Supporting Information Appendix

SI Materials and Methods

Patient treatments

The patients were treated with standard first-line "3+7" induction regimens consisting of idarubicin (12 mg/m^2 for days 1-3) and Ara-C (100 mg/m^2, Days 1-7) at the first cycle of induction. Of the 101 patients, 15 (14.9%) received second cycle of induction. The patients received 2-4 courses of high-dose cytarabine-based therapy (2 g/m^2 every 12 h for days 1-3, total of 6 doses) or hematopoietic stem cell transplantation (HSCT), as consolidation therapy. Of the 101 patients, 18 (17.8%) underwent HSCT. A minority of patients received tyrosine kinase inhibitors during consolidation. Detailed information of the patients' treatment is listed in the Dataset S1.

Morphologic, Immunophenotypic and Clinical Molecular Analysis

The morphology of the bone marrow smear or Cytospin with Wright-Giemsa staining was reviewed by two independent pathologists at the Shanghai Institute of Hematology. The clinical flow cytometry was performed using 10-color multiparametric flow cytometry. The hotspot of the *KIT* gene mutation was detected by polymerase chain reaction (PCR) and Sanger sequencing using the sense primer 5'- TATTGTGAATCTACTTGGAGCC-3' and the antisense primer 5'- AATCCCATAGGACCAGACG-3'. The molecular detection of *RUNX1-RUNX1T1* transcripts was carried out using reverse transcription-PCR using the sense primer 5'- CTACCGCAGCCATGAAGAACC-3' and the antisense primer 5'- AGAGGAAGGCCCATTGCTGAA-3'.

Fluorescence-Activated Cell sorting (FACS)

FACS was performed on a FACSAria^{TM} III sorter (BD Biosciences). The following antibodies were used for cell staining: anti-CD34-FITC (IM1870U, Beckman Coulter), anti-CD34-APC (340441, BD Biosciences), anti-CD117-APC (341096, BD Biosciences), anti-CD117-PE (340529, BD Biosciences), anti-HLA-DR-APC-Cy7 (307618, Biolegend), anti-CD15-PE (IM1954U, Beckman Coulter) and anti-CD11b-FITC (IM0530, Beckman Coulter). FACSDiva (BD) software was used for data analysis.

Cell Cycle Analysis

Surface marker-stained BMMCs were fixed with BD Cytofix/Cytoperm Fixation for 1 hour at 4°C in the dark and were then washed with BD Perm/Wash buffer twice. Two microliters of anti-Ki67 antibodies (≤ 0.5 μg per million cells, 652404, Biolegnd) were added to each tube, and the cells were incubated at 4°C in the dark. Thirty minutes later, 5 μl Hoechst

33342 (10mg/mL, Life Technologies) was added to each tube, and the cells were immediately analyzed by flow cytometry. Flow cytometry was performed on an LSR Fortessa^{TM} X-20 (BD Biosciences) and was analyzed by FlowJo Software.

Migration Assay

A total of 1×10^5 of sorted cells were seeded in 200 μl of serum-free medium in the upper chamber of an 8-μm-pore-size Corning Costar Transwell plate (Corning). The lower chamber was filled with 800 μL of complete culture medium (RPMI 1640, Gibco) containing 20% FBS (Gibco). Cells were allowed to migrate towards the lower chamber for 12 hours at 37°C in a 5% CO_2 incubator. Then, cells were collected and resuspended in 200 μL PBS. Viable cells were counted using a Countstar®BioTech (ALIT Life Science) according to the manufacturer's instructions.

Colony-forming Unit (CFU) Assay

The clonogenic capacity of sorted $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM cells was evaluated by a colony-forming unit assay. Briefly, 5×10^4 cells from each sorted group were seeded in semisolid methylcellulose media (Methocult H4435; STEMCELL Technologies) in 35-mm tissue culture dishes in duplicate. Cells were incubated at 37°C in 5% CO_2. Seven days later, colonies consisting of more than 50 cells were counted according to the StemCell Protocol.

Cellular Drug Sensitivity Assay

Cytarabine, daunorubicin, venetoclax and dasatinib were purchased from Selleck and was stored at -20°C. Isolated $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM cells were plated at 2×10^4 cells per well in triplicate in 96-well plates in RPMI 1640 medium (Gibco) supplemented with 20% FBS (Gibco). Cells were treated with cytarabine, daunorubicin, venetoclax or dasatinib for 48 hours. CellTiter-Glo luminescent cell viability assays (G9242, Promega) were performed according to the manufacturer's instructions, and luminescence was measured on a VarioCskan® Flash Spectral Scanning Multimode Reader (Thermo Scientific).

Droplet Digital Polymerase Chain Reaction (ddPCR) Assay

Total RNA samples from t(8;21) AML specimens were extracted using an AllPrep DNA/RNA Mini Kit (Qiagen) and were converted to cDNA with a PrimeScript^{TM} RT Reagent Kit with gDNA Eraser (RR047A, TAKARA). For the *RUNX1-RUNX1T1* transcript, the sense primer was 5'-CACCTACCACAGAGCCATCAAA-3', the reverse primer was 5'-ATCCACAGGTGAGTCTGGCATT-3', and the probe was 5'-FAM-

AACCTCGAAATCGTACTGAGAAGCACTCCA-BHQ1-3'. For the *RUNX1-RUNX1T19a* transcript, the sense primer was 5'-TGAGCATTGCTGTCCTGGGTCATA-3', the reverse primer was 5'-TTGGATACTAGATACTGCAAGGGCCG-3', and the probe was 5'-FAM-TGAGGTCACATTGCTTCTCCAAAGGC-BHQ1-3'. The copy number of *ABL1* was used as an internal reference. The sense primer of *ABL1* was 5'-TGGAGATAACACTCTAAGCATAACTAAAGGT-3', the reverse primer was 5'-GATGTAGTTGCTTGGGACCCA-3', and the probe was 5'-FAM-CCATTTTTGGTTTGGGCTTCACACCATT-BHQ1-3'. The primer and probe for each target sequence were at final concentrations of 900 nM and 250 nM, respectively. A 20 μl reaction mixture consisted of input cDNA (50 ng), ddPCR Supermix (Bio-Rad laboratories) and primer/probes. The mixture was partitioned into droplets using a QX200 Droplet Generator (Bio-Rad Laboratories). PCR was performed in duplicate according to the manufacturer's recommended protocol, and the reaction products were analyzed using a QX200 Droplet Reader (Bio-Rad Laboratories). The transcript levels were assessed using QuantaSoft Analysis Pro software (Bio-Rad Laboratories).

Genomic DNA, Total RNA Extraction and Next-generation Sequencing

Genomic DNA and total RNA samples were extracted using an AllPrep DNA/RNA Mini Kit (Qiagen). DNA/RNA quality and quantity were assessed on Agilent DNA/RNA 6000 chips (Agilent Technologies) and Qubit (Life Technologies) before next-generation sequencing library preparation, respectively. RNA-seq libraries of isolated $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$ and AM were constructed using a TruSeq RNA Sample Prep Kit (Illumina), followed by sequencing on an Illumina MiSeq platform according to a 2 × 250 bp protocol. Total RNA-seq libraries of 62 t(8;21) AML patients were constructed using a KAPA RNA Hyper Kit (Roche) and were subjected to paired-end (2 × 150 bp) sequencing on a NovaSeq platform (Illumina). The libraries of whole-exome sequencing of patient AML-016 at different time points were constructed using KAPA Hyper Prep Kit and SeqCap EZ Human Exome v3.0 (Roche) according to the manufacturer's instructions. WES was performed on NovaSeq platform (Illumina) according to the manufacturer's protocol.

RNA-seq Alignment and Count Matrix Generation

The raw RNA-Seq reads were aligned to the human reference genome hg19 using STAR (v2.7.0d)[1], which was downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/). Name-sorted and indexed BAM files were generated by Samtools (v1.8-47)[2]. Table files with transcript counts were generated by the HTSeq

(v0.6.1)[3]. Afterward, the count matrix was obtained using DESeq2[4]. Fragments per kilobase million (FPKM) was used to evaluate expression levels of individual genes by normalizing the length of genes using the count matrix. Additionally, to quantify transcript expression levels, transcripts per kilobase million (TPM) value were determined using Salmon[5]. The R package limma[6] was used to identify differentially expressed genes (DEGs) using normalized read counts. An in-house R package gvmap (https://github.com/ytdai/gvmap) was used to generate a heatmap of the expression profile.

Functional and Pathway Enrichment Analysis of Gene Transcriptional Expression Data

Gene Ontology (GO) enrichment analysis of differentially expressed genes was performed using both Ingenuity Pathway Analysis (IPA) software (IPA®, Qiagen Redwood City) and DAVID (https://david.ncifcrf.gov/) with the default parameters. Hierarchical clustering and principal component analysis (PCA) were performed to identify the differences between the samples. Gene set enrichment analysis (GSEA)[7] was performed with a preranked algorithm using the clusterProfiler R package[8]. Normalized RNA-seq data were rank-ordered by the fold change. Gene sets were download from the Molecular Signatures Database (MSigDB, v7.0) of the Broad Institute. HALLMARK gene sets (H) and MSigDB curated gene sets (C2) were used to perform GSEA analysis in a 1,000-gene-set with a two-sided permutation[7]. R package enrichplot (https://github.com/GuangchuangYu/enrichplot) was used to interpret enrichment results for visualization of GSEA. Gene signatures of hematopoietic cells in different stages were downloaded from Blueprint project in xCell database[9].

Mutation Calling Using RNA-seq Data

In RNA-seq data, raw read counts were aligned to the human reference genome hg19 using STAR (v2.7.0d)[1]. The mutation calling steps mainly followed the Genome Analysis Toolkit (GATK, v4.0.3)[10] forum recommended best-practice pipeline. RNA variant calling datasets were annotated by ANNOVAR[11].The screening for variant identification followed our previously reported pipeline[12]: 1) > 10x depth in the variants site; 2) ≥ 5% variant allele frequency (VAF); 3) > 3 individual mutant reads; 4) Filter variants both observed on positive-strand and negative-strand; 5) Mutated genes occurred in database of AML with t(8;21), which referred to the results of genomic landscape AML with t(8;21)[13]; 6) Filter sites annotated in dbSNP (v147)[14] database but not found in COSMIC (v81)[15] database;

7) Checked by the Samtools[2] mpileup module and Integrative Genomics Viewer (IGV)[16].

Single-cell RNA Sequencing

Viably frozen Ficoll-isolated BMMCs from patient samples were thawed in a 37°C incubator and were suspended in RPMI 1640 medium supplemented with 10% FBS. Cells were washed three times with 1× PBS supplemented with 0.04% bovine serum albumin, counted by a Countstar®BioTech (ALIT Life Science), and were diluted to a concentration of 1 million cells/ml. Single-cell libraries were generated using a GemCode Single-cell Instrument and a Single Cell 3' Library & Gel Bead Kit v2 and Chip Kit (10x Genomics) according to the manufacturer's instructions. The scRNA-seq libraries were subjected to paired-end (2 × 150 bp) sequencing on a NovaSeq platform (Illumina).

10x Genomics Single-cell RNA-sequencing Data Analysis

Raw reads in a FASTQ format were aligned to a human reference (hg19, v3.0.0) using Cell Ranger software (v3.0.2). The alignment reference and software were both provided by 10x Genomics (https://support.10xgenomics.com). The same software was used for barcode assignment and unique molecular identifier (UMI) counting using the parameter --expect-cells 10000. Raw count data were then analyzed with the R package Seurat (v3.1.2)[17]. Cells that expressed less than 800 genes or over 10% mitochondrial RNA were filtered out, as these samples might represent doublets. Filtered count matrix from different patients or time points were merged together in Seurat. Expression data was normalized using a global-scaling normalization method, which was provided by Seurat package with default parameters. Subsequently, 2000 variable genes were identified. Batch effects were removed using ComBat[18,19]. Principal component analysis (PCA) was performed on the variable genes and the resolution parameter to identify clusters was set to 0.8. For visualization purposes, two nonlinear dimensional reduction methods were performed, namely, t-distributed stochastic neighbor embedding (tSNE)[20] and uniform manifold approximation and projection (UMAP)[21]. UMAP was chosen to further cell visualization. Clusters were identified that referred to the top markers found in each cluster with an adjusted *p*-value (p_val_adj) ≤ 0.05 and an average log fold change (avg_logFC) ≥ 0.5. In the analysis of gene signature similarity of the three cell populations between bulk RNA-seq and single-cell RNA-sequencing data, we chose top 200 DEGs from bulk RNA-seq data according to the ascending order of adjusted *P*-value with fold change over 2. GSEA was performed using the top 200 DEGs as the input functional gene-sets. Cell cycle phase scores were calculated using the built-in function CellCycleScoring in Seurat

with default parameters[22]. 17-gene leukemic stem cell (LSC17) score was calculated based on the equation provided by Stanley et al[23].

Differential Trajectory Analysis

After all cell populations were identified, Monocle (v2.14)[24] was used to reconstruct the differentiation trajectory. The expression data and cell type to build the Monocle object was built directly from the Seurat object. Top 100 variable genes that were differentially expressed in each cell type with a q-value of less than 0.01 were selected as “order genes” in the next analysis. The DDRTree method provided in Monocle was used for the dimensionality reduction.

Whole-Exome Sequencing Analysis and Mutation Calling

WES with an average read depth of 219.67× was performed for AML-016 with samples from the four stages of diagnosis, complete remission, relapse and pot-relapse. BWA (v0.7.17-r1188)[25] was used to align raw WES sequence reads to the human reference hg19. Samtools (v1.8-47)[2] was used to generate chromosomal-coordinated, sorted and indexed BAM files. The preprocessing steps were carried out mainly according to the Genome Analysis Toolkit (GATK, v4.0.3)[10] forum recommended best-practice pipeline. The results from GATK HaplotypeCaller were used as the high-confidence sites. Several extra variant callers, such as GATK UnifiedGenotyper and VarScan2 (v2.4.3)[26], were used to prevent overly strict filtration by GATK HaplotypeCaller. WES data from complete remission was used as a normal control, and Mutect2 in GATK was used to call somatic mutations.

Clonal Evolution

Screening of gene mutations from the raw VCF file followed these analysis criteria: 1) > 10x coverage in the variant sites; 2) variant allele frequency (VAF) ≥ 5% and at least three individual mutant reads; 3) filter reads of the base quality < 13; 4) filter variants observed only in the positive-strand or negative-strand; 5) ≤ 0.1% VAF in complete remission samples; and 6) passed the artificial check using the Samtools[2] mpileup module and Integrative Genomics Viewer (IGV)[16]. All filtered variant sites from all time points were confirmed in IGV. Variants with a depth > 100× were enrolled in clonal evolution analysis. Clonal evolution was evaluated using variants that were detected at diagnosis, complete remission, relapse and post-relapse and was visualized by the R package fishplot[27].

Validation using GEO and TCGA cohorts

To validate the prognostic impact of genes significantly differentially expressed in

$CD34^+CD117^{dim}$-high group in public cohorts, we used The Cancer Genome Atlas (TCGA-LAML)[28] and Gene Expression Omnibus (GEO, GSE37642)[29] datasets. In the analysis of TCGA-LAML cohort, the raw expression matrix of FPKM and clinical data were downloaded using TCGAbiolinks[30]. 142 cases (142/151) in TCGA-LAML cohort had available overall survival data. For GSE37642 cohort, the raw expression matrix and clinical data were downloaded via GEOquery[31]. 553 cases (553/562) in GSE37642 cohort had available overall survival data and 30 cases were with *RUNX1-RUNX1T1* fusion. We calculated the enrichment score of the 215 up-regulated genes in $CD34^+CD117^{dim}$-high group using ssGSEA algorithm in GSVA package[32]. The mean value of enrichment score was used as the cutoff to separate cases into two groups ($CD34^+CD117^{dim}$ gene-set-low group and $CD34^+CD117^{dim}$ gene-set-high group).

Statistical Analysis

The clinical characteristics of the two groups were analyzed using the x^2-test or Fisher's exact test for categorical parameters and the Mann-Whitney *U* test for continuous variables. ROC curve analysis was performed using the survivalROC package[33]. Overall survival (OS) was measured from the date of disease diagnosis to the date of death (failure) or the last follow-up time (censored). Relapse-free survival (RFS) was defined as the time from the documentation of CR to treatment failure, such as relapse, refractory disease, death, or survival in CR at the last follow-up (censored). A Cox regression model was used for the univariate and multivariate analyses of OS and RFS. The Kaplan-Meier method was used to estimate the probabilities of OS and RFS and the log-rank test was used to compare the *P* values. Major molecular remission (MMR) was based on the *RUNX1-RUNX1T1* transcript level as previously described[34-36]. Statistical analyses were performed with SPSS 25.0 (IBM) and GraphPad Prism 6.0.

SI Figures

Fig. S1. Clinical Immunophenotypic Characteristics and Morphological Features of Distinct Leukemic Cell Populations in t(8;21) AML Patients.

(**A**) The representative flow cytometry images of forward-scatter (FSC) and side-scatter (SSC) and the FSC and SSC value of the $CD34^{+}CD117^{dim}$ and $CD34^{+}CD117^{bri}$ populations (mean ± s.d., n = 3). *, $P < 0.05$; **, $P < 0.01$. Statistical significance was determined using two-sided Student's t test. (**B**) Representative Wright-Giemsa-stained bone marrow smear of primary t(8;21) AML patients. (**C** and **D**) Representative flow cytometry images and corresponding morphologic differential counts of different cell populations on Cytospin were presented (mean ± s.d., n = 4). The red arrow denoted the predicted differentiation trajectory of cells from myeloblasts to differentiated myeloid cells.

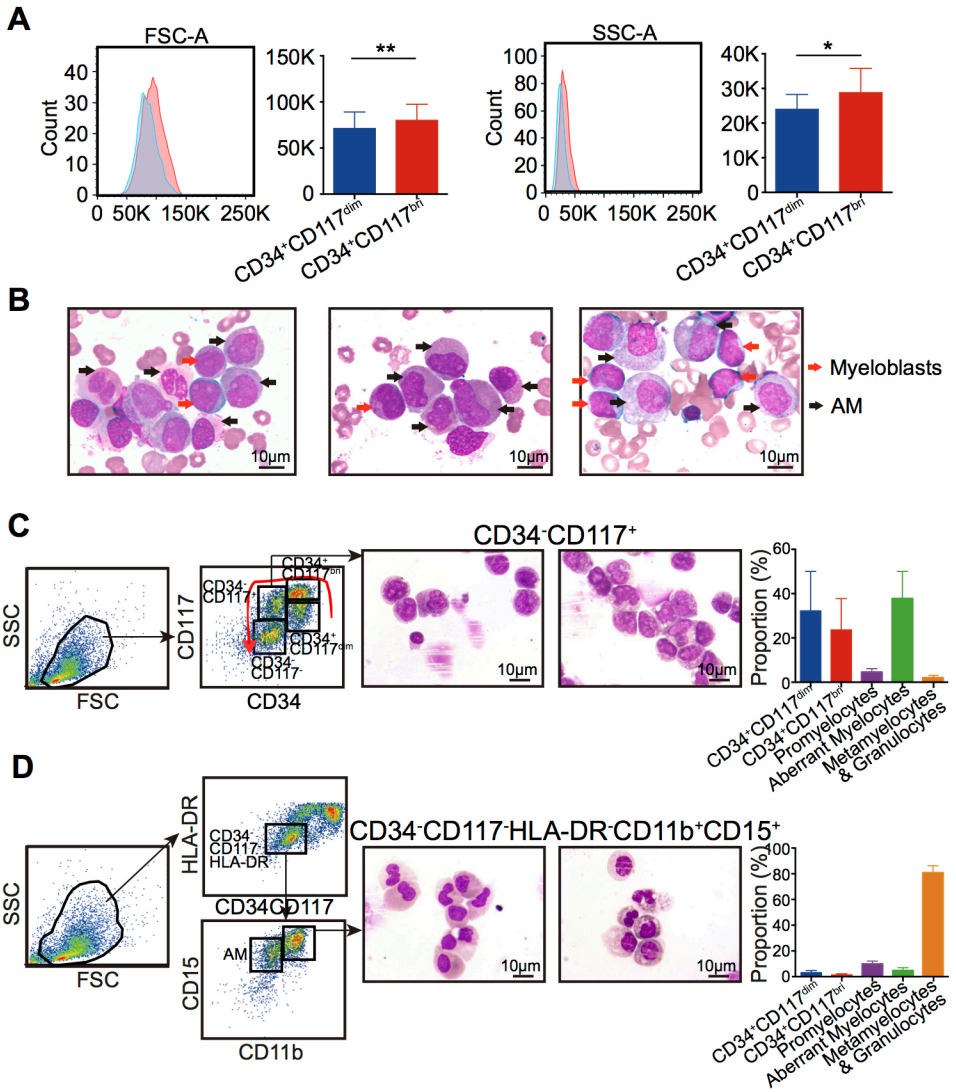


Fig. S2. Gene Expression Profiles and Biological Characteristics of $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM Populations.

(**A**) Gene ontology (GO) analysis of the upregulated genes in $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM populations.

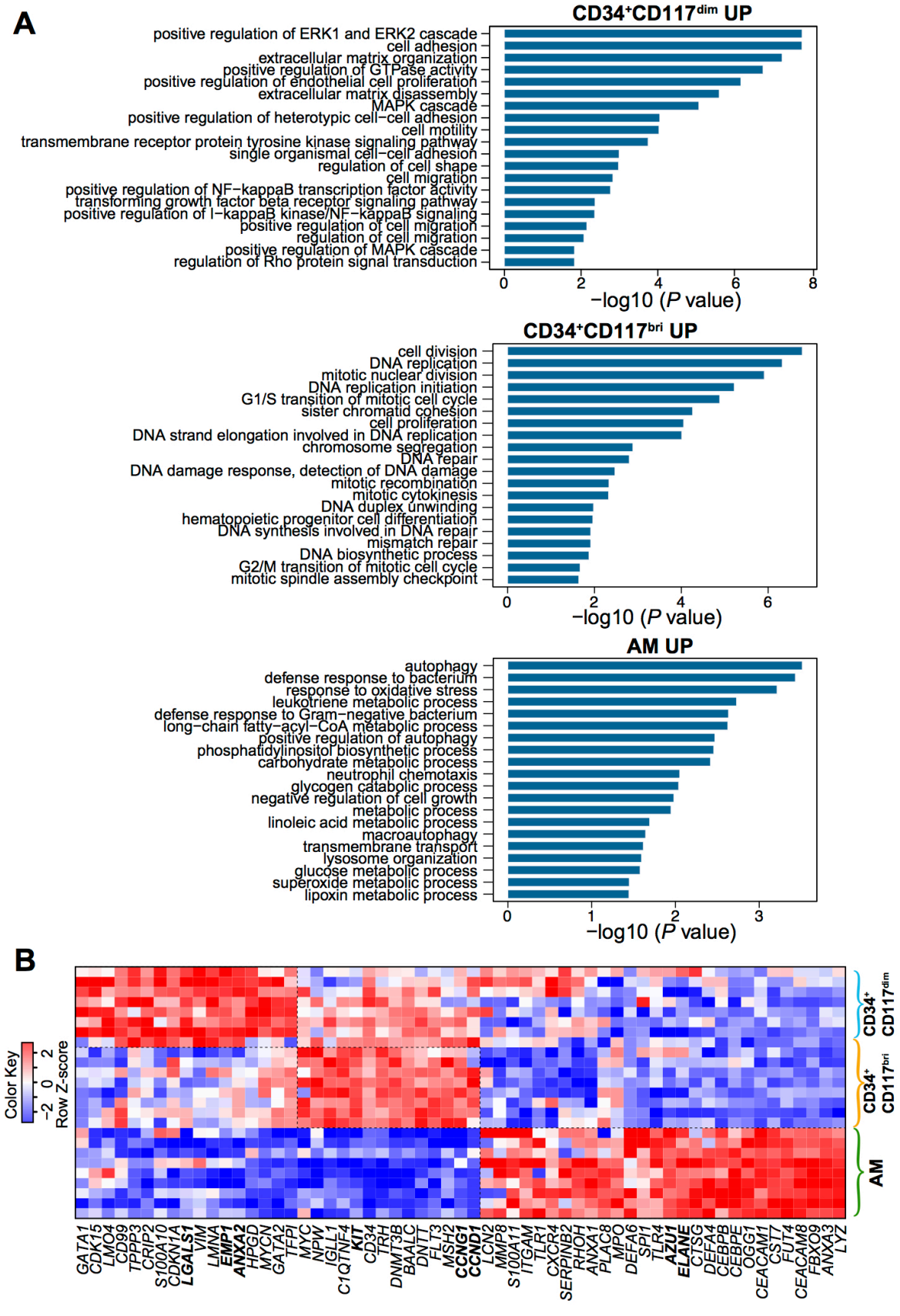
(**B**) Heatmap of highly expressed genes in isolated $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM populations.

(**C**) Gene expression levels of the representative drug resistance related genes in isolated $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM populations. FPKM, fragments per kilobase million. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Statistical significance was determined using two-sided Wilcoxon test.

(**D**) The alternatively spliced 9a isoform of RUNX1T1 was detected in all cases in our transcriptome analysis. The location of the 9a exon is indicated by the red box.

(**E**) ddPCR showed the quantification of the *RUNX1-RUNX1T1* and *RUNX1-RUNX1T19a* transcripts in isolated $CD34^+CD117^{dim}$, $CD34^+CD117^{bri}$ and AM populations. Results were normalized, and the copy number of *ABL1* was used as the internal reference (mean ± s.d., $n = 6$ subjects with duplicates). *, $P < 0.05$. Statistical significance was determined using two-sided Student's *t* test.

(**F**) Flow cytometry analysis of the BMMC samples from t(8;21) AML showed that $CD34^+CD117^{dim}$ and $CD34^+CD117^{bri}$ populations presented granulocyte-monocyte progenitor (GMP) markers ($CD34^+CD38^+CD7^-CD10^-CD45RA^+$).



Continued on the next page

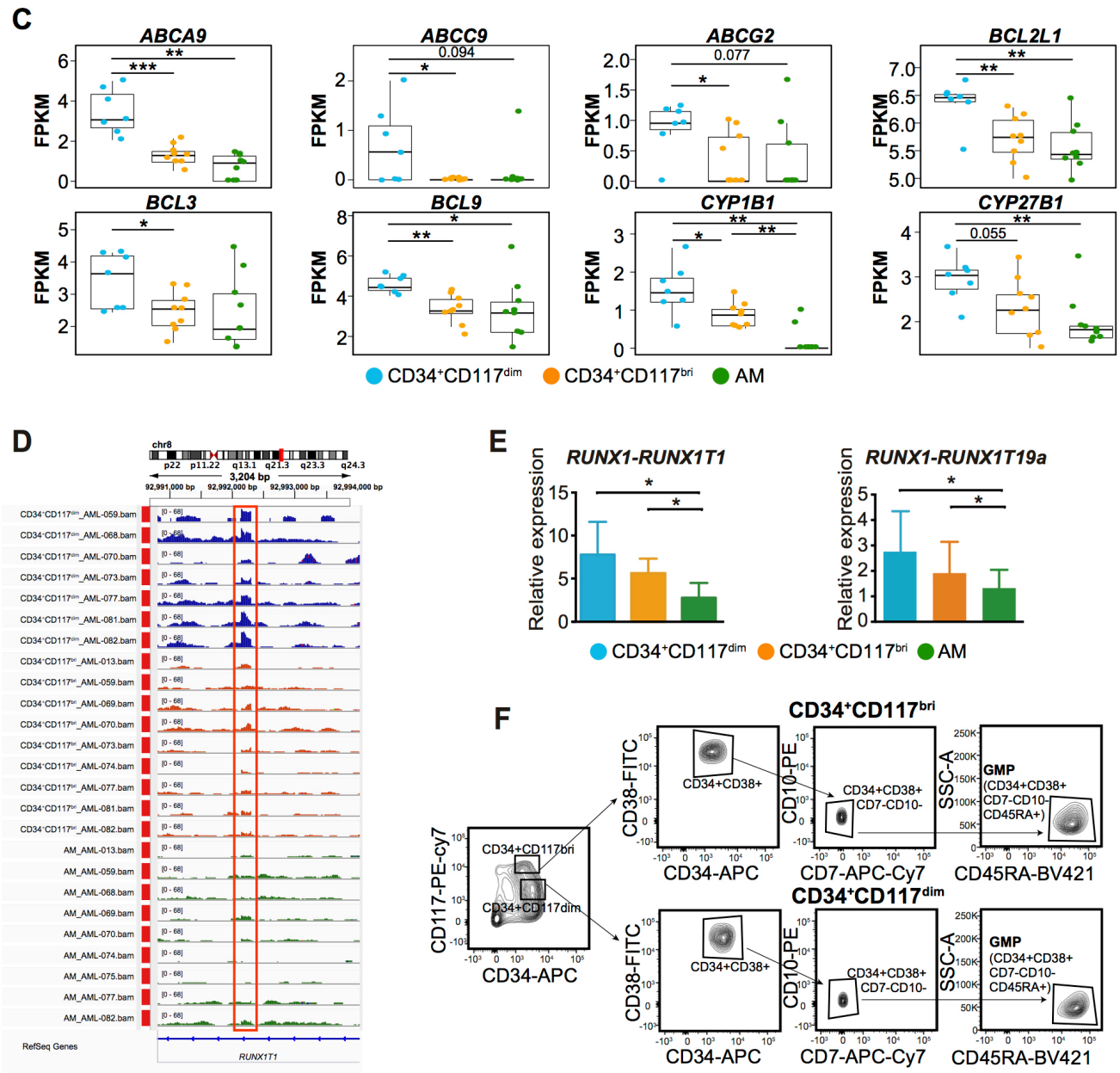

C
ABCA9
ABCC9
ABCG2
BCL2L1
BCL3
BCL9
CYP1B1
CYP27B1
FPKM
0.094
0.077
0.055
CD34+CD117dim
CD34+CD117bri
AM
D
chr8
3,204 bp
RefSeq Genes
RUNX1T1
E
RUNX1-RUNX1T1
RUNX1-RUNX1T19a
Relative expression
F
CD34+CD117bri
CD34+CD117dim
CD117-PE-cy7
CD34-APC
CD38-FITC
CD10-PE
CD7-APC-Cy7
SSC-A
CD45RA-BV421
CD34+CD38+
CD34+CD38+ CD7-CD10-
GMP (CD34+CD38+ CD7-CD10- CD45RA+)

Fig. S3. Single-cell Transcriptomic Analysis of t(8;21) AML at Diagnosis.

(**A**) Overview of the t(8;21) AML patients subjected to scRNA-seq. *KIT*-mut represents the *KIT* mutation status in t(8;21) AML patients. Cell numbers represent the single-cell transcriptomes that passed quality control. For each patient, Cell % (pie chart) indicates the cell count proportion of blasts and AM cells in Ficoll-isolated BMMCs.

(**B**) Comparison of the percentages of leukemic cells including blasts and AM cells between the clinical morphological counts at diagnosis and the morphological counts of BMMC Cytospin preparations after isolation the BMMCs (left panel). Representative Wright-Giemsa-stained Cytospin preparations of Ficoll-isolated BMMCs from the bone marrow of t(8;21) AML patients (upper right). Representative fluorescence in situ hybridization (FISH) plots of Ficoll-isolated BMMC cytospin preparations showing the t(8;21) translocation in most cells by using the probes specific to *RUNX1* (green) and *RUNX1T1* (red, lower right).

(**C**) UMAP analysis of each t(8;21) AML patients. Each dot represents a cell, and the colors represent different cell clusters. DC, dendritic cells; Immature Ery, Immature erythroid cells.

(**D**) Comparison of the percentages of major cell types between scRNA-seq results and morphologic counts of Wright-Giemsa-stained BMMC Cytospin preparations.

(**E**) Violin plot showing the *CD34* transcript (left) and *KIT* transcript (right) expression level in distinct cell types at diagnosis. The Y-axis shows the normalized read counts.

(**F**) GSEA analysis of the highly expressed gene-sets of the $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$ and AM clusters identified in scRNA-seq compared with the RNA-seq data. Normalized enrichment score (NES) and nominal *P* value are given.

(**G**) IPA analysis showing the enrichment pathways of highly expressed genes in $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$ and AM clusters based on the scRNA-seq data.

.

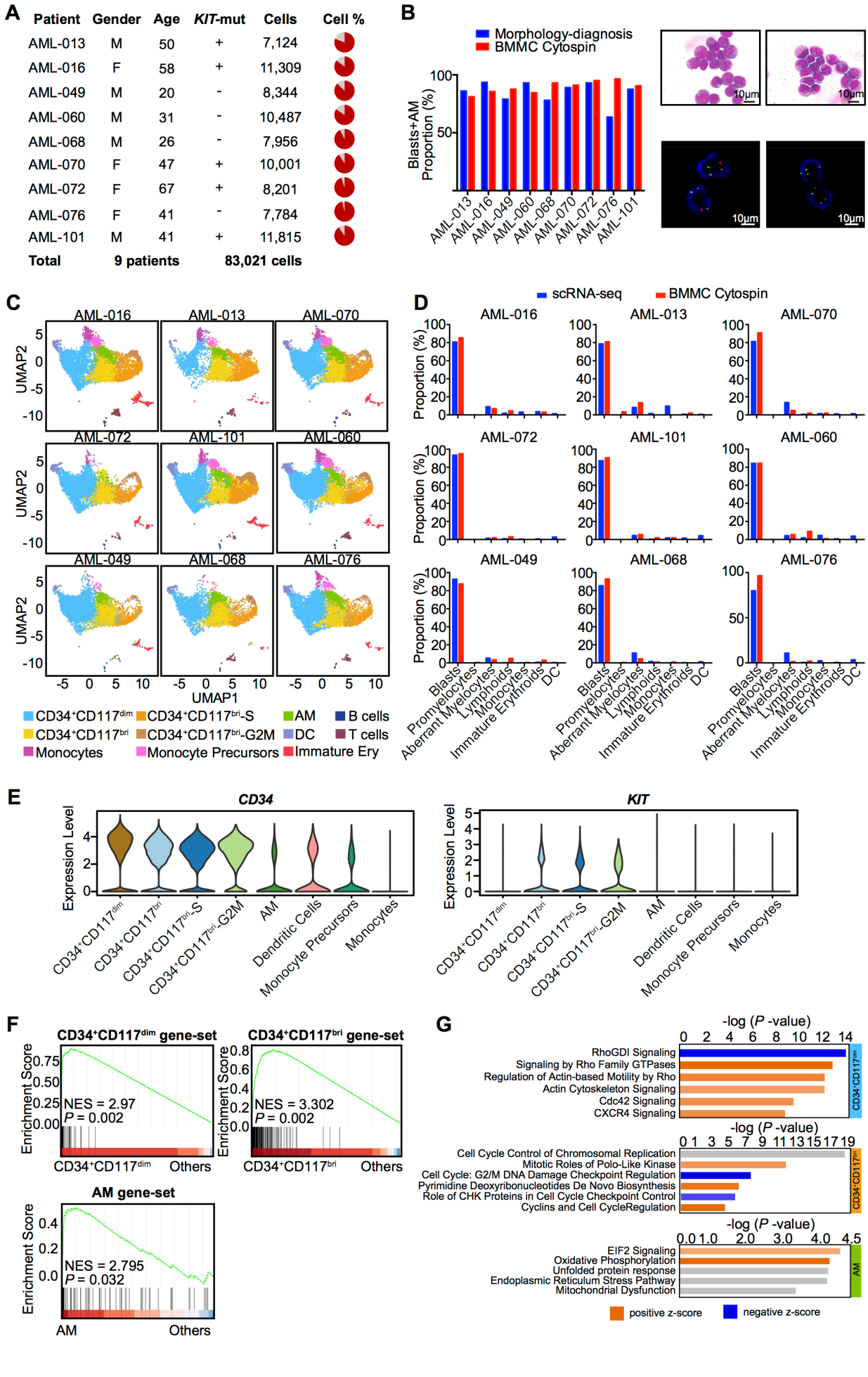

A
Patient Gender Age KIT-mut Cells Cell %
AML-013 M 50 + 7,124
AML-016 F 58 + 11,309
AML-049 M 20 - 8,344
AML-060 M 31 - 10,487
AML-068 M 26 - 7,956
AML-070 F 47 + 10,001
AML-072 F 67 + 8,201
AML-076 F 41 - 7,784
AML-101 M 41 + 11,815
Total 9 patients 83,021 cells
B
Morphology-diagnosis
BMMC Cytospin
Blasts+AM Proportion (%)
AML-013 AML-016 AML-049 AML-060 AML-068 AML-070 AML-072 AML-076 AML-101
10μm
C
AML-016 AML-013 AML-070 AML-072 AML-101 AML-060 AML-049 AML-068 AML-076
UMAP2
UMAP1
CD34+CD117dim
CD34+CD117bri
Monocytes
CD34+CD117bri-S
CD34+CD117bri-G2M
Monocyte Precursors
AM
DC
Immature Ery
B cells
T cells
D
scRNA-seq
BMMC Cytospin
Proportion (%)
Blasts
Promyelocytes
Aberrant Myelocytes
Lymphoids
Monocytes
Immature Erythroids
DC
E
CD34
KIT
Expression Level
CD34+CD117dim
CD34+CD117bri
CD34+CD117bri-S
CD34+CD117bri-G2M
AM
Dendritic Cells
Monocyte Precursors
Monocytes
F
CD34+CD117dim gene-set
CD34+CD117bri gene-set
AM gene-set
Enrichment Score
NES = 2.97
P = 0.002
NES = 3.302
P = 0.002
NES = 2.795
P = 0.032
Others
G
-log (P -value)
RhoGDI Signaling
Signaling by Rho Family GTPases
Regulation of Actin-based Motility by Rho
Actin Cytoskeleton Signaling
Cdc42 Signaling
CXCR4 Signaling
Cell Cycle Control of Chromosomal Replication
Mitotic Roles of Polo-Like Kinase
Cell Cycle: G2/M DNA Damage Checkpoint Regulation
Pyrimidine Deoxyribonucleotides De Novo Biosynthesis
Role of CHK Proteins in Cell Cycle Checkpoint Control
Cyclins and Cell CycleRegulation
EIF2 Signaling
Oxidative Phosphorylation
Unfolded protein response
Endoplasmic Reticulum Stress Pathway
Mitochondrial Dysfunction
positive z-score
negative z-score

Fig. S4. Single-cell Transcriptomic Analysis and Clonal Evolution Analysis of t(8;21) AML Along Disease Progression.

(**A**) Representative FISH plots of cytospin preparations of bulk cells, $CD34^{+}CD117^{dim}$ and $CD34^{+}CD117^{bri}$ populations obtained at different time points in patient AML-016 showing the t(8;21) translocation in most cells (left panel). Representative Wright-Giemsa-stained Cytospin preparations of bulk cells, $CD34^{+}CD117^{dim}$ and $CD34^{+}CD117^{bri}$ populations obtained at different time points in patient AML-016 showing the morphology was dominated by blast cells (right panel).

(**B**) UMAP analysis of cells from different samples after removing batch effects. Cells are colored by cell clusters (left), samples (middle) and cell-cycle states (right). AML-016-CR, BMMC sample obtained from patient AML-016 at complete remission. Two scRNA-seq data from BMMC of healthy donors were downloaded from database (https://support.10xgenomics.com/single-cell-gene-expression/datasets).

(**C**) Heatmap of the highly expressed genes in $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$, $CD34^{+}CD117^{bri}$-S, $CD34^{+}CD117^{bri}$-G2M and AM clusters at different time points (diagnosis, relapse and post-relapse). The relative expression level of genes (rows) across cells (columns) is shown.

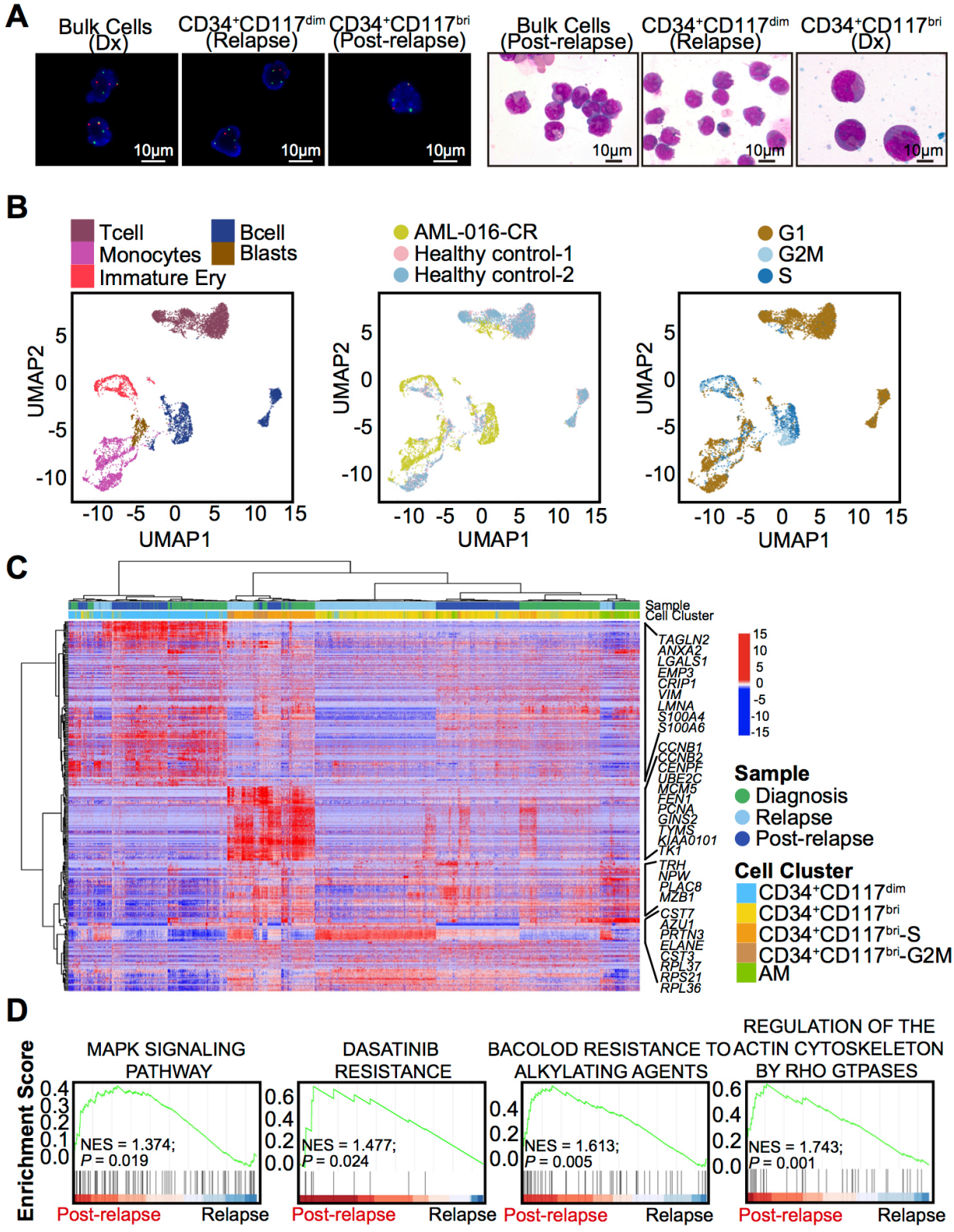
(**D**) GSEA showing the activated pathways at post-relapse stage. Nominal *P* value and normalized enrichment score (NES) are given.

(**E**) Violin plot showing the LSC17 score of $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$, $CD34^{+}CD117^{bri}$-S and $CD34^{+}CD117^{bri}$-G2M clusters at diagnosis (Dx), relapse (R) and post-relapse (P-R). **, $P < 0.01$; ****, $P < 0.0001$. Statistical significance was determined using two-sided Wilcoxon test.

(**F**) Violin plot showing the *CD34* transcript expression level in $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$, $CD34^{+}CD117^{bri}$-S, $CD34^{+}CD117^{bri}$-G2M and AM clusters at diagnosis (Dx), relapse (R) and post-relapse (P-R). The Y-axis shows the normalized read counts. **, $P < 0.01$; ****, $P < 0.0001$. Statistical significance was determined using two-sided Wilcoxon test.

(**G**) The line chart shows the $CD34^{+}CD117^{dim}$ proportion among $CD34^{+}$ cells at relapse and post-relapse time points in ten t(8;21) AML patients, including AML-016. Statistical significance was determined using two-sided Student's *t* test (upper left). The clinical flow cytometry data show the distribution of $CD34^{+}$

myeloblasts by antigen CD34 and CD117 in ten t(8;21) AML patients at relapse and post-relapse time points. Cells were gated according to the cell distribution characteristics of each patient.



Continued on the next page

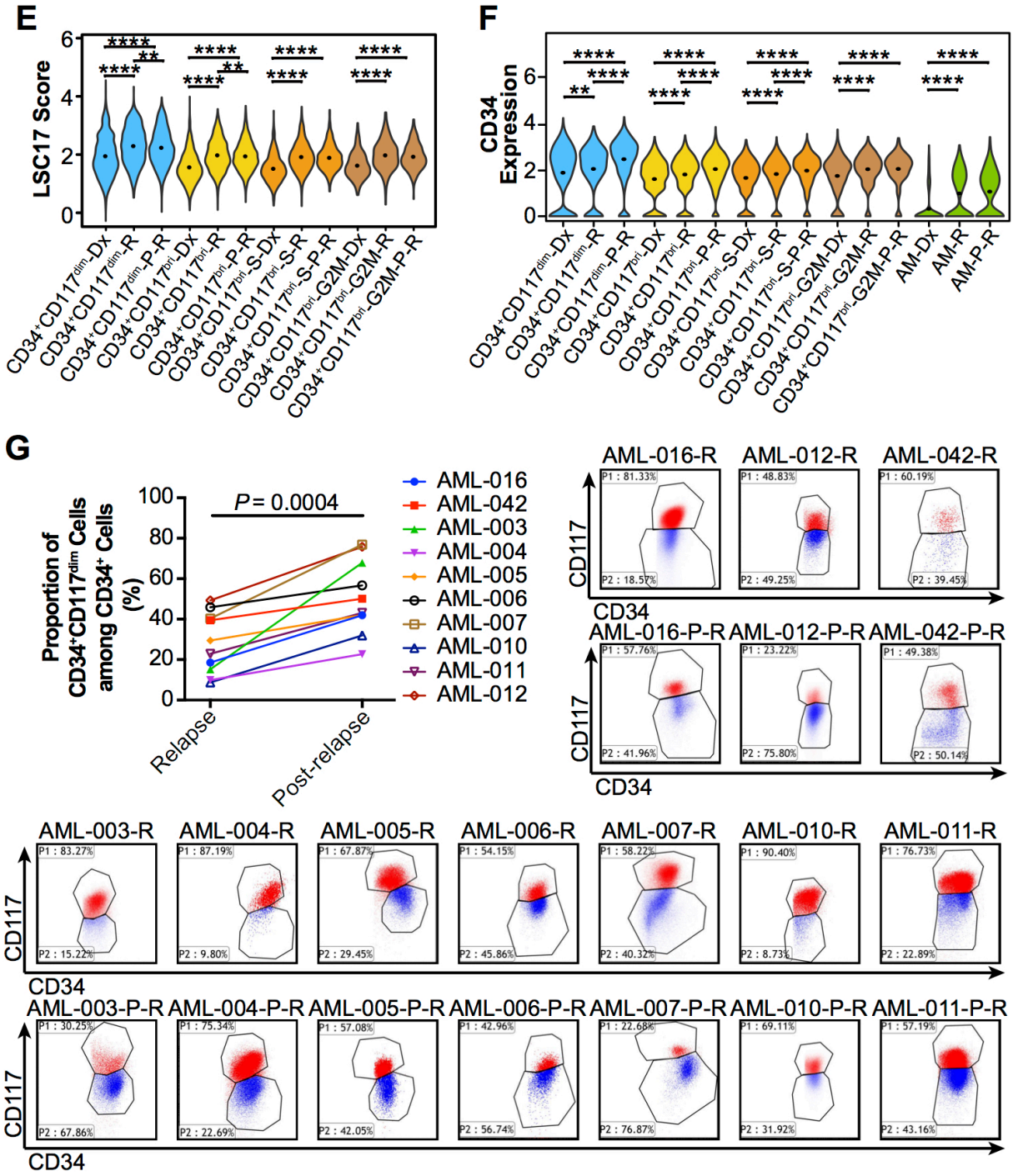

E
LSC17 Score
F
CD34 Expression
G
Proportion of CD34+CD117dim Cells among CD34+ Cells (%)
P = 0.0004
Relapse
Post-relapse
AML-016
AML-042
AML-003
AML-004
AML-005
AML-006
AML-007
AML-010
AML-011
AML-012
AML-016-R
AML-012-R
AML-042-R
AML-016-P-R
AML-012-P-R
AML-042-P-R
AML-003-R
AML-004-R
AML-005-R
AML-006-R
AML-007-R
AML-010-R
AML-011-R
AML-003-P-R
AML-004-P-R
AML-005-P-R
AML-006-P-R
AML-007-P-R
AML-010-P-R
AML-011-P-R
CD117
CD34

Fig. S5. Gene Expression Profiles and Clinical Features of t(8;21) AML Patients with Different $CD34^{+}CD117^{dim}$ Cell Proportions

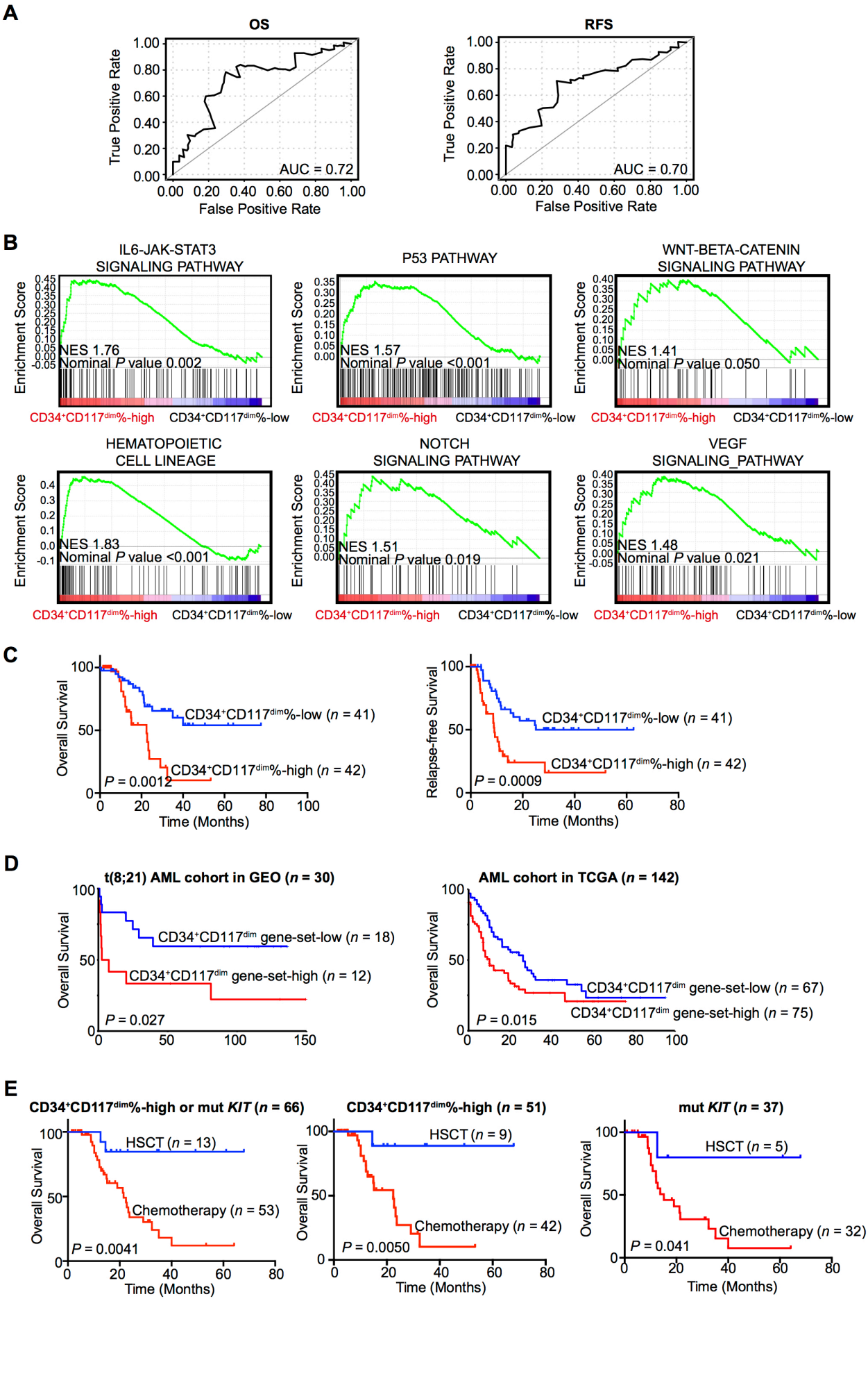
(**A**) Receiver operating characteristic (ROC) curves of $CD34^{+}CD117^{dim}$ proportion in terms of overall survival (OS) and relapse-free survival (RFS). Area under curves (AUC): OS, 0.72; RFS, 0.70.

(**B**) GSEA results showing the activated pathways of IL6-JAK-STAT3, P53, WNT-β-catenin, hematopoietic cell lineage, NOTCH and VEGF in $CD34^{+}CD117^{dim}$%-high group. Normalized enrichment score (NES) and nominal *P* value were given.

(**C**) OS and RFS stratified by $CD34^{+}CD117^{dim}$ proportion in 83 t(8;21) AML patients who did not receive HSCT.

(**D**) We established a gene-set based on the highly expressed genes of $CD34^{+}CD117^{dim}$%-high group ($CD34^{+}CD117^{dim}$ gene-set). $CD34^{+}CD117^{dim}$ gene-set-low and $CD34^{+}CD117^{dim}$ gene-set-high represent patients with low expression of $CD34^{+}CD117^{dim}$ gene-set and patients with high expression of $CD34^{+}CD117^{dim}$ gene-set, respectively. Kaplan-Meier survival curves for 30 t(8;21) AML patients in GEO and 142 AML patients in TCGA were shown.

(**E**) Kaplan-Meier survival curves stratified by different therapeutic strategies in t(8;21) AML patients with high $CD34^{+}CD117^{dim}$% or *KIT* mutations (*n* = 66, left), with high $CD34^{+}CD117^{dim}$% (*n* = 51, middle), and with *KIT* mutations (*n* = 37, right).


A
OS
True Positive Rate
False Positive Rate
AUC = 0.72
RFS
AUC = 0.70
B
IL6-JAK-STAT3 SIGNALING PATHWAY
NES 1.76
Nominal P value 0.002
P53 PATHWAY
NES 1.57
Nominal P value <0.001
WNT-BETA-CATENIN SIGNALING PATHWAY
NES 1.41
Nominal P value 0.050
HEMATOPOIETIC CELL LINEAGE
NES 1.83
Nominal P value <0.001
NOTCH SIGNALING PATHWAY
NES 1.51
Nominal P value 0.019
VEGF SIGNALING_PATHWAY
NES 1.48
Nominal P value 0.021
Enrichment Score
CD34+CD117dim%-high
CD34+CD117dim%-low
C
Overall Survival
Relapse-free Survival
Time (Months)
CD34+CD117dim%-low (n = 41)
CD34+CD117dim%-high (n = 42)
P = 0.0012
P = 0.0009
D
t(8;21) AML cohort in GEO (n = 30)
CD34+CD117dim gene-set-low (n = 18)
CD34+CD117dim gene-set-high (n = 12)
P = 0.027
AML cohort in TCGA (n = 142)
CD34+CD117dim gene-set-low (n = 67)
CD34+CD117dim gene-set-high (n = 75)
P = 0.015
E
CD34+CD117dim%-high or mut KIT (n = 66)
HSCT (n = 13)
Chemotherapy (n = 53)
P = 0.0041
CD34+CD117dim%-high (n = 51)
HSCT (n = 9)
Chemotherapy (n = 42)
P = 0.0050
mut KIT (n = 37)
HSCT (n = 5)
Chemotherapy (n = 32)
P = 0.041

SI Tables

Table S1. Summary of single-cell RNA sequencing metrics of nine t(8;21) AML patients

Patient ID	Stage	Sequencer	Number of Cells Recovered	Median Genes per cell	Median unique molecular identifier counts per cell	Fraction Reads in Cells	Reads mapped confidently to exonic regions	Reads mapped confidently to intronic regions	Reads mapped confidently to intergenic regions	Valid barcodes
AML-013	Diagnosis	Illumina novaseq 6000 System	7124	1304.5	4707	85.70%	65.10%	20.00%	5.40%	97.20%
AML-068	Diagnosis	Illumina novaseq 6000 System	7956	1291	4946.5	82.70%	60.10%	27.40%	4.80%	97.80%
AML-060	Diagnosis	Illumina novaseq 6000 System	10487	1890	7365	91.40%	68.10%	20.40%	3.80%	98.00%
AML-072	Diagnosis	Illumina novaseq 6000 System	8201	2382	9149	95.50%	69.70%	17.40%	4.20%	98.00%
AML-049	Diagnosis	Illumina novaseq 6000 System	8344	1328	3765	82.00%	57.40%	26.70%	4.80%	97.70%
AML-070	Diagnosis	Illumina novaseq 6000 System	10001	1949	6585	78.80%	69.00%	18.70%	4.50%	97.90%
AML-076	Diagnosis	Illumina novaseq 6000 System	7784	1948	7478.5	86.40%	68.90%	17.40%	3.80%	98.00%
AML-101	Diagnosis	Illumina novaseq 6000 System	11815	2110	8150	96.20%	70.10%	16.50%	3.80%	98.00%
AML-016	Diagnosis	Illumina novaseq 6000 System	11309	1906	7925	92.30%	64.50%	19.50%	4.10%	97.70%
AML-016	Complete Remission	Illumina novaseq 6000 System	6536	763	2616.5	75.30%	69.10%	14.70%	3.70%	96.80%
AML-016	Relapse	Illumina novaseq 6000 System	8249	1687	8911	93.20%	71.50%	15.10%	3.80%	96.80%
AML-016	post-relapse refractory disease stage	Illumina novaseq 6000 System	7474	2348.5	11290	93.90%	66.00%	18.70%	3.90%	97.20%

Table S2. Specific immunophenotypic markers and gene markers of distinct cell clusters

Cluster	ID	Immunophenotype	Gene Marker
0	$CD34^{+}CD117^{dim}$	$CD34^{+}CD38^{+}CD7^{-}CD10^{-}$ $CD45RA^{+}$ (GMP) $CD117^{dim}$	*CRIP1, CRIP2, LGALS1, ANXA2, EMP3, VIM*
1	$CD34^{+}CD117^{bri}$	$CD34^{+}CD38^{+}CD7^{-}CD10^{-}$ $CD45RA^{+}$ (GMP) $CD117^{bri}$	*NPW, C1QTNF4, TRH, PLAC8, KIT, DNTT*
2	$CD34^{+}CD117^{bri}$-S		*TYMS, TK1, PCNA, CCNB1, CENPF, UBE2C*
3	$CD34^{+}CD117^{bri}$-G2M		
4	AM	$CD34^{-}CD117^{-}$HLA-DR^{-} $CD15^{+}CD11b^{-}$	*PRTN3, ELANE, AZU1, CST7*
5	DC	$CD45^{+}$	*FCER1A, HPGD, MS4A3, IL5RA*
6	Monocyte precursor	$CD34^{dim/-}CD14^{-}$	*MS4A6A, LYZ, CST3, S100A9, CCL3*
7	Monocytes	$CD34^{-}CD117^{-}CD64^{+}CD14^{+}$	*FCN1, CCL3, S100A8, S100A9, LYZ, CST3*
8	B cells	$CD19^{+}CD79A^{+}$	*CD79A, CD79B, ZCCHC7*
9	T cells	$CD3^{+}CD7^{+}$	*CD3D, CD3E, CD7, GZMA*
10	Immature erythroid cells	$CD71^{+}$	*CA1, CA2, HBM, HBD*

Table S3. Top 200 differentially expressed genes of $CD34^{+}CD117^{dim}$, $CD34^{+}CD117^{bri}$ and AM cell populations from bulk RNA-seq data

$CD34^{+}CD117^{dim}$ *vs.* other signature		$CD34^{+}CD117^{bri}$ *vs.* other signature		AM *vs.* other signature	
MS4A4E	*COX6CP17*	*ZRANB3*	*RRP1*	*MCU*	*ANKRD22*
HPGD	*TGM2*	*LAPTM4B*	*TFAP4*	*MYO7B*	*WIPI1*
DNAH8	*AXL*	*TRAP1*	*NDC1*	*TACSTD2*	*PIP4P2*
SIGLEC6	*SYTL3*	*PPAT*	*GEMIN5*	*CAPN3*	*IPCEF1*
MAF	*MOB3B*	*DPY19L2P2*	*IDH1*	*GLRX*	*DMXL2*
CAMK1D	*ZFP36L1*	*TERT*	*NUP205*	*ABCA13*	*RILPL1*
NTRK1	*MS4A2*	*NPW*	*IARS*	*ANO10*	*ARFGEF1*
ABCA9	*CDC42EP3*	*C1QTNF4*	*NLE1*	*ENTPD1*	*CALM1*
TPSAB1	*FBXL2*	*KIT*	*NPM3*	*GYG1*	*LRRC25*
SEMA7A	*EBF1*	*PALB2*	*NETO2*	*PADI2*	*DHCR7*
ADAM8	*MYCNOS*	*CENPV*	*ARMC6*	*ALDOC*	*PIGB*
CD226	*ZDHHC11B*	*RTN4R*	*CDCA7*	*DIRC2*	*PSTPIP1*
KIFC3	*POU2F2*	*GATB*	*POLD2*	*COL17A1*	*UGCG*
C2orf66	*ARHGAP31*	*MRPS26*	*PGBD4P1*	*GCA*	*KRT8P26*
SLC10A5	*NLRP1*	*MDFI*	*USP51*	*MSRB1*	*GADD45A*
FCER1A	*RARA*	*SAMD11*	*DNAJC12*	*FBXO9*	*TMEM87A*
STAB1	*LMO4*	*AADAT*	*ALDH18A1*	*NQO2*	*CPNE2*
TNIK	*CD209*	*DYDC2*	*ALDH1B1*	*LTB4R*	*CEBPE*
IL5RA	*GDF11*	*GPSM2*	*NUP35*	*CFLAR*	*CPNE3*
GLOD5	*CSPG4*	*TRUB2*	*KLRG2*	*LRG1*	*NADK*
NR4A3	*TPPP3*	*ANKRD16*	*DANCR*	*SNX18*	*ZFAND5*
UGT2B11	*SLC6A6*	*ANAPC1*	*DUSP27*	*BST1*	*BPI*
SFXN3	*CNTN4-AS1*	*NOG*	*BMP1*	*C3AR1*	*NCF2*
LGALS1	*ANXA2*	*KNOP1*	*NOP56*	*RGL4*	*SPAG4*
AHNAK	*ISPD-AS1*	*SLC27A5*	*GPT2*	*CITED2*	*OCRL*
FOSL2	*PARP15*	*THOC3*	*INPP5J*	*SLCO4C1*	*PLIN5*
RGS9	*FAM83F*	*MMAB*	*UMPS*	*ANXA3*	*SORT1*
S1PR1	*CCDC141*	*KLHL23*	*HNRNPA1P10*	*AMPD3*	*PSTPIP2*
TPSD1	*NRBP2*	*HADH*	*ECI2*	*VILL*	*LRRC75A*
GATA1	*EFNA4*	*VSIG10*	*SUV39H2*	*ABHD5*	*WAC-AS1*
CYSLTR2	*NMT2*	*ADA*	*TRBV28*	*CLTCL1*	*CALML4*
MICALCL	*SPTBN2*	*PRELID3A*	*EARS2*	*NLRC4*	*CSNK1A1L*
CRISPLD2	*ITGB5*	*CCDC8*	*B9D1*	*AKTIP*	*GFOD1*
IGLON5	*RNU6-509P*	*MSH2*	*MATK*	*DAPK2*	*MLNR*
CCR4	*PHLPP2*	*KRT17P3*	*SLC25A19*	*SRPK1*	*DZIP1L*
LINC01878	*DDIT4*	*PYCR3*	*RPGRIP1L*	*SLC26A8*	*LILRA2*
ACVRL1	*MEIS2*	*AIF1L*	*LRPPRC*	*PIWIL4*	*AFF2*
DOK6	*ARID5B*	*AGMAT*	*CHCHD4*	*PPM1M*	*NFAM1*
EMP1	*VIM*	*STAP2*	*NME1*	*FAR2*	*MIF4GD*

CRIP2	*LINC01121*	*FUT10*	*PAICS*	*CEACAM1*	*PTPN22*
MICAL2	*DMWD*	*RPUSD3*	*B4GAT1*	*STXBP5-*	*AGTRAP*
NR4A1	*AP1B1*	*FAM86C2P*	*LINC00920*	*CEACAM8*	*TRIB3*
GRAP2	*FLT1*	*MSH6*	*SIGMAR1*	*TTPAL*	*CPEB2*
CSF1	*PDE4A*	*UNG*	*ZNF730*	*FAM200B*	*SERPINB1*
SORBS1	*ADGRE2*	*IMPDH2*	*NOC3L*	*ALPK1*	*NCBP3*
ITGB8	*TAL1*	*FARSA*	*RUVBL1*	*JAG1*	*TESMIN*
CD22	*SLC2A6*	*NR2C2AP*	*FKBP4*	*ALAS1*	*ST7L*
CTNNBL1	*MTCL1*	*THOP1*	*CHEK2*	*ZFP92*	*ICA1*
CDK15	*RRAD*	*NUDT8*	*AGPAT5*	*FBXL5*	*G6PD*
LPP	*IDS*	*DNPH1*	*CEP41*	*GFOD2*	*COL18A1*
PPP1R16B	*PHLDA1*	*NME4*	*DCTPP1*	*MPP7*	*TECPR2*
TRAF1	*CCR3*	*FAHD2B*	*IFITM3*	*OGG1*	*PLB1*
CD44	*ATF3*	*RAD54B*	*IGFBP6*	*FRMD3*	*SYK*
LMNA	*ZNF609*	*MAST4-AS1*	*ATP6V0E2-*	*PDE6H*	*B4GALT5*
IL1RL1	*HLA-L*	*IFRD2*	*HLTF*	*PTH2R*	*NOCT*
SLC35E4	*GRASP*	*RPP40*	*ADRA2C*	*CLEC9A*	*NXF3*
TIMP3	*RHBDD2*	*TMEM97*	*BOLA3*	*FUT4*	*IFT20*
CACNA2D	*PAOX*	*CD320*	*EBPL*	*SUCO*	*ANKRD42*
GAS7	*TOX*	*F7*	*CDC7*	*NSUN5*	*FSTL3*
MTSS1	*KCNH8*	*USP27X-AS1*	*LSM14B*	*DNAJC5*	*PNP*
BAIAP3	*AHR*	*TMIE*	*C19orf48*	*PIK3CG*	*GANC*
CHDH	*MMP19*	*HPDL*	*MYO5C*	*BRI3*	*KIAA1211L*
DUSP5	*AGBL4*	*MFAP4*	*WDR3*	*S100P*	*AZU1*
PHACTR1	*METRNL*	*FOLH1*	*SLC39A14*	*SPINK8*	*FZD9*
EEF2K	*CHST3*	*AKR1A1*	*BLMH*	*ENTPD7*	*PRPSAP1*
RIPOR3	*NRCAM*	*POLR1E*	*SLC25A15*	*ARG1*	*VPS8*
SRC	*GAB2*	*PRSS57*	*SKA3*	*ZNF33A*	*FCAR*
GBP2	*LPXN*	*ALDH7A1*	*ZNF835*	*ATP8A1*	*C1orf162*
SLC43A2	*S100A10*	*ZNF724*	*ZNF519*	*BCL2L15*	*C9orf66*
CD80	*SYNGAP1*	*GAMT*	*MRPL11*	*TARM1*	*ITGAE*
CALCRL	*N4BP3*	*JADE3*	*TRH*	*CST7*	*STOM*
SIPA1L1	*NT5DC2*	*PM20D2*	*CENPI*	*LYZ*	*VSTM1*
HVCN1	*CYP1B1*	*PDCD11*	*CSRP2*	*CYSTM1*	*NDUFB4P2*
AR	*EHD2*	*KTN1-AS1*	*DSCC1*	*TESK2*	*GNS*
DAPK1-IT1	*PPP4R1L*	*MACROD1*	*TMIGD2*	*GGTA1P*	*WNT5B*
CD109	*DHRS3*	*AGK*	*CACNA1C-*	*CEACAM6*	*SLC35C2*
RCAN3	*FCMR*	*PCCB*	*BUB1*	*MNDA*	*ATP8B4*
LINC02458	*DMD*	*TEDC2*	*SERPINE2*	*NPL*	*CTSD*
TPPP	*ZNF521*	*IGLL1*	*ERCC6L*	*OLR1*	*ELANE*
TTLL6	*CALHM2*	*TRBV20-1*	*WDR18*	*MGAT5*	*TUBA4A*
EMP3	*SLC9A9*	*ITPRIPL1*	*STMN1*	*MAPK14*	*AP4B1*
CNRIP1	*MYADM*	*TSEN2*	*MRPL17*	*SLC45A4*	*RHOU*
SERPINI2	*SLC12A3*	*COQ3*	*C1QBP*	*CDADC1*	*VAT1*

MYCN	*FAM49A*	*CCDC138*	*ACTL6A*	*APMAP*	*C5AR2*
CHST2	*RASL10A*	*CD200*	*GPATCH4*	*ROGDI*	*NKG7*
ITPR3	*PTGDR2*	*CYB5A*	*LDHB*	*ARHGAP24*	*ZNF189*
LPAR5	*GMPR*	*TLN2*	*SNHG1*	*RGCC*	*S100A8*
MAFF	*RARG*	*SAPCD2*	*MREG*	*FAM46A*	*KCNE1*
BMP6	*LINC00578*	*NARS2*	*TRMT11*	*ACTN1*	*GPR160*
TPSB2	*LINC01218*	*ANKLE1*	*GINS1*	*CD24P4*	*FAM107B*
IFNK	*SIDT1*	*FIRRE*	*ZNF550*	*SLA*	*TBC1D22B*
AQP1	*IL13*	*CPXM1*	*KDELC2*	*DNAJC13*	*DPY19L1P1*
CRTC3	*TMOD1*	*DKC1*	*HSPC324*	*PLEKHA2*	*RASA1*
SPINK4	*IL17RB*	*GNPTAB*	*MAD2L1*	*SLC25A37*	*PFKFB3*
TGFB1I1	*DUSP8*	*KYAT1*	*CCNG1*	*SRGN*	*MFSD14B*
DMPK	*ZMIZ1*	*MYOZ3*	*MYT1*	*S100A9*	*CD82*
MS4A1	*NOS3*	*PACSIN3*	*ANKRD26*	*SLC22A4*	*KCNH7*
ENPP3	*IFFO2*	*SLC25A1*	*PARP1*	*ZNF438*	*YWHAH*
GPRC6A	*ELK3*	*POLD1*	*PGF*	*MLKL*	*P2RY13*
TNPO1P1	*PLD2*	*DNMT3B*	*C20orf197*	*ZBTB47*	*TMEM170B*

Table S4. Clinical characteristics of patient AML-016 performing scRNA-seq and WES at different time points

ID	Gender	Age	Days	Diagnosis	WBC ($x10^9$/L)	Morphology		Cytogenetics	*RUNX1-RUNX1T1* (RT-PCR)			*KIT* Mutation	LAIP %	Treatment
						Myeloblasts %	AM %		*RUNX1-RUNX1T1*	*ABL*	Ratio			
AML-016	Female	59	0	Diagnosis	22.8	57.5	29.5	46,XX,der(8)t(8;21)(q22;q22),-13,-21[1]/46,XX[4]/[6]	2.33E+06	7.05E+05	3.3	+	51.4	Idarubicin x 3 days + Ara-C x 7 days
			163	CR (5 months)	4.84	<5	-	/	2.34E+03	9.10E+05	2.57E-03	-	<0.01	Intermediate-dose Ara-C
			401	Relapse	2.64	85	-	NA	1.14E+06	5.70E+05	1.99	+	72.5	Idarubicin x 3 days + Ara-C x 7 days + Dasatinib x 14 days
			585	Refractory disease post-relapse	4.56	51	-	44~45,XX,der(8)t(8;21)(q22;q22),-11,-14,+M1~M2[CP2]/45,X,-X,+der(8)t(8;21)(q22;q22),-11,-14,+M1~M2[CP3]/44~45,XX,der(8)t(8;21)(q22;q22),add(11q23),-14,+M1~M2[CP2]/46,XX[2]/[11]	NA			+	68.4	CAG regimen (Aclarubicin x 4 days, Ara-C x 14 days, G-CSF x 14 days)

Table S5. Clinical characteristics of 101 t(8;21) AML patients

Characteristic	Total	$CD34^+CD117^{dim}$% -low	$CD34^+CD117^{dim}$% -high	*P* value
Patients (N)	101	50	51	
Age, median (range) (years)	41 (17-74)	42.5 (17-67)	41 (18-74)	0.433
Gender (Male/Female)	55/46	26/24	29/22	0.624
WBC, median (range) (× 10^9/L)	10.5 (2.3-94.6)	7.8 (2.4-94.6)	12.8 (2.3-69.4)	0.009
Hemoglobin, median (range) (g/L)	76.0 (41.0-153.0)	79.5 (42.0-125.0)	74.0 (41.0-153.0)	0.187
Platelet, median (range) (× 10^9/L)	32.0 (4.0-221.0)	33.0 (4.0-221.0)	32.0 (6.0-118.0)	0.974
Marrow blasts, median (range) (%)	44.5 (9.0-93.0)	43.25 (9.0-93.0)	45.5 (12.0-92.0)	0.772
AM, median (range) (%)	12.5 (0-45.0)	14.0 (0-39.0)	11.5 (0-45.0)	0.296
Karyotype[1]				
t (8;21) alone/all patients	30/91	18/46	12/45	0.206
Loss of X or Y chromosome/all patients	44/91	19/46	25/45	0.174
Molecular mutations				
KIT mutation/all patients[2]	38/100	15/49	23/51	0.136
Immunophenotype				
CD34 (positive/negative)[3]	100/1	49/1	51/0	0.495
CD117 (positive/negative)	101/0	50/0	51/0	-
CD19 (positive/negative)	63/38	33/17	30/21	0.457
CD56 (positive/negative)	77/24	37/13	40/11	0.601
CD11b (positive/negative)[4]	6/95	3/47	3/48	1.000
CD15 (positive/negative)	19/82	12/38	7/44	0.186
Induction cycles to attain CR				
1 cycle/ > 1 cycle	86/15	45/5	41/10	0.175

Note:

1. 91 patients were available for cytogenetic analysis.
2. Out of the 101 patients, *KIT* mutation was unknown in 1 patient.
3. CD34 was tested by Fisher's Exact Test.
4. CD11b was tested by Continuity Correction. The rest binary variables were tested by Person Chi-Square.

Reference

1. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013; 29(1):15-21.
2. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25(16):2078-9.
3. Anders S, Pyl PT & Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015; 31(2):166-9.
4. Love MI, Huber W & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; 15(12):550.
5. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017; 14(4):417-419.
6. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015; 43(7):e47.
7. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005; 102(43):15545-50.
8. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012; 16(5):284-7.
9. Aran D, Hu Z & Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 2017; 18(1):220.
10. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20(9):1297-303.
11. Wang K, Li M & Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010; 38(16):e164.
12. Li JF, Dai YT, Lilljebjorn H, et al. Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. *Proc Natl Acad Sci U S A.* 2018; 115(50):E11711-E11720.
13. Christen F, Hoyer K, Yoshida K, et al. Genomic landscape and clonal evolution of acute myeloid leukemia with t(8;21): an international study

on 331 patients. *Blood.* 2019; 133(10):1140-1151.

14. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29(1):308-11.
15. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 2017; 45(D1):D777-D783.
16. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011; 29(1):24-6.
17. Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018; 36(5):411-420.
18. Leek JT, Johnson WE, Parker HS, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012; 28(6):882-3.
19. Buttner M, Miao Z, Wolf FA, et al. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods.* 2019; 16(1):43-49.
20. Amir el AD, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol.* 2013; 31(6):545-52.
21. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol.* 2018;
22. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science.* 2016; 352(6282):189-96.
23. Ng SW, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature.* 2016; 540(7633):433-437.
24. Qiu X, Hill A, Packer J, et al. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods.* 2017; 14(3):309-315.
25. Li H & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25(14):1754-60.
26. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012; 22(3):568-76.

27. Miller CA, McMichael J, Dang HX, et al. Visualizing tumor evolution with the fishplot package for R. *BMC Genomics.* 2016; 17(1):880.
28. Cancer Genome Atlas Research N, Ley TJ, Miller C, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med.* 2013; 368(22):2059-74.
29. Herold T, Metzeler KH, Vosberg S, et al. Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood.* 2014; 124(8):1304-11.
30. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 2016; 44(8):e71.
31. Davis S & Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007; 23(14):1846-7.
32. Hanzelmann S, Castelo R & Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics.* 2013; 14(7.
33. Heagerty PJ, Lumley T & Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics.* 2000; 56(2):337-44.
34. Zhu HH, Zhang XH, Qin YZ, et al. MRD-directed risk stratification treatment may improve outcomes of t(8;21) AML in the first complete remission: results from the AML05 multicenter trial. *Blood.* 2013; 121(20):4056-62.
35. Schnittger S, Weisser M, Schoch C, et al. New score predicting for prognosis in PML-RARA+, AML1-ETO+, or CBFBMYH11+ acute myeloid leukemia based on quantification of fusion transcripts. *Blood.* 2003; 102(8):2746-55.
36. Leroy H, de Botton S, Grardel-Duflos N, et al. Prognostic value of real-time quantitative PCR (RQ-PCR) in AML with t(8;21). *Leukemia.* 2005; 19(3):367-72.