with an unsupervised clustering analysis using two types of data: somatic copy number alteration (CNA), and gene expression levels from RNA-seq measurements. We then identified differences among the three resulting clusters in their risk stratification, and in overall survival, using datasets with information on mutations, putative copy number alterations from GISTIC (Genomic Identification of Significant Targets in Cancer), with matched clinical data. Next, we performed pathway analyses to find differences in which molecular pathways were enriched in the three molecular subgroups we found. Further analyses focused on molecular characterization of the one cluster with the worst outcomes under the current treatment regimen.

Dataset preparation

We downloaded the TCGA adult AML datasets directly from https://www.cbioportal.org/study/clinicalData?id=laml_tcga_pub. We used the total of 166 samples with transcriptomic, copy number alteration, mutation, and clinical datasets. These samples were obtained from peripheral blood and represented the major morphologic and cytogenetic subgroups of AML [3]. We used two different CNA datasets: CNA segmentation and discrete CNA values datasets.

Calculation of CNA segment

We estimated gene level CNA as the segment mean of copy numbers of the genomic region of a gene by using TCGA-Assembler 2 [4] downloaded from https://github.com/compgenome365/TCGA-Assembler-2 (version 2.0.6). Degree of CNA was calculated as log2 (tumor values/normal values). Across samples, CNA of all genes had a standard deviation greater than the median. Therefore, in order to exclude near normal (very low) CNA values, only genes with a sum of CNA values across samples greater than zero were used for analysis, resulting in 13,019 genes total. Hg19 annotation was used to obtain gene position.

Discrete CNA values

Putative copy-number calls determined using GISTIC 2.0 were used to obtain the information for six genes known to be important for AML: *FLT3*, *NPM1*, *DNMT3A*, *CEBPA*, *RUNX1*, and *RUNX1T1*. Patients with CNA values greater than or equal to 1 were classified as copy number amplifications, while patients with values less than or equal to -1 were classified as copy number deletions. Patients with zero values were classified as unchanged.

RNA-seq expression

We used RSEM (RNA-Seq by Expectation Maximization) expected raw count expression dataset. Genes without at least 1 count-per-million reads in at least 50% of the total samples were filtered out. The resulting RNA dataset was log2 transformed and quantile normalized. A total of 12,934 genes were retained for analysis.

Mutation dataset

The AML genes *RUNX1*, *RUNX1T1*, *CEBPA*, *FLT3*, *NPM1*, *DNMT3A*, and *TP53* were analyzed across 166 samples.

Clinical dataset

The clinical dataset provided information such as cytogenetic abnormalities and the risk stratification.

Pathway database

We downloaded the gmt file of MSigDB hallmark gene set collection (version 7.1) from https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp for annotation. The 50 hallmarks in this collection represent biological state or process [5].

Multiple omics data integrative clustering and gene set analysis (MOGSA)

MOGSA is an R software package for multivariate single sample gene set analysis [6]. Using this package (version 1.22.1), we integrated transcriptomic data and gene level CNA over the same set of samples. Based on the chosen PCs from multiple factorial analysis (MFA), we performed unsupervised clustering to identify subgroups and calculated Hall Mark gene set pathway scores for individual samples.

Determine the number of principle components

MFA [7] was incorporated into the moa function from MOGSA to perform principal component analysis (PCA) for CNA and RNA-seq Expression.

Figure 1. Distribution of variances explained by top 20 principle components (PCs). The first six PCs were used to identify subgroups by clustering. These contained a total of 43.6% of total variance, with CNA and RNA-seq expression contributing equally.

Identify of subgroups

We used ConsensusClusterPlus (version 1.52.0) [8] to identify subgroups based on the first six PCs (Figure 1). We used these parameter settings: maxK=6, reps=10000, pItem=0.8, clusterAlg="hc",

finalLinkage= “ward.D2”, distance=”pearson”. We named 3 chosen clusters (Figure 2A & 2B) as follows: C1 or ‘Intermediate Risk Molecular Subgroup’; C2 or ‘Low Risk Molecular Subgroup’; and C3 or ‘High Risk Molecular Subgroup’ (Figure 2C). Descriptive names were based on our survival analysis (Figure 5).

Figure 2. (A) Delta area shows the numbers of clusters (k) (X axis) and their relative change in area under CDF curve (Y-axis). (B) Silhouette plot of chosen clusters with k=3. (C) The separation of 3 subgroups: Low Risk Molecular Subgroup, Intermediate Risk Molecular Subgroup, and High Risk Molecular Subgroup.

Survival analysis

We used the R modules Survfit and coxph [9] to perform overall survival analysis based on the three subgroups resulting from the total of 166 TCGA adult samples (Figure 2).

Calculate gene-set pathway scores

We used the MOGSA (the Integrative Single Sample Gene-set Analysis of Multiple Omics Data) package, to identify the MSigDB hallmark pathways’ gene set scores (GSS). We used these parameter settings: nf=6, proc.row=”center_ssq1”, w.data= “lambda1”, and statis=FALSE.

Select representative molecular pathways for three subtypes

In order to choose representative molecular pathways, we firstly selected the pathways resulting from the MOGSA function with GSS FDR (false discovery rate) values smaller than 0.01 in 50% of all samples. Secondly, we used the R functions, fitting generalized linear models (GLM), and t test to calculate the difference of GSSs in each subgroup vs. that in the rest and selected the top 5 and bottom 5 representative pathways ranked by GLM T values, resulting in 16 unique representative pathways total with GLM FDR <0.01 and t test P values < 0.01. Lastly, we visualized z-score scaled GSSs in a heatmap (Figure 5).

Cytogenetics information of the total of 166 patients

The patients’ information on cytogenetic risk and their genetic abnormalities is shown in supplemental table 1.

Results

Three primary clusters of AML cancer cases were chosen due to the higher relative changes in the delta area graph (Figure 2A) and the best separation of the subgroups (Figure 2B & 2C). These three putative AML subgroups differed in in several aspects of their molecular makeup (Figure 3), in molecular traits (Figure 4), and also differed in prognosis (Figure 5).

Among the six AML genes we examined, the High Risk Molecular Subgroup had significantly fewer gene mutations than the other patients. In contrast, among these six AML genes, the High Risk Molecular Subgroup had a higher frequency than the other patients of copy number alterations (CNAs) (Figure 3). When separated by CNA type, this difference was statistically significant for copy-number amplifications (Fisher exact test, one $p = 0.03$), but not for copy-number deletions ($p = 0.08$).

The patients in our High Risk Molecular Subgroup had significantly lower overall survival than did other patients (Figure 5). This was largely consistent with their risk stratification based on cytogenetics (Table 1).

The three new molecular subgroups were significantly associated with established cytogenetic risk stratifications from clinical data (Table 1) with P value < 10-14 by Fisher's exact test. Among our samples, most patients with a 'poor' cytogenetic risk classification fell within our multi-omics High Risk Molecular Subgroup, while all those with a 'good' cytogenetic risk classification fell into other subgroups. As expected, based on this association with established poor cytogenetic risk stratification, High Risk Molecular Subgroup patients had poor overall survival. However, High Risk Molecular Subgroup included only a subset of the poor cytogenetic risk group patients in our dataset (23 of 35 total, see Table 1), and this subset had even worse survival than did cytogenetic poor risk patients as a whole.

Table 1. Counts of patients in our three molecular subgroups, as classified by established cytogenic risk levels. Details of the cytogenetic abnormalities associated with each cytogenetic risk category are provided in Supplemental Table 1.

Cytogenic Risk Stratification	Low Risk Molecular Subgroup	Intermediate Risk Molecular Subgroup	High Risk Molecular Subgroup
Good	19	13	0
Intermediate	46	48	2
Not determined	2	1	0
Poor	8	4	23

In a previous analysis, overall survival of patients in the poor cytogenetic risk group at two years was reportedly about 20% [3] (we replicated this survival analysis with our subset of 166 of the 200 patients used in the earlier study, see Supplemental Figure 1). In contrast, among the patients in our High Risk Molecular Subgroup, overall survival at two years was worse, at only about 10% (Figure 5).

Because patients in our High Risk Molecular Subgroup had significantly worse clinical outcomes than other patients, we focused on this molecular subgroup for further molecular characterization.

Multi-omics data (RNA-seq and CNA), revealed significant differences among the molecular subgroups in the activation scores of various molecular pathways. The High Risk Molecular Subgroup showed higher activation than other patients of molecular pathways related to immune function, cell proliferation, and DNA damage (Figure 4).

Gene mutations

Among genes known to be important in AML, mutation frequencies differed in our High Risk Molecular Subgroup versus other patients (Table 2). Among the seven AML genes in our mutation data set, most (4/7) had lower mutation frequencies in High Risk Molecular Subgroup than in other patients, but these differences were not statistically significant. In contrast to the other AML genes, *TP53* was mutated in most patients in the High Risk Molecular Subgroup, but not in any other patients, constituting a highly significant difference (Table 2).

Table 2. Frequencies of mutation in AML genes in High Risk Molecular Subgroup, versus other patients. P-values are from Fisher's exact tests on counts of mutant and wild-type genes.

Gene	High Risk frequency	Non-High Risk frequency	p value (<0.01)
RUNX1	0%	11%	NS
RUNX1T1	0	1.4%	NS
CEBPA	8%	7.1%	NS
FLT3	12%	31%	NS
NPM1	4%	33%	0.003 **
DNMT3A	20%	24%	NS
TP53	56%	0	10^{-10} **

Note: NS: nonsignificant, and **: significant

Gene over-expression

The dataset for RNA-seq included seven genes known to be important in AML: *FLT3*, *DMT3A*, *NPM1*, *CEPBA*, *RUNX1*, *E2F4*, and *TP53*. To broaden our scope, we also looked at an additional 200 AML genes derived from the literature. We examined all these genes for differences among our 3 subgroups.

Six of the seven, excluding *CEPBA*, varied significantly among clusters (ANOVA test, $p < 0.01$). Only one of these genes, E2F4 had elevated expression in the High Risk Molecular Subgroup (Figure 6). This difference was highly significant ($p < 10\text{-}5$).

For the list of 200 genes from the literature, we identified thirty-seven genes that were the most over-expressed in the high risk molecular subgroup relative to other patients. Each of these genes differed significantly among clusters (ANOVA test, $p < 0.01$). Among these genes, the top five genes for this difference, ranked by magnitude of the over-expression, were *CD34*, *CD109*, *MN1*, *MLLT3*, and *CD200* (Figure 3).

Discussion

Our findings indicate several candidates for drug targets specific to the extremely high-risk patients of our High-Risk Molecular Subgroup. These candidate targets include mutations of gene *TP53,* which was mutated in most High-Risk Molecular Subgroup patients (Table 2), as well as overexpression of six genes that were highly over-expressed in the High Risk Molecular Subgroup: *CD34*, *CD109*, *CD200*, *E2F4*, *MN1*, and *MLLT3*. Other potential targets may be found in the molecular pathways that are highly activated in our High Risk Molecular Subgroup (Figure 4).

One of the strongest molecular associations with our High Risk Molecular Subgroup is mutations in *TP53*. It has long been established that *TP53* mutations are associated with resistance to chemotherapy and with short survival in hematologic malignancies [10]. The importance of *TP53* mutations specifically for our High Risk Molecular Subgroup is also consistent with the guidelines of the National Comprehensive Cancer Network, which classify AML patients with normal cytogenetics in the poor/adverse risk category if they harbor *TP53* mutations [11]. In AML, mutations in *TP53* are associated with poor responses to chemotherapy, and with very poor prognosis [12]. These authors [12] suggested that it was important to test whether other pathways activated by *TP53* mutations could be therapeutically targeted.

Overexpression of *MN*1 is known to confer resistance to chemotherapy, and a worse AML prognosis. Pardee [13] investigated the mechanisms for this and suggested that therapies directed at increasing *TP53* function may be useful for such patients.

Another of the genes most over-expressed in our high risk subgroup was *E2F4*. This is unsurprising, as it is known that *TP53* mutations can drive the expression of E2Fs, including *E2F4* [14]. The over-expression of *E2F4* in our high-risk molecular subgroup was also consistent with a recent report that, *E2F4* over-expression was associated with poor prognosis in AML patients, and that in a mouse model, depleting E2F4 inhibited proliferation, and induced the differentiation and suppressed the growth of AML cells [15]. These authors suggested *E2F4* as a potential therapeutic target [15], and here we concur by showing the importance of this gene specifically in the subgroup of patients expected to fare worst under untargeted chemotherapy.

Other molecular characteristics of our High Risk Molecular Subgroup include highly activated molecular pathways in the categories of immune function, DNA damage, and cell proliferation, all three of which are consistent with previous reports. A high level of DNA damage has been reported for cells of AML patients categorized as having high-risk cytogenetics and is accompanied by activation of DNA damage pathway [16]. Our results show that inflammatory response and IL6 JAK STAT signaling pathways were highly activated in High Risk Molecular Subgroup. This is consistent with the findings that the inflammatory pathway leads to an activation of the JAK/STAT signaling in AML which fosters leukemia proliferation [17].

Our results suggest that pathways activated by mutations in *TP53* might be targeted therapeutically. We found that the pathways highly activated in our High Risk Molecular Subgroup are in the proliferation category, including, E2F targets, G2M checkpoint, and Myc targets V2 (Supplemental Figure 2). Activation of these proliferation pathways can be promoted by the overexpression of the *E2F4* gene.

Conclusions

An identifiable subset of AML patients is not successfully treated with the current standard of care chemotherapy, and urgently need targeted therapeutics. Potential drug targets include over-expressed genes *E2F4*, and *MN1*, as well as mutations in *TP53,* and several molecular pathways.

In this High Risk Molecular Subgroup, we have identified several over-expressed genes including E2F4. In addition , the presence of TP53 mutations in most of the samples in this subgroup probably contributed to poor responses to chemotherapy, and to poor prognosis. TP53 mutations may drive the elevated transcriptomic expression of E2F4. We have found that E2F targets, G2M checkpoint, and Myc targets V2 are highly activated in this High Risk Molecular Subgroup. We suggested that some potential therapeutic targets include over-expressed genes E2F4, and MN1, as well as mutations in TP53.

Abbreviations

AML: acute myeloid leukemia; TCGA: the cancer genome atlas; CNA: copy number alteration; RNA-Seq: transcriptomics sequencing; RSEM: RNA-Seq by Expectation Maximization; MOGSA: multiple omics data integrative clustering and gene set analysis; MFA: multiple factorial analysis; PCA: principal component analysis; GSS: gene set scores; FDR: false discovery rate.

Declarations

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable.

Competing interests:

The authors have no competing interests to report.

Funding:

There is no funding.

Availability of data and material:

All data used in this study can be accessed from https://www.cbioportal.org/study/clinicalData?id=laml_tcga_pub.

Authors' contributions:

T.N planned and carried out the analysis and wrote the manuscript. C.N helped with the input data. J.P., C.N., Y.F., C.Y., Q.R., D.M., provided advice for carrying out the analysis and for the manuscript. All authors reviewed the manuscript.

Acknowledgements:

Not applicable.

References

1. Green SD, Konig H. Treatment of Acute Myeloid Leukemia in the Era of Genomics-Achievements and Persisting Challenges. Frontiers in Genetics **2020**;11
2. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A*, et al.* The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood **2009**;114:937-51
3. Ley TJ. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia (vol 368, pg 2059, 2013). New England Journal of Medicine **2013**;369:98-
4. Wei L, Jin ZL, Yang SJ, Xu YX, Zhu YT, Ji Y. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. Bioinformatics **2018**;34:1615-7
5. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems **2015**;1:417-25
6. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data. Molecular & Cellular Proteomics **2019**;18:S153-S68
7. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. Wiley Interdisciplinary Reviews-Computational Statistics **2013**;5:149-79

8. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics **2010**;26:1572-3
9. Li JCA. Modeling survival data: Extending the Cox model. Sociological Methods & Research **2003**;32:117-20
10. Wattel E, Preudhomme C, Hecquet B, Vanrumbeke M, Quesnel B, Dervite I*, et al.* P53 MUTATIONS ARE ASSOCIATED WITH RESISTANCE TO CHEMOTHERAPY AND SHORT SURVIVAL IN HEMATOLOGIC MALIGNANCIES. Blood **1994**;84:3148-57
11. Daver N, Schlenk RF, Russell NH, Levis MJ. Targeting FLT3 mutations in AML: review of current knowledge and evidence. Leukemia **2019**;33:299-312
12. Wang Y, Liu Y, Bailey C, Zhang HX, He M, Sun DX*, et al.* Therapeutic targeting of TP53-mutated acute myeloid leukemia by inhibiting HIF-1 alpha with echinomycin. Oncogene **2020**;39:3015-27
13. Pardee TS. Overexpression of MN1 Confers Resistance to Chemotherapy, Accelerates Leukemia Onset, and Suppresses p53 and Bim Induction. Plos One **2012**;7
14. Blandino G, Di Agostino S. New therapeutic strategies to treat human cancers expressing mutant p53 proteins. Journal of Experimental & Clinical Cancer Research **2018**;37
15. Feng YB, Li LL, Du Y, Peng XQ, Chen FH. E2F4 functions as a tumour suppressor in acute myeloid leukaemia via inhibition of the MAPK signalling pathway by binding to EZH2. Journal of Cellular and Molecular Medicine **2020**;24:2157-68
16. Cavelier C, Didier C, Prade N, Mas VMD, Manenti S, Recher C*, et al.* Constitutive Activation of the DNA Damage Signaling Pathway in Acute Myeloid Leukemia with Complex Karyotype: Potential Importance for Checkpoint Targeting Therapy. Cancer Research **2009**;69:8652-61
17. Habbel J, Arnold L, Chen YY, Mollmann M, Bruderek K, Brandau S*, et al.* Inflammation-driven activation of JAK/STAT signaling reversibly accelerates acute myeloid leukemia in vitro. Blood Advances **2020**;4:3000-10

Supplementary Files

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

- SupFigure1.pdf
- SupFigure2.pdf
- SupTable1.pdf