SUPPLEMENTAL DATA

Single-cell multi-omics reveals increased plasticity, upfront resistant subpopulations and stem-cell-like blasts in KMT2A-rearranged leukemia

Supplemental Methods

Human biospecimens

Peripheral blood or bone marrow samples were from patients with infant ALL enrolled on the Children’s Oncology Group (COG) clinical trial AALL15P1 or they were collected under the Children’s Hospital of Philadelphia (CHOP) Institutional Review Board Protocol (IRB) Protocol 10-007767 (CHOP Center for Childhood Cancer Research (CCCR) Biorepository), or under CHOP IRB Protocol 94-000771. All samples were obtained with parental informed consent according to the Declaration of Helsinki with IRB approval from all participating centers. Mononuclear cells from patients were prepared using a ficoll gradient and stored frozen in 10% DMSO. Sample information is listed in Supplemental Table 1.

Cell sorting

Leukemia samples were thawed and stained with appropriate antibodies and DAPI (Cat #: 1246530100, Sigma) and immediately subjected to FACS sorting (FACSAria Fusion, BD). Sorted CD19+ and CD19- populations were subjected to scRNA-Seq and scATAC-Seq. For snmC-Seq2 protocol, sorted CD19+ were used. For healthy donor samples, DAPI-, DAPI-Lin-CD34+ and DAPI-Lin-CD34+CD38- populations were sorted and
subjected to scRNA-Seq and scATAC-Seq. Antibody information is listed in Supplemental Table 3.

**scRNA-Seq**

Sorted cells were immediately processed using the 10x Genomics Chromium controller and the Chromium Single Cell 3’ Reagent Kits V3 protocol. 7,000-16,000 cells were loaded for each sample. Cells were partitioned into gel beads, lysed, and barcoded through reverse transcription. cDNA were purified and amplified using appropriate cycle number following the manufacturer’s protocol. Libraries were constructed using 10x Genomics Library Prep Kit. Library quality was checked using Agilent High Sensitivity DNA Kit and Bioanalyzer 2100. Libraries were quantified using dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen) on Qubit fluorometer and the qPCR-based KAPA quantification kit. Libraries were sequenced on an Illumina Nova-Seq 6000 with 28:8:0:87 paired-end format.

**scATAC-Seq**

Sorted cells were centrifuged at 300g for 5 min at 4°C. 45uL of chilled lysis buffer was added to cell pellets and mixed by pipetting gently three times, and incubated 3 min on ice. After incubation, 50uL of pre-chilled wash buffer was added without mixing and centrifuged immediately at 300g for 5 min at 4°C. 95uL supernatant was carefully discarded and 45uL pre-chilled diluted nuclei buffer (10x Genomics) was added without mixing and sample was centrifuged at 300g for 5 min at 4°C. The nuclei pellet was then resuspended in 7uL pre-chilled diluted nuclei buffer and nuclei concentration was determined using a Countess II cell counter (Invitrogen). 7,000-16,000 nuclei were used
for the transposition reaction in bulk, and then loaded to the 10x Genomics Chromium controller. Cells were partitioned into gel beads, lysed, and barcoded followed by the Chromium Single Cell ATAC Reagent Kits protocol. The barcoded and transposed DNA was amplified with appropriate sample index. Library quality was checked using Agilent High Sensitivity DNA Kit and Bioanalyzer 2100. Libraries were quantified using dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen) on Qubit fluorometer and the qPCR-based KAPA quantification kit. Libraries were sequenced on an Illumina Nova-Seq 6000 with 49:8:16:49 paired-end format.

**snmC-Seq2**

The plates were prepared in an AirClean®600 PCR Workstation to minimize environmental DNA contamination. 4 µL lysis mix (containing 2 µL M-digestion buffer, 0.2 µL Proteinase K (20 mg/mL, from Zymo kit, cat. #D5023), 1.8 µL diluted lambda DNA (0.01 pg/well)) was added to each well of 96-well plates before covering by sealing film (Thermo Scientific, cat. #ab-0558). Single DAPI- CD19+ cell was sorted into each well. After cell sorting, the plates were covered by sealing film and incubated at 50°C for 20 min (PK-digestion). All steps of library preparation were performed in an AirClean®600 PCR Workstation to minimize environmental DNA contamination. The snmC-seq2 library preparation was performed as previously described with minor modifications ¹. Briefly, bisulfite conversion of single cells was carried out using Zymo EZ-96 DNA Methylation-Direct™ Kit (Deep Well Format, cat. #D5023). The DNA was eluted in 9 µL Low EDTA TE buffer (Swift Biosciences) plus 1 µL of an assigned
random primer (final concentration 0.5 µM) (Supplemental Table 4) to allow for later multiplexing (8-plex) reactions.

Before random-priming, plates were denatured by incubating at 95°C for 3min and were immediately chilled on ice for 2min. 10 µL enzyme mix was added to each well. The following program was used for random-priming: 4°C for 5min, ramp up to 25°C at 0.1°C/sec, 25°C for 5min, ramp up to 37°C at 0.1°C/sec, 37°C for 60min. Then 2 µL Exonuclease I (20 U/µL, NEB cat. # M0293L) and 1 µL Shrimp Alkaline Phosphatase (rSAP) (1 U/µL, NEB cat. # M0371L) was added to each reaction and followed by incubation in a thermocycler at 37°C for 30min. The pooled samples were purified once with 0.8x SPRI beads and eluted with 10 µL Low EDTA TE buffer. Eluted samples were transferred to new PCR strips for downstream steps.

Before Swift Adaptase reaction, samples were denatured in a thermocycler at 95°C for 3min and subsequently chilled on ice for 2min. Then 10.5 µL Adaptase master mix was added to each reaction. Reactions were incubated in a thermocycler at 37°C for 30min. Subsequently, 30 µL PCR mix was added to each well, followed by mixing with pipetting. The following program was used for indexing-PCR: 95°C for 2min, 98°C for 30sec, 15 cycles of (98°C for 15 sec, 64°C for 30 sec, 72°C for 2min), 72°C for 5 min, and hold at 4°C. PCR products were cleaned with two rounds of 0.8x SPRI beads. Library quality was then checked using Agilent High Sensitivity DNA Kit and Bioanalyzer 2100. Libraries were quantified using dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen) on Qubit fluorometer and the qPCR-based KAPA quantification kit. Libraries were sequenced on an Illumina Nova-Seq 6000 with 150:8:8:150 paired-end format.
**Single-cell targeted long-read sequencing with Oxford Nanopore technology**

Barcoded full-length cDNA library from 10x Genomics protocol was amplified using the primer that targets *KMT2A* exon 7 (Supplemental Table 4) and KAPA HiFi HotStart ReadyMix (KAPA, Cat: KK2600). The following program was used for PCR: 95°C for 3 min, 12 cycles of (98°C for 20 sec, 60°C for 15 sec, 72°C for 5 min), 72°C for 5 min, and hold at 4°C. Samples were cleaned with Dynabeads Kilobase BINDER (Thermo Fisher, Cat: 6010) to enriched for biotinylated amplified cDNA. After Dynabeads cleanup, samples were appended with Illumina P5 and P7 adapter sequences using additional PCR reaction: 95°C for 3 min, 15 cycles of (98°C for 20 sec, 60°C for 15 sec, 72°C for 5 min), 72°C for 5 min, and hold at 4°C. PCR products were cleaned with one round of 0.7x SPRI beads. PCR products were quantified using dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen) on Qubit fluorometer. Nanopore libraries were prepared using Nanopore Ligation Sequencing Kit (Nanopore, Cat #: SQK-LSK109). Samples were sequenced on Nanopore GridION.

**Generation of CRISPR/Cas9 knockout cell lines**

Guide RNA sequences targeting human *KLF9* exon 1 (p5-

CACCGCACCGCAGCGGTCGGAAA-3 and p5-

CACCGCGACGTGACCACCGAATCT-3) and human *NR3C1* exon 1 (p5-

CACCGGCTGAACT CTTGGGGTTCTC-3 and 5-CACCGTTACATCTGGTCTCATGCT-

3) were cloned into the lentiviral vector LRG2.1 (Addgene) using a BsmBI restriction site. The LentiV_Cas9_puro vector (Addgene) was used for generating Cas9-expressing cells. 10 µg LRG2.1GFP sgRNA or LentiV_Cas9_puro plasmid were
cotransfected with 9 µg p8.9 and 1 µg VSVg plasmids (Addgene) using 25 µL Lipofectamine to 293T cells. Media were collected every 24h for three days. Supernatant containing the virus was concentrated using polyethylene glycol (PEG). 

*KMT2A-r* cell lines (SEMK2, KOPN8 and RS4;11) first were transduced with LentiV_Cas9_puro virus and Cas9-expressing cells were enriched by plating in puromycin from 24h to 72h after transduction. Cells were then sorted for Cas9-EGFP expression by flow cytometry. Knockout was confirmed by Western Blot.

**Cell culture, drug treatment and viability assay**

The human *KMT2A-r* cell lines were cultured in RPMI 1640 supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin under 5% CO2 humidified atmosphere at 37°C. Knockout and wild type cells were exposed to different concentrations of dexamethasone and prednisone for specific time and viability of cells was monitored by the colorimetric XTT-based assay (Roche).

**DNA FISH**

Primary patient cells were sorted as Lin-CD19-CD34+CD63+. Cells were subjected to cytospin and fixed with 3:1 methanol:acetic acid for 10 min on slides. Vysis LSI MLL Dual Color probe (Abbott Molecular, Cat #: 05J90-001) was used in this study. Probe was denatured at 73°C for 5 min in the heat block and applied to the slides. Slides with the coverslips were denatured at 73°C for 8-10 min on a slide warmer. Hybridization was performed in a dark humidity chamber for 16 hr. Slides were washed with 0.4X SSC/0.3% NP-40 (SSC buffer, S6639, Sigma) at 75°C for 2 min and 2X SSC/0.1% NP-
40 at RT for 1 min, followed by staining with DAPI (P36935, Invitrogen). Slides were stored at -20°C or imaged immediately.

**NK cell isolation, ex vivo expansion, and co-culture assay**

CD56+CD3- NK cells were isolated from fresh human PBMC samples using NK Cell Isolation Kit (130-092-657, Miltenyi Biotec). NK cells were expanded *ex vivo* in complete NK MACS medium (130-114-429, Miltenyi Biotec) with 500IU/mL IL-2 (200-02, Pepro Tech) and 140IU/mL IL-15 (200-15, Pepro Tech) for 14 days. HSPC-like cells and CD19+ blasts were sorted from primary apheresis samples, and then stained with 0.1 uM carboxyfluorescein diacetate succinimidyl ester (CFSE, C1157, Invitrogen) in 96-well plate at 37°C for 15 min. Cells were centrifuged at 300g for 5 min at RT and resuspended in the prewarmed NK MACS medium. CFSE-labeled cells were incubated at 37°C for another 30 min and washed one more time with prewarmed NK MACS medium. NK cells were then plated with the target cells by the ratio of 5:1 (NK:target) and centrifuged at 300g for 3 min. The plate was incubated for overnight (~12-hr) and then analyzed by flow. The percentage of dead cells in CFSE-positive population were calculated by subtracting the baseline dead cell percentage without NK co-culture.

**Transplantation experiment**

NOD/SCID/IL2Rγ-/- (NSG) and NOD-SGM3 (NSGS) mice were purchased from the Jackson laboratory and maintained at the Animal Research Facility at the Children’s Hospital of Philadelphia. To test whether HSPC-like cells are able to initiate leukemia, we sorted CD19+ blasts and HSPC like cells (Lin-, CD19-, CD34+, CD63+) from
peripheral blood of infant ALL patient and transplanted them into busulfan-conditioned NSG mice via tail vein injection. To test whether myeloid primed blasts can give rise to AML, we sorted CD19+/CD33-, CD19+/CD33+, and CD19+/CD34-CD38- cells from infant ALL patient with lineage switch and transplanted $150 \times 10^3$ cells of each population into busulfan-conditioned NSG and NSGS mice via tail vein injection.

Mice were euthanized when they exhibited signs of disease, such as increased leukemic burden, weight loss, or hind limb paralysis. Following sacrifice, leukemic cells were harvested from the bone marrow and spleen, checked for human CD45, and subjected to further flow analysis. All experiments were performed under protocols approved by the Institutional Animal Care and Use Committee (IACUC).

**scRNA-Seq data processing**

scRNA-Seq data for each patient were first processed using cellranger (v3.1.0) followed by additional processing using customized R scripts. Specifically, we filtered out cells with fewer than 1500 UMIs or greater than 50000 UMIs. Cells with more than 15% of UMI from mitochondrial genes were removed for downstream analyses. We further removed red blood cells with UMI for $HBB$ gene greater than or equal to 3.

DoubletFinder$^2$ was used to remove cell doublets with 5% of expected rate of doublets. The filtered gene-by-cell count matrix was used for downstream analyses.

**Integrated analysis of scRNA-Seq from 18 infant KMT2A-r patients**

The 18 filtered gene-by-cell count matrices were combined into a bigger gene-by-cell count matrix and converted to a Seurat object, with the fraction of mitochondrial UMI
and sample information added as the metadata. We computed cell cycle S phase and G/M2 phase scores using the CellCycleScoring function in Seurat, with cell cycle genes downloaded from https://satijalab.org/seurat/v3.2/cell_cycle_vignette.html. We additionally computed a heat shock gene signature score using the AddModuleScore function in Seurat using genes associated with heat shock proteins from HGNC (https://www.genenames.org/data/genegroup#!/group/#!/group/582). Next, we normalized the data using the NormalizeData function in Seurat. We selected the top 3,000 variably expressed genes (VEGs) using the FindVariableFeatures function in Seurat with the default setting. The VEGs were further scaled by the ScaleData function in Seurat. We also regressed out confounding factors including the fraction of mitochondrial UMIs, the total number of UMIs per cell, S phase score, G2/M phase score and heat shock score. We then performed principal component analysis (PCA) using the RunPCA function in Seurat with 50 principal components (PCs). The data were clustered using the FindNeighbors and FindClusters functions, with the resolution=0.4. To reduce batch effect, we re-selected VEGs across the clusters. To do so, we summarized the count matrix into a gene-by-cluster count matrix and normalized it using the cpm function in the edgeR package with the parameter prior.count = 1. The top 3,000 VEGs across the normalized gene-by-cluster count matrix were used. Data were scaled, regressed out for confounding factors followed by PCA and clustering analyses. VEGs that were expressed in fewer than 1% of cells were removed from the analysis. This process of re-selecting VEGs was repeated once and the resulting top 50 PCs were used in Uniform Manifold Approximation and Projection (UMAP) using the RunUMAP function in Seurat for visualization with the default setting.
Construction of normal hematopoietic development trajectory using scRNA-Seq data from healthy donors

The 27 filtered (9 donors each with 3 sorted samples) gene-by-cell count matrices were combined into a bigger gene-by-cell count matrix. To prevent sorted progenitor cells from dominating the cell composition, we downsampled the sorted cells to 12,000, about half of the live cells. The combined matrix was converted to a Seurat object, with the fraction of mitochondrial UMI and sample information added as the metadata. We next applied the same procedure as for integrating patient scRNA-Seq data, except using top 1,500 VEGs, the first 20 PCs, and only regressing out confounding factors of the total UMI count per cell and the fraction of mitochondrial UMIs per cell.

Cell type annotation for scRNA-Seq data

The cells of integrated patient data were annotated into Blasts, Monocytes, Mature B, T and NK cells based on marker gene expression. The normal hematopoietic trajectory built using healthy donor samples was further annotated into 17 cell types. The detailed cell types and corresponding marker genes used are summarized in Supplemental Table 2.

Projection of patient scRNA-Seq data onto the normal developmental trajectory

To project the patient cells onto the normal reference trajectory, we needed to transform the raw patient data as we did for constructing the normal developmental trajectory. Specifically, for each of the VEGs used for constructing the normal trajectory, we
extracted the linear regression model and saved mean and standard deviation of the regression residuals. Then, for each corresponding gene in patient data, we subtracted the predicted score using the linear regression model and scaled the residuals with the corresponding mean and standard deviation. This transformed gene-by-cell matrix was then multiplied by the PCA feature loading matrix from normal trajectory to calculate the projected PC scores for each patient cell. Finally, the \textit{umap\_transfer} function from the \textit{uwot} R package was used to generate the projected patient UMAP coordinates. The projected cell type of each patient cell is defined as the cell type of the closest (in Euclidean distance) healthy donor cell in the UMAP embedding.

\textbf{Identification of differentially expressed genes}

For each comparison, we identified the differential expressed genes between two groups using the \textit{FindMarkers} function in Seurat. We performed LR test on the raw counts and regressed out the total UMI count per cell as a confounding factor.

\textbf{scATAC-Seq data processing}

scATAC-Seq data for each sample was first demultiplexed using the \texttt{mkfastq} function of \texttt{cellranger-atac} (v.1.1.0) tool. The fastq files were then processed using the scATAC-pro 3 package with the default settings for most modules. Specifically, the cell barcode was written to the name of each read in the fastq files using the \texttt{demplx\_fastq} module and then the reads were aligned to the hg38 genome assembly using the \texttt{mapping} module with bwa as the sequence aligner. Peaks were called using the \texttt{call\_peak} module using the ‘COMBINED’ option. Peaks were called using MACS2 for each cell cluster. Peaks
identified from different clusters were further merged if they were within 200bp of each other. With the called peaks, we then constructed the raw peak-by-barcode count matrix. Using the call_cell module, barcodes with more than 3,000 total fragments and the fraction of reads in peaks (FRiP) greater than 50% were identified as cells. The peak-by-cell count matrix was used for downstream analyses.

**Integrated analysis of scATAC-Seq from 18 infant KMT2A-r ALL patients**

To integrate data from all patients, we first merged peaks called from all patients using the mergePeaks module of scATAC-pro. With this union set of peaks, we reconstructed the peak-by-cell count matrix for each patient using the reConstMtx module in scATAC-pro. The 18 peak-by-cell matrices were combined into a bigger peak-by-cell matrix and converted to a Seurat object. The data were next normalized using the TF-IDF normalization in scATAC-pro. Top 10,000 variably accessible peaks (VAPs) were identified using the FindVariableFeatures function in Seurat. We further filtered peaks that were accessible in fewer than 1% of the cells. TF-IDF normalization was then performed on the count matrix with only the VAPs. The PCA was performed on the VAPs and the cells were clustered using the first 50 PCs and the FindClusters function in Seurat with the resolution=0.4. To reduce batch effect, we reselect the VAPs across the clusters. To do so, we summarized the count matrix into a peak-by-cluster count matrix and normalized it using the cpm function in the edgeR package with parameter prior.count = 1. The top 10,000 VAPs across the normalized peak-by-cluster count matrix were selected and scaled, followed by PCA and clustering analyses. This process of re-selecting VAPs was repeated twice and the resulting first 50 PCs were
used in Uniform Manifold Approximation and Projection (UMAP) using the RunUMAP function in Seurat for visualization with the default setting.

**Construction of normal hematopoietic development trajectory using scATAC-Seq data from healthy donors**

Similar to integrating data from 18 patients, we first merged the peaks from all healthy donors using the mergePeaks module in scATAC-pro and reconstructed the peak-by-cell count matrix using combined peaks and the reConstMtx module in scATAC-pro. The 21 (7 healthy donors each with 3 sorted samples) peak-by-cell count matrices were then combined into a bigger peak-by-cell count matrix and converted to a Seurat object. We next applied the same procedure as integrating patient scATAC-Seq data, except using the first 30 PCs.

**Projection of patient scATAC-Seq data onto the normal developmental trajectory**

For each patient, we first reconstructed the peak-by-cell matrix using the VAPs used for constructing normal reference trajectory, using the reConstMtx module in scATAC-pro. The reconstructed matrix for each patient was normalized using the TF-IDF module in scATAC-pro. For each peak, we scaled the data using the mean and standard deviation calculated based on that peak in the healthy donor data. The transformed peak-by-cell matrix was then multiplied with the PCA feature loading matrix from normal reference to calculate the projected PC scores for each of the patient cells. Finally, the umap_transfer function from uwot R package was used to generate the projected patient UMAP coordinates. The projected cell type of each patient cell is defined as the
cell type of the closest (in Euclidean distance) normal control cell in the UMAP embedding.

**Cell type annotation for scATAC-Seq data**

The cells in the integrated patient data were annotated as Blasts, Monocytes, Mature B, T/NK cells based on TF enrichment score and gene activity score of marker genes. The TF enrichment score was computed using *chromVAR* and the gene activity score was computed by summarizing fragments overlapped with the gene body or gene promoter (+/- 2kb around the TSS). The normal scATAC-Seq data was annotated for cell types by label transfer from normal scRNA-seq data using Seurat with the default setting.

**Identification of transcription factor motifs with differential genome-wide chromatin accessibility**

We performed differential motif enrichment analysis based on the following comparisons: 1) between HSPC-like blasts and normal HSPCs; 2) between blasts arrested at a given development stage and corresponding normal cells at that stage; 3) between blasts of the same developmental stage in younger and older patients; 4) between blasts of the B-cell lineage and the myeloid lineage. For each cell group, we first used the chromVar algorithm to identify TF motifs that had enriched genome-wide chromatin accessibility in that cell group. We then tested differential enrichment between two cell groups using Wilcoxon test and the chromVar deviation scores. P-values were adjusted for multiple testing using the Benjamini-Hochberg method. Significant TF motifs were removed if the corresponding TF genes were expressed in
fewer than 10% (20% for the comparison between HSPC-like blast and normal HSPC) of cells in the group.

**snmC-Seq2 data processing**

Sequencing reads were demultiplexed using the 6 bp indices in the fastq files, followed by adapter trimming by cutadapt with the parameter settings: `-f fastq -q 20 -u 10 -m 30 -a AGATCGGAAGAGCACACGTCTGAAC`. Demultiplexed and trimmed reads were aligned using Bismark \(^4\) with the default settings.

Aligned reads were post-processed to remove ambiguous reads with mapping quality < 10, clonal reads, and reads containing three consecutive non-CpG methylation (indicating failure of bisulfite conversion). Aligned and filtered reads were counted using the countBam function and mapping rate was computed using the pileup function, both of which were implemented in the Rsamtools package \(^5\). For downstream analysis, we only used cells that had 200,000 filtered aligned reads and 20% mapping rate. We computed the methylation level of a region by calculating the ratio of methylated CpG sites to all methylation calls in that region. At least 10 overlapping CpG methylation calls were required to estimate the methylation levels of a region for each cell.

To integrate snmC-Seq and scATAC-Seq data, we first constructed a gene-by-cell matrix that consisted of the methylation level of the gene body and 2kb flanking region for each gene. For the missing value of a gene in a cell, we imputed it as the mean of the observed methylation level of the rest of genes in the same cell. Next, we co-embedded snmC-Seq data with the scATAC-Seq data of CD19+ blast population from the same 11 patients using Seurat. We first created a Seurat object that contained
1-log2(mtx+0.5), where the mtx was the imputed gene-by-cell methylation rate matrix. We then used `FindVariableFeatures`, `ScaleData` and `RunPCA` functions in Seurat to process the snmC-Seq data. For scATAC-Seq data, we constructed a gene activity-by-cell matrix using accessibility of gene body and 2kb flanking region and constructed a Seurat ‘assay’ called ACTIVITY in the original scATAC-Seq Seurat object. This assay was subjected to normalization, scaling and dimensionality reduction using Seurat. The snmC Seurat object was then co-embedded with the scATAC-Seq Seurat object treating the scATAC-Seq data as the reference. A cell in snmC-Seq data was annotated based on its nearest cell in the scATAC-Seq data in the UMAP space.

For each developmental stage, we performed differential methylation analysis using MethPipe and aggregated cells from younger and older patients as the input. We then calculated the methylation rate for each CpG, differentially methylated CpGs (using p-value=0.05 as the cutoff) and differentially methylated regions (DMRs) using `methcounts`, `hmr`, `methdiff` and `dmr` functions in MethPipe, respectively. We further filtered DMRs to have at least 4 CpGs and at least 2 differentially methylated CpGs in the DMR. We computed the methylation entropy for each CpG using the `methentropy` function of MethPipe and the default sliding window size covering 4 CpGs. Because low sequencing coverage at a locus can cause inaccurate estimates of methylation entropy, we only calculated methylation entropy using CpGs having more than 100 reads in the sliding window.

**Analysis of Nanopore sequencing data**
Oxford nanopore sequencing reads were demultiplexed using AmpBinner (https://github.com/WGLab/AmpBinner). AmpBinner first identified the p5 and read1 sequence of each read. The 20 bp sequence downstream of the read1 is expected to contain the barcode sequence and was extracted from each read and aligned to all possible cell barcodes in the sample. The alignment was performed using minimap2 with optimized parameters for short sequences (-k 3 -w 2 -n 1 -m 10 -s 40). The number of base edits (including mismatches, insertions, and deletions) were counted from the ‘CIGAR’ string. A read was confidently assigned to a cell barcode if the number of base edits was fewer than 3 and the second-best barcode had 3 or more base edits.

Demultiplexed sequencing reads were aligned to the human reference genome GRCh38 using minimap2 with parameters allowing detection of splicing isoforms (-ax splice -uf -k14). Fusion reads were called from the bam file using LongGF (https://github.com/WGLab/longgf) and AlignQC. A read was identified as a fusion read if it was aligned to both the primer region and the gene body of the expected fusion partner. The read was identified as wild-type read if it was aligned to both the primer region and exon 15 of the KMT2A gene. For downstream analysis, only reads consistently called by both methods were considered fusion or wild-type reads.

**Prediction of enhancer-promoter interactions**

Enhancer-Promoter (EP) interactions were predicted using a regression-based method described in 9. A linear regression was conducted for each gene, with the gene expression in each cell as the dependent variable, and the normalized accessibility of the peaks within +/- 500kb of the gene promoter as the independent variables. If the
regression coefficient of a peak is greater than 0.1 and the Benjamini-Hochberg adjusted p-value is smaller than 0.05, the EP pair was called significant. Cell-type-specific EP pairs were called if the enhancer peak had higher chromatin accessible (log$_2$FC $>$ 1) in the given cell type than the compared cell type(s).

**Construction of transcriptional regulatory network**

The transcriptional regulatory network was constructed using differentially expressed genes and enriched TFs in HSPC-like blast compared to normal HSPC. A gene was defined as a target of a TF if there was a predicted HSPC-like-blast specific enhancer-promoter interaction and a TF motif hit at the enhancer. The color of the gene node is proportional to the -log10(FDR) in the DEG analysis, and the size of the TF node is proportional to the -log10(FDR) in the TF enrichment analysis. The thickness of an edge is proportional to the -log10(FDR) of the linear regression coefficient in the EP prediction analysis.

**NK cell tumor killing potential induced by IFNG signaling**

We computed a NK cell cytotoxicity score induced by IFNG signaling in the CD19+ blasts and HSPC-like cells based on the expression levels of the two classes of genes, cell adhesion genes (*ICAM1, VCAM1, NCAM1*, and *CD44*) and MHC class I genes (*HLA-A, HLA-B, HLA-C, HLA-E, HLA-F, HLA-G* and *B2M*). Specifically, for each NK cell, the cytotoxicity score is the sum of normalized expression of cell adhesion genes minus the sum of normalized expression of MHC class I genes.
Analysis of ligand-receptor interactions of signaling pathways

We identified ligand-receptor (L-R) interactions between blasts and cytotoxic lymphocytes using CellphoneDB 10 and Kumar’s method 11. Normalized gene-by-cell expression matrix (downsampled to 30K cells) and cell type annotation were used as the input to both methods. L-R interactions predicted by CellphoneDB with a p-value < 0.01 were considered significant. For Kumar’s method, we used the top 15% of predictions for each cell-type pair as significant interactions. Interactions found by both methods were considered as final predictions. For downstream analysis of comparing HSPC-like cells and CD19+ blasts, among significant L-R pairs, we only considered those in which the ligand/receptor gene was differentially expressed (Wilcoxon test, FDR < 0.05 ) in one of the two blast types. We computed an overall lymphocyte cytotoxicity score for NK or CD8 cells based on the overall expression of activating and suppressive L-R interactions. Specifically, the score was computed as the sum of mean gene expression of activating L-R pairs minus the sum of mean gene expression of suppressive L-R pairs.

Pathway enrichment analysis

To compute the pathway enrichment score for each cell, we used a modified approach based on AUCell. Instead of using all genes for ranking, we used an approach called IFF selection 9 to remove housekeeping genes and keep cell-type specific genes. We then performed stage-wise Student’s t-test on the enrichment score for comparison between healthy donor HSPCs and patient HSPC-like blasts. Enriched pathways were identified using q-value <= 0.01.
GO enrichment analysis

We used Metascape \(^{12}\) to perform GO enrichment analysis for differentially expressed genes using the default setting.

Random forest classifier for annotating patient cells

We first splitted cells from healthy donor samples into a training and a test set. For each cell type, two thirds of the cells were used for the training set and the rest for the test set. A random forest classifier was trained using the *randomForest* R package\(^{13}\) and all genes that are expressed in more than 1% of the cells. Applying the trained classifier to the test set, it achieved 92.9% in accuracy. We then selected 1000 genes with the highest importance scores to train another random forest classifier. The prediction accuracy for the test set was increased to 94.7% using the second classifier, and the 5-fold cross validation accuracy was 93.4%. Therefore, the prediction for patient cells was performed using the second classifier.
Supplemental Tables

**Supplemental Table 1.** Patient and healthy donor clinical and biospecimen information.

| Patient ID | Diagnosis | Fusion partner | Gender | Age at diagnosis | Stage of diagnosis | Therapy | Specimen type | HTAN ID |
|------------|-----------|----------------|--------|------------------|--------------------|---------|---------------|---------|
| PAYWJZ     | BCP ALL   | KMT2A-MLLT1    | female | 5 d              | initial diagnosis  | AALL15P1 | PB            | HTA4_1  |
| PAZGKI     | BCP ALL   | KMT2A-MLLT3    | female | 1 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_2  |
| PAYUZM     | BCP ALL   | KMT2A-MLLT10   | female | 1 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_3  |
| 1154       | BCP ALL   | KMT2A-MLLT3    | female | 2 mo             | initial diagnosis  | AALL0631 | Pheresate     | HTA4_4  |
| PAYKGI     | BCP ALL   | KMT2A-AFF1     | female | 2 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_5  |
| PAZBSZ     | BCP ALL   | KMT2A-MLLT1    | female | 3 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_6  |
| PAYWKI     | BCP ALL   | KMT2A-AFF1     | male   | 3 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_7  |
| PAYSBA     | BCP ALL   | KMT2A-AFF1     | male   | 4 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_8  |
| PAZBLA     | BCP ALL   | KMT2A-MLLT10   | male   | 4 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_9  |
| PAYZLC     | BCP ALL   | KMT2A-EPS15    | female | 4 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_10 |
| PAYLNH     | BCP ALL   | KMT2A-EPS15    | male   | 6 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_11 |
| PAYUZJ     | BCP ALL   | KMT2A-MLLT1    | female | 7 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_12 |
| PAYYBG     | BCP ALL   | KMT2A-MLLT1    | female | 8 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_13 |
| PAZFPH     | BCP ALL   | KMT2A-MLLT1    | male   | 9 mo             | initial diagnosis  | AALL15P1 | PB            | HTA4_14 |
| PAYYNY     | BCP ALL   | KMT2A-AFF1     | male   | 10 mo            | initial diagnosis  | AALL15P1 | PB            | HTA4_15 |
| PAZBGV     | BCP ALL   | KMT2A-MLLT3    | male   | 10 mo            | initial diagnosis  | AALL15P1 | PB            | HTA4_16 |
| PAYZVY     | BCP ALL   | KMT2A-MLLT3    | male   | 10 mo            | initial diagnosis  | AALL15P1 | PB            | HTA4_17 |
| PAYZWN     | BCP ALL   | KMT2A-AFF1     | female | 11 mo            | initial diagnosis  | AALL15P1 | PB            | HTA4_18 |

| Patient ID | Diagnosis | Fusion partner | Gender | Age at diagnosis | Stage of diagnosis | Therapy | Specimen type | HTAN ID |
|------------|-----------|----------------|--------|------------------|--------------------|---------|---------------|---------|
| 2639       | ALAL      | KMT2A-MLLT4    | male   | 147 mo           | initial diagnosis  | AALL1131 | -like         | HTA4_22 |

**Infant KMT2A-rearranged leukemia**

**Other leukemia**
| Donor ID | Cell population | gender | Age      | Specimen type | HTAN ID |
|---------|-----------------|--------|----------|---------------|---------|
| 2117    | DAPI- Live      | NA     | 1-5 yr   | BM            | HTA4_26 |
|         | CD34+           |        |          |               |         |
|         | Lin-CD34+CD38-  |        |          |               |         |
| 2471    | DAPI- Live      | NA     | 1 yr     | BM            | HTA4_27 |
|         | CD34+           |        |          |               |         |
|         | Lin-CD34+CD38-  |        |          |               |         |
| 2153    | DAPI- Live      | NA     | >20 yr   | BM            | HTA4_28 |
|         | CD34+           |        |          |               |         |
|         | Lin-CD34+CD38-  |        |          |               |         |
| 2111    | DAPI- Live      | NA     | 15-20 yr | BM            | HTA4_29 |
|         | CD34+           |        |          |               |         |
|         | Lin-CD34+CD38-  |        |          |               |         |
Supplemental Table 2. Marker genes used for annotating cell types in the single-cell data.

| Cell type       | Marker genes                                      |
|-----------------|---------------------------------------------------|
| CD19+ blast     | CD19, HOXA9, MEIS1, MEF2C                         |
| HSPC            | CD34, HGMA2                                       |
| LMPP            | CD34, CD38                                        |
| MEP             | CD34, GATA1                                       |
| GMP             | MPO, ELANE                                        |
| CLP             | CD38, IGLL1                                       |
| Pre-pro-B       | CD38, IGLL1, MME, DNTT                           |
| Pro-B           | CD38, IGLL1, MME, DNTT, CD19                     |
| Pre-B           | CD38, IGLL1, MME, CD19                           |
| Immature B      | CD19, MME, MS4A1                                  |
| Mature B        | CD19, MS4A1                                       |
| Plasma B        | CD27                                              |
| pDC             | IL3RA, IRF8                                       |
| cDC             | CD1C, FCER1A                                      |
| T cell          | CD3D, CD3E, CD3G                                 |
| NK cell         | GNLY                                              |

Supplemental Table 3. List of antibodies used in this study.

| Antibody   | Clone   | Lot #  | Cat #  | Fluorophore |
|------------|---------|--------|--------|-------------|
| anti-CD19  | J3-119  | 200131 | IM3628U| PC7         |
| anti-CD45  | 2D1     | 9217025| 641408 | APC-H7      |
| anti-CD3   | HIT3a   | 8162560| 555340 | PE          |
| anti-CD34  | 581     | 200060 | IM1870U| FITC        |
| anti-CD38  | HIT2 (RUO) | 8353764| 560980 | APC         |
| anti-CD8   | B9.11   | 76     | IM0452U| PE          |
| anti-CD4   | 13B8.2  | 108    | IM0449U| PE          |
| anti-CD20  | B9E9    | 49     | IM1451U| PE          |
| anti-CD56  | NCAM16.2| 37243  | 340363 | PE          |
| anti-CD14  | RMO52   | 69     | IM0650U| PE          |
| anti-CD11b | ICRF44  | 7017973| 555388 | PE          |
| anti-CD63  | H5C6    | B297329| 353007 | APC         |
**Supplemental Table 4.** List of primers used in this study.

| Sequence | Purpose |
|----------|---------|
| KMT2A_exon7 | /5Biosg/GTGACTGGAGTTTACAGACGTGCTCTTCCGATCTTACCCCATACAGCAAGAGAG | scTLR-Seq first PCR |
| Universal 3' Primer | CTACACGAGCTCTTTCCGATCT | |
| P5 PCR2 | AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAACAGACGCTC | scTLR-Seq second PCR |
| P7 PCR2 | CAAGCAGAAGACGGCATAACAGAGATGAGTGAACGGAGGTTCAACAGGTC | |
| P5L_AD002_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTGATTGT(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | snmC-Seq2 |
| P5L_AD006_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTGCTGCAAT(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |
| P5L_AD008_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTAcTCTTGA(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |
| P5L_AD010_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTTAGCTT(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |
| P5L_AD001_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTACACG(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |
| P5L_AD004_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTTGACCA(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |
| P5L_AD007_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTCACTACG(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |
| P5L_AD012_H | /5SpC3/TTCCCTACACGACGCTCTTTCCGATCTTCTGTA(H1:33340033)(H1)(H1)(H1)(H1)(H1)(H1)(H1)(H1) | |

**Supplemental Table 5.** List of differentially expressed genes (abs(log2FC) > 0.5 & FDR < 0.05) of blasts arrested across B-cell developmental stages between younger (< 6 months old) and older patients (> 6 months old). DEGs are organized into clusters identified by hierarchical clustering.

**Supplemental Table 6.** List of TF motifs with differential accessibility in blasts arrested across B-cell developmental stages between younger and older patients.
**Supplemental Table 7.** List of differentially methylated regions in blasts across B-cell developmental stages between younger and older patients.

**Supplemental Table 8.** List of differentially expressed genes between five projected subpopulations of HSPC-like cells and their corresponding normal populations.

**Supplemental Table 9.** Full transcriptional regulatory network for HSC/MPP-like blasts vs normal HSC/MPP comparison.

**Supplemental Table 10.** List of predicted ligand-receptor pairs between HSPC-like cells/CD19+ blasts and three types of cytotoxic lymphocytes.
Supplemental Figures

Supplemental Figure 1. Patient clinic data and quality assessment of single-cell omics data (related to Figure 1).

A) Summary of patient clinical data, including leukemia subtype, fusion partner, age, and gender.

B) Representative FACS plot for sorting of whole bone marrow hematopoietic compartment (CD45+), lineage- CD34+ progenitors, and lineage-, CD34+, CD38- stem cell enriched population.

C) Summary of scRNA-Seq data after QC and removal of low-quality cells. Top, number of cells sequenced per sample; Bottom, number of genes identified per cell. Red horizontal line, average value across all samples.

D) Summary of scATAC-Seq data after QC and removal of low-quality cells. Top, number of cells sequenced per sample; Middle, number of uniquely mapped fragments per cell; Bottom, Fraction of accessible fragment reads located in ATAC-Seq peaks (FRIP). Red horizontal line, average value across all samples.

E) QC metrics of snmC-Seq2 data. Top left, number of total reads sequenced per cell by sample; top right, percentage of uniquely aligned reads per cell by sample; middle left, percentage of genome coverage per cell by sample; middle right, percentage of methylated CpG sites per cell by sample; bottom left, bisulfite conversion rate per cell by sample; bottom right, number of total CpG sites recovered per cell by sample. Red horizontal lines, average values across all samples.
F) UMAP of all scRNA-Seq cells (left panel) and all scATAC-Seq cells (right panel) of 18 infant ALL samples, colored by samples. Assigned cell populations were labeled on the UMAPs.
Supplemental Figure 2. Normal hematopoietic trajectory and projection evaluation (related to Figure 2).

A) Genome browser tracks for aggregated ATAC-Seq signals at representative cell-type-specific marker genes in healthy donors.

B) Motif accessibility enrichment computed by chromVAR for known lineage specific TFs, including myeloid (CEBPA), B lineage (EBF1), T lineage (EOMES), and hematopoietic stem and progenitor cells (GATA1).

C) Scatter plot for fraction of CD19+ blasts projected to cell types in the normal hematopoiesis developmental trajectory. X-axis, projected fraction in scRNA-Seq data. Y-axis, projected fraction in scATAC-Seq data.

D) UMAP of normal hematopoietic trajectory generated using scRNA-Seq data from two younger donors (one infant and one young child aged between 1 and 5 years).

E) Scatter plot for the composition of projected cell types using the reference built on all healthy donors versus that using younger donors in panel D.

F) Scatter plot for the composition of projected cell types versus that based on prediction of cell types using a random forest classifier. Pearson correlation coefficients are shown on panels E and F.

G) Dot plots of age vs entropy and age vs percentage of HSPC-like cells (each dot represents a patient sample, related to Figure 2G).
Supplemental Figure 3. Transcriptomic and epigenomic signatures of blasts in younger patients (related to Figure 3).

A) Enriched pathways among DEGs in additional clusters in Figure 3A.

B) Heatmap of differentially methylated regions (DMRs) overlapping with promoters of DEGs in blasts across B-cell developmental stages between younger and older patients. DMRs were identified using MethPipe (p-value < 0.05 & total CpG sites > 4 & differential CpG sites > 2). Red, up-regulated DEGs overlapped with hypomethylated DMRs in younger patients; Blue, down-regulated DEGs with hypermethylated DMRs in younger patients.

C) Example DEGs that overlapped with DMRs in pre-pro-B-like blasts. Top, NR3C1; Middle, KLF9; Bottom, RUNX2. Left panels, aggregated scATAC-Seq signals and methylation rates of blasts in younger and older patients. Right panels, normalized scRNA-Seq expression values of blasts in younger and older patients. Promoters are highlighted in yellow.

D) Western blots showing knockout of NR3C1 and KLF9 using two different sgRNAs per gene.

E) Viability of wild type and NR3C1/KLF9 KO cell lines after dexamethasone and prednisone treatment with different doses. Error bar, standard deviation of two biological replicates. P-values by t-test for KO versus control are shown: *, p < 0.05; **, p < 0.01; ***, p < 0.001, n.s., not significant.

F) RNA expression levels of NR3C1 and KLF9 in three KMT2A-r cell lines. Data was downloaded from Cancer Cell Line Encyclopedia (CCLE).
Supplemental Figure 4. A rare progenitor-like blast population in younger patients (related to Figure 4).

A) Genome browser tracks for aggregated scATAC-Seq signals at representative HSPC-specific (\textit{GATA2, CD34, HMGA2}) and B lineage-specific (\textit{PAX5, VPREB1, DNTT}) marker genes.

B) Representative FACS plot for sorting of HSPC-like cells.

C) Percentages of CD34+, CD19+, CD34+CD63+ cells in the engrafted leukemia cells in BM. P-values by t-test for mice transplanted with HSPC-like cells (\(n = 8\)) versus CD19+ blasts (\(n = 8\)) are shown: *, \(p < 0.05\); n.s., not significant.

D) The cell type composition of HSPC-like cells based on projection.

E) Volcano plot of DEGs between HSC/MPP-like and HSC/MPP cells. Genes with absolute log2 fold change > 0.5 and FDR < 0.05 are defined as DEGs and top 20 DEGs (based on fold change) in each direction are highlighted.

F) Enriched pathways based on GO analysis in each of the five major subpopulations (identified in panel D) of HSPC-like cells, compared with its normal counterpart.
Supplemental Figure 5. Normal immune cell types in 18 KMT2A-r patients and blast-NK cell co-culture experiment (related to Figure 5).

A) UMAP of scATAC-Seq data for normal immune cells. Number of sequenced cells is indicated.

B) Frequency of normal immune cell types in younger and older patients.

C) Scatter plot for the fraction of NK T cells expressing interferon-γ vs the fraction of HSPC-like cells expressing IFNGR2. Each dot represents a patient sample that has both HSPC-like cells and NKT cells. Pearson correlation coefficient is shown.

D) Representative flow plots for dead target cells in the blast-NK cell co-culture experiment. Top row, analysis of dead HSPC-like cells; bottom row, analysis of dead CD19+ blasts. Fractions of dead cells were indicated in the plots.
Supplemental Figure 6. Characterization of longitudinal samples of two patients who had lineage switch after relapse and immunotherapy (related to Figure 6).

A) Genome browser tracks for aggregated scATAC-Seq signals at representative B lineage (CD19, VPREB1, CD79A) and myeloid lineage-specific (CD33, LYZ, CEBPA) marker genes.

B) FACS plots showing CD19 and CD33 expression in bone marrow blasts of patient 1 before (top, sample ID 2184) and after lineage switch (bottom, sample ID 2263).

C) FACS plots showing CD19 expression in patient 2 at four time points. Top left, initial diagnosis ALL (blasts CD19+ on clinical flow, sample ID 1979); Top right, first relapsed ALL after chemotherapy and CD19 directed immunotherapy (blasts CD19 negative on clinical flow, sample ID 2524); Bottom left, second relapsed ALL (blasts CD19 negative on clinical flow, sample ID 2578); Bottom right, lineage switch to AML after CD22 targeted immunotherapy (AML, sample ID 2741).

D) & E) UMAPs of scRNA-Seq (D) and scATAC-Seq (E) data for a pediatric ALL KMT2A-r patient with leukemic samples at four time points, initial diagnosis, relapse 1 post chemotherapy and CAR19 therapy, relapse 2 post inotuzumab therapy, and after lineage switch to AML post CAR22 therapy. Left panel, UMAP of longitudinal samples, colored by assigned cell types. Total numbers of sequenced cells are indicated. Right panel, projection of patient cells to the normal hematopoietic trajectory. Grey dots, cells from healthy donors; colored dots, patient cells.

F) Fraction of B-, myeloid-lineage, and LMPP blasts at four time points. Top panel, fraction based on scRNA-Seq data; Bottom panel, fraction based on scATAC-Seq data.

G) Violin plots for representative signature genes in blasts from the four time points.
H) Transplantation experiment using NSG and NSGS mice. The following subpopulations were isolated from the initial diagnostic sample of patient #1979: immature (CD19+CD34+CD38-), Ly (pure lymphoid, CD19+CD34-CD33-), Ly/My (Lymphoid/Myeloid co-expression, CD19+CD33+). Shown is the % CD33 expression in the bone marrow (upper panel) and spleen (lower panel) of NSG (blue) or NSGS mice (red, NSGS mice are NSG expressing human IL3 GM-CSF and SCF) mice transplanted with the respective subpopulations. P-values by t-test are shown: *, p < 0.05.
Supplemental Figure 7. LMPP-like blasts in KMT2A-r leukemia patients (related to Figure 6).

A) Fraction of LMPP-like blasts in KMT2A-r leukemia samples based on scRNA-Seq data.

B) Fraction of LMPP-like blasts in KMT2A-r leukemia samples based on scATAC-Seq data.

C) Scatter plot for frequencies of M-lineage blasts and LMPP-like blasts in KMT2A-r leukemia samples. Spearman correlation is shown.

D) Enriched pathways among differentially expressed genes between normal LMPPs from healthy donors and LMPP-like blasts in patient samples post lineage switch.
References

1. Luo C, Rivkin A, Zhou J, et al. Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* 2018;9(1):3824.
2. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* 2019;8(4):329-337.e4.
3. Yu W, Uzun Y, Zhu Q, Chen C, Tan K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol.* 2020;21(1):94.
4. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics.* 2011;27(11):1571–1572.
5. Morgan M, Pages H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. *R package version.* 2016;1(0):677–689.
6. Song Q, Decato B, Hong EE, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One.* 2013;8(12):e81148.
7. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–3100.
8. Weirather JL, de Cesare M, Wang Y, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research.* 2017;6:100.
9. Zhu Q, Gao P, Tober J, et al. Developmental trajectory of prehematopoietic stem cell formation from endothelium. *Blood.* 2020;136(7):845–856.
10. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* 2020;15(4):1484–1506.
11. Kumar MP, Du J, Lagoudas G, et al. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep.* 2018;25(6):1458-1468 e4.
12. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 2019;10(1):1523.
13. Liaw A, Wiener M, Others. Classification and regression by randomForest. *R news.* 2002;2(3):18–22.