PNAS

Supplementary Information for

Transcriptome-based molecular subtypes and differentiation hierarchies improve the classification framework of acute myeloid leukemia

Wen-Yan Cheng,[a,1] Jian-Feng Li,[a,b,1] Yong-Mei Zhu,[a,1] Xiang-Jie Lin,[c,d,1] Li-Jun Wen,[e,f,1] Fan Zhang,[a,1] Yu-Liang Zhang,[a] Ming Zhao,[a] Hai Fang,[a] Sheng-Yue Wang,[a] Xiao-Jing Lin,[a] Niu Qiao,[a] Wei Yin,[a] Jia-Nan Zhang,[a] Yu-Ting Dai,[a] Lu Jiang,[a] Xiao-Jian Sun,[a] Yi Xu,[e,f] Tong-Tong Zhang,[e,f] Su-Ning Chen,[e,f] Hong-Hu Zhu,[c,d] Zhu Chen,[a,*] Jie Jin,[c,d,g,*] De-Pei Wu,[e,f,*] Yang Shen,[a,*] and Sai-Juan Chen[a,*]

Corresponding authors

Saijuan Chen, Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Er Road, Shanghai, 200025, China; Tel: +86 021-64370045; e-mail: sjchen@stn.sh.cn

Yang Shen, Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Er Road, Shanghai, 200025, China; Tel: +86 021-64370045; e-mail: yang_shen@sjtu.edu.cn

Depei Wu, Jiangsu Institute of Hematology, The First Affiliated Hospital of Soochow University, Shizi St 188, Suzhou, 215006, China; Tel: +86 512-67780390; e-mail: wudepei@suda.edu.cn

Jie Jin, The First Affiliated Hospital, College of Medicine, Zhejiang University, No. 79 Qingchun Road, Hangzhou, 310003, Zhejiang, China; Tel: +86 571-97236898; e-mail: jiej0503@zju.edu.cn

Zhu Chen, Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Er Road, Shanghai, 200025, China; Tel: +86 021-64370045; e-mail: zchen@stn.sh.cn

Leading contact: Saijuan Chen

This PDF file includes:

Supplementary Information Text
Figures S1 to S26
Tables S1 to S4
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S11

Supplementary Information Text

SI Materials and Methods

Treatment protocols

For non-M3 AML, young patients (< 60 years) were given standard intensive “3+7” IA-based regimens as initial induction, which contained idarubicin/daunorubicin (10–12/45–60 mg/m^2, D1–3) and cytarabine (100 mg/m^2, D1–7). When CR was achieved, 4 cycles of high-dose cytarabine (HDAC, 2g/m^2 q12h×6, D1–3) were delivered as consolidation. Elderly patients (≥ 60 years) were evaluated by the treating physician. Fit patients received reduced IA/DA-based induction chemotherapy comprising idarubicin (6 mg/m^2 D1–3) and cytarabine (100 mg/m^2, D1–7), and reduced the consolidation to 2 cycles of HDAC (2 g/m^2 q12h×6, D1–3). While unfit patients were assigned to other less intensive therapies, e.g., demethylation agents at the discretion of the physician.

For patients with acute promyelocytic leukemia (APL), the combination of All-trans retinoic acid (ATRA) and Arsenic trioxide (ATO) with or without chemotherapy was administered based on Sanz risk stratification.

Nucleic acid extraction and next generation sequencing

Bone marrow (BM) mononuclear cells were isolated by Ficoll density gradient centrifugation, from which genomic DNA and total RNA were extracted by using the AllPrep DNA/RNA Mini Kit (Qiagen) or TRIzol reagent (Invitrogen) according to the manufacturer’s instructions. The quality and quantity of DNA/RNA were respectively evaluated by the Agilent 2100 Bioanalyzer system (Agilent Technologies) and Qubit (Life Technologies) before library preparation. For samples from SIH (n = 442), RNA sequencing (RNA-Seq) libraries were constructed using the KAPA RNA HyperPrep kit (Roche), followed by sequencing on the NovaSeq 6000 platform (Illumina) per manufacturer’s protocol. Libraries for whole exome sequencing (WES) were constructed using SeqCap EZ Human Exome v3.0 kit (Roche) and were sequenced on the NovaSeq 6000 platform (Illumina). Hybrid capture-based targeted exome sequencing (TES) was performed on the consensus coding sequence of 100 genes involved in acute leukemia. Library enrichment for TES was carried out using the NadPrep EZ DNA Library Preparation Kit (Nanodigmbio), and sequencing was performed on a NextSeq 550 platform (Illumina). For samples from JIH (n = 110), RNA-seq libraries were prepared through using the TruSeq RNA Sample Preparation Kit (Illumina), and were sequenced on the HiSeq 2500 platform (Illumina). Targeted sequencing of the entire coding sequences of 88 gene targets in myeloid neoplasms was performed with a custom amplicon-based targeted enrichment assay (Agilent). TES libraries were prepared using the TruSeq DNA Sample Preparation Kit (Illumina), and sequencing was performed using the MiSeq instrument (Illumina). Finally, for samples from ZIH (n = 103), libraries for RNA-Seq were prepared utilizing the KAPA RNA HyperPrep Kit with RiboErase (Roche), and were sequenced on a NovaSeq 6000 platform (Illumina). TES was performed using a KAPA Library Amplification Kit (Roche), with a NimbleGen kit (Roche) used to capture the target region. Then the hybridized captured samples were subjected to sequencing on the NovaSeq 6000 system (Illumina).

A subset of RNA-seq data used in this study have been published by us (1) and another research group from SIH (2), which is denoted in Dataset S1. Relevant raw RNA-Seq data have been deposited to the Gene Expression Omnibus (GEO) database with the accession number GSE172057 and GSE201492, respectively.

Gene expression quantification and consensus clustering

For alignment-based gene expression quantification methods, the raw reads of RNA-Seq were aligned to the human hg38 reference genome by using STAR (v2.7.9a) (3) two-pass mode and the gene model of GENECODE v38. The Featurecounts (v2.0.1) (4) and Htseq subprogram htseq-count (v0.11.3) (5) were used to generate the transcript or gene counts table using the aligned BAM files, while two genome alignment-free methods including salmon (v1.2.1) (6) and kallisto (v0.46.2) (7) were used to quantify the transcript read counts using the raw FASTQ files. The DESeq2 (v1.28.0) (8) was utilized to conduct the internal normalization and to generate the gene expression matrix with variance-stabilizing transformation based on the count table files, which were also converted to Transcripts Per Kilobase Million (TPM), another popular and normalized gene expression matrix format.

In the unsupervised clustering pre-process steps, the principal component analysis was conducted to check the potential batch effect in the normalized expression matrix by using the gmodels (v2.18.1) function ‘fast.prcomp’. Then, the identified batch effect was adjusted by using the ComBat function in the R sva package (v3.40.0) (9), which uses empirical Bayes frameworks for adjusting data batch effect. To avoid potential mask of gender factors, we filtered all genes in the gender-related X and Y chromosomes. As one of the extra gene filters, we used the ‘adjust_matrix’ function in the cola (10) R package, a framework conducting consensus clustering, for removing rows with low variance.

The R package ComplexHeatmap (11) was used to carry out the unsupervised clustering based on the ward.D method and ‘1-cor(t(x)))/2’ distance measure. First, we conducted the unsupervised clustering based on top 2000 variance protein-coding genes. Potential unstable-related gene clusters were identified and removed. Then, gradient top variance of protein-coding genes was respectively selected to perform unsupervised clustering. The correlation between WHO classification and defined subgroups were used to determine the clustered gene set for conducting consensus clustering. Finally, the top variance of protein-coding genes (n = 859) was included in the hierarchical and consensus clustering, which were generated from the STAR hg38 alignments using the Featurecounts as the reads counter with batch effect adjusting. The cola package was used to conduct the sampling of patients and features. Totally, twenty top-genes/clustering methods were used to calculate the label-probability of patients and stability of defined gene expression subgroups based on the preselect gene sets.

The gene enrichment analysis (GSEA) of gene expression profiling (GEP)-defined subgroups was conduct by the Broad GSEA command-line program (v4.3.0) (12). The 17-gene leukemia stem cells (LSCs) signature was utilized to assess the stemness score in each GEP-defined subgroup (13), and the immune infiltration with specific cell type abundance in each subgroup was determined by CIBERSORTx (14). An emerging AML prognostic score (APS) was also included in the comparison of parameters in multivariate regression models (15). Both LSC17 and APS scores were generated using log2 (TPM +1) gene expression measure and the original coefficients. Gene co-expression networks were constructed based on Weighted Gene Co-Expression Network Analysis (WGCNA) algorithm (v.1.70-3) (16) following the official standard workflow. Unsigned mode and the Pearson correlation were used to get nodes and edges of unscaled networks. Core networks were visualized in Cytoscape (v3.9.1) (17) based on 0.05 threshold subset of 'exportNetworkToCytoscape' function output.

Arrest stage analysis of AML patients

In order to resolve the cell arrest stage of each molecular subgroups in AML, we performed quantitative computations based on the single-sample GSEA (ssGSEA) algorithm and published gene sets (18). The ssGSEA analysis was implemented in the GSVA tool (v1.42.0) and the scores were normalized using the absolute difference between the minimum and maximum values. The gene set of cellular arrest stages (top 30 signature genes per types) was extracted from single-cell dataset of normal bone marrow and AML patient to characterize the three major developmental stages (HSPC-like, GMP-like, and monocyte-like) (18). Meanwhile, diffusion map (19), the uniform manifold approximation and projection (UMAP, http://github.com/lmcinnes/umap) and hierarchical clustering algorithms were used for dimensionality reduction visualization showing the quantitative scores and developmental branches.

Establishment and validation of predictive models for GEP-defined subgroups

The automatic machine learning (AutoML) technique, Autogluon (v0.4.2) (https://arxiv.org/abs/2003.06505), was used to establish a pool of prediction models for recognizing gene expression subgroups in AML. The prediction accuracy was evaluated in the 10-fold internal test datasets. The data sampling with replacement were performed using the 'createDataPartition' in caret R package (v6.0-88), which split 90% of samples as the training dataset and 10% as the internal test dataset. To further validate the reliability of the GEP-defined gene expression subgroups, we applied the prediction models using the same gene sets (overlap only) or all protein-coding genes on the TCGA LAML and Beat AML cohort. The gene expression counts of TCGA LAML and Beat AML cohort were downloaded from public database. It was normalized followed the same process (DESeq2 VST and TPM). The TPM-based consistently predicted patients of TCGA LAML and Beat AML cohort were included for down-stream survival/drug resistant analysis. The Rtsne (v0.15) (https://github.com/jkrijthe/Rtsne) method and DESeq2 VST data matrix were used to conduct dimensionality reduction for visualization of predicted GEP-defined subgroups via using the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm and the features of unsupervised clustering. The labels of two (TCGA-AB-2830, TCGA-AB-2939) and eight Beat samples (BA2044R, BA2660R, BA2760R, BA2816R, BA2866R, BA2955R, and BA3149R) were adjusted according to the outliers of groups, which were partially validated by fusion genes status.

Identification of fusion genes and tandem duplications

Fusion genes were detected by RNA-Seq, karyotyping, and/or fluorescence in situ hybridization (FISH). Results from TES/WES and Sanger sequencing were established as the gold-standard positive reference of *FLT3*-ITD/*KMT2A*-PTD, and for *FLT3*-ITD only variant allele frequency (VAF) ≥ 5% at the DNA level was considered. Four computational methods including Fusioncatcher (v1.20) (20), STAR-Fusion (v1.9.0) (21), and two recently described Arriba (v2.0.0) (22) and CICERO (v0.3.0) (23) were used to detect fusions from RNA-Seq data. Apart from the Arriba and CICERO, three newly published RNA-Seq-based pipelines, HAMLET (v1.0.1) (24), KM (v2.0.2) (25), and RNAmut (v1.1) (26) were used to identify the *FLT3*-ITD/*KMT2A*-PTD. Novel gene fusions and inconsistent *FLT3*-ITD/*KMT2A*-PTD events were validated by RT-PCR/Sanger sequencing.

Calling and annotation of small sequence variants from RNA-Seq

Raw sequence reads of RNA-Seq were aligned to human hg19 reference genome by using STAR (v2.7.9a) (3) two-pass mode. The Genome Analysis Toolkit (GATK, v4.1.7.0) (27) was used to mark the duplication reads in aligned BAM files. Samtools (v1.7) (28) was applied to generate the MD tag in the marked duplications in BAM files. The MD field is designed for the small sequence variants calling without mapping to the reference. Rnaindel (v2.2.2) (29) was used to detect Indels in the processed BAM files, which with marked duplications were then processed by following GATK pre-processing steps: SplitNCigarReads, BaseRecalibrator, and ApplyBQSR. The GATK HaplotypeCaller (v4.1.7.0), GATK UnifiedGenoTyper (v3.8.0), Lofreq (v2.1.2) (30), Freebayes (v.1.3.2) (arXiv:1207.3907v2), and Varscan2 (v2.4.4) (31) were used to detect SNVs and/or Indels. The generated VCF files were annotated and converted to MAF format files by using the VEP (v100) (32) and vcf2maf (v1.6.18) (https://github.com/ckandoth/vcf2maf), which contained numerous basic annotation information, such as the variant allele frequency (VAF), existing sites, population frequency, ClinVar (33), and variant effect. Besides, we merged multiple MAF files generated by different methods and added the variant caller field annotation (e.g., HaplotypeCaller and Freebayes), in which the maximum variant allele frequency was selected. The R package anor (https://github.com/clindet/anor) was used to annotate other databases, such as the 1,285 RNA-Seq calling sets from patients with BCP-ALL, the RNA-editing database including DARNED (34), RADAR (35), and REDIportal (36). Besides, the variant sites reported by targeted sequencing or WES before and those validated as false-negative events were introduced. Finally, the KM and seed sequence match script were used to double-check the events that cannot be identified by the genome alignment-based methods.

Screen of small sequence variants from RNA-Seq

A progressive screen strategy for small sequence variants from RNA-Seq was proposed in this study. To avoid repeated upstream calling in subsequent analyses, we did not limit the genome regions when utilizing the variant calling methods. By intersecting all VCF files with a dynamic and targeted genome region, irrelevant genes and regions could be excluded. To build the targeted genome region, we selected 32 genes and used the exon with an extra 10 bp length at 5' and 3' end, which could speed up the annotation of mutations, especially when the original results were very large. Variant sites with a population frequency of more than 1% were excluded (gnomAD_AF and AF fields), and those involved in 3' UTR, 3' flank, 5' flank, 5' UTR, intron, silent, intergenic region (IGR), and splice region translation start site were also filtered. The recurrent counts of variant sites in different samples were regarded as an important feature of false-positive events, such as single base duplications with significant higher recurrence and lower quality scores. Collectively, three different filter modes were used to select the variant list for the down-stream check, which included the strict mode: depth ≥ 10, supporting counts ≥ 3, and VAF > 0.04, the intermediate mode: depth ≥ 7, supporting counts ≥ 2, and VAF > 0.02, and the loose mode: depth ≥ 5, supporting counts ≥ 1, and VAF > 0.01. To facilitate checking the variant list, variant calling results were divided into nine categorizes: 1) All sites in genes that are top mutant or with clinical significance in AML. 2) Sites that have been reported positive in AML. 3) Sites that have been reported positive in leukemia. 4) Genes that have been reported as germline origin. 5) Sites that are associated with any clinical significance, such as in ClinVar (33). 6) Truncated or damaging sites including the frameshift, nonsense, and splicing variants 7) Missense variants with damaging scores. 8) Known SNP sites that have not been reported as germline or somatic mutations in tumor cohort. 9) Other sites. In the comparison between RNA-Seq and DNA-based sequencing data, we labelled and classified all variant sites that were reported by DNA-based methods. The variants were simultaneously checked in VCF, MAF, and BAM files in order to reduce the false-negative rate. The distribution of the detection rate in the combination of different variant calling methods using RNA-Seq data was also calculated.

Variant calling from DNA sequencing

Paired-end reads were aligned to the hg19 reference genome. SNVs and indels were obtained by synthetically utilizing three callers, namely GATK4 Mutect2, VarDict (v1.5.8) (37), and MuTect (v1.1.7), with default parameters or authors' recommendations used. All mutations were annotated by snpEff (v4.2) (38) and ANNOVAR. All the functional mutations, including missense, nonsense, splicing, and nonstop SNVs, as well as indels were obtained. Homemade pipelines were used to filter SNVs and indels detected by the aforementioned software, according to the following analysis standards to screen raw variants sites: 1) Mutations that were called more than one software. 2) Mutations with VAF of more than 5% and at least 4 individual mutant reads. 3) A normal control variation database, including B-cell acute lymphoblastic leukemia, T-cell acute lymphoblastic leukemia, diffuse large B-cell lymphoma, and natural killer T cell lymphoma was built based on our previous publications. 4) Population related variants that were reported in dbSNP (v151) (39) but absent in the catalogue of somatic mutations in cancer (COSMIC) (v85) (40) were excluded. 5) Variant frequency in 1000 Genome Project, Genome Aggregation Database (gnomAD), NHLBI Exome Sequencing Project (ESP6500), and UK10K database (https://www.uk10k.org/) was less than 10^{-4}. The WES calling set was obtained from our previously published APL paper (1).

Drug sensitivity analysis of gene expression subgroups in AML

The drug screening data of AML patients were accessed from the Beat AML project, which covered the different molecular subtypes of AML. First, we predicted the label of gene expression subgroups (G1-G8) in patients of the Beat AML cohort based on gene expression features that were normalized by TPM transformation. Only the consistent cases in different models were retained in the down-stream comparison. The median of area-under-the-curve (AUC) values was used as the indicator for predicting inhibitors sensitivity of patients with AML. The p-value of drug response was calculated using the Wilcoxon signed-rank test. We also generate the combination of different subgroups to reduce the false negative rate.

Zebrafish breeding

The zebrafish wild-type Tübingen strain (ZFIN ID: ZDB-GENO-990623-3) used in this study was maintained under standard conditions as previously described (41). Embryos were maintained in egg water at 28.5 ℃ and 1-phenyl-2-thiourea (PTU; Sigma) was used to prevent pigmentation. All animal experimental procedures were performed according to the guidelines of the Committee on Animal Care of Shanghai, China, and were approved by the Institutional Animal Care and Use Committee (IACUC) of Shanghai Jiao Tong University.

Plasmid construction, mRNA synthesis and microinjection

The coding sequences of human *CYB5A::DYM*, *MX1::FAM3B* and *NUP98::TNRC18* fusion genes were synthesized and cloned into an pCS2+ plasmid. Then, mRNAs of these three novel fusions were transcribed through mMessage mMachine SP6 Transcription kit (Thermo Fisher Scientific; AM1340) and purified by Nucaway Spin Columns (Ambion; 10070). Fusion mRNA was injected into 1-cell stage embryos separately at a final concentration of 100-120 ng/uL.

Whole-mount mRNA in situ hybridization (WISH)

Probes of myeloid markers *lyz*, *mpx*, and *lcp1* were transcribed with T7 or T3 polymerase. WISH was performed as described previously (42) by using NBT/BCIP Alkaline Phosphatase Substrate Kit (Vector Laboratories, SK-5400). Images were captured through Nikon SMZ1500 microscope.

Other visualization and statistical analysis

The one-dimensional fusion protein diagram was visualized using the ProteinPaint tool (https://pecan.stjude.cloud/proteinpaint) (43). The circlize (v0.4.13) (44) was used to draw the genomic circle diagram with fusion genes links. The venn (v1.10) (https://github.com/dusadrian/venn) and ComplexHeatmap (v2.8.0) (11) function UpSet was respectively used to display the interaction between data sets. The ggplot2 (v3.3.5), ggpubr (v0.4.0) and the ggstatsplot (v0.8.0) (https://doi.org/10.21105/joss.03167) were used to draw basic statistical graphics. The coexistence and mutual exclusion analysis of mutant genes was performed using the tool DISCOVER (Discrete Independence Statistic Controlling for Observations with Varying Event Rates, v0.9.3) (45), which provides statistical testing with a lower false-positive rate.

Categorical variables were compared by Pearson's Chi-square or Fisher's exact test, and continuous data by t-test or Wilcoxon rank sum test. The R package survival (v3.2-11) was used to construct the Kaplan-Meier (KM) model, and the log-rank test was used to calculate estimates of survival probabilities and hazard ratios. Multivariable Cox analysis of overall survival (OS) in non-M3 AML patients was applied, with backward elimination used for model selection. The survminer (v0.4.9) was used to plot the KM survival curve and draw the forest visualization of Cox regression models. Most statistical analyses were performed using the R 4.0.2 software package. Other descriptions of public datasets, analysis strategies, and visualizations are provided in Supplementary Methods.

SI Figures

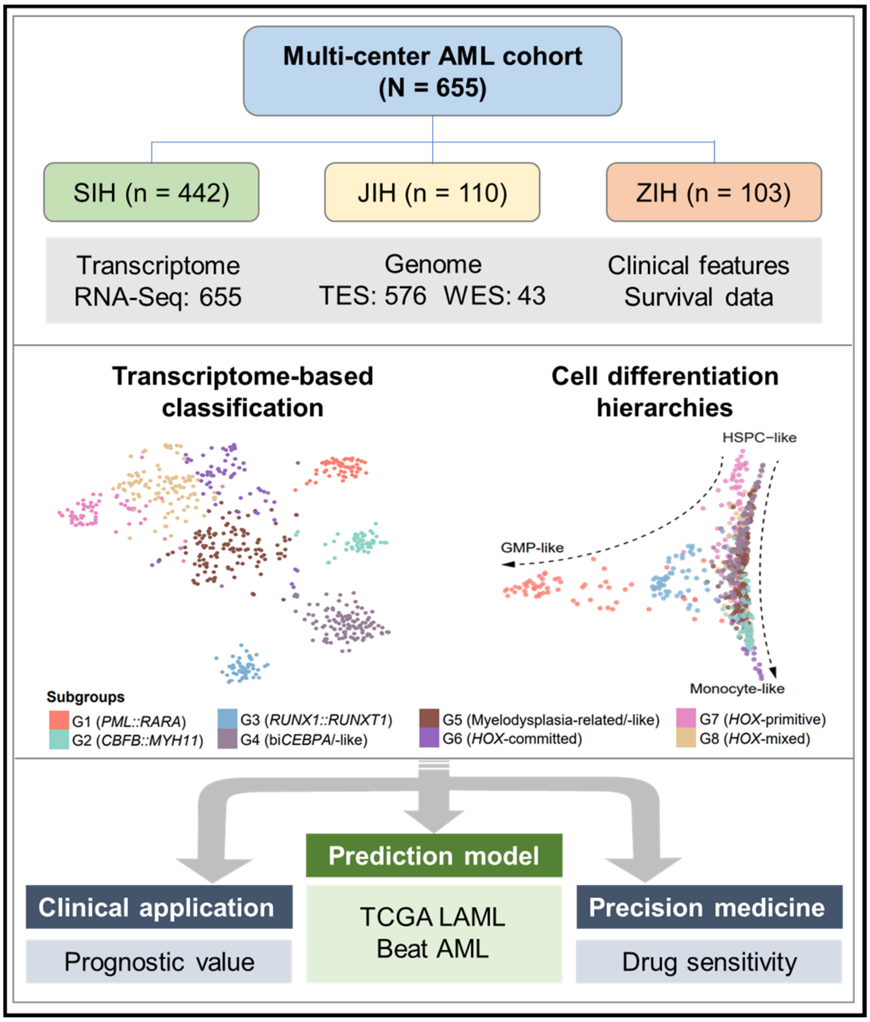


Fig. S1. Graphical abstract of the multi-center AML study. The whole study cohort consists of 655 primary AML patients from three centers in China. All patients were subjected to RNA-Seq, among which, 619 cases (94.5%) harbored both RNA-Seq and TES/WES data. The main findings of this study include two aspects. Firstly, we established eight transcriptome-based molecular subtypes (G1–G8) with distinct biological and clinical features through enhanced consensus clustering. On the other hand, these molecular subgroups demonstrated different stages of cell differentiation, including HSPC-like (G5, G7, and G8), GMP-like (G1, G3), and monocyte-like (G2, G6) signatures. Through development of prediction models, the eight gene expression subgroups could be convincingly reproduced in both Beat AML and TCGA AML cohorts, showing different prognostic value and drug sensitivity. The robust transcriptome-based molecular subtypes hold great potential in clinical application, and may facilitate the implement of precision medicine in AML. GMP, granulocyte-monocyte precursor; HSPC, hematopoietic stem/progenitor cell-like; RNA-Seq, RNA sequencing; TES, targeted exome sequencing; WES, whole exome sequencing.

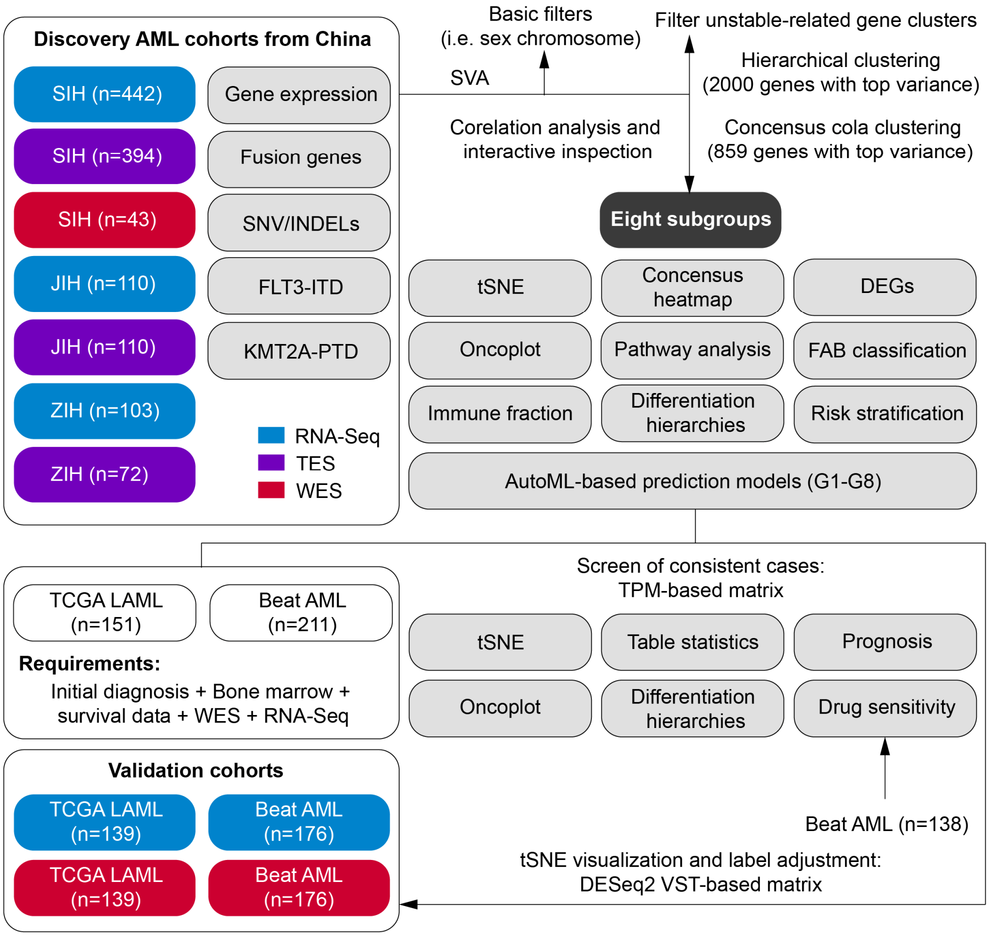


Fig. S2. Flow diagram of the whole study. Flow diagram depicting the relationship of multiple datasets, modeling procedures and analytical steps in this study. Discovery cohorts of China consist of three centers including SIH (n = 442), JIH (n = 110), and ZIH (n = 103). The uniformly screened TCGA LAML and Beat AML patients are included in validation cohorts. Multiple preprocess steps, i.e., batch effect adjustment and pre-filters of genes were adopted to reduce the impact of bias factors. Meanwhile, the consensus clustering strategy of the cola package and the pre-screened features defined the stable gene expression subgroups. A pool of AutoML-based prediction models predicts the label of validation cohorts. TPM-based gene expression matrix drops fuzzy samples. The DESeq2 VST normalization and its tSNE visualization help to further refine the labels. Multidimensional features including genetic mutations, clinical prognosis, and differentiation hierarchies are clarified both in discovery and validation cohorts. A portion of available Beat AML (n = 138) was included in the drug sensitivity analysis step. SIH, Shanghai Institute of Hematology. ZIH, Zhejiang Institute of Hematology. JIH, Jiangsu Institute of Hematology. AutoML, automatic machine learning. VST, variance stable transformation. tSNE, t-distributed stochastic neighbor embedding.

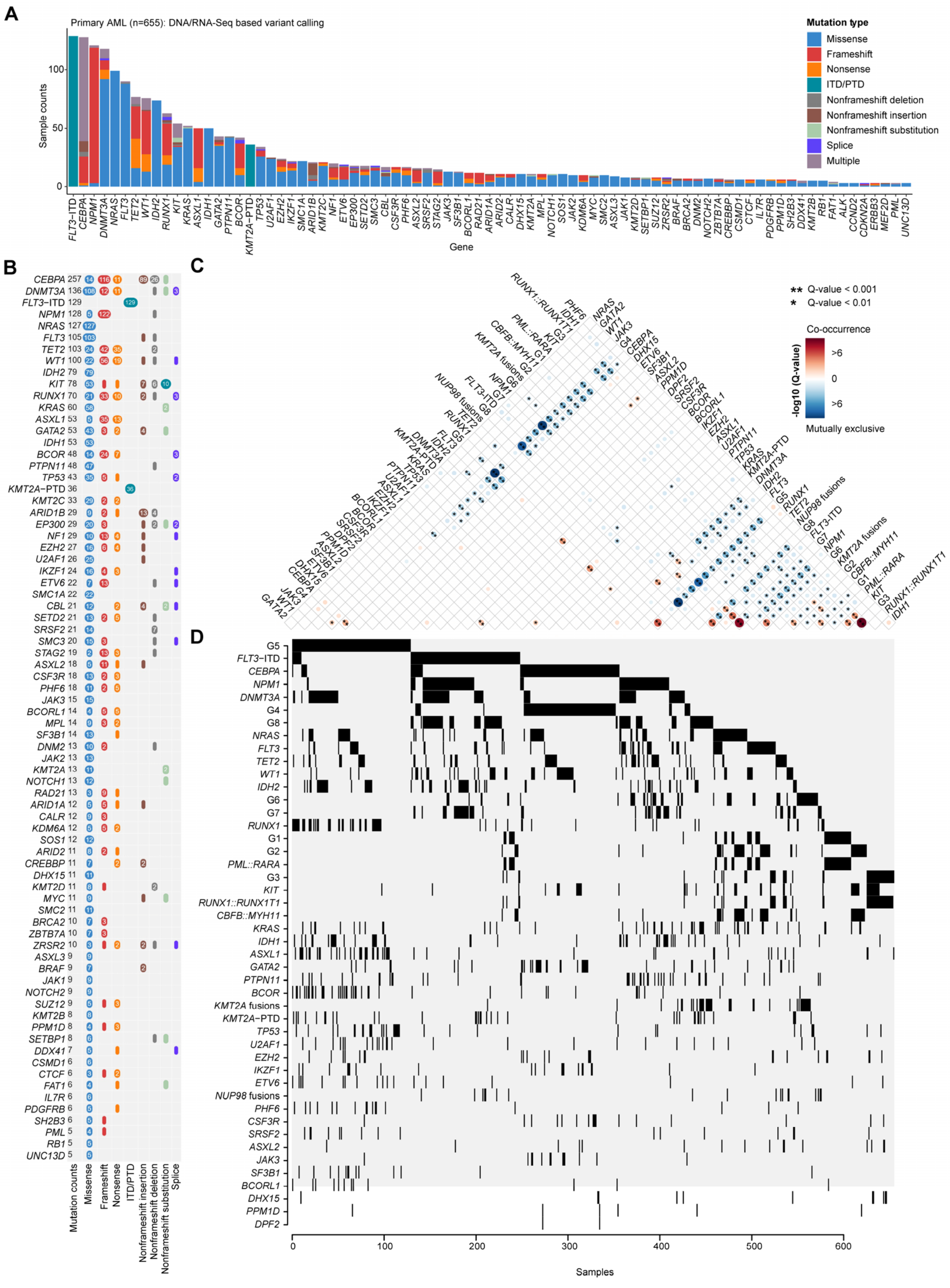


Fig. S3. Genomic landscape of 655 patients with AML in China. (A) Bar plot indicates the top-mutant genes of AML. Mutation types are show in different color. Multi-mutations of same genes in one patient are merged. (B) Colored table of top mutations. (C, D) Co-occurrence and mutually exclusive between molecular subgroups (G1-G8), fusion genes, small sequencing variants, and internal tandem duplication (ITD)/partial tandem duplication (PTD).

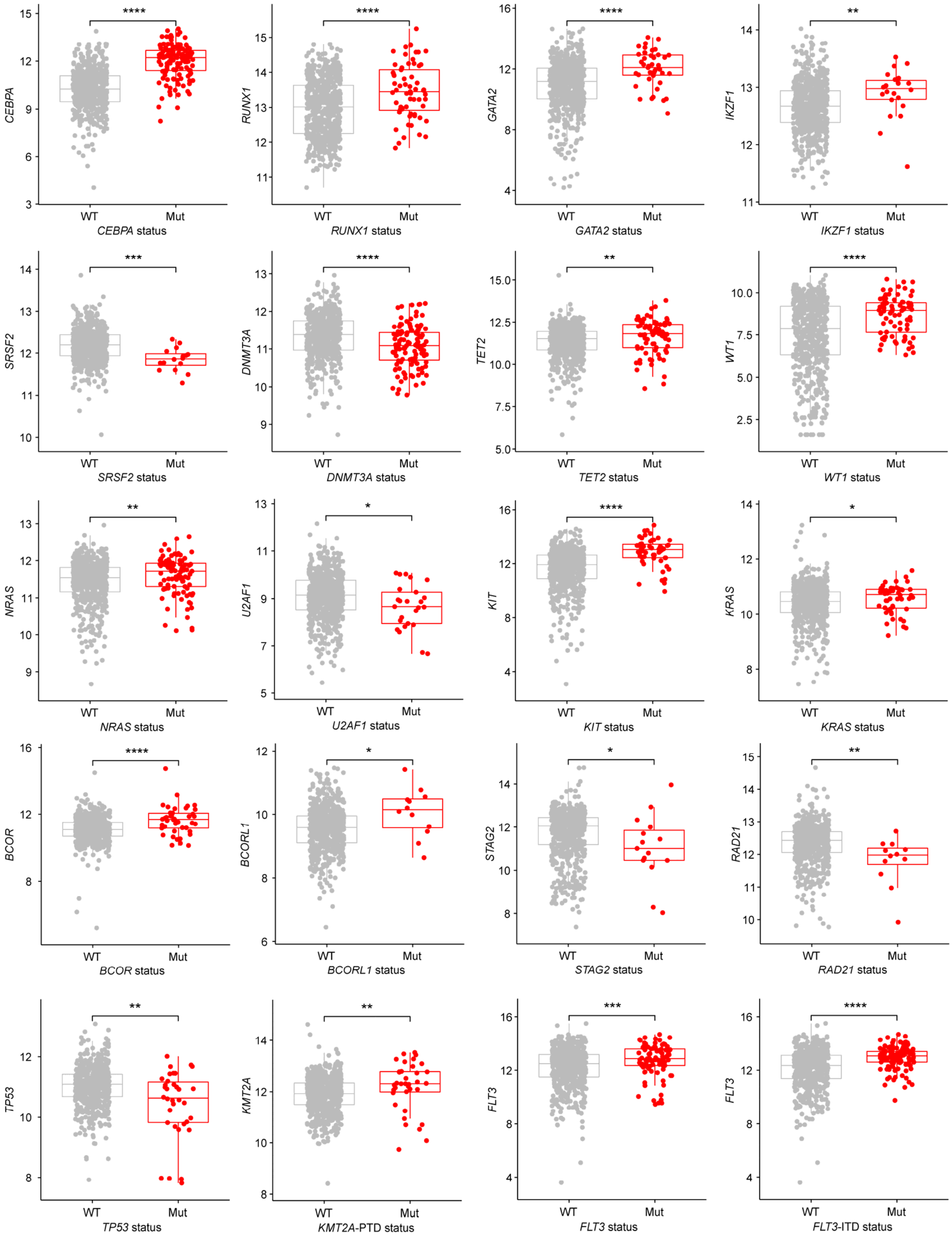


Fig. S4. Genetic mutations and its self-regulation. Gene expression levels of mutated genes in AML, with wild-type (WT) and mutant (Mut) genes denoted in gray and red dots, respectively.

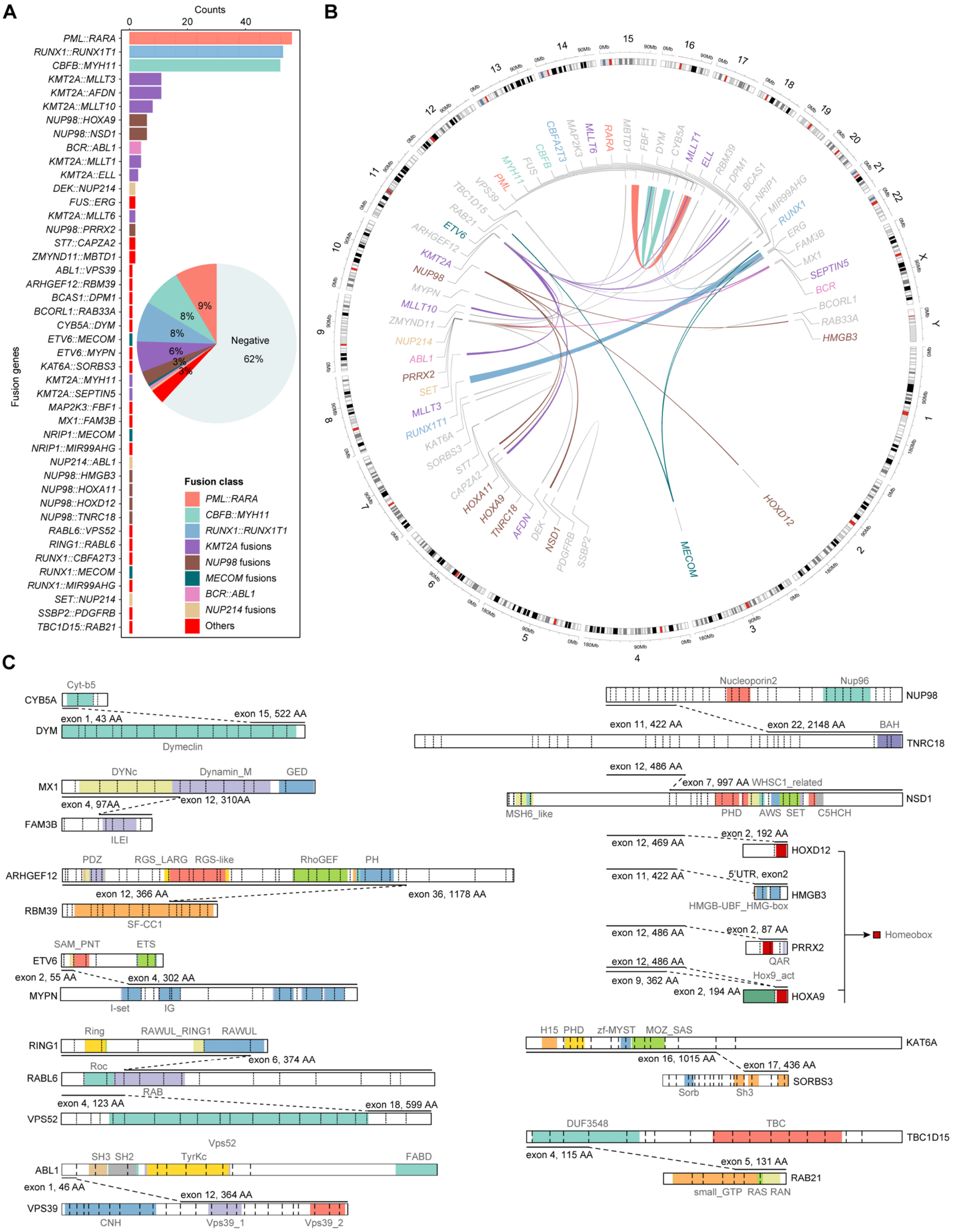


Fig. S5. Fusion genes detected in this study. (*A*) The frequency and proportion of validated fusion genes in 655 primary AML patients. (*B*) Circos plot of validated fusion genes, with partner genes linked using ribbons. Ribbon width indicates the count of fusion events. (*C*) Schematic of novel and *NUP98* fusion transcripts. Structural domains in partner genes and junction points in the novel fusions are depicted, and the name of each domain is listed.

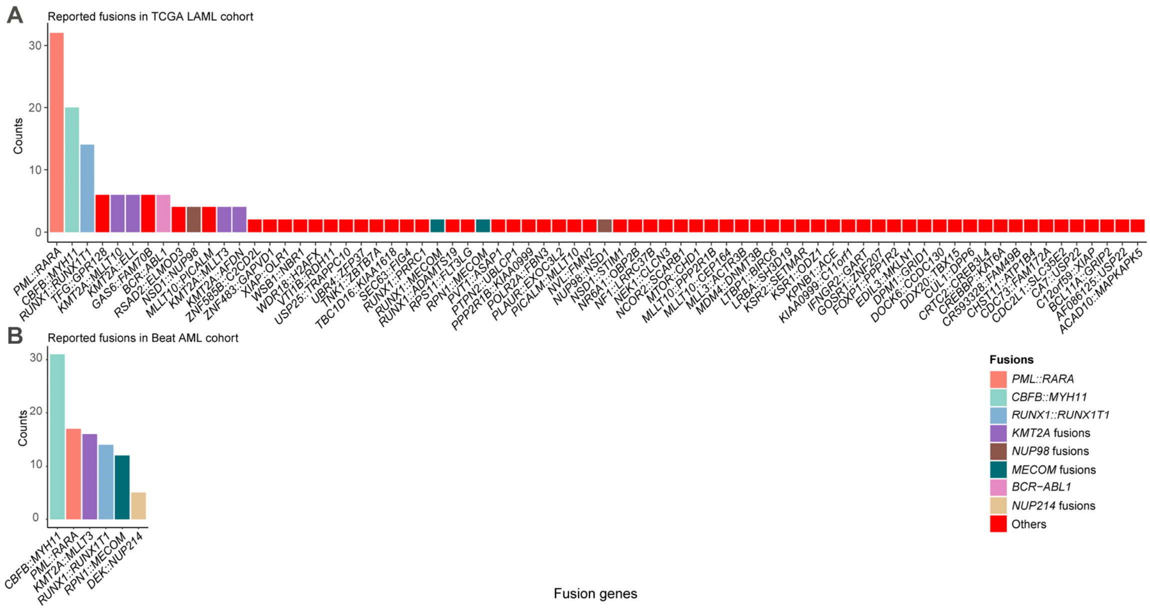


Fig. S6. Previously reported fusion genes in large AML studies. Reported fusions in the TCGA LAML cohort (2013) (*Upper*) and Beat AML cohort (2018) (*Lower*).

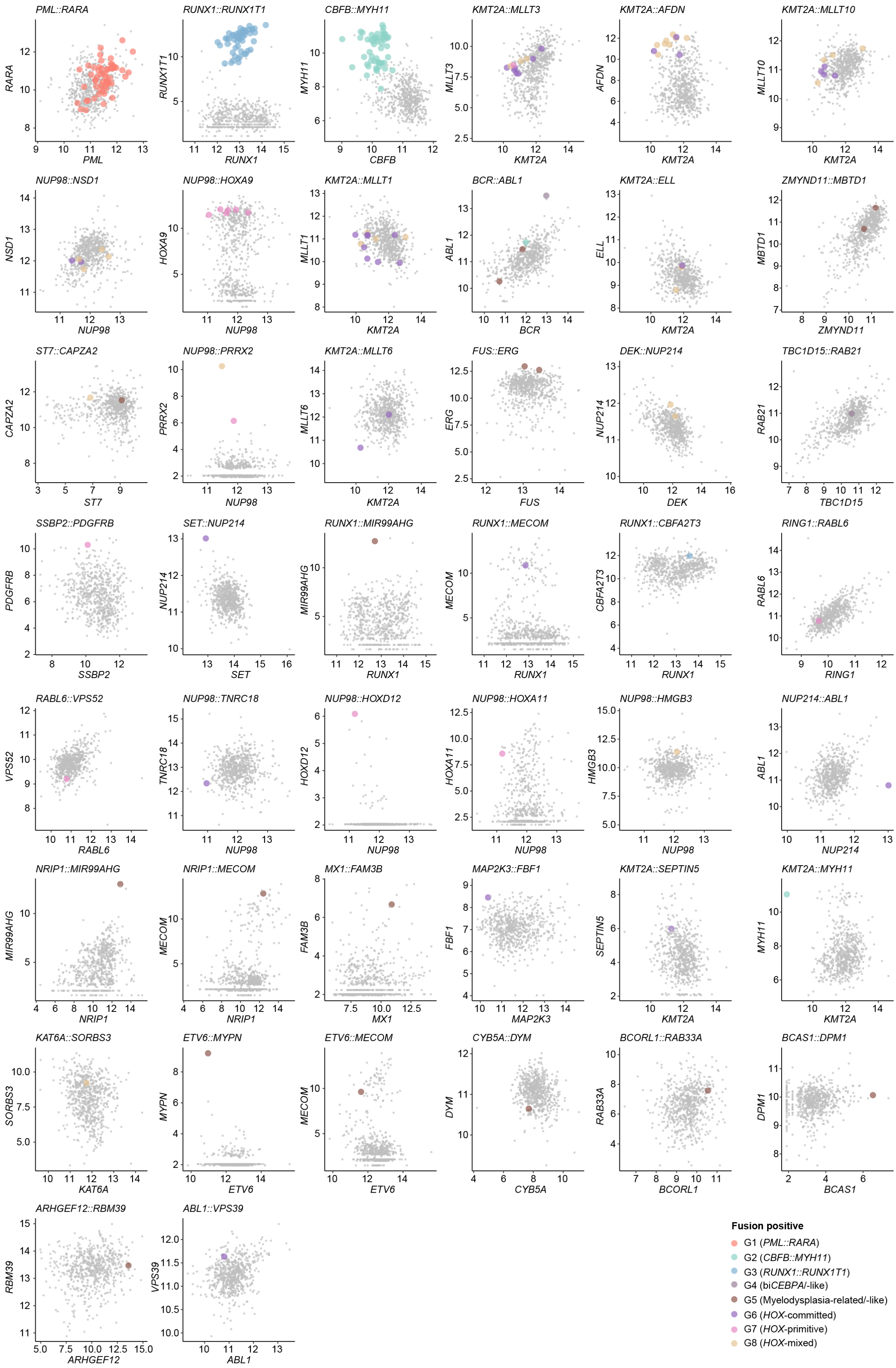

PML::RARA
RARA
PML
RUNX1::RUNX1T1
RUNX1T1
RUNX1
CBFB::MYH11
MYH11
CBFB
KMT2A::MLLT3
MLLT3
KMT2A
KMT2A::AFDN
AFDN
KMT2A
KMT2A::MLLT10
MLLT10
KMT2A
NUP98::NSD1
NSD1
NUP98
NUP98::HOXA9
HOXA9
NUP98
KMT2A::MLLT1
MLLT1
KMT2A
BCR::ABL1
ABL1
BCR
KMT2A::ELL
ELL
KMT2A
ZMYND11::MBTD1
MBTD1
ZMYND11
ST7::CAPZA2
CAPZA2
ST7
NUP98::PRRX2
PRRX2
NUP98
KMT2A::MLLT6
MLLT6
KMT2A
FUS::ERG
ERG
FUS
DEK::NUP214
NUP214
DEK
TBC1D15::RAB21
RAB21
TBC1D15
SSBP2::PDGFRB
PDGFRB
SSBP2
SET::NUP214
NUP214
SET
RUNX1::MIR99AHG
MIR99AHG
RUNX1
RUNX1::MECOM
MECOM
RUNX1
RUNX1::CBFA2T3
CBFA2T3
RUNX1
RING1::RABL6
RABL6
RING1
RABL6::VPS52
VPS52
RABL6
NUP98::TNRC18
TNRC18
NUP98
NUP98::HOXD12
HOXD12
NUP98
NUP98::HOXA11
HOXA11
NUP98
NUP98::HMGB3
HMGB3
NUP98
NUP214::ABL1
ABL1
NUP214
NRIP1::MIR99AHG
MIR99AHG
NRIP1
NRIP1::MECOM
MECOM
NRIP1
MX1::FAM3B
FAM3B
MX1
MAP2K3::FBF1
FBF1
MAP2K3
KMT2A::SEPTIN5
SEPTIN5
KMT2A
KMT2A::MYH11
MYH11
KMT2A
KAT6A::SORBS3
SORBS3
KAT6A
ETV6::MYPN
MYPN
ETV6
ETV6::MECOM
MECOM
ETV6
CYB5A::DYM
DYM
CYB5A
BCORL1::RAB33A
RAB33A
BCORL1
BCAS1::DPM1
DPM1
BCAS1
ARHGEF12::RBM39
RBM39
ARHGEF12
ABL1::VPS39
VPS39
ABL1
Fusion positive
G1 (PML::RARA)
G2 (CBFB::MYH11)
G3 (RUNX1::RUNX1T1)
G4 (biCEBPA/-like)
G5 (Myelodysplasia-related/-like)
G6 (HOX-committed)
G7 (HOX-primitive)
G8 (HOX-mixed)

Fig. S7. Fusion genes and its self-regulation. Gene expression levels of the 5’ (x-axis) and 3’ (y-axis) partner genes involved in fusion genes, and samples positive for corresponding fusions are depicted in red dots.

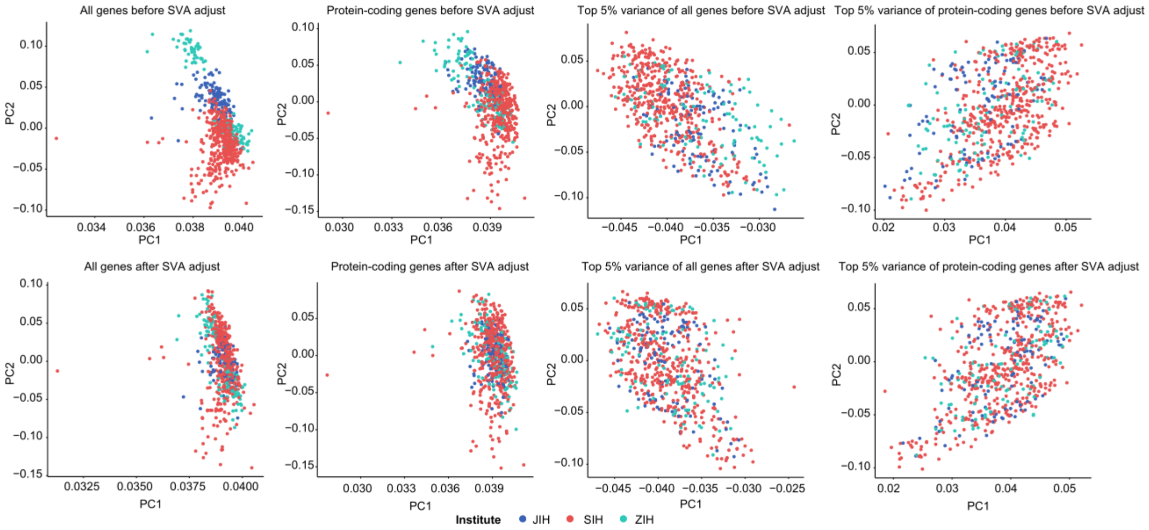


Fig. S8. Batch effect removal of multi-center RNA-Seq data. PCA results in protein-coding genes, top 5% high variance protein-coding genes, all genes, and top 5% of all genes with the largest variance before and after SVA adjust. PCA, principal component analysis; SVA, surrogate variable analysis.

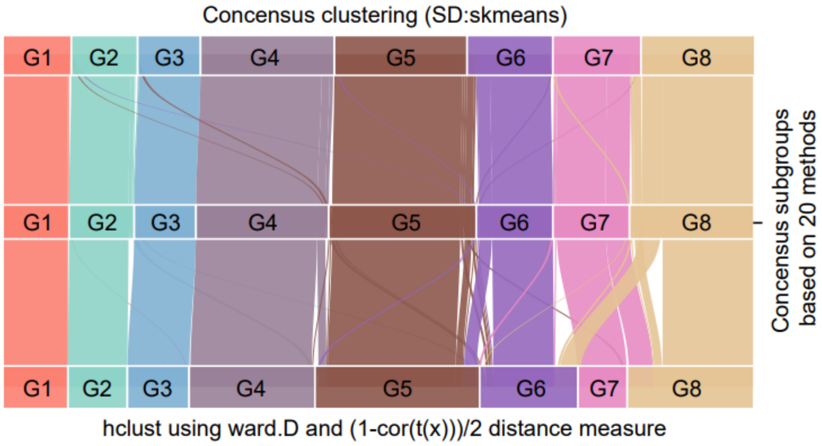


Fig. S9. Comparison of different clustering methods. Sankey plot shows the comparison between consensus clustering and unsupervised hierarchical clustering.

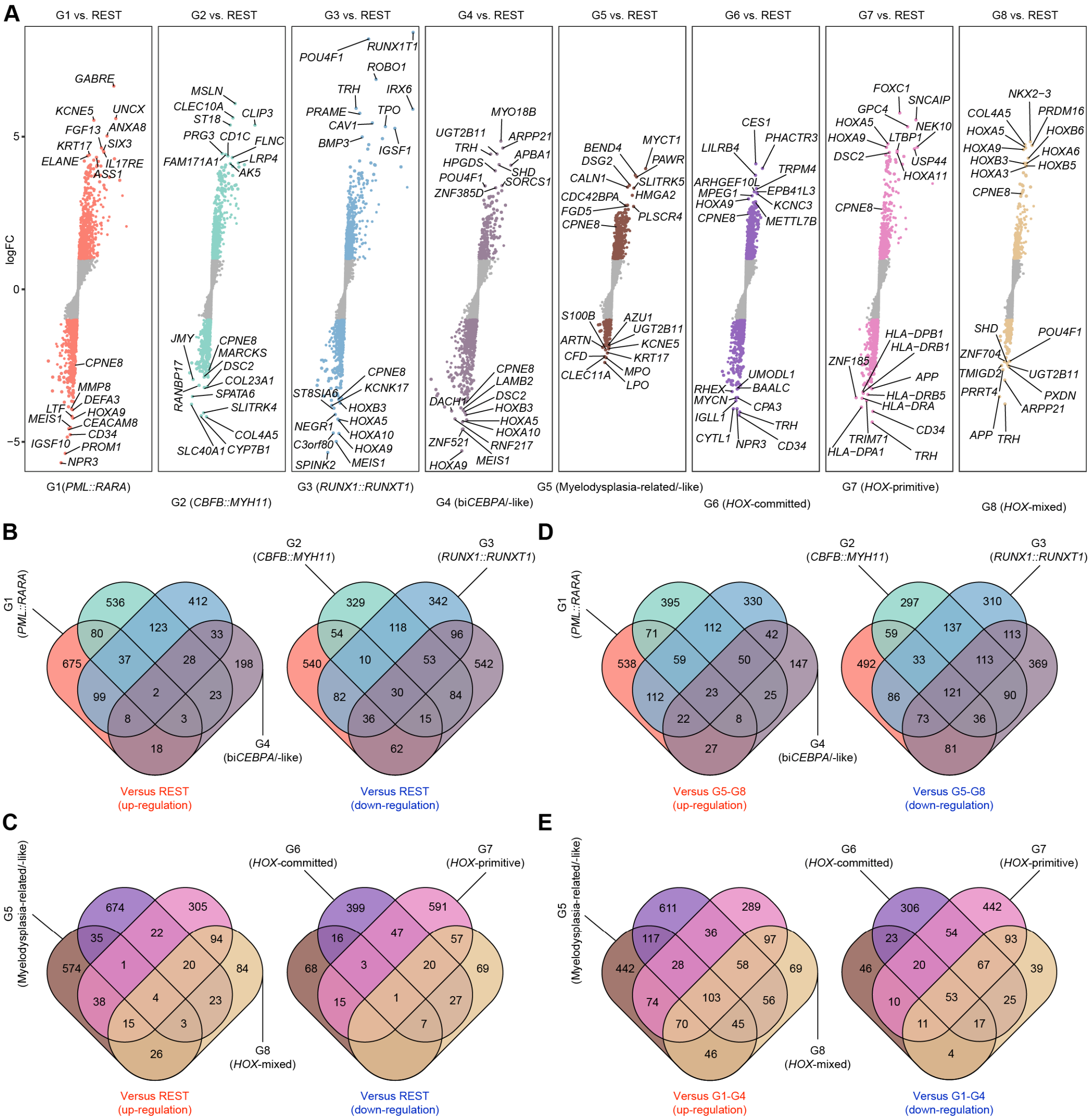


Fig. S10. Extracted gene expression features and overlap between gene expression subgroups. (A) Scatter plots indicates the top up-regulated and down-regulated genes in AML subgroups. Different subgroups of AML are labeled by different colors. The Y-axis indicates $\log_2$ (fold change) of gene expressions versus rest samples. (B-D) respectively shows the gene sets intersections of G1-G4 or G5-G8 versus rest or combined G5-G8/G1-G4 patients.

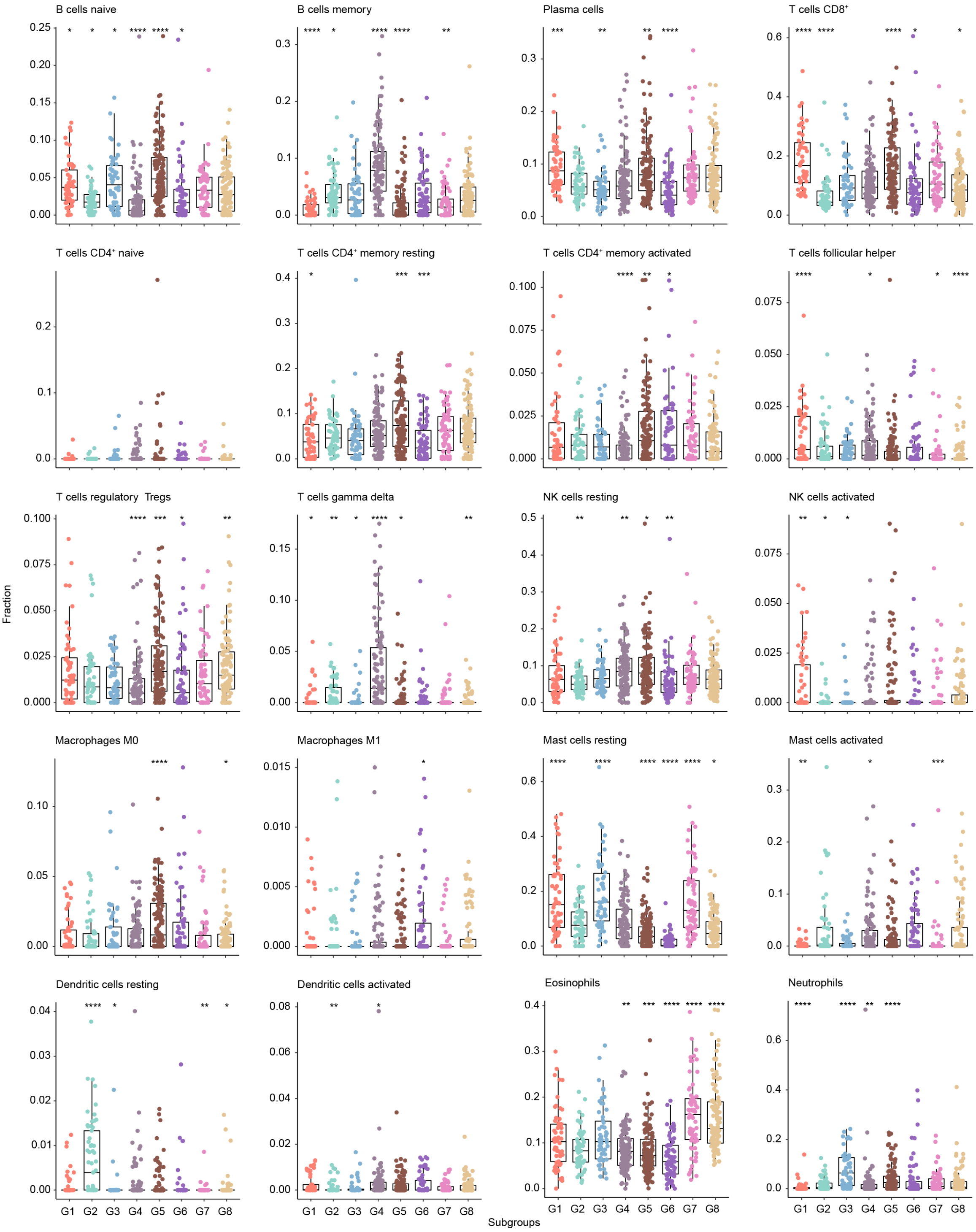


Fig. S11. Other comparison of immune fractions among defined molecular subgroups of AML.

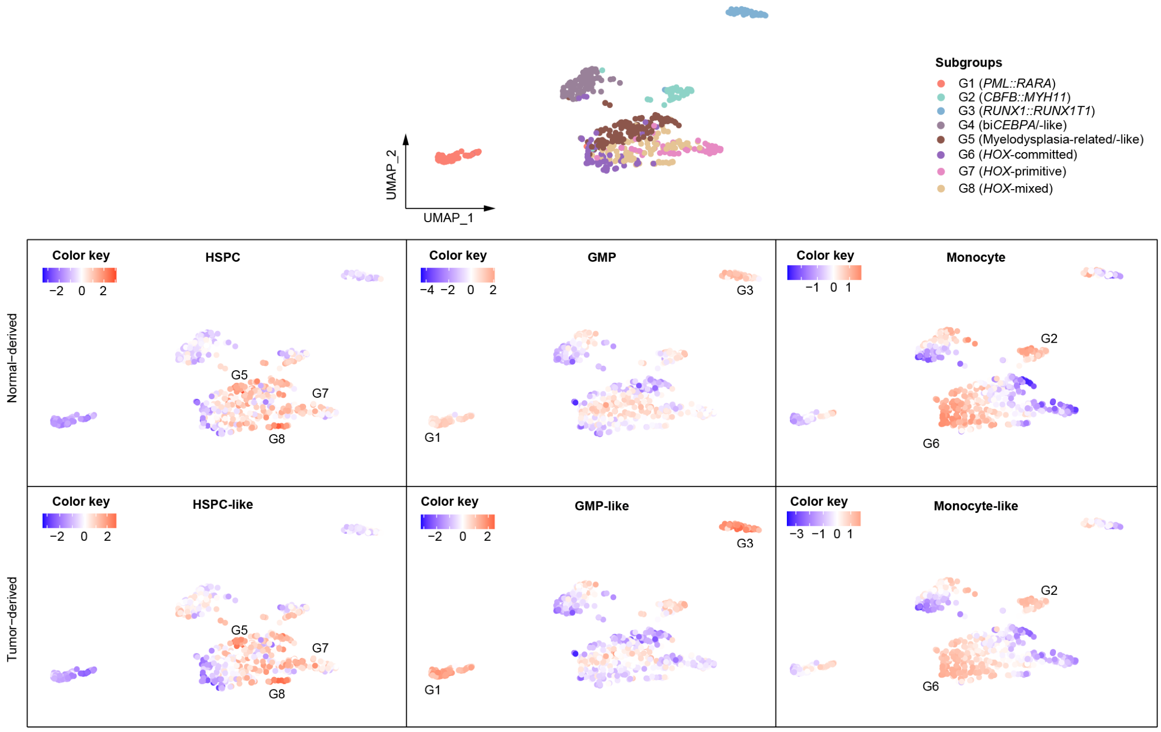


Fig. S12. UMAP visualization of predicted arrest stages of AML patients. Each point represents a patient. Top panel labeled points with defined subgroups. Bottom panel shows the scaled scores calculated based on ssGSEA method and sc-RNASeq extracted gene sets.

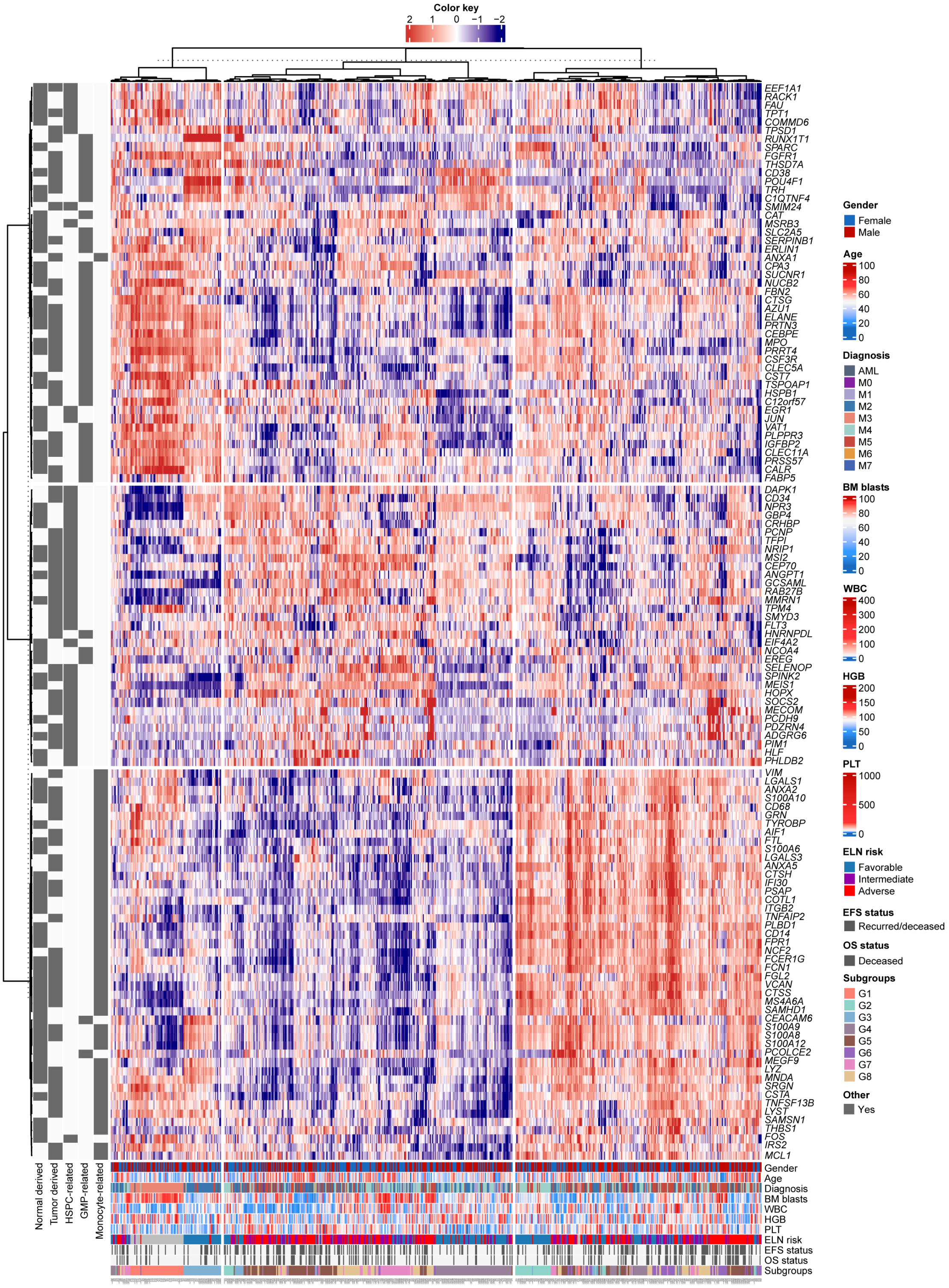


Fig. S13. Hierarchical clustering of signatures related to arrest stages of AML cells. Three clusters of gene and patients could be determined (HSPC, GMP, Monocyte or -like).

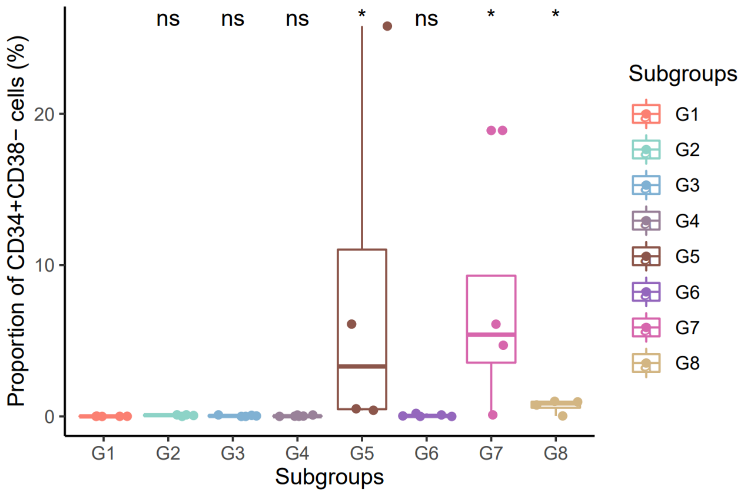


Fig. S14. Comparison of the proportion of CD34$^+$CD38$^-$ cells. Immunophenotypes of 36 AML cases randomly selected from G1–G8 subgroups. The proportion of CD34$^+$CD38$^-$ cells was compared among the eight gene expression subgroups.

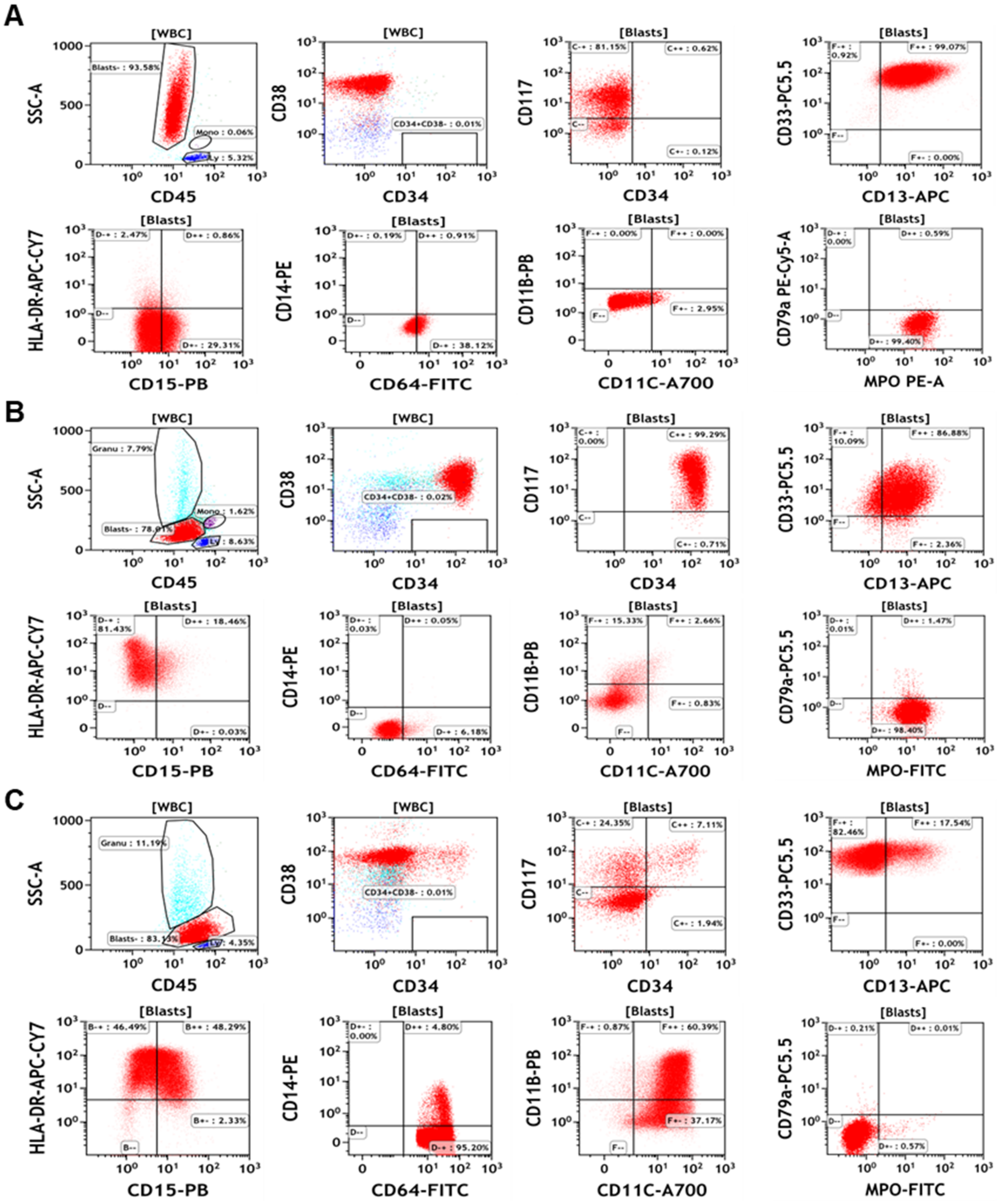


Fig. S15. Representative immunophenotypes detected by flow cytometry. (A) G1 (*PML::RARA*) subgroup exhibits the distinctive immunophenotype of CD34$^-$HLA-DR$^-$ CD117$^+$MPO^{st+}. (B) G3 (*RUNX1::RUNX1T1*) and G4 (bi*CEBPA*/-like) subgroups show CD34$^+$CD38$^+$CD117$^+$MPOst immunophenotype. (C) G2 (*CBFB::MYH11*) and G6 (*HOX*-committed) subgroups present the typical monocytic differentiation phenotype. One representative sample from each immunophenotype group is shown.

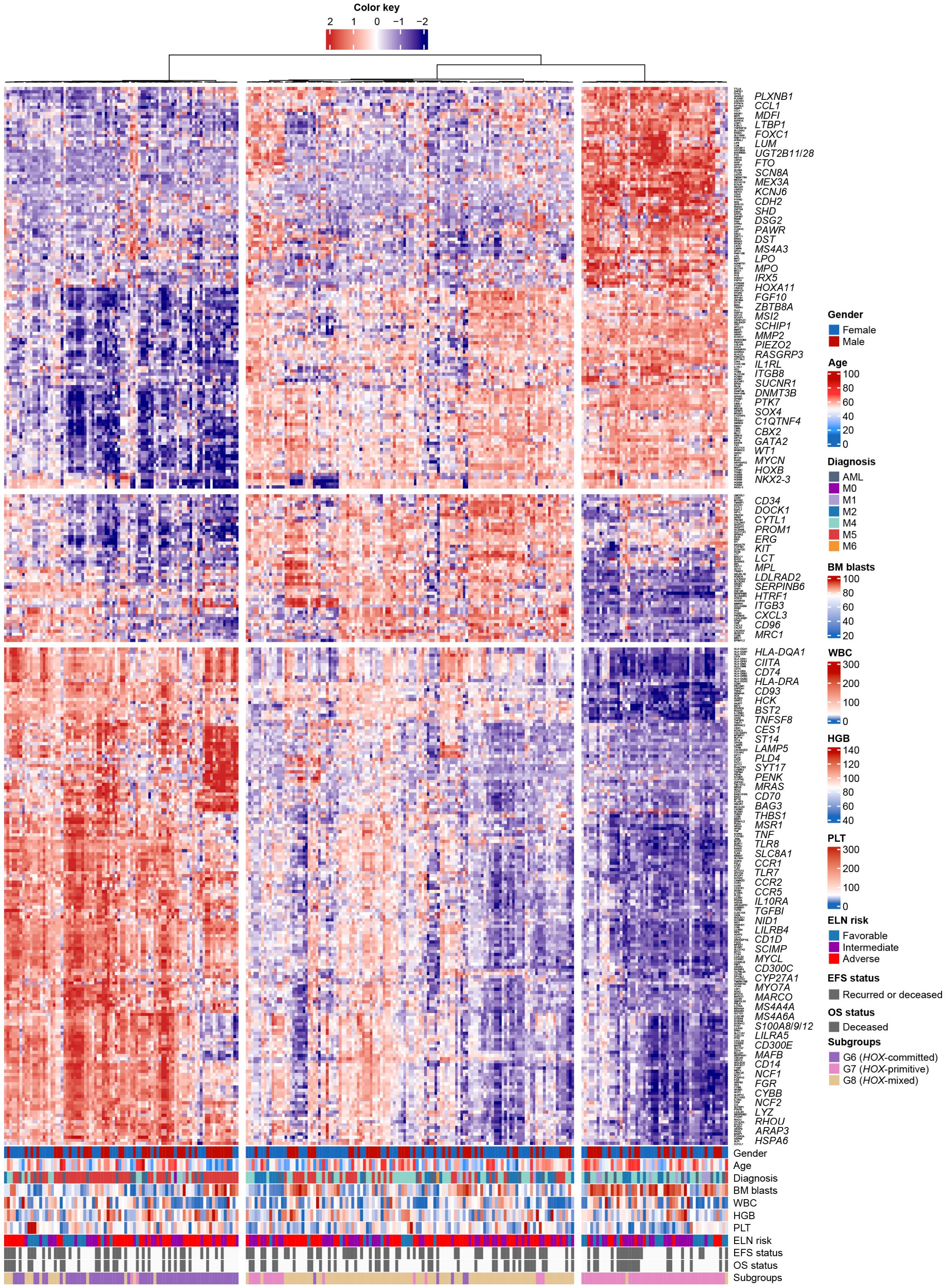


Fig. S16. Hierarchical clustering of differentially expressed genes in G6 to G8 subgroups.

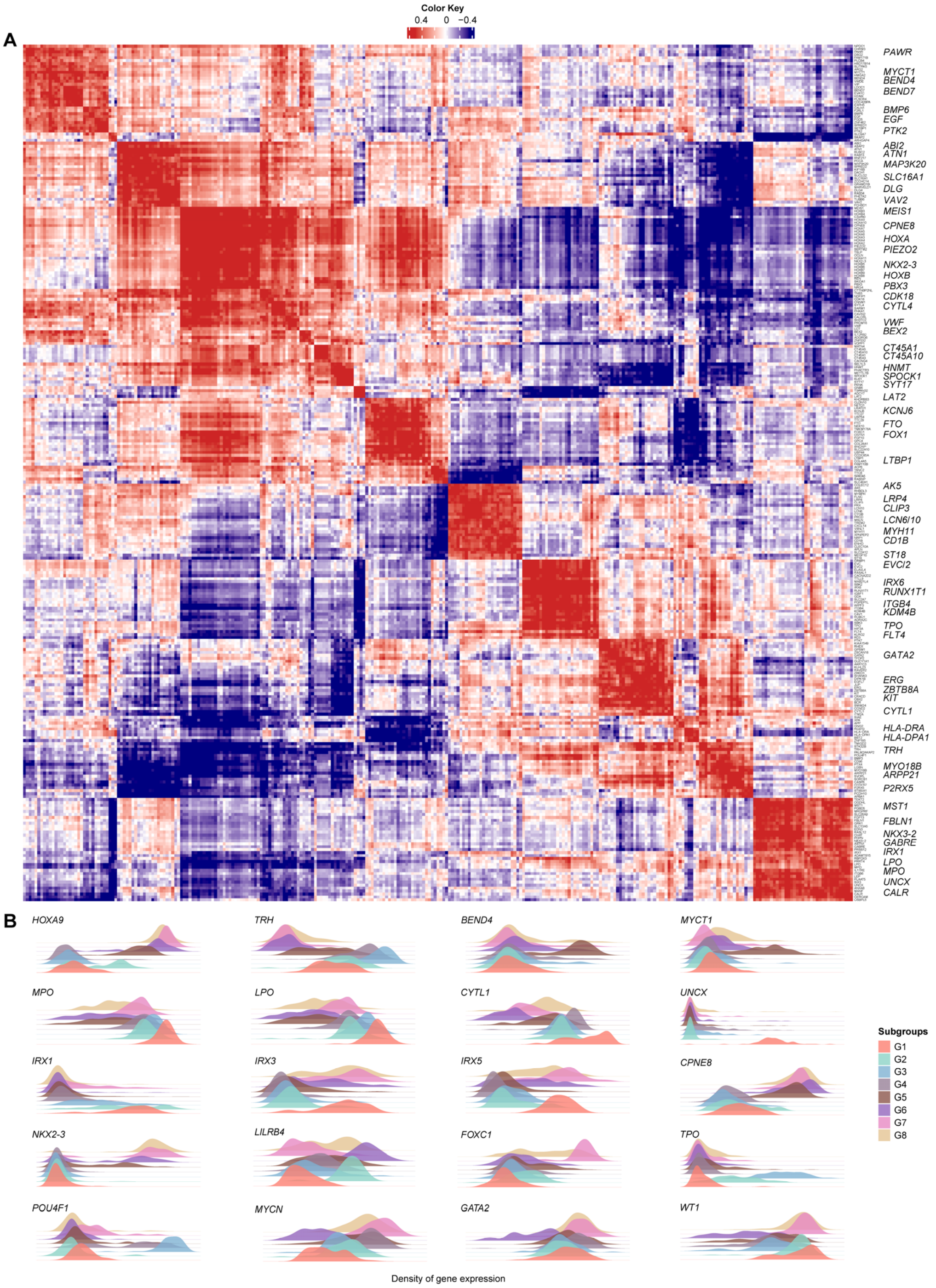
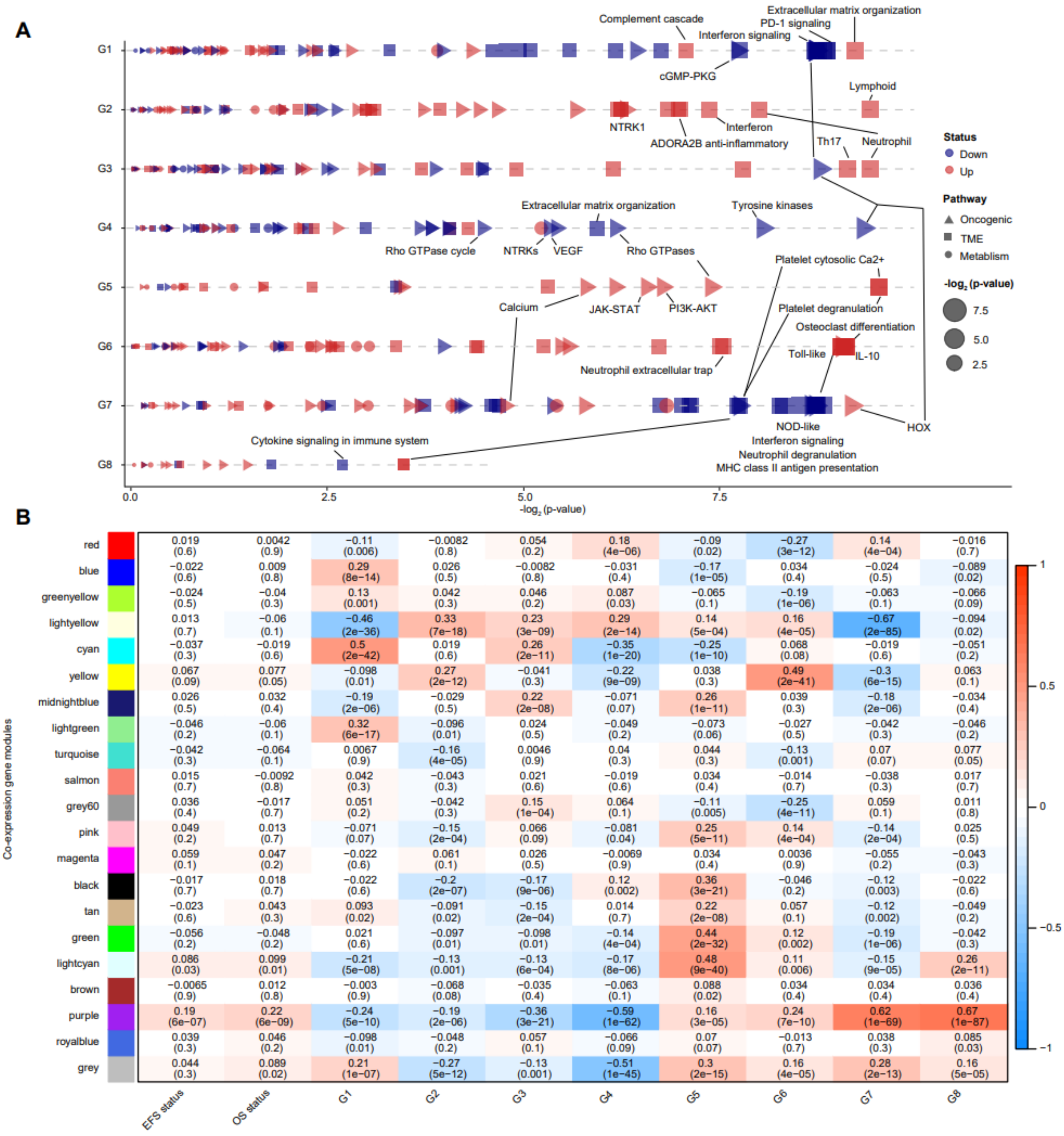


Fig. S17. Co-expression and distribution of hallmark genes in defined AML subgroups. (A) Correlation heatmap of top de-regulated genes in G1-G8 subgroups of AML. (B) Ridge plots shows the gene expression distribution of hallmark genes of G1-G8 subgroups. A higher expression of myeloid differentiation markers *MPO*, *LPO*, and thyrotropin-releasing hormone family gene *TRH* could be seen in G1 to G4 subgroups. In contrast, the significant overexpression of HSPC-related genes *PAWR*, *MYCT1* in G5, embryonic development markers such as *NKX2-3*, *WT1*, *GATA2* and *MYCN* in G7–G8, and while immunoregulatory

factor *LILRB4* and metabolic enzyme *CES1* in G6 were observed, which was concordant with the differentiation stage of each subgroup.



	EFS status	OS status	G1	G2	G3	G4	G5	G6	G7	G8
red	0.019 (0.6)	0.0042 (0.9)	−0.11 (0.006)	−0.0082 (0.8)	0.054 (0.2)	0.18 (4e−06)	−0.09 (0.02)	−0.27 (3e−12)	0.14 (4e−04)	−0.016 (0.7)
blue	−0.022 (0.6)	0.009 (0.8)	0.29 (8e−14)	0.026 (0.5)	−0.0082 (0.8)	−0.031 (0.4)	−0.17 (1e−05)	0.034 (0.4)	−0.024 (0.5)	−0.089 (0.02)
greenyellow	−0.024 (0.5)	−0.04 (0.3)	0.13 (0.001)	0.042 (0.3)	0.046 (0.2)	0.087 (0.03)	−0.065 (0.1)	−0.19 (1e−06)	−0.063 (0.1)	−0.066 (0.09)
lightyellow	0.013 (0.7)	−0.06 (0.1)	−0.46 (2e−36)	0.33 (7e−18)	0.23 (3e−09)	0.29 (2e−14)	0.14 (5e−04)	0.16 (4e−05)	−0.67 (2e−85)	−0.094 (0.02)
cyan	−0.037 (0.3)	−0.019 (0.6)	0.5 (2e−42)	0.019 (0.6)	0.26 (2e−11)	−0.35 (1e−20)	−0.25 (1e−10)	0.068 (0.08)	−0.019 (0.6)	−0.051 (0.2)
yellow	0.067 (0.09)	0.077 (0.05)	−0.098 (0.01)	0.27 (2e−12)	−0.041 (0.3)	−0.22 (9e−09)	0.038 (0.3)	0.49 (2e−41)	−0.3 (6e−15)	0.063 (0.1)
midnightblue	0.026 (0.5)	0.032 (0.4)	−0.19 (2e−06)	−0.029 (0.5)	0.22 (2e−08)	−0.071 (0.07)	0.26 (1e−11)	0.039 (0.3)	−0.18 (2e−06)	−0.034 (0.4)
lightgreen	−0.046 (0.2)	−0.06 (0.1)	0.32 (6e−17)	−0.096 (0.01)	0.024 (0.5)	−0.049 (0.2)	−0.073 (0.06)	−0.027 (0.5)	−0.042 (0.3)	−0.046 (0.2)
turquoise	−0.042 (0.3)	−0.064 (0.1)	0.0067 (0.9)	−0.16 (4e−05)	0.0046 (0.9)	0.04 (0.3)	0.044 (0.3)	−0.13 (0.001)	0.07 (0.07)	0.077 (0.05)
salmon	0.015 (0.7)	−0.0092 (0.8)	0.042 (0.3)	−0.043 (0.3)	0.021 (0.6)	−0.019 (0.6)	0.034 (0.4)	−0.014 (0.7)	−0.038 (0.3)	0.017 (0.7)
grey60	0.036 (0.4)	−0.017 (0.7)	0.051 (0.2)	−0.042 (0.3)	0.15 (1e−04)	0.064 (0.1)	−0.11 (0.005)	−0.25 (4e−11)	0.059 (0.1)	0.011 (0.8)
pink	0.049 (0.2)	0.013 (0.7)	−0.071 (0.07)	−0.15 (2e−04)	0.066 (0.09)	−0.081 (0.04)	0.25 (5e−11)	0.14 (4e−04)	−0.14 (2e−04)	0.025 (0.5)
magenta	0.059 (0.1)	0.047 (0.2)	−0.022 (0.6)	0.061 (0.1)	0.026 (0.5)	−0.0069 (0.9)	0.034 (0.4)	0.0036 (0.9)	−0.055 (0.2)	−0.043 (0.3)
black	−0.017 (0.7)	0.018 (0.7)	−0.022 (0.6)	−0.2 (2e−07)	−0.17 (9e−06)	0.12 (0.002)	0.36 (3e−21)	−0.046 (0.2)	−0.12 (0.003)	−0.022 (0.6)
tan	−0.023 (0.6)	0.043 (0.3)	0.093 (0.02)	−0.091 (0.02)	−0.15 (2e−04)	0.014 (0.7)	0.22 (2e−08)	0.057 (0.1)	−0.12 (0.002)	−0.049 (0.2)
green	−0.056 (0.2)	−0.048 (0.2)	0.021 (0.6)	−0.097 (0.01)	−0.098 (0.01)	−0.14 (4e−04)	0.44 (2e−32)	0.12 (0.002)	−0.19 (1e−06)	−0.042 (0.3)
lightcyan	0.086 (0.03)	0.099 (0.01)	−0.21 (5e−08)	−0.13 (0.001)	−0.13 (6e−04)	−0.17 (8e−06)	0.48 (9e−40)	0.11 (0.006)	−0.15 (9e−05)	0.26 (2e−11)
brown	−0.0065 (0.9)	0.012 (0.8)	−0.003 (0.9)	−0.068 (0.08)	−0.035 (0.4)	−0.063 (0.1)	0.088 (0.02)	0.034 (0.4)	0.034 (0.4)	0.036 (0.4)
purple	0.19 (6e−07)	0.22 (6e−09)	−0.24 (5e−10)	−0.19 (2e−06)	−0.36 (3e−21)	−0.59 (1e−62)	0.16 (3e−05)	0.24 (7e−10)	0.62 (1e−69)	0.67 (1e−87)
royalblue	0.039 (0.3)	0.046 (0.2)	−0.098 (0.01)	−0.048 (0.2)	0.057 (0.1)	−0.066 (0.09)	0.07 (0.07)	−0.013 (0.7)	0.038 (0.3)	0.085 (0.03)
grey	0.044 (0.3)	0.089 (0.02)	0.21 (1e−07)	−0.27 (5e−12)	−0.13 (0.001)	−0.51 (1e−45)	0.3 (2e−15)	0.16 (4e−05)	0.28 (2e−13)	0.16 (5e−05)

Fig. S18. Enriched pathways and correlation between co-expression gene modules in eight gene expression subgroups. (A) The color, shape, and size of each point represents regulatory status, pathway class, and enrichment significance, respectively. (B) Heatmap shows the identified co-expression gene modules and its correlation with OS, EFS and G1–G8 subgroups. Red and blue indicates the positive and negative correlation trend. Correlation coefficient and significance between gene modules and traits are labeled in the box. Selected core networks including purple (G5–G8)/lightcyan (G5/G8) and yellow (G6) are shown in Fig.S18 and Fig.S19, respectively. Gene sets were obtained from the KEGG, GO, and Reactome database. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

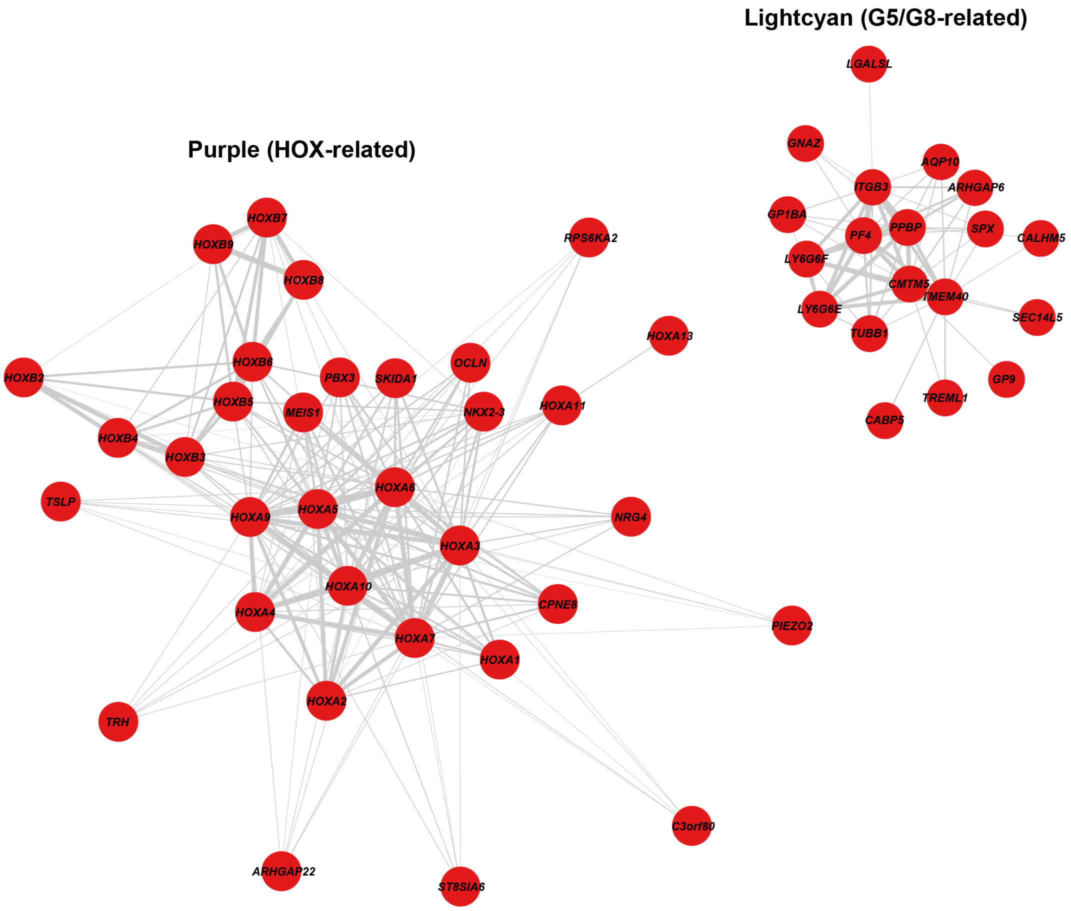


Fig. S19. *HOX*- and G5/G8-related co-expression gene modules. Each node represents a gene. Gene module purple and lightcyan are two core networks that could predict poor prognosis in AML. Lightcyan is associated with the platelet Ca^{2+} signaling pathway.

Yellow (Monocyte-related)

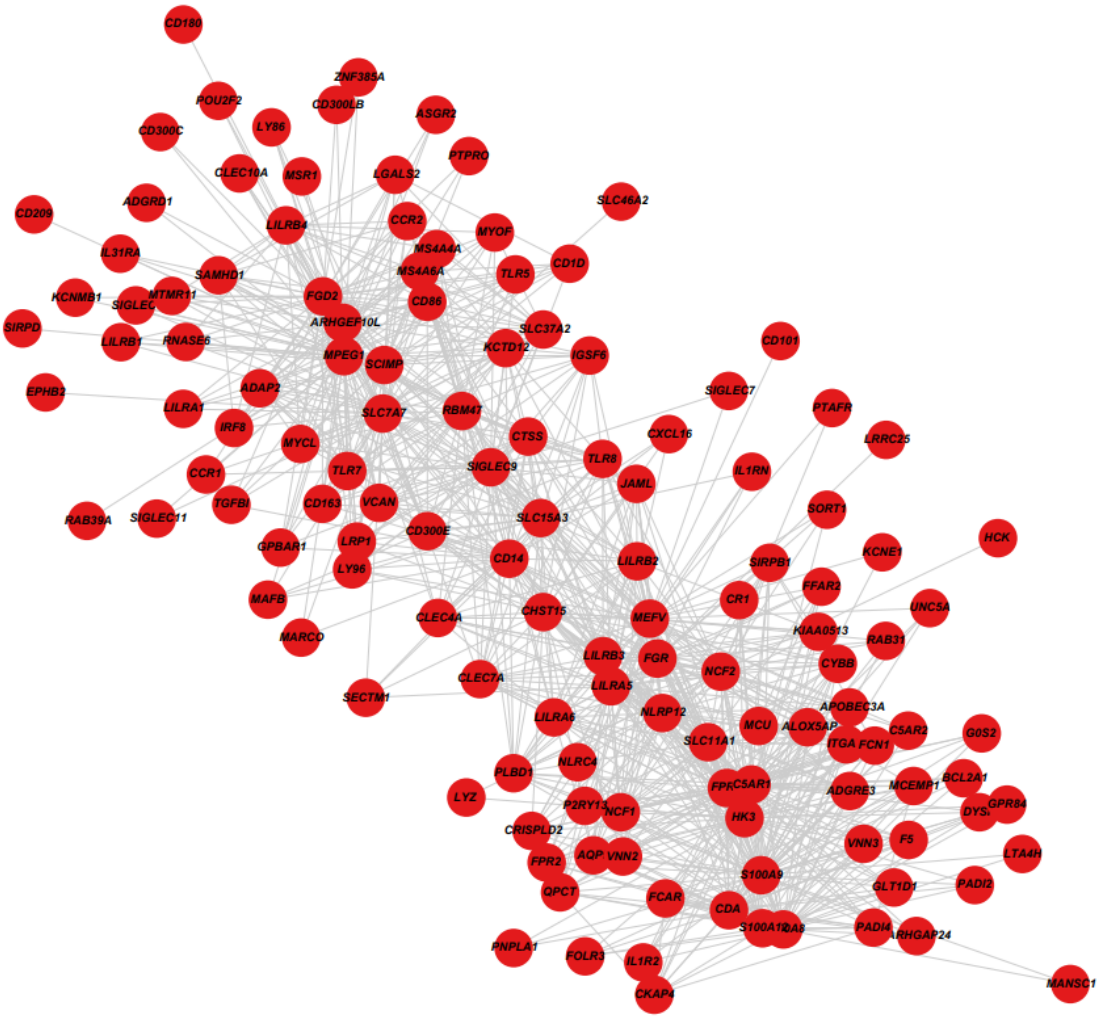
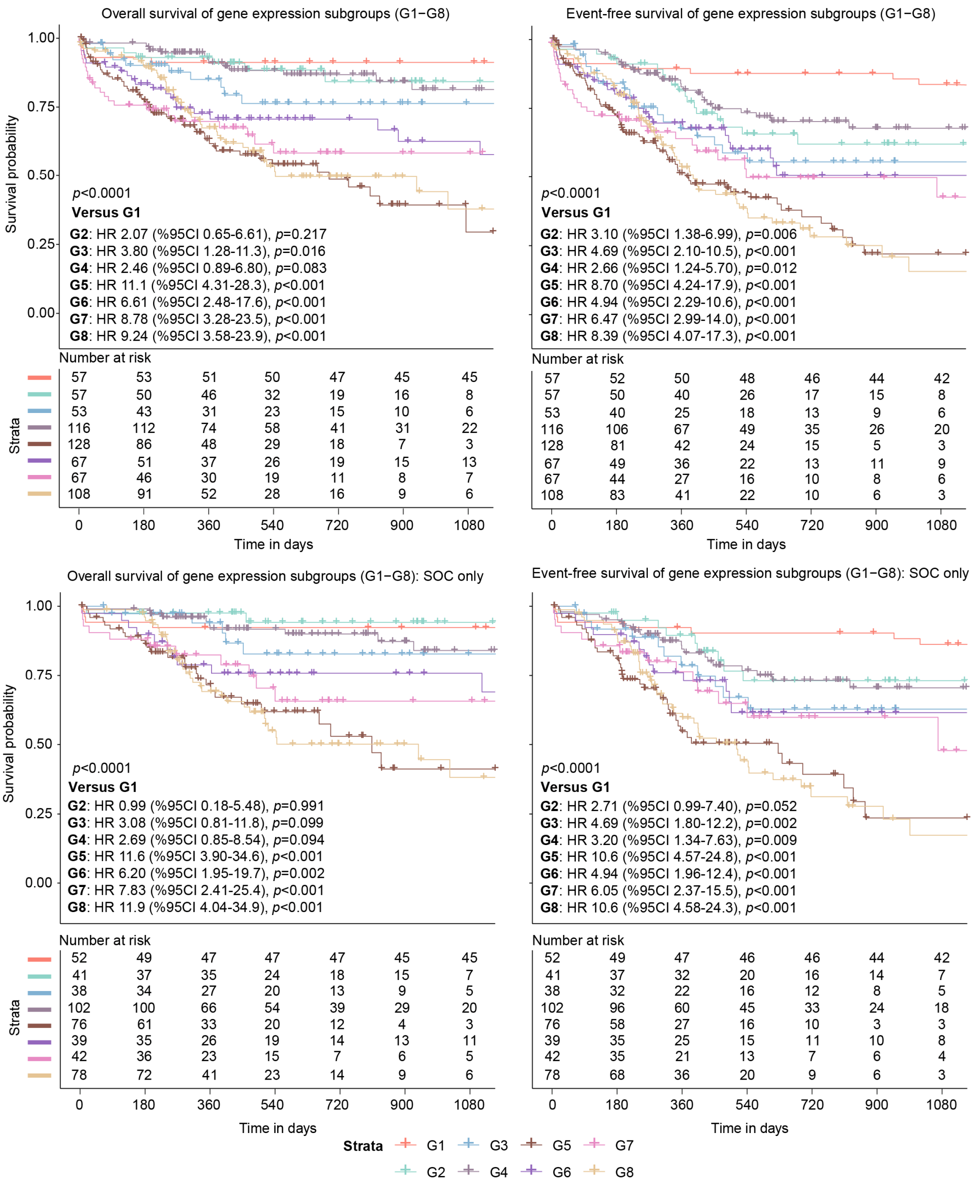


Fig. S20. Monocyte-related co-expression gene network. The gene module (yellow) can be used to predict the resistance to Venetoclax and T-cell inhibitions.



Overall survival of gene expression subgroups (G1−G8)

Event-free survival of gene expression subgroups (G1−G8)

Overall survival of gene expression subgroups (G1−G8): SOC only

Event-free survival of gene expression subgroups (G1−G8): SOC only

Number at risk (Overall survival)

Strata	0	180	360	540	720	900	1080
G1	57	53	51	50	47	45	45
G2	57	50	46	32	19	16	8
G3	53	43	31	23	15	10	6
G4	116	112	74	58	41	31	22
G5	128	86	48	29	18	7	3
G6	67	51	37	26	19	15	13
G7	67	46	30	19	11	8	7
G8	108	91	52	28	16	9	6

Number at risk (Event-free survival)

Strata	0	180	360	540	720	900	1080
G1	57	52	50	48	46	44	42
G2	57	50	40	26	17	15	8
G3	53	40	25	18	13	9	6
G4	116	106	67	49	35	26	20
G5	128	81	42	24	15	5	3
G6	67	49	36	22	13	11	9
G7	67	44	27	16	10	8	6
G8	108	83	41	22	10	6	3

Number at risk (Overall survival: SOC only)

Strata	0	180	360	540	720	900	1080
G1	52	49	47	47	47	45	45
G2	41	37	35	24	18	15	7
G3	38	34	27	20	13	9	5
G4	102	100	66	54	39	29	20
G5	76	61	33	20	12	4	3
G6	39	35	26	19	14	13	11
G7	42	36	23	15	7	6	5
G8	78	72	41	23	14	9	6

Number at risk (Event-free survival: SOC only)

Strata	0	180	360	540	720	900	1080
G1	52	49	47	46	46	44	42
G2	41	37	32	20	16	14	7
G3	38	32	22	16	12	8	5
G4	102	96	60	45	33	24	18
G5	76	58	27	16	10	3	3
G6	39	35	25	15	11	10	8
G7	42	35	21	13	7	6	4
G8	78	68	36	20	9	6	3

Fig. S21. Survival analysis of defined AML subgroups. Top panel respectively shows overall survival (OS) and event-free survival (EFS) of AML subgroups. Bottom panel only keep the patients that were treated with standard 3+7 chemotherapy.

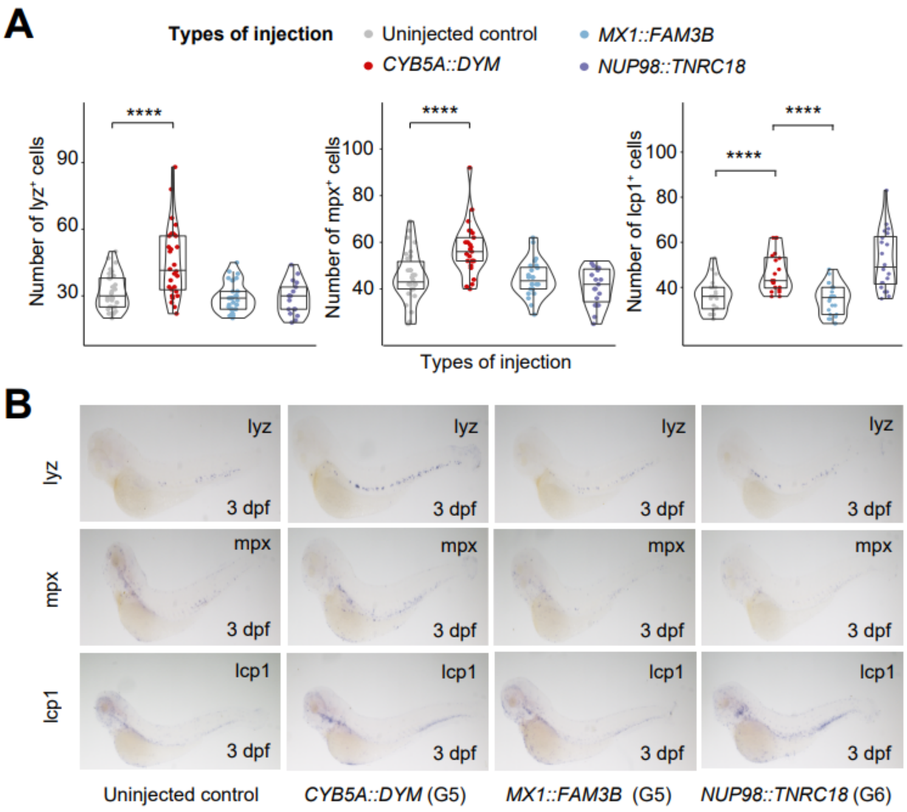


Fig. S22. Functional exploration of novel fusion genes in zebrafish. Three newly identified in-frame fusion genes (*CYB5A::DYM*, *MX1::FAM3B* in G5, and *NUP98::TNRC1* in G6) from patients who had a grim prognosis but lacked known strong leukemogenic mutations were selected. (A) Significantly increased lyz+, mpx+, and lcp1+ cells in zebrafish embryos injected with *CYB5A::DYM* mRNAs at 3 dpf, and increased lcp1+ cells in *NUP98::TNRC18* as compared with control zebrafish. (B) WISH assays conducted in 3-dpf zebrafish embryos injected with three novel fusions, with uninjected zebrafish embryos being the control. dpf, days postfertilization; WISH, whole mount RNA in situ hybridization.

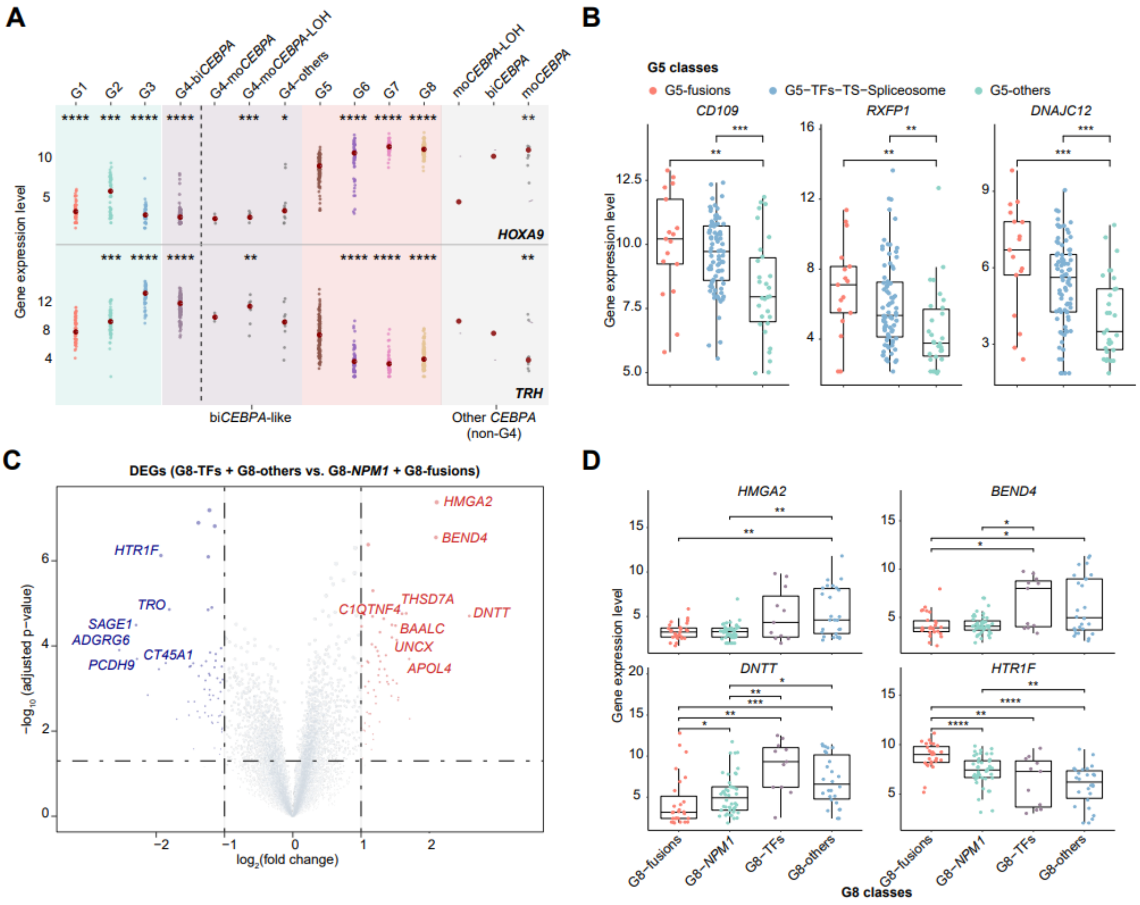


Fig. S23. Use case of gene expression classification: screen of prognostic genes from inter- and intra-subgroups. (A) Dot plot shows the gene expression level of *HOXA9* and *TRH* between different molecular subtypes including *CEBPA* mutant status. The gene expression level of *HOXA9* and *TRH* are highly similar in G4-bi*CEBPA* and bi*CEBPA*-like clusters, which were reversed in other *CEBPA* mutations. (B–D) Some differentially expressed genes with different molecular mutations within G5 and G8. The *CD109* was successfully screened in one of the emerging prognostic risk scores of AML (15).

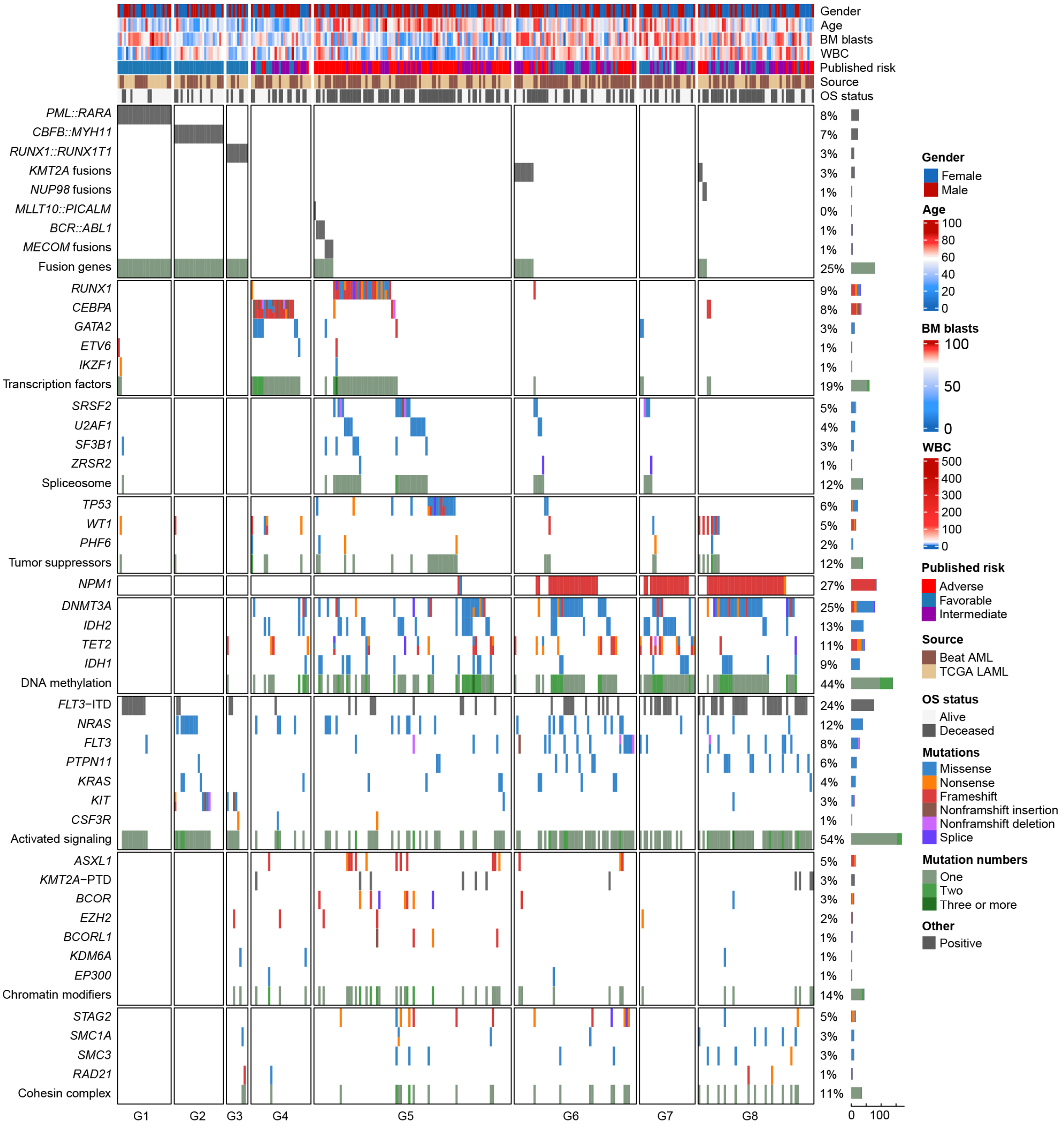


Fig. S24. The oncogenic landscape of G1-G8 subgroups in the merged TCGA LAML and Beat AML cohorts. Each row represents a genomic feature. The clinical annotations are shown in the top panel. Each column is a patient. Different types of mutations are labeled with different colors. The genomic landscape can validate the classification of G1-G8 enriched genomic events including *PML::RARA*, *CBFB::MYH11*, *RUNX1::RUNX1T1*, *CEBPA*, *RUNX1*, *TP53*, *DNMT3A*, and *NPM1* mutations.

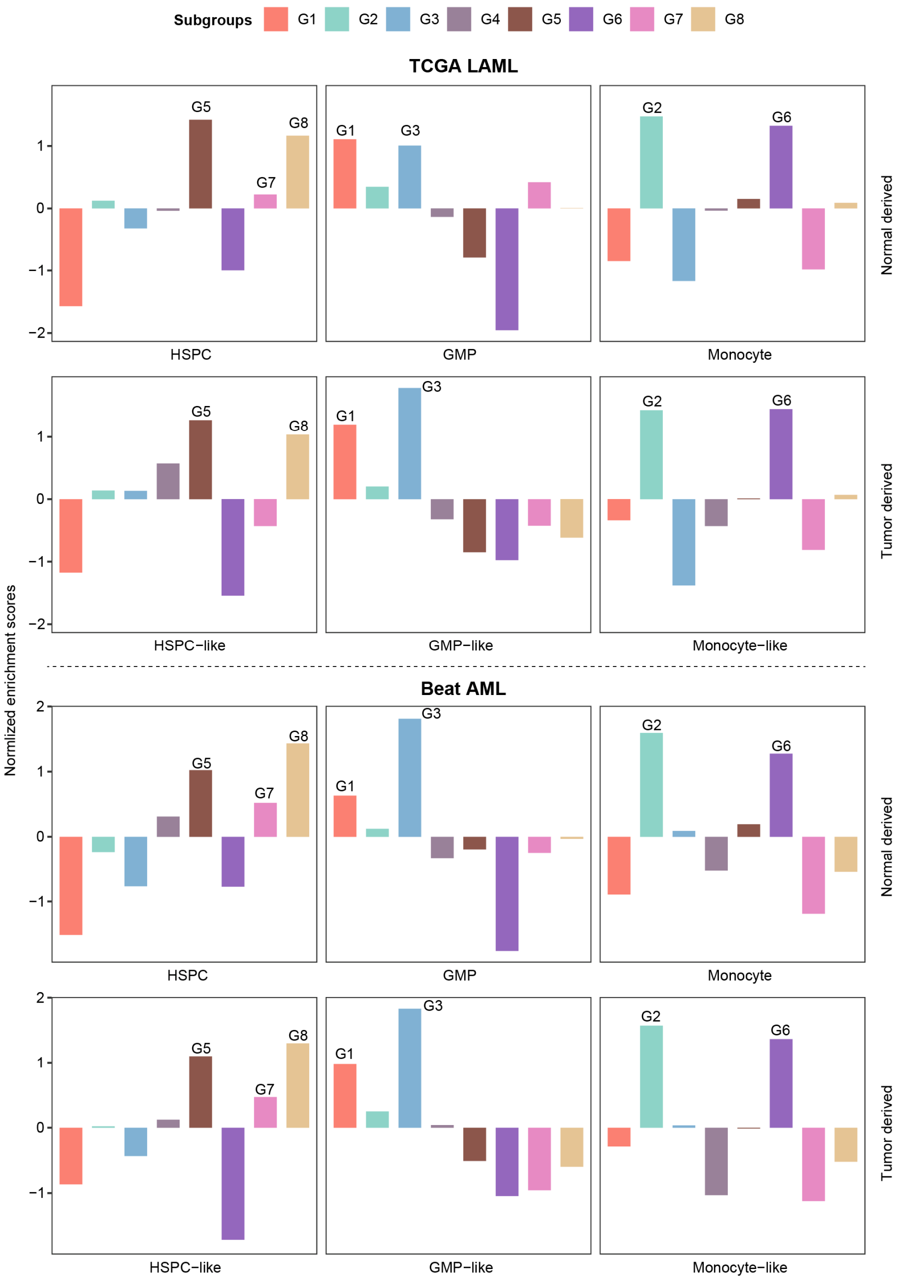


Fig. S25. Validation of differentiation signatures of G1-G8 based on TCGA LAML and Beat AML cohorts. The bar plots of normalized enrichment score of TCGA LAML and Beat AML cohorts are shown in the top and bottom panels. The distribution pattern of TCGA LAML and Beat AML cohorts are trend consistent, i.e., HSPC/-like of G5/G8, GMP/-like of G1/G3 and Monocyte/-like of G2 and G6.

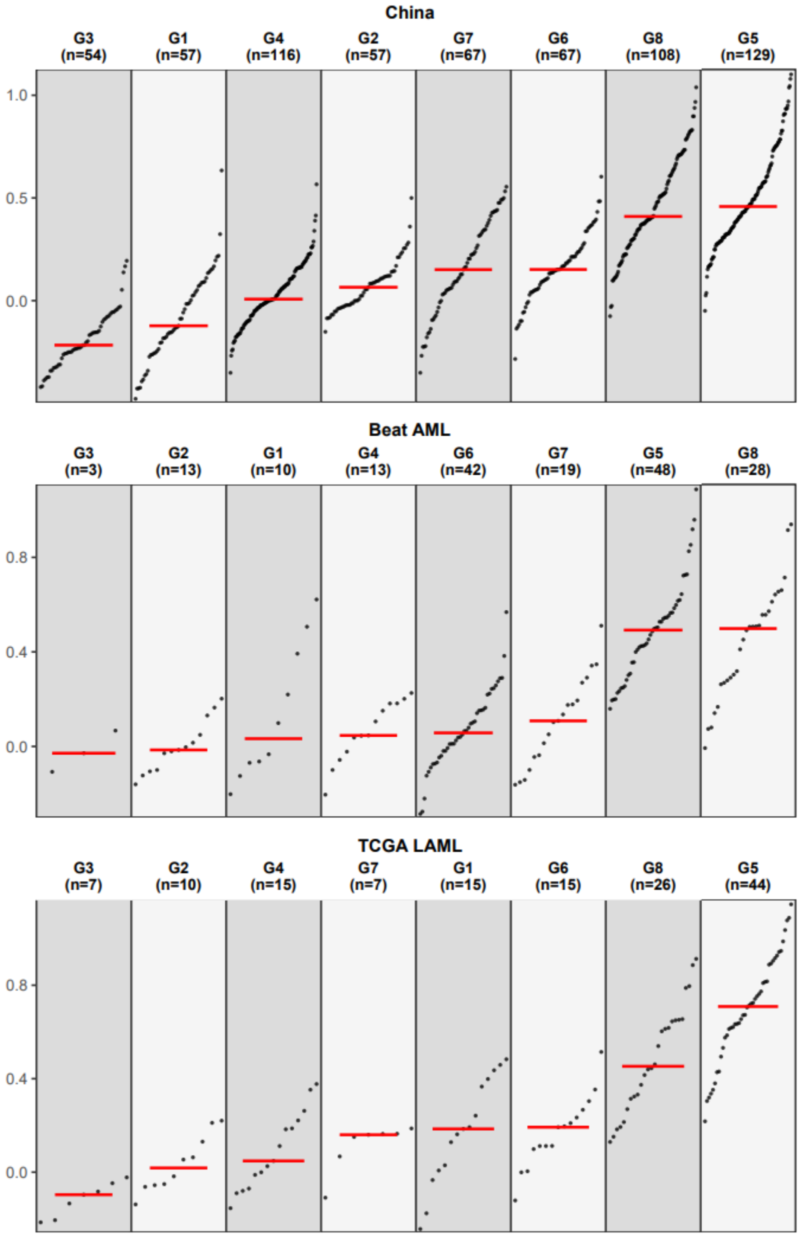


Fig. S26. LSC17 risk scores in three independent cohorts including China, Beat AML and TCGA LAML. The subgroups are ordered by median of LSC17 from low to high**.** The high LSC17 risk scores are strongly associated with G5 and G8 subgroups in three independent cohorts, which also confirmed the HSPC/-like gene signatures.

SI Tables

Table S1. Clinical characteristics of newly diagnosed AML patients in the whole cohort and three participating centers.

Factor	Overall (n = 655)	SIH (n = 442)	JIH (n = 110)	ZIH (n = 103)
Age (year)				
Median (IQR)	48 (34–60)	50 (36–62)	37 (26–45.8)	57 (42.5–65.5)
Male gender, n (%)	339 (51.8)	225 (50.9)	65 (59.1)	49 (47.6)
WBC, × 10^9/L				
Median (IQR)	12.4 (3.5–44)	9.9 (3–40.5)	35.1 (11.8–63.6)	8.8 (2.7–31.6)
HGB, g/L				
Median (IQR)	86 (68–106.5)	86 (67–106)	87 (71–107)	88 (67–104)
PLT, × 10^9/L				
Median (IQR)	40 (23–78)	41 (23–81)	32 (18.5–61.5)	51 (26–78.5)
Bone marrow blasts, %				
Median (IQR)	68 (45–84)	68 (45.6–84.5)	69.80 (48–85)	68 (42–80.8)
WHO category, n (%)				
AML with defining genetic abnormalities				
APL with *PML::RARA* fusion	56 (8.5)	52 (11.8)	0 (0.0)	4 (3.9)
AML with *RUNX1::RUNX1T1* fusion	53 (8.1)	33 (7.5)	8 (7.3)	12 (11.7)
AML with *CBFB::MYH11* fusion	51 (7.8)	29 (6.6)	15 (13.6)	7 (6.8)
AML with *DEK::NUP214* fusion	2 (0.3)	2 (0.5)	0 (0.0)	0 (0.0)
AML with *BCR::ABL1* fusion	3 (0.5)	0 (0.0)	1 (0.9)	2 (1.9)
AML with *KMT2A* rearrangement	40 (6.1)	18 (4.1)	12 (10.9)	10 (9.7)
AML with *MECOM* rearrangement	3 (0.5)	2 (0.5)	1 (0.9)	0 (0.0)
AML with *NUP98* rearrangement	18 (2.7)	10 (2.3)	7 (6.4)	1 (1.0)
AML with *NPM1* mutation	120 (18.3)	91 (20.6)	11 (10.0)	18 (17.5)

(Continued on next page…)

Factor	Overall (n = 655)	SIH (n = 442)	JIH (n = 110)	ZIH (n = 103)
AML with *CEBPA* mutation	99 (15.1)	65 (14.7)	26 (23.6)	8 (7.8)
AML, myelodysplasia-related	126 (19.2)	91 (20.6)	13 (11.8)	22 (21.4)
FAB category, n (%)				
AML with minimal differentiation	2 (0.3)	0	0	2 (1.9)
AML without maturation	33 (5)	7 (1.6)	23 (20.9)	3 (2.9)
AML with maturation	136 (20.8)	62 (14.0)	36 (32.7)	38 (36.9)
Acute promyelocytic leukemia	56 (8.5)	52 (11.8)	0 (0.0)	4 (3.9)
Acute myelomonocytic leukemia	189 (28.9)	164 (37.1)	17 (15.5)	8 (7.8)
Acute monoblastic or monocytic leukemia	165 (25.2)	92 (20.8)	28 (25.5)	45 (43.7)
Pure erythroid leukemia	1 (0.2)	0	1 (0.9)	0
Acute megakaryoblastic leukemia	1 (0.2)	0	1 (0.9)	0
AML, undefined	72 (11)	65 (14.7)	4 (3.6)	3 (2.9)
Karyotype, n (%)				
Normal karyotype	285 (43.5)	173 (39.1)	60 (54.5)	52 (50.5)
Complex karyotype	54 (8.2)	42 (9.5)	4 (3.6)	8 (7.8)
Other abnormal karyotype	295 (45)	219 (49.6)	46 (41.9)	30 (29.1)
Unknown	21 (3.2)	8 (1.8)	0	13 (12.6)
ELN risk, n (%)				
Favorable	225 (34.4)	149 (33.7)	45 (40.9)	31 (30.1)
Intermediate	143 (21.8)	88 (19.9)	26 (23.6)	29 (28.2)
Adverse	230 (35.1)	152 (34.4)	39 (35.5)	39 (37.9)
Unknown	57 (8.7)	53 (12.0)	0	4 (3.9)

ELN, European LeukmiaNet; HGB, hemoglobin; IQR, interquartile range; NOS, not otherwise specified; PLT platelet; WBC, white blood count.

Table S2. Novel fusions identified by RNA–Seq.

Sample	Sex/Age	Dx	BM blast (%)	Fusion genes	Reading frame	Kayotype	Gene mutations (VAF)	Prognosis
SIH035	M/72	AML	46.5	*BCAS1::DPM1*	out-of-frame	44~48, X, -Y, -13, -19, -20, +dm, M1~M6[cp4]/45, X, -Y/46, XY	*TP53* p.L93Rfs*29 (48.2%)	Dead
SIH053	M/52	M4	57.5	*ST7::CAPZA2*	out-of-frame	46, XY	*FLT3*-ITD (19.2%); *DNMT3A* p.R882C (42.5%); *NPM1* p.W288Cfs*11 (37.6%); *ZBTB7A* p.N153S (38.9%)	Dead
SIH128	M/60	M5	75	*MX1::FAM3B*	in-frame	43~45, XY, -7[cp9]/43~45, XY, inv(3)(p21q21), -7, [cp7]/46, XY	*DNMT3A* p.R544Hfs*30 (42.5%); *DNMT3A* p.R320* (45.4%); *IDH2* p.R140Q (44.9%); *KANSL1* p.K692Qfs*4 (20.7%); *KANSL1* p.K267Sfs*12 (21.8%)	Dead
SIH135	F/64	M5	28	*ARHGEF12::RBM39*	in-frame	46, XX	*KMT2A*-PTD; *PTPN11* p.E76K (21.2%); *RUNX1* p.F416Lfs*185 (37.4%); *U2AF1* p.Q157R (40.2%); *BCOR* p.E1611Mfs*8 (27.4%)	Relapsed
SIH172	M/42	M5	48	*RUNX1::MIR99AHG*	out-of-frame	43~45, XY, -18[CP4]/46, XY	*FLT3* p.D835Y (21%); *SETD2* p.P241A (45%)	Alive
SIH181	M/84	M5	72.5	*MAP2K3::FBF1*	out-of-frame	36~46, XY, +6, +8, -11, -12, -16, -17, +M2~M4[cp7]/46, XY	*TP53* p.V203E (43%); *FAT1* p.I2236T (37.8%); *FAT1* p.Y500* (21%); *SSTR5* p.A159V (24%)	Dead
SIH202	M/63	AML	86.5	*RABL6::VPS52* *RING1::RABL6*	both in-frame	46, XY	*KMT2A*-PTD	Dead
SIH219	M/66	AML	38	*CYB5A::DYM*	in-frame	73~78,XXYY,+add(10q24)x2,+M1~M7[cp4]/74-76,XXYY,+add(10q24),+add(14q32)+M5~M6[cp2]/46,XY	*TP53* p.R248W (81%)	Dead
SIH225	M/54	M4	79.5	*NUP98::TNRC18*	in-frame	47, XY, +8/46, XY[1/25]	*NRAS* p.Q61H (21.7%); *IDH1* p.R132C (42%); *SMC1A* p.R586Q (83%)	Dead
SIH246	F/61	AML	29.5	*ETV6::MYPN*	out-of-frame	ND	*KRAS* p.G13D (35%); *TP53* p.Y327* (43%); *TP53* p.S215N (45%)	Dead

(Continued on next page…)

Sample	Sex/Age	Dx	BM blast (%)	Fusion genes	Reading frame	Kayotype	Gene mutations (VAF)	Prognosis
SIH280	M/34	M4	26.5	*ST7::CAPZA2*	out-of-frame	46, XY	*ETV6* p.R369Q (44%); *KRAS* p.Q61R (25%); *ASXL1* p.T822fs (47%); *U2AF1* p.S34Y (45%)	No remission
SIH302	M/68	M4	67.5	*BCORL1::RAB33A*	out-of-frame	47, XY, +8/46, XY	*DNMT3A* p.R882S (48.6%); *TET2* p.K1254fs (45.3%); *KRAS* p.G60V (3.4%); *KRAS* p.T58I (12.2%); *RUNX1* p.N448fs (82.8%); *U2AF1* p.S34F (43.6%); *BCOR* p.S1077fs (91.4%); *BCORL1* p.V1389fs (8.1%)	Relapsed
JIH107	M/16	M4	68	*NUP98::HOXD12*	in-frame	46, XY, t(2, 11)(q31; p15)	*EZH2* p.G630R (20%); *PTPN11* p.T73I (42%); *RUNX1* p.L102delinsPPFVL (23%)	Alive
SIH381	F/67	M4	21	*TBC1D15::RAB21*		47, XX, +8/46, XX	*DNMT3A* p.I407T (24.1%); *IDH2* p.M397V (44.3%); *CEBPA* p.Q311fs (21.25%);*CEBPA* p.E89fs (26.4%);*SMC1A* p.N40D (27.3%)	Alive
SIH441	M/22	M5	93	*SET::NUP214, ABL1::VPS39, NUP214::ABL1*	all in-frame	85~92,XXYY[cp7]/46,XY	*BRAF* p.Q257R (42.1%); *NOTCH1* p.P2512L (44.4%); *NF1* p.M2569fs (72.7%); *RUNX1* p.G168fs (35.3%); *PHF6* p.Y303D (84.5%)	Alive
SIH458	M/57	M4	37	*KAT6A::SORBS3*	in-frame	46, XY, del(8)(p11p21)/46, XY	*DNMT3A* p.R882C (40.1%); *IDH1* p.R132H (39.4%)	Alive

BM, bone marrow; Dx, diagnosis; ND, not determined; VAF, variant allele frequency

Table S3. Clinical and molecular features of the robust gene expression subgroups (G1–G8).

Factor	**G1 (n = 57, 8.7%)**	**G2 (n = 57, 8.7%)**	**G3 (n = 54, 8.2%)**	**G4 (n = 116, 17.7%)**	**G5 (n = 129, 19.7%)**	**G6 (n = 67, 10.2%)**	**G7 (n = 67, 10.2%)**	**G8 (n = 108, 16.5%)**
Age (year)								
Median (IQR)	39 (28–48)	42 (32–55)	46 (28.3–55)	43.5 (34–56)	58 (45–66)	48 (31.5–59)	55 (40–63.5)	51 (37.8, 60.8)
Male gender, n (%)	28 (49.1)	32 (56.1)	27 (50.0)	73 (62.9)	74 (57.4)	36 (53.7)	26 (38.8)	43 (39.8)
WBC, × 10^9/L								
Median (IQR)	4.1 (1.6–16.2)	38 (11.9–60.2)	9.5 (4.8–17.1)	12.9 (5.2–41.9)	3.6 (1.6–15.3)	28.2 (4.7–54.8)	31.6 (3.7–79.3)	30.9 (6.8–69.4)
HGB, g/L								
Median (IQR)	95 (73–117)	85 (77–107)	77.5 (57–93.8)	99 (75.5–117)	75 (60–94.8)	87 (72–110)	81 (67–95)	83 (70.5–103)
PLT, × 10^9/L								
Median (IQR)	33 (16–64)	31 (24–57)	26.5 (17.8–39)	24 (13–42)	57.5 (27.3–108)	69 (46.8–111.5)	48 (26–81)	49 (30–86)
BM blasts, %								
Median (IQR)	87.5 (81–91.5)	66.8 (55.5–75.1)	61.3 (46.3–76.3)	63.5 (47–75.3)	44 (28.5–72)	72.8 (53.8–84.6)	82 (50.8–92)	72.3 (47.4–85)
FAB subtype								
M0	0	0	0	0	1 (0.8)	1 (1.5)	0	0
M1	0	1 (1.8)	0	16 (13.8)	4 (3.1)	2 (3.0)	8 (11.9)	2 (1.9)
M2	0	2 (3.5)	39 (72.2)	40 (34.5)	19 (14.7)	2 (3.0)	19 (28.4)	15 (13.9)
M3	56 (98.2)	0	0	0	0	0	0	0
M4	0	40 (70.2)	3 (5.6)	42 (36.2)	25 (19.4)	6 (9.0)	28 (41.8)	45 (41.7)
M5	0	13 (22.8)	8 (14.8)	13 (11.2)	47 (36.4)	49 (73.1)	1 (1.5)	34 (31.5)
M6	0	0	0	0	0	1 (1.5)	0	0
M7	0	0	0	0	1 (0.8)	0	0	0
AML	1 (1.8)	1 (1.8)	4 (7.4)	5 (4.3)	32 (24.8)	6 (9.0)	11 (16.4)	12 (11.1)
Molecular events, n (%)								
PML::RARA	56 (98.2)	0	0	0	0	0	0	0
RUNX1::RUNX1T1	0	0	53 (98.1)	0	0	0	0	0
CBFB::MYH11	0	52 (91.2)	0	0	0	0	0	0
KMT2A::MLLT3	0	0	0	0	0	7 (10.4)	1 (1.5)	3 (2.8)

(Continued on next page…)

Factor	G1 (n = 57, 8.7%)	G2 (n = 57, 8.7%)	G3 (n = 54, 8.2%)	G4 (n = 116, 17.7%)	G5 (n = 129, 19.7%)	G6 (n = 67, 10.2%)	G7 (n = 67, 10.2%)	G8 (n = 108, 16.5%)
Other *KMT2A* fusions	0	1 (1.8)	0	0	0	15 (22.4)	0	14 (13.0)
NUP98 fusions	0	0	0	0	2 (1.6)	3 (4.5)	9 (13.4)	6 (5.6)
Biallelic *CEBPA*	0	0	0	95 (81.9)	1 (0.8)	0	0	0
Monoallelic *CEBPA*	0	0	0	2 (1.7)	3 (2.3)	1 (1.5)	5 (7.5)	4 (3.7)
Monoallelic *CEBPA* with LOH	0	0	0	8 (6.9)	1 (0.8)	0	0	0
NPM1	0	1 (1.8)	0	0	2 (1.6)	21 (31.3)	48 (71.6)	49 (45.4)
FLT3-ITD	9 (15.8)	5 (8.8)	4 (7.4)	7 (6.0)	10 (7.8)	8 (11.9)	29 (43.3)	57 (52.8)
KMT2A-PTD	0	0	0	0	11 (8.5)	4 (6.0)	6 (9.0)	15 (13.9)
DNMT3A+NPM1+ FLT3-ITD	0	0	0	0	0	3 (4.5)	4 (6.0)	24 (22.2)
TET2 or *IDH2+ NPM1+FLT3*-ITD	0	0	0	0	0	0	19 (28.4)	8 (7.4)
Karyotype, n (%)								
Normal karyotype	1 (1.8)	14 (24.6)	2 (3.7)	74 (63.8)	61 (47.3)	28 (41.8)	45 (67.2)	60 (55.6)
Complex karyotype	0	0	0	9 (7.8)	24 (18.6)	10 (14.9)	4 (6.0)	7 (6.5)
Monosomal karyotype	0	0	0	10 (8.6)	26 (20.2)	9 (13.4)	5 (7.5)	5 (4.6)
+8	0	3 (5.3)	0	3 (2.6)	13 (10.1)	8 (11.9)	1 (1.5)	4 (3.7)
-5/5q-	0	0	0	1 (0.9)	7 (5.4)	2 (3.0)	0	1 (0.9)
-7/7q-	0	0	2 (3.7)	2 (1.7)	18 (14.0)	2 (3.0)	0	1 (0.9)
-17/abn(17p)	0	0	2 (3.7)	2 (1.7)	10 (7.8)	4 (6.0)	3 (4.5)	2 (1.9)
Unknown	0	2 (3.5)	2 (3.7)	1 (0.9)	3 (2.3)	1 (1.5)	3 (4.5)	9 (8.3)
ELN risk, n (%)								
Favorable	0	52 (91.2)	49 (90.7)	78 (67.2)	1 (0.8)	12 (17.9)	18 (26.9)	15 (13.9)
Intermediate	1 (1.8)	2 (3.5)	0	18 (15.5)	30 (23.3)	18 (26.9)	32 (47.8)	42 (38.9)
Adverse	0	3 (5.3)	5 (9.3)	20 (17.2)	97 (75.2)	37 (55.2)	17 (25.4)	51 (47.2)
Unknown	56 (98.2)	0	0	0	1 (0.8)	0	0	0

BM, bone marrow; ELN, European LeukmiaNet; HGB, hemoglobin; IQR, interquartile range; NOS, not otherwise specified; PLT platelet; WBC, white blood count.

Table S4. Comparison of clinical features of newly diagnosed AML patients incorporated into analysis in the TCGA LAML, Beat AML and our cohort.

Factor	Our cohort (n = 655)	TCGA LAML (n = 139)	Beat AML (n = 176)	P value
Age (year)				
Median (IQR)	48 (34–60)	55 (41.5–65.5)	59.5 (46–66.3)	<0.001
Male gender, n (%)	339 (51.8)	72 (51.8)	92 (52.3)	0.992
WBC, × 10^9/L				
Median (IQR)	12.4 (3.52–44)	15.1 (4.1–46)	26.6 (6.3–61.1)	0.155
Bone marrow blasts, %				
Median (IQR)	68 (45–84)	72 (51.5–85.5)	75 (48–90)	0.043
WHO category, n (%)				
AML with defining genetic abnormalities				<0.001
APL with *PML::RARA* fusion	56 (8.5)	15 (10.8)	10 (5.7)	
AML with *RUNX1::RUNX1T1* fusion	53 (8.1)	7 (5.0)	3 (1.7)	
AML with *CBFB::MYH11* fusion	51 (7.8)	10 (7.2)	13 (7.4)	
AML with *DEK::NUP214* fusion	2 (0.3)	0 (0.0)	0 (0.0)	
AML with *BCR::ABL1* fusion	3 (0.5)	3 (2.2)	1 (0.6)	
AML with *KMT2A* rearrangement	40 (6.1)	7 (5.0)	4 (2.3)	
AML with *MECOM* rearrangement	3 (0.5)	1 (0.7)	3 (1.7)	
AML with *NUP98* rearrangement	18 (2.7)	2 (1.4)	0 (0.0)	
AML with *NPM1* mutation	120 (18.3)	34 (24.5)	50 (28.4)	
AML with *CEBPA* mutation	99 (15.1)	9 (6.5)	10 (5.7)	
AML, myelodysplasia-related	126 (19.2)	21 (15.1)	43 (24.4)	
ELN risk, n (%)				<0.001
Favorable	225 (34.4)	32 (23.0)	80 (45.5)	
Intermediate	143 (21.8)	66 (47.5)	36 (20.5)	
Adverse	230 (35.1)	38 (27.3)	60 (34.1)	
Unknown	57 (8.7)	3 (2.2)	0 (0.0)	

(Continued on next page…)

Factor	Our cohort (n = 655)	TCGA LAML (n = 139)	Beat AML (n = 176)	P value
Gene expression subgroups, n (%)				<0.001
G1	57 (8.7)	15 (10.8)	10 (5.7)	
G2	57 (8.7)	10 (7.2)	13 (7.4)	
G3	54 (8.2)	7 (5.0)	3 (1.7)	
G4	116 (17.7)	15 (10.8)	13 (7.4)	
G5	129 (19.7)	44 (31.7)	48 (27.3)	
G6	67 (10.2)	15 (10.8)	42 (23.9)	
G7	67 (10.2)	7 (5.0)	19 (10.8)	
G8	108 (16.5)	26 (18.7)	28 (15.9)	
Median follow-up time (months)	20.2	48.3	48.5	<0.001
Survival probability (%)				<0.001
1-year OS rate	78.1	59.5	60.7	
2-year OS rate	69.7	50.6	50	
3-year OS rate	65	42.5	45.9	

ELN, European LeukmiaNet; IQR, interquartile range; OS, overall survival.

Dataset S1. Clinical information of 655 primary AML.

Dataset S2. Small sequence variants of AML patients based on TES/WES and RNA-Seq.

Dataset S3. Pre-screened gene features for unsupervised classification of AML.

Dataset S4. Cola-based consensus classification of 655 AML patients.

Dataset S5. Differentially expressed genes in G1-G8.

Dataset S6. DISCOVER-based test for mutually exclusive and co-occurrence analysis.

Dataset S7. Enrichment score of AML patients in three differentiation hierarchies.

Dataset S8. Gene co-expression networks and pathway analyses of 655 AML patients.

Dataset S9. Predicted label of gene expression subgroups in TCGA LAML and Beat AML cohort.

Dataset S10. Molecular summary of TCGA LAML and Beat AML (G1-G8).

Dataset S11. Drug response prediction of gene expression subgroups of AML.

SI References

1. X. Lin *et al.*, Integration of Genomic and Transcriptomic Markers Improves the Prognosis Prediction of Acute Promyelocytic Leukemia. *Clin Cancer Res* **27**, 3683-3694 (2021).
2. P. Jin *et al.*, Large-scale in vitro and in vivo CRISPR-Cas9 knockout screens identify a 16-gene fitness score for improved risk assessment in acute myeloid leukemia. *Clin Cancer Res* 10.1158/1078-0432.Ccr-22-1618 (2022).
3. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
4. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
5. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
6. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
7. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527 (2016).
8. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
9. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).

10. Z. Gu, M. Schlesner, D. Hübschmann, cola: an R/Bioconductor package for consensus partitioning through a general framework. *Nucleic Acids Res* **49**, e15 (2021).

11. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

12. A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. **102**, 15545-15550 (2005).

13. S. W. Ng *et al.*, A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**, 433-437 (2016).

14. A. M. Newman *et al.*, Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782 (2019).

15. T. R. Docking *et al.*, A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat Commun* **12**, 2474 (2021).

16. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

17. R. Saito *et al.*, A travel guide to Cytoscape plugins. *Nat Methods* **9**, 1069-1076 (2012).

18. P. van Galen *et al.*, Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e1224 (2019).

19. P. Angerer *et al.*, destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241-1243 (2016).

20. D. Nicorici *et al.*, <strong>FusionCatcher</strong> – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. 10.1101/011650 %J bioRxiv, 011650 (2014).

21. B. J. Haas *et al.*, Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* **20**, 213 (2019).

22. S. Uhrig *et al.*, Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res* **31**, 448-460 (2021).

23. L. Tian *et al.*, CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol* **21**, 126 (2020).

24. W. Arindrarto *et al.*, Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. *Leukemia* 10.1038/s41375-020-0762-8 (2020).

25. E. O. Audemard *et al.*, Targeted variant detection using unaligned RNA-Seq reads. *Life Sci Alliance* **2** (2019).

26. M. Gu *et al.*, RNAmut: robust identification of somatic mutations in acute myeloid leukemia using RNA-seq. *Haematologica* 10.3324/haematol.2019.230821 (2019).

27. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

28. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

29. K. Hagiwara *et al.*, RNAIndel: discovering somatic coding indels from tumor RNA-Seq data. *Bioinformatics* 10.1093/bioinformatics/btz753 (2019).

30. A. Wilm *et al.*, LoFreq: a sequence-quality aware, ultra-sensitive variant caller for

uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40**, 11189-11201 (2012).

31. D. C. Koboldt *et al.*, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).
32. W. McLaren *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
33. M. J. Landrum *et al.*, ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-d1067 (2018).
34. A. Kiran, P. V. Baranov, DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**, 1772-1776 (2010).
35. G. Ramaswami, J. B. Li, RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* **42**, D109-113 (2014).
36. E. Picardi, A. M. D'Erchia, C. Lo Giudice, G. Pesole, REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res* **45**, D750-d757 (2017).
37. Z. Lai *et al.*, VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108 (2016).
38. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
39. S. T. Sherry *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).
40. J. G. Tate *et al.*, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-d947 (2019).
41. C. B. Kimmel, W. W. Ballard, S. R. Kimmel, B. Ullmann, T. F. Schilling, Stages of embryonic development of the zebrafish. *Dev Dyn* **203**, 253-310 (1995).
42. C. Thisse, B. Thisse, High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nat Protoc* **3**, 59-69 (2008).
43. X. Zhou *et al.*, Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* **48**, 4-6 (2016).
44. Z. Gu, L. Gu, R. Eils, M. Schlesner, B. Brors, circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).
45. S. Canisius, J. W. Martens, L. F. Wessels, A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol* **17**, 261 (2016).