Identification of prognostic signature based on the copy number variation (CNV) and expression in acute myeloid leukemia

Changchun Niu (✉ bright_star2000@163.com)

Chongqing General Hospital <https://orcid.org/0000-0002-8847-7369>

Di Wu

The University of Chicago Medical Center

Alexander J. Li

The University of Chicago Medical Center

Kevin H. Qin

The University of Chicago Medical Center

Daniel A. Hu

The University of Chicago Medical Center

Eric J. Wang

The University of Chicago Medical Centre

Andrew Blake Tucker

The University of Chicago Medical Centre

Fang He

The University of Chicago Medical Centre

Linjuan Huang

The University of Chicago Medical Centre

Hao Wang

The University of Chicago Medical Centre

Qing Liu

The University of Chicago Medical Centre

Na Ni

The University of Chicago Medical Centre

Deyao Shi

The University of Chicago Medical Centre

Xia Zhao

The University of Chicago Medical Centre

Yafang Wan

Chongqing General Hospital

Tian Li

Chongqing General Hospital

Tong-Chuan He

The University of Chicago Medical Centre

Pu Liao

Chongqing General Hospital

Research article

Keywords: Acute myeloid leukemia, copy number variation, gene expression, prognosis

Posted Date: January 5th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-64413/v2>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Acute myeloid leukemia (AML) is caused by multiple genetic alterations in the hematopoietic progenitors, and molecular genetic analysis has provided useful information for AML diagnosis and prognosis. However, an integrative understanding of the prognostic value of specific copy number variation (CNV) and CNV-modulated gene expression has been limited.

Methods We conducted an integrative analysis of CNV profiling and gene expression using data from the TARGET and TCGA AML cohorts. The CNV data from TCGA were analyzed using the GISTIC and all CNV data by genes on every patient were obtained. CNV survival analysis and mRNA survival analysis were conducted with the Multivariate Cox proportional hazards regression model using R software with “survminer” and “survival” packages. KEGG cancer panel genes were extracted from the cancer-related pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG). The R package “circlize” was used for mapping the CNV genes to chromosomes.

Results From this investigation, we observed distinct CNV patterns in the AML risk groups as well as the expression of 251 genes significantly modulated by CNV in both cohorts. There were 102 CNV genes (located at 7q31-34, 16q24) associated with clinical outcomes in AML, which were identified in the TARGET cohort and validated in the TCGA cohort, three of which being miRNA genes (*MIR29A*, *MIR183*, *MIR335*) that overlapped with a KEGG cancer panel. Five genes were identified whose expression was modulated by CNV and significantly associated with clinical outcomes, and among them, the deletion of *SEMA4D* and *CBFB* were found to potentially have protective effects against AML. The result was also validated with patient marrow samples. Moreover, the distribution of CNV in these five CNV-modulated genes was independent of the risk groups, which suggests that they are independent prognosis factors.

Conclusion Overall, this study identified 102 CNV genes and five CNV-modulated gene expression that is crucial for developing new modes of prognosis evaluation and target therapy for AML.

Background

In 2008, it was newly reported for the first time that whole genome sequencing (WGS) was applied for an acute myeloid leukemia (AML) patient [1]. Today, molecular analysis has provided large amounts of useful information for AML diagnosis, prognosis, and treatment[2]. The prognosis study provided the information about potential outcomes and survival rates, which is crucial for understanding the disease and improving its treatment[3]. Copy number variation (CNV) is a potential valuable prognosis marker. It is a type of structural variation that is larger than 50bp and involves unbalanced rearrangements that result in gains or losses of DNA content[4]. CNV drives genome evolution and alters the expression of genes[5], which influence the progress of diseases such as cancers [6]. For example, in acute lymphoblastic leukemia (ALL) *IKZF* CNV was identified as a poor prognosis marker[7].

Though the prognosis and diagnosis of AML have shifted from histological analysis to a more comprehensive analysis such as genetic alteration[8], prognosis needs to be further improved, as about

60% of AML patients are with intermediate risk and respond to therapy variously [9], and the AML patients with rarer abnormalities need be evaluated even more properly[10].

Today, cytomorphology, cytochemistry, immunophenotyping, cytogenetics and molecular genetics play crucial roles in AML diagnosis, and prognosis and risk stratification[11]. Cytogenetics abnormality of CNV was included in WHO classification(2016)[12] or risk stratifications[13-15], and CNV analyzing has a higher revolution than cytogenetics analyzing[16]. Though many studies concerning the CNV in AML[9, 17-19] were conducted, none of them integrative evaluated the roles of gene expression derived by CNV. Moreover, the role of CNV functioning as a prognostic marker by modulating gene expression in AML remains elusive. Therefore, it is crucial to investigate the association between CNV and the prognosis of AML to better understand the genetic disorders of the disease. In turn, more novel, reliable predictions of outcomes for AML patients could arise.

To thoroughly understand the association between CNV and AML, we collected the CNV whole genome sequencing (WGS) data and gene expression data from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) AML cohort, and the CNV Array analysis data and gene expression data from The Cancer Genome Atlas (TCGA) consortium AML cohort. Through a comprehensive analysis of CNVs and CNV-modulated gene expression-alterations, we show the essential CNV genes needed for the outcome and survival of AML patients.

Methods

1. CNV data and gene expression data

In this study, the TARGET and TCGA AML cohorts were used to analyze the prognostic roles of CNVs and CNV modulated gene expression. TARGET CNV (WGS data) and mRNA (mRNA and miRNA sequencing) data were downloaded from target-data.nci.nih.gov (<https://target-data.nci.nih.gov/Public/AML/>). The clinical data, including patient identification, gender, risk group, age, overall survival time, and vital status, were downloaded and pre-processed. The CNV data (Affymetrix Genome-Wide SNP array 6.0 data), mRNA data (mRNA and miRNA sequencing), and clinical data of the TCGA AML cohort were downloaded from gdac.broadinstitute.org, and overall survival time = "patient.days_to_death" + "patient.days_to_last_followup".

Bone marrow samples were collected from AML patients (n = 121, 53 females and 68 males, median age 39 years) from Xinqiao Hospital and Chongqing General Hospital, Chongqing, China since November 2016. The AML patients were diagnosed according to the French–American–British (FAB) and WHO classifications, and had not received bone marrow transplantation. The use of clinical samples was approved by the Ethics Committee of Chongqing General Hospital.

2. Database preparation

To reduce the influence of different synonyms of genes, the gene names were transferred to “symbol” based on the file “gene_info.gz” downloaded from <ftp.ncbi.nlm.nih.gov>.

Cancer-related pathways in KEGG were obtained using an API with Biopython package “Bio.KEGG”, from <https://www.genome.jp/kegg/>, and KEGG cancer panel genes(including 483 genes) were extracted from these pathways.

3. Statistical analysis

3.1. The CNV data from TCGA were divided into two groups, (the sample from the primary blood-derived cancer and solid normal tissue) and analyzed separately, using the GISTIC (version 2.0.23) downloaded from <ftp.broadinstitute.org>, and parameters were set up as TCGA suggested: the threshold for copy number amplifications was 0.1, the threshold for copy number deletions was 0.1, the maximum number of segments was 2000, and the significance threshold for q-values was 0.25 (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/).

The data in “all_thresholded.by_genes.txt” were used for the following analysis, and greater than or equal to 1 were defined as amplification and less than or equal to -1 was defined as deletion.

3.2 CNV survival analysis

The Multivariate Cox proportional hazards regression model (Cox model) was used for identifying survival-related CNVs in AML patients.

The R package “circlize” (function “[circos.genomicTrack\(\)](#)”) was used for mapping the CNV genes to chromosomes.

In TARGET cohorts, 193 patients (96 females and 97 males, median age 9 years) who had integrated CNV data (WGS data) and clinical data were included in survival analysis. In TCGA cohorts, 191 patients (87 females and 104 males, median age 58 years) who had integrated CNV data and clinical data were included for validating the survival analysis result from TARGET cohorts.

We removed those CNVs occurring in less than 4 samples. For every gene, patients were divided into two groups, named “normal” or “CNV,” and analyzed with the Cox model adjusted for gender and age. The p -values were adjusted with “HDR”. $p < 0.05$, and adjusted $p < 0.05$ were considered significant. And the Kaplan-Meier curve was used for visualizing the survival. We performed all analyses using R software with “survminer” and “survival” packages.

3.3 mRNA survival analysis

The gene expression data from the 294 TARGET AML cohort (137 females and 157 males, median age 10 years, the recurring patients were excluded) were used for exploring the relationship between gene expression and survival. The expression of mRNA (RPKM) was log2-transformed. The patients who had integrated mRNA sequencing data and clinical data were included and analyzed with multivariate Cox

proportional hazards regression models adjusted for gender and age, and univariate Cox proportional hazards regression models for gender or age. We used R “survminer” and “survival” packages to perform the analysis of Cox proportional hazards regression models, and “HDR” to adjust the p values afterward. For statistical tests, $p < 0.05$, and adjusted $p < 0.05$ were considered significant, and upper and lower 95% confidence intervals were reported. In the TCGA AML cohorts, 179 patients (84 females and 95 males, median age 58 years) who had integrated CNV data and clinical data were included for validating the survival analysis results from the TARGET cohorts, and multivariate Cox proportional hazards regression models adjusted for gender and age was used for analyzing, and $p < 0.05$ was considered significant

3.4 Integrative analysis of gene expression and CNV data

Integrative analysis of gene expression and CNV was performed on 156 patients in the TARGET cohort and 171 patients in the TCGA cohort. Correspondence between gene CNV and expression was analyzed as follows: The patients were divided into groups with different CNV statuses (“Normal”, “Duplication” and “Deletion”), and the Kruskal test was used for comparing more than two groups, Wilcox test for comparing two groups with R. The p values were adjusted with “HDR,” and $p < 0.05$ and adjusted $p < 0.2$ (Adjusted with “HDR”) was considered to be statistically significant.

4. Patient samples validation analyses

The use of clinical samples was approved by the Institutional Research Ethics Committee of the Affiliated Chongqing Hospital of the University of Chinese Academy of Sciences, and Chongqing General Hospital, Chongqing, China. DNA and RNA of patient bone marrow samples were extracted using the Tiangen DNA/RNA kit (Beijing, China) according to the manufacture’s instruction.

Extracted DNA was for CNV analyzing using AccuCopy™ method developed by Genesky Biotechnologies Inc. (Shanghai, China) as described previously[20]. The primers were as follows: *SEMA4D*, 5'-GGATGAAACTTGCCACGTGAA -3', 5'-GGAAATGCCTTGCCCTAAACC-3'; *DNMT1*, 5'-GATCAGGCAGCTCAATAATTTGTGT-3', 5'-TGACCTCAAATATGGGCAGCA-3'; *CBFB*, 5'-GTCATTGCAGGCAAGAAGACAAC-3', 5'-GAGAACAGCGACAAACACCTA-3'; *CHAF1B*, 5'-TAAATGGCTCCTGGCCCCTAT-3', 5'-TCTTCCACGGACGGTTACTGCT-3'. For each gene, two primers were used to increase the accuracy of CNV analyzing.

Rever Tra Acea-First Strand cDNA Synthesis Kit (Toyobo) was used to generate cDNA using the extracted RNA, and real-time quantitative PCR with SYBR Green using PCR System7500 (ABI) was performed to determine the gene expression, and The primers for real-time PCR were as follows: *CBFB*, forward: 5'-ACTGGATGGTATGGGCTGTC-3', reverse: 5'-AAGGCCTGTTGTGCTAATGC-3'; *CHAF1B*: forward: 5'-CTGGGCAACTGATGGGAATT-3', reverse: 5'-GCAGCACCCCTGTCACAGCT-3'; *DNMT1*: forward: 5'-GTTCTTCCTCCTGGAGAATGTCA-3', reverse: 5'-GGGCCACGCCGTACTG-3'; *SAE1*: forward: 5'-AGGACTGACCATGCTGGATCAC-3', reverse: 5'-CTCAGTGTCCACCTTCACATCC-3'; *SEMA4D*: forward: 5'-GTCTTCAAAGAAGGGCAACAGG-3', reverse: 5'-GAGCATTTTCAGTTCCGCTGTG-3'; *β-actin*(intern control)

forward:5'-AGTTGCGTTACACCCTTTC-3', reverse: 5'-CCTTCACCGTTCCAGTTT-3'. The PCR reaction was conducted in triplicate for each sample. Allgene expression was normalized to that of *GAPDH* using the $2^{-\Delta\Delta Ct}$ method.

The patient samples were divided into groups based on their CNV statuses ("Normal", "Duplication" and "Deletion"). The Kruskal test was used for comparing more than two groups, Wilcox test for comparing two groups with $R. p < 0.05$ was considered statistically significant.

The gene expression was normalized, and multivariate Cox proportional hazards regression models adjusted for gender and age were used for survival analysis using the R "survminer" and "survival" packages.

5. VENN graphic presentations

VENN diagrams were plotted with the R package "VennDiagram".

Results

1. The distribution of the CNV

The CNV data (WGS) from 193 TARGET AML cohort patients (96 females and 97 males, median age 9 years), and the CNV data (SNP array) from 191 TCGA cohort patients (87 females and 104 males, median age 58 years) were used to investigate the distribution of CNV in AML.

The results of the TARGET AML cohort showed that the distribution of CNV in 99.6% (17,507/17,586) of said genes were below 6%. However, there were 35 CNV genes that were above 90%. By contrast, the distribution of CNVs in the TCGA cohort was at most 14.1% (Fig 1A).

Consistent with a previous report[9], this study showed that CNVs were not randomly distributed in the chromosomes. Amplifications were more frequent in chromosomes 1,8,19, 21, and 22 of the TARGET and TCGA cohorts, and deletions were more frequent in chromosomes 7, 16, and X. There were few CNVs in the normal tissues compared to the AML samples (Fig1 B).

The results also showed the distinct CNV distribution patterns between the TARGET and TCGA cohorts in three risk groups. There were different patterns in two cohorts: CNVs occurred more frequently in the standard risk group in the TARGET cohort, and more frequently in the high-risk group in the TCGA cohort. In the high-risk groups, there were more frequent amplifications in chromosome 7 and more frequent deletions in chromosome 19 in TARGET cohort, and deletions in chromosome 7 frequently occurred in the standard and low risk groups. In the TCGA cohort, CNVs occurred less frequently in standard and low risk groups compared with in high-risk group.

2. Integrative analysis of gene expression in concordance with CNV

It has been demonstrated that modulating gene expression is one of the most important ways for CNV to play diversified roles in diseases[21-24]. In this study, the results showed that 5022 gene expression in TCGA was significantly modulated by CNV ($p < 0.05$, adjusted $p < 0.2$), and 577 genes in TARGET were significantly associated with CNV ($p < 0.05$, adjusted $p < 0.2$). 251 genes overlapped in both cohorts (supplementary table 1), and among them 52.5% (132/251) gene-specific CNVs were located on chromosome 19 (19q13, 19p13) (Fig 2A), and four genes overlapped with the KEGG cancer panel, including KEAP1, SIRT6, RHEB and DNMT1 (Fig 2B).

3. The prognostic value of CNV in AML

To further explore the prognostic values of CNVs, a Multivariate Cox proportional hazards regression model analysis was performed on the TARGET cohort, adjusted with age and gender ($p < 0.05$, adjusted $p < 0.05$). There were 758 CNV genes found to be significantly associated with patient survival. Among them, 102 CNV genes overlapped with the TCGA AML cohorts ($p < 0.05$) (supplementary table 2), including 7 lncRNA and 10 miRNA genes, implicating potential gene-regulating capabilities.

Moreover, the CNV genes with the largest probability of high risk were located on chromosome 7 and 16 (7q31,7q32, 7q33, 7q34, 16q24.1) (Fig 3A). Besides, it was observed that 7.8% (8/102) of these CNVs modulated gene expression in the TARGET cohort and 52.0% (53/102) of these CNVs modulated gene expression in the TCGA cohort respectively (Fig 3C).

We also compared these 102 genes with the KEGG cancer panel and found that three miRNA genes (*MIR29A*, *MIR183*, *MIR335*) overlapped with the KEGG cancer panel (Fig 2B). COX model analysis adjusted for age and gender additionally showed that *MIR335* expression (miRNA sequencing data) was associated with survival in the TARGET cohort ($n = 300$, $p = 0.015$, HR: 0.911 95%: 0.844-0.982, Fig 3DE).

4. Identification of CNV-modulated gene expression associated with survival

The results showed that 2058 gene expression was associated with survival in the TARGET AML cohort, 268 of which were validated in the TCGA cohort. Among these 268 genes were 7 CNV-regulated genes associated with survival in the TARGET cohort, and 87 CNV-regulated genes associated with survival in the TCGA AML cohort (Fig 4A). Furthermore, there were 5 genes in both cohorts whose expression was associated with survival and also overlapped with concordant CNV directions, including *CBFB* (16q22), *CHAF1B* (21q22), *DNMT1* (19p13), *SAE1* (19q13), and *SEMA4D* (9q22) (Fig4 B). The effects of CNV on the expression of these five genes were shown in Fig4 C-L, and deletions of *CBFB* and *SEMA4D* downregulated the gene expression, which implicated that CNVs of *CBFB* and *SEMA4D* may be protective, and amplification of *CHAF1B*, *SAE1*, and *DNMT1* upregulated the gene expression.

Additionally, these CNVs were independent of the risk groups; in the TARGET cohort, they occurred more frequently in standard-risk patients, and in the TCGA cohort, they occurred more frequently in high-risk patients (Fig4 M).

We also validated the analysis using the patient bone marrow samples at our hospital in Chongqing, China. The results showed that the expression of four genes (*SEMA4D*, *CBFB*, *CHAF1B*, and *DNMT1*) was associated with the survival of AML patients significantly, and the CNV of these four genes regulated the gene expression. Because fewer CNVs (1 duplication and 1 deletion occurred) were detected in these patients, we did not validate the *SAE1* CNV result.

Discussion

Using a large number of subjects from the TARGET and TCGA AML cohorts, this study demonstrated not only the association between many gene-specific SNVs and AML prognosis but also provided valuable potential AML prognosis markers. Our study showed that the CNVs occurred with distinct patterns in AML risk groups; 251 genes whose expression was modulated by CNV were identified in both cohorts, and among them were 4 genes that overlapped with the KEGG cancer panel. Additionally, 102 CNV genes directly associated with the survival of AML were identified and located on chromosomes 7 and 16, and among them, three miRNA genes overlapped with the KEGG cancer panel; We also identified the five gene expression modulated by CNV that have vital effects on the survival of subjects and had no obvious relationship with risk groups in the two cohorts (Figure 6). The identification of these CNVs is crucial for developing new modes of prognosis evaluation and target therapy.

The array data were used in CNV analysis in previous AML studies [9, 17, 25, 26]. In this study, we used GWS data from TARGET to discover valuable CNV markers and validated them with the array data from TCGA. The CNV in the TARGET cohort showed a different distribution pattern than the TCGA cohort, which may be due to different approaches. A previous study indicated that GWS is the most powerful approach because of its high sensitivity and ability to provide accurate breakpoint information[27], whereas the resolution capacity of Array is limited[4]. Concerning the distribution of CNV frequency in the chromosomes, our study showed more frequent amplifications on chromosomes 8, 11, and 21, and more frequent deletions on chromosome 7, which is consistent with O Nibourel et al's study(In Nibourel et al's study, more frequent amplifications on chromosomes 8, 11 and 21, and deletions on chromosomes 7, 12,17 and 21 were observed in AML samples)[9]. Comparing the TCGA cohort with the normal tissues of AML subjects, CNVs occurred in AML tissues more frequently, supporting the crucial role of CNVs in cancer[28]. Distinct distribution patterns of CNV in risk-groups suggests that the CNV is a potential risk stratification marker, which needs to be further studied and validated. However, there were different CNV patterns between TCGA and TARGET cohorts, which may also be due to the different approaches, but also needs to be further studied.

Modulating gene expression is the major approach for CNVs to play biological roles. In this study, it was observed that a large part of gene expression (28.5% genes (5022/17627) in the TCGA cohort, 7.4% genes (577/7839) in the TARGET cohort) was modulated by CNV. And there were 251 overlapping genes in both cohorts, and the distribution of these genes in the chromosomes showed a similar pattern with the distribution of CNV, suggesting that CNV modulating expression is widespread in AML. The different percentages of modulating in the two cohorts may be due to different CNV analysis approaches used.

Through survival analysis using the COX model, this study identified the 102 CNV genes that locate at 7q31,7q32, 7q33, 7q34, and 16q24.1. Deletion in 7q33-34 has been previously reported[25]. Moreover, these 102 CNV genes included 7 lncRNA and 10 miRNA genes, which implicated potential gene-regulating capability, as lncRNA and miRNA function as master regulators of gene expression[29, 30]. Furthermore, three miRNA genes (*MIR29A*, *MIR183*, *MIR335*) overlapped with cancer (KEGG cancer). It has been reported that miR-29a is a key epigenetic regulator and has crucial roles in cancers, including AML[31], and that miR-335 is involved in diverse cancers, including AML[32], and our result indicated that the high expression level of miR-335 is a protective factor.

This study also identified 5 CNV genes that, through regulating gene expression, indirectly affected the outcomes of patients; among them, deletions of *CBFB* and *SEMA4D* induced down-regulating expression that may be protective. Moreover, these CNV distributions were independent of risk groups, which suggests that they may be independent prognostic markers. Previous studies have shown that some of them were associated with AML. *CBFB* forms fusion genes *CBFB/MYH11*, which block differentiation in AML[33], but the deletion of *CBFB* usually was accompanied by the deletion of *MYH11*, and the deletion of *MYH11/CBFB* had no negative effect on the prognosis[34]. Our findings suggested deletion-induced downregulating *CBFB* expression has positive effects on the prognosis. Elevated expression of *CHAF1B* was observed in leukemia and are critical for leukemogenesis and maintaining the undifferentiated state of leukemia cells, and heterozygous deletions of *Chaf1b* in mice block leukemogenesis[35], and the results of this study showed that in both cohorts, CNV amplification of *CHAF1B* led to expression upregulations and a negative prognosis. *DNMT1*, a member of DNA methyltransferases, was up-regulated and considered to be a potential oncogene in AML through down-regulating p15 expression, and our result showed that CNV-modulated upregulating *DNMT1* expression may lead to bad outcomes in AML patients. *SAE1* is a key molecule in the small ubiquitin (Ub)-like modifier (SUMO) pathway, and promyelocytic leukemia protein (PML) and fusion protein PML–RAR α (retinoic acid receptor- α) fusion oncoprotein are substrates of the SUMO pathway[36], but no direct association between *SAE1* and AML was reported, and our study found that upregulated *SAE1* expression derived from CNV amplifications were associated with poor outcomes. *SEMA4D* was considered a good predictor for worse clinical outcomes [37], and our findings showed that the deletion of *SEMA4D* induced downregulated expression was protective.

Nonetheless, there were some limitations in this study. First, the median ages of two AML cohorts were different (10 years in TARGET vs. 58 years in TCGA). Interestingly, the results from our own sample set showed that age was not associated with outcomes of AML patients (Figure 4B) when the survival analysis was adjusted for age. Second, the distinct CNV distribution patterns between the TARGET and TCGA cohorts in three risk groups may be resulted from the different approaches of CNV analyses and different populations of patients. Therefore, for possible widespread clinical use of these prognostic signatures, more investigations of the biological roles of these CNV genes in AML are needed.

Conclusion

In conclusion, we identified 102 CNV genes (7q31-34, 16q24) associated with a higher risk of worse outcomes, 5 CNV-modulating gene expression associated with the risk of the patients' outcomes, and found that deletion of *SEMA4D* and *CBFB* might have protective effects.

Abbreviations

CNV: copy number variation

AML: Acute myeloid leukemia

KEGG: Kyoto Encyclopedia of Genes and Genomes

TCGA: The Cancer Genome Atlas

TARGET: Therapeutically Applicable Research To Generate Effective Treatments

WGS: Whole genome sequencing

WHO: World Health Organization

API: Application Programming Interface

Declarations

Funding

This work was supported by the National Natural Science Foundation of China (81702075) and the Natural Science Foundation of Chongqing (cstc2019jcyj-msxmX0276), China. The funder provided support in study design, data analysis, and experimental validation.

Competing interests /Competing interests

The authors declare no competing conflicts of interest.

Ethics approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The study was approved by the Ethics Committee of Chongqing General Hospital.

Consent to participate

All patients had approved for the use of clinical tissues for research purposes and signed informed consent.

Consent for publication

The authors declare the consent for publication.

Availability of data and material

The dataset supporting the conclusions of this article is included within the article and its additional files.

Code availability

Not applicable.

Authors' contributions

Conceived and designed the study: CCN, PL, and TCH. Performed the study: CCN, DW, AJL, FH, KHQ, LJH. Analyzed the data: CCN, DAH, EJW, ABT, HW, QL. Wrote the paper: CCN, NN, DYS, XZ, WYF, and TL. All authors read and approved the final manuscript.

Acknowledgments

The authors thank Therapeutically Applicable Research To Generate Effective Treatments (TARGET) and The Cancer Genome Atlas (TCGA) for providing the data. The reported work was supported in part by research grants from the National Natural Science Foundation of China (81702075) and the Natural Science Foundation of Chongqing, China (cstc2019jcyj-msxmX0276).

References

1. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M *et al*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456**(7218):66-72.
2. Cai SF, Levine RL: **Genetic and epigenetic determinants of AML pathogenesis.** *Seminars in hematology* 2019, **56**(2):84-89.
3. Wee Y, Liu Y, Lu J, Li X, Zhao M: **Identification of novel prognosis-related genes associated with cancer using integrative network analysis.** *Scientific reports* 2018, **8**(1):3233.
4. Zarrei M, MacDonald JR, Merico D, Scherer SW: **A copy number variation map of the human genome.** *Nature reviews Genetics* 2015, **16**(3):172-183.
5. Liang L, Fang JY, Xu J: **Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy.** *Oncogene* 2016, **35**(12):1475-1482.
6. Lauer S, Gresham D: **An evolving view of copy number variants.** *Current genetics* 2019, **65**(6):1287-1295.
7. Medeiros BC: **Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia.** *The New England journal of medicine* 2009, **360**(17):1787; author reply 1787-1788.

8. Arber DA: **The 2016 WHO classification of acute myeloid leukemia: What the practicing clinician needs to know.** *Seminars in hematology* 2019, **56**(2):90-95.
9. Nibourel O, Guihard S, Roumier C, Pottier N, Terre C, Paquet A, Peyrouze P, Geffroy S, Quentin S, Alberdi A *et al.* **Copy-number analysis identified new prognostic marker in acute myeloid leukemia.** *Leukemia* 2017, **31**(3):555-564.
10. Grimwade D, Hills RK, Moorman AV, Walker H, Chatters S, Goldstone AH, Wheatley K, Harrison CJ, Burnett AK: **Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials.** *Blood* 2010, **116**(3):354-365.
11. Haferlach T, Schmidts I: **The power and potential of integrated diagnostics in acute myeloid leukaemia.** *British journal of haematology* 2020, **188**(1):36-48.
12. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, Bloomfield CD, Cazzola M, Vardiman JW: **The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia.** *Blood* 2016, **127**(20):2391-2405.
13. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, Dombret H, Ebert BL, Fenaux P, Larson RA *et al.* **Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel.** *Blood* 2017, **129**(4):424-447.
14. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, Potter NE, Heuser M, Thol F, Bolli N *et al.* **Genomic Classification and Prognosis in Acute Myeloid Leukemia.** *The New England journal of medicine* 2016, **374**(23):2209-2221.
15. Grimwade D, Ivey A, Huntly BJ: **Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance.** 2016, **127**(1):29-41.
16. Mehrotra M, Luthra R, Ravandi F, Sargent RL, Barkoh BA, Abraham R, Mishra BM, Medeiros LJ, Patel KP: **Identification of clinically important chromosomal aberrations in acute myeloid leukemia by array-based comparative genomic hybridization.** *Leukemia & lymphoma* 2014, **55**(11):2538-2548.
17. Radtke I, Mullighan CG, Ishii M, Su X, Cheng J, Ma J, Ganti R, Cai Z, Goorha S, Pounds SB *et al.* **Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(31):12944-12949.
18. Vosberg S, Herold T, Hartmann L, Neumann M, Opatz S, Metzeler KH, Schneider S, Graf A, Krebs S, Blum H *et al.* **Close correlation of copy number aberrations detected by next-generation sequencing with results from routine cytogenetics in acute myeloid leukemia.** *Genes, chromosomes & cancer* 2016, **55**(7):553-567.
19. Bolouri H, Farrar JE: **The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions.** 2018, **24**(1):103-112.
20. Du R, Lu C, Jiang Z, Li S, Ma R, An H, Xu M, An Y, Xia Y, Jin L *et al.* **Efficient typing of copy number variations in a segmental duplication-mediated rearrangement hotspot using multiplex competitive amplification.** *Journal of human genetics* 2012, **57**(8):545-551.

21. Berg KCG, Sveen A, Høland M, Alagaratnam S, Berg M, Danielsen SA, Nesbakken A, Søreide K, Lothe RA: **Gene expression profiles of CMS2-epithelial/canonical colorectal cancers are largely driven by DNA copy number gains.** *Oncogene* 2019, **38**(33):6109-6122.
22. Zhou B, Guo R: **Integrative Analysis of Genomic and Clinical Data Reveals Intrinsic Characteristics of Bladder Urothelial Carcinoma Progression.** *Genes* 2019, **10**(6).
23. Henrichsen CN, Chaignat E, Reymond A: **Copy number variants, diseases and gene expression.** *Human molecular genetics* 2009, **18**(R1):R1-8.
24. Shao X, Lv N, Liao J, Long J, Xue R, Ai N, Xu D, Fan X: **Copy number variation is highly correlated with differential gene expression: a pan-cancer study.** *BMC Medical Genetics* 2019, **20**(1):175.
25. Kühn MW, Radtke I, Bullinger L, Goorha S, Cheng J, Edelmann J, Gohlke J, Su X, Paschka P, Pounds S *et al*: **High-resolution genomic profiling of adult and pediatric core-binding factor acute myeloid leukemia reveals new recurrent genomic alterations.** *Blood* 2012, **119**(10):e67-75.
26. Vujkovic M, Attiyeh EF, Ries RE, Goodman EK, Ding Y, Kavcic M, Alonzo TA, Wang YC, Gerbing RB, Sung L *et al*: **Genomic architecture and treatment outcome in pediatric acute myeloid leukemia: a Children's Oncology Group report.** *Blood* 2017, **129**(23):3051-3058.
27. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, Liu S: **Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges.** *Oncotarget* 2013, **4**(11):1868-1881.
28. Shlien A, Malkin D: **Copy number variations and cancer.** *Genome medicine* 2009, **1**(6):62.
29. Peng WX, Koirala P, Mo YY: **LncRNA-mediated regulation of cell signaling in cancer.** *Oncogene* 2017, **36**(41):5661-5667.
30. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136**(2):215-233.
31. Gado MM, Mousa NO, Badawy MA, El Taweel MA, Osman A: **Assessment of the Diagnostic Potential of miR-29a-3p and miR-92a-3p as Circulatory Biomarkers in Acute Myeloid Leukemia.** *Asian Pacific journal of cancer prevention : APJCP* 2019, **20**(12):3625-3633.
32. Zhou JD, Li XX, Zhang TJ, Xu ZJ, Zhang ZH, Gu Y, Wen XM, Zhang W, Ji RB, Deng ZQ *et al*: **MicroRNA-335/ID4 dysregulation predicts clinical outcome and facilitates leukemogenesis by activating PI3K/Akt signaling pathway in acute myeloid leukemia.** *Aging* 2019, **11**(10):3376-3391.
33. Opatz S, Bamopoulos SA, Metzeler KH, Herold T, Ksienzyk B, Bräundl K, Tschuri S, Vosberg S, Konstandin NP, Wang C *et al*: **The clinical mutafome of core binding factor leukemia.** *Leukemia* 2020, **34**(6):1553-1562.
34. Lv L, Yu J, Qi Z: **Acute myeloid leukemia with inv(16)(p13.1q22) and deletion of the 5'MYH11/3'CBFB gene fusion: a report of two cases and literature review.** *Molecular Cytogenetics* 2020, **13**(1):4.
35. Volk A, Liang K, Suraneni P, Li X, Zhao J, Bulic M, Marshall S, Pulakanti K, Malinge S, Taub J *et al*: **A CHAF1B-Dependent Molecular Switch in Hematopoiesis and Leukemia Pathogenesis.** *Cancer cell* 2018, **34**(5):707-723.e707.

36. Seeler JS, Dejean A: **SUMO and the robustness of cancer.** *Nature reviews Cancer* 2017, **17**(3):184-197.
37. de Boer B, Prick J, Puis MG, Keane P, Imperato MR, Jaques J, Brouwers-Vos AZ, Hogeling SM, Woolthuis CM, Nijk MT *et al.* **Prospective Isolation and Characterization of Genetically and Functionally Distinct AML Subclones.** *Cancer cell* 2018, **34**(4):674-689.e678.

Figures

Figure 1

Genome-wide distribution of CNV specific genes in two AML cohorts. (a) Distribution of CNV percentages in the TARGET AML cohort (n= 193), TCGA AML cohort (n=191) and TCGA-normal tissue (n=189). (b) Frequency of cumulative CNV counts (total CNV counts of all patients) in chromosome map locations. (c,d) Frequency of cumulative CNV counts (mean CNV counts of every patient) in chromosome map locations in TARGET and TCGA cohorts.

Figure 2

Distribution of genes modulated by CNV. (a) The distribution in chromosome map locations of the genes (n=251) whose expressions were significantly modulated by CNV in both cohorts. (b) Venn diagram showed the relationship of the genes modulated by CNV and the KEGG cancer panel.

Figure 3

The distribution of CNV genes associated with the survival of AML patients. (b) Zooming diagram showed the chromosome map locations of CNV genes associated with the outcomes of patients. (b) Venn diagram showed the relationship of the CNV genes associated with survival and the KEGG cancer panel. (c) Venn diagram showed the relationship of CNV genes associated with survival and the genes whose expressions were modulated by CNV. (d, e) Kaplan–Meier curve showed the patient survival rates according to the MIR335 gene CNV status.

Figure 4

The CNV-modulated genes associated with survival in AML. (a) Venn diagram showed the relationship of genes whose expressions were associated with clinical outcomes in AML and the genes whose expressions were modulated by CNV. (b) Forest plot showed the results of Cox proportional hazards regression models in the TARGET AML cohort; univariate Cox proportional hazards regression models for Gender and Age. (c-l) The gene expressions modulated by CNV; c, e, g, i, k were the results in the TARGET

cohort; d, f, h, j, k were the results in the TCGA cohort. (m) The distribution of 5 CNV genes in the risk groups in the two AML cohorts.

Figure 5

The validation of CNV genes in clinical samples. (a) Forest plot showed the results of Cox proportional hazards regression models in the validation AML cohort; univariate Cox proportional hazards regression models for Gender and Age. (B-F) The gene expressions modulated by CNV; the mRNA expressions were normalized with β -actin, and followed by subtracting the mean of every group.

Figure 6

Pipeline for the discovery of prognostic signature based on the CNV and expression in AML.

Supplementary Files

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

- [supplementarytable1.docx](#)
- [supplementarytable2.docx](#)