SUPPLEMENTAL DATA

SUPPLEMENTAL METHODS

Human biospecimens

Peripheral blood or bone marrow samples were collected from patients with: 1) AML enrolled on the Children's Hospital of Philadelphia (CHOP) Institutional Review Board (IRB) Protocol 10-007767 (CHOP Center for Childhood Cancer Research (CCCR) Biorepository); 2) B-ALL enrolled on the Children's Oncology Group (COG) clinical trial AALL15P1 (NCT02828358); 3) T-ALL enrolled on the Children's Oncology Group (COG) clinical trial AALL0434 (NCT04408005) and AALL1231 (NCT02112916). All samples were obtained with parental informed consent according to the Declaration of Helsinki with IRB approval from all participating centers. Detailed clinical data is listed in **Table S2**.

Cell preparation for single-cell assays and PDX injection

Peripheral blood or bone marrow samples were quickly thawed at 37°C, treated with 1/10 volume of 1mg/mL DNase I (Sigma, Cat #: D4513) for 90 sec at 37°C, washed once with 10mL IMDM + 2% FBS. Samples were then treated with DNase I (1/10 volume) and washed one more time with 10mL IMDM + 2% FBS and resuspended in PBS + 1% BSA buffer for FACS sorting. Primary patient samples were subjected to DAPI- live cell sorting. Healthy bone marrow samples were subjected to DAPI-, DAPI-Lin-CD34+, and DAPI-Lin-CD34+CD38- sorting to enrich enough stem/progenitor cells. After sorting, cells were divided into two aliquots for scRNA-Seq and scATAC-Seq, respectively. For each patient sample, more than one million live cells were aliquoted for tail vein injection into NSG mice for ALL and NSGS mice for AML/MPAL, with the remaining cells stained with DAPI (Invitrogen, Cat #: D1306) and immediately subjected to FACS sorting (FACSAria Fusion, BD) for DAPI- live cells.

scRNA-Seq

For primary patient samples, sorted cells were resuspended in PBS + 0.04% BSA and counted using Countess II cell counter. ~17,000 cells per sample were loaded onto the 10x Genomics Chromium controller and processed with the Chromium NEXT GEM Single Cell 3' reagent kits V3.1. GEX libraries were constructed using 10x Genomics library preparation kit following manufacturer's instruction. For PDX samples, DAPI-hCD45+ cells were sorted and resuspended in PBS + 0.04% BSA and counted using Countess II cell counter. 500k~1M cells per sample were then stained with 10x Genomics CellPlex for multiplexing. After staining and washing, equal number (10k~20k) of cells was pooled together for each of six PDX mice and ~60k-115k cells per pool were loaded onto 10x Genomics Chromium X and processed with the Chromium NEXT GEM Single Cell 3' HT kits V3.1. GEX libraries were constructed using 10x Genomics library preparation kit following manufacturer's instruction. Library quality was checked using Agilent High Sensitivity DNA kit (Agilent, Cat #: 5067-4626) and Bioanalyzer 2100. Libraries were quantified using dsDNA High-Sensitivity (HS) assay kit (Invitrogen, Cat #: Q33231) on Qubit fluorometer and quantified using the qPCR-based KAPA quantification kit (Kapa Biosystems, Cat #: KK4844). Libraries were sequenced on an Illumina Nova-Seq 6000 with 28:8:0:87 paired-end format with 50k-100k read depth per cell.

scATAC-Seq

scATAC-Seq was performed using the protocol we previous described[1] with minor modifications. Briefly, 40k-100k sorted live cells were resuspended in 45uL lysis buffer and incubated 3 min on ice. Next, 50µL of prechilled wash buffer was added without pipetting and centrifuged immediately at 400g 5 min at 4°C. Cells were washed once with 45µL diluted nuclei buffer (10x Genomics). The nuclei pellet was then resuspended in 7µL prechilled diluted nuclei buffer, and nuclei

concentration was determined using a Countess II cell counter. 7,000-20,000 nuclei per sample were used for the transposition reaction and then loaded onto the 10x Genomics Chromium controller and processed with the Chromium NEXT GEM Single Cell ATAC Reagent Kit V1.1. Library quality was checked using Agilent High Sensitivity DNA Kit and Bioanalyzer 2100. Libraries were quantified using dsDNA High-Sensitivity (HS) assay kit 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 with 50k-100k read depth per cell.

Genotyping of Transcriptomes (GoT) library preparation and sequencing

Primers for variants for Genotyping of Transcriptomes are listed in **Table S4**. cDNA libraries were amplified and circularized following the protocol described previously with a few modifications [2]. In brief, 10ng of 10x Genomics barcoded full-length cDNA from each patient sample was amplified using primer set 1 and KAPA HiFi Ready mix (KAPA, Cat: KK2600). PCR1 was conducted using the following the program: 98°C for 3 minutes, 15 cycles of (98°C for 20 seconds, 65°C for 30 seconds, 72°C for 2 minutes), 72°C for 5 minutes, and held at 4°C. The PCR1 product was purified with 1.3x SPRI beads, followed by the first circularization reaction using Gibson assembly (NEB, Cat: E2611S). PCR2 was conducted as previously described using primer set 2, but with only 10 cycles. A second circularization reaction was then performed on the PCR2 product, and the final linearization reaction (PCR3) was carried out as previously described with 7 cycles. Illumina adapter and index were added via a final PCR using a generic forward PCR primer (P5_generic) and an RPI-x primer, following the program: 95°C for 3 minutes; 5 cycles of (98°C for 20 seconds, 67°C for 30 seconds, and 72°C for 30 seconds); 72°C for 5 minutes, and held at 4°C. The libraries were sequenced on an Illumina NextSeq using a 28:6:0:130 format with 10k-50k read depth per variant per cell.

Expansion of leukemic blasts in PDX model

NOD/SCID/IL2Rγ-/- (NSG for ALL) and NOD-SGM3 (NSGS for AML and MPAL) mice were purchased from the Jackson laboratory and maintained at the Animal Research Facility at the Children's Hospital of Philadelphia. Sorted HSPC-like and lineage-like blasts were transplanted into busulfan-conditioned mice via tail vein injection (**Supplemental Figure 4K-M**). For B-ALL case, 1,000 cells were injected into each mouse for both sorted populations. For T-ALL and MPAL, 100,000 cells were injected into each mouse for both sorted populations. 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 using flow cytometer, and subjected to further flow analysis and single-cell assays. All experiments were performed under protocols approved by the Institutional Animal Care and Use Committee (IACUC) of Children's Hospital of Philadelphia.

***In vitro* treatment experiment**

Drug response of ALL cells was evaluated using a mesenchymal stem cells (MSC) co-culture system and determined by an imaging-based assay, as described previously[3,4]. Briefly, hTERT-immortalized MSCs were first seeded one day prior to the assay, followed by plating leukemia cells with serial drug solution diluted in the AIM VTM serum free medium (Gibco, #12055083). HSPC-like derived PDX cells and lineage-like derived PDX cells from a single patient (PAVYVY, ETP-ALL) were used in this treatment experiment. Duplicates were included for each of the drug concentrations. After 96 hours of incubation at 37°C with 5% CO2, leukemia cells were subject to the cell viability assay, where drug-induced death was estimated by comparing leukemia cells treated with medium alone. Cell viability data were analyzed by Harmony high-content imaging and analysis software (version 4.9). The concentrations of drug lethal to 50% of the cells (LC50) were calculated by a four-parameter dose-response model. For cases in which even the lowest

drug concentration killed >50% of leukemia cells, LC50 was assigned as half of the minimum tested concentration. Conversely, for cases with >50% viability even at the highest drug concentration, LC50 values were assigned as twice of the highest tested concentration.

scRNA-Seq data pre-processing

For primary samples, scRNA-Seq data for each patient were first processed using cellranger (V3.1.0) followed by additional processing using customized R scripts with Seurat V4.0.5[5]. Specifically, we removed cells with fewer than 1,500 UMIs or greater than 50,000 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[6] was used to remove cell doublets with 5% of expected rate of doublets. For PDX CellPlex samples, data was first demultiplexed using cellranger (V6.1.2) with "multi" function. The samples were further cleaned and filtered by Demultiplex2 with default parameters[7]. We then removed cells using the same cutoffs as primary sample data. The filtered gene-by-cell count matrix was used for downstream analyses.

scATAC-Seq data pre-processing

scATAC-Seq data for individual sample was demultiplexed using the mkfastq function of cellranger-atac (v.1.1.0). Raw fastq data was then processed using the scATAC-pro package[8] with the default settings for most modules. First, peaks were called using MACS2 at bulk level for each sample with the following parameters: -q 0.01 -g hs --nomodel --extsize 200 --shift -100. Second, cells were called using the following cutoffs: 1) minimum uniquely mapped fragments per barcode: 3,000; 2) maximum uniquely mapped fragment per barcode: 50,000; 3) minimum fraction of reads in peaks per barcode: 0.4; and 4) maximum fraction of mitochondrial reads per barcode: 0.2. The peak-by-cell count matrix was generated for each sample and used for downstream analyses.

Integrated analysis of scRNA-Seq from 96 primary acute leukemia patients

The 96 high-quality gene-by-cell count matrices were combined into a bigger gene-by-cell count matrix and converted to a Seurat object (V4.0.5)[5] with the fraction of mitochondrial UMI and sample information added as 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/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 gene 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 of 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 total 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 downstream global UMAP visualization with the default setting.

Integrated analysis of scRNA-Seq from PDX samples
For primary-PDX pairs and sorted PDX analysis (**Supplemental Figure 5**), we combined the gene-by-cell count matrices and converted to a Seurat object (V4.0.5). The combined matrix was normalized and 2,000 VEGs were selected using the FindVariableFeatures function in Seurat with the default setting. VEGs that were expressed in fewer than 1% of total cells were removed from the analysis. Filtered VEGs were scaled by the ScaleData function in Seurat and 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 gene 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 of 0.4. Top 50 PCs were used in Uniform Manifold Approximation and Projection (UMAP) using the RunUMAP function in Seurat for visualization with the default setting.

Integrated analysis of scATAC-Seq from 96 primary acute leukemia patients
We first merged peaks called from all 96 patient samples to generate a master peak set of acute pediatric leukemias. 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 96 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 LSI was performed on the VAPs and the cells were clustered using the first 50 LSIs and the FindClusters function in Seurat with the resolution of 0.4. To reduce batch effect, we reselected 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 LSI and clustering analyses. This process of re-selecting VAPs was repeated twice and the resulting first 50 LSCs were used in Uniform Manifold Approximation and Projection (UMAP) using the RunUMAP function in Seurat for visualization with the default setting.

Construction of healthy reference atlas using scRNA-Seq and scATAC-Seq data
The healthy reference atlases were constructed using the pipeline described previously[1,9] with a few modifications, especially for scATAC-Seq to generate a more balanced trajectory among myeloid, T-, and B-lineages. In brief, 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 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 for using the first 30 LSIs for Uniform Manifold Approximation and Projection (UMAP) using the RunUMAP function in Seurat for visualization with the default setting.

Cell type annotation and identification of neoplastic cells
We first annotated our scRNA-Seq data based on canonical surface markers (**Figure 1D, Supplementary Figure 1D**). Distinguishing malignant cells (AML, B-ALL, T-ALL, and MPAL) from non-malignant cells (major cell types including T/NK, B, and myeloid cells) in scRNA-seq data involved a process designed to use multiple independent lines of evidence for annotation. Three lines of evidence supported identification and annotation of non-malignant cells. Firstly, we leveraged the fact that non-malignant cells from different patients would be more similar than malignant cells arising from different patients. We utilized a cluster-based statistic, Shannon

Entropy, to identify clusters of cells at multiple clustering resolutions (k = 1.2, k = 2, k = 3) to identify 4 cell populations which had contribution from every patient (**Supplementary Figure 2A-B**). The Shannon Entropy statistic, calculated using the formula $-\sum \mathrm{p(x)} * \log \mathrm{p(x)}$, where $\mathrm{p(x)}$ is the frequency of cells arising from any one patient in any one cluster, ranges from 0 to 1, where 1 represents equal contribution from every patient, and 0 represents 100% contribution from a single patient. Second, we applied a single-cell quantitative metric by calculating a similarity score to healthy controls across all patient derived cells. Patient data and healthy control data were co-embedded into a low-dimensional space using the default anchor-based CCA method in Seurat 4.0.5 (30 dimensions, 2,000 anchor features), and k=30 mutual-nearest neighbor score (representing distance between each patient cell and its k=30 nearest anchors in PC space) was assigned for each cell using the TransferData function. In support of cluster-based scoring, we observed that all non-malignant cell types had high concordance with healthy controls (mean similarity score > 0.95), as compared to malignant blasts (mean similarity score < 0.7) (**Supplementary Figure 2C-D**). Third, we then utilized several lines of evidence to validate preliminary malignant cell annotations obtained above. As malignant blasts are defined by the genomic aberrations they harbor, we next utilized InferCNV v1.6.011[10] on a randomly downsampled (to 20,000 cells of each leukemia type) subset of patient data to infer putative, genome-wide copy number variations based on scRNA-Seq expression data. InferCNV results showed gross genomic aberrations specifically in blasts across different subtypes of leukemia, with non-malignant B, myeloid and T/NK cells showing no clear patterns of CNV acquisition (**Supplemental Figure 2E-H**). We successfully identified known CNVs in various leukemia subtypes, such as gain-of-expression in chromosome 21 in iAMP21 B-ALL (**Supplementary Figure 2G**).

Projection of patient scRNA-Seq data onto healthy reference maps

To project the patient cells onto the healthy reference trajectory, we needed to transform the raw patient data as we did for constructing the healthy developmental trajectory. Specifically, for each of the VEGs used for constructing the healthy 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 healthy trajectory to calculate the projected PC scores for each patient cell. Finally, the umap_transfer function from the 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.

Projection of patient scATAC-Seq data onto healthy reference maps

For each patient, we first reconstructed the peak-by-cell matrix using the VAPs used for constructing healthy 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 healthy reference to calculate the projected PC scores for each patient cell. 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) healthy reference cell in the UMAP embedding.

Integration of HSPC-like blasts across different leukemias

Based on the projection results, we assigned cell state to each leukemic blast on the developmental trajectories. To integrate heterogeneous HSPC-like blasts and remove the patient-

specific batch effect, we performed integration using reciprocal PCA (RPCA) in Seurat V5[11]. Briefly, we subset HSPC-like blasts from combined global data matrix. For better data integration, we excluded samples with fewer than 30 HSPC-like cells (resulting in 59 patient samples remaining in this analysis). We re-normalized each sample and utilized the *SelectIntegrationFeatures* function in Seurat to select top 2000 VEGs. Then each patient sample was scaled with those top 2000 integration features using the *ScaleData* function in Seurat and PCA analysis was performed using *RunPCA* function in Seurat. Top 30 were used for the following integration anchors identification using *FindIntegrationAnchor* function with "reduction = rpca" and datasets were integrated by *IntegrateData* with default parameter settings except "k.weight = 30". The integrated matrix was then scaled and confounding factors were regressed out (fraction of mitochondrial UMIs, total number of UMIs per cell, S phase score, G2/M phase score and heat shock gene score). The resulting top 30 PCs were used in Uniform Manifold Approximation and Projection (UMAP) using the *RunUMAP* function in Seurat with the default setting. Clusters were identified with the *FindClusters* function with "resolution = 0.5". The integrated Seurat object was used for downstream analysis and visualization **(Supplementary Figure S4E-G, Supplementary Figure S7D-E).**

Identification of differentially expressed genes

For each leukemia subtype, we identified the differential expressed genes between HSPC-like and lineage-like blasts using the FindMarkers function in Seurat (V4.0.5). We performed LR test on the raw counts and regressed out the total UMI count per cell as a confounding factor. Differentially expressed genes (DEGs) were called if abs(log2(FC)) > 0.25 and adjusted p-value < 0.05. DEGs from each leukemia subtype were pooled and clustered by k-means clustering for HSPC-like blast upregulated genes (k=9) and HSPC-like blast downregulated genes (k=5) based on their $\log_2$FC. We used elbow curve method to determine the best k values. For each k, we calculated the total within-cluster sum of squares (WSS) and then plot these points to find the k where the WSS values falls suddenly.

GoT Data Analysis

Samples were demultiplexed into FASTQ files using bcl2fastq. FASTQ files were then processed using Ironthrone v2.1 with the default parameters and inputs for 10x v3.1 scRNA-seq data. Specifically, for each variant, Ironthrone was run in circularization mode (--run = circ) with UMI length 12 (--umilen 12) and cell barcodes from each sample's cellranger output (--whitelist sample.specific.barcodes.tsv). For each sample, custom configuration files were created based on primer, mutant, and WT sequences to parse R1 and R2 FASTQ files, following the configuration set within IronThrone v2.1 documentation (https://github.com/dan-landau/IronThrone-GoT). UMI concatenated summary tables outputted by IronThrone v2.1 were further filtered based on cell barcode and UMI combinations seen in 10x Cellranger output. Filtered cell x variant tables were then imported into R and intersected with high quality cells retained in scRNA-seq analysis. We established a stringent threshold for identifying mutant cells for each variant, taking cells exceeding 90th percentile of mutant UMI noted in non-malignant populations (T/NK, B, and myeloid). For each cell barcode, mutant calls were then concatenated, and the number of unique mutations summed. The cells with > 2 unique mutant UMIs were called as malignant cells.

Deconvolution and subpopulation fraction analysis of bulk RNA-Seq

Bulk RNA-seq data (**Table S1**) were obtained from public databases. First, the bulk RNA-Seq matrix was normalized for gene length and sequencing depth using the edgeR package. Second, the scRNA-Seq data was re-normalized to align with the bulk RNA-Seq data, utilizing the Seurat function NormalizeData with method parameter set to "RC" and a scale factor of 1E6. The cell annotation of scRNA-Seq was subsequently integrated into the normalized matrix. Both bulk and

single-cell matrices were then uploaded to the CIBERSORTx website (https://cibersortx.stanford.edu/) and run CIBERSORTx[12,13] with S-mode batch correction and set the number of permutations to 100. The resulting outputs, consisting of the fraction of each projected population, were used for downstream analysis.

MRD and survival analysis

We analyzed the relationship between the gene signature scores and clinical outcomes. The HSPC-like gene sets were identified using scRNA-Seq data in **Figure 4** and integrated scRNA-Seq and scATAC-Seq data in Fig. 6. We then used these gene sets to compute gene signature score in the bulk RNA-Seq cohorts. Briefly, a z-score was calculated for each feature in the gene set and an average z-score was obtained to represent a sample-level signature score. Patients were then divided into 1) MRD+ and MRD- based on their value of BM residual blast percentage at Day 29 after induction chemotherapy, and 2) induction failure (% of MRD > 5%), MRD5 (1% < % of MRD <= 5%), MRD1 (0.01% < % of MRD <= 1%), and MRD-. For survival analysis, we used the same gene signature scores to stratify overall survival. Briefly, we binarized patients based on signature score and utilized Cox Proportional Hazard model with Day 29 MRD taken as a covariate using the survfit function from survival v3.2-1322: survfit(Surv(time.OS, status.OS) ~ high.HSPC + D29.MRD.

Regression analysis of MRD, LC50 and leukemic blast proportion

MRD levels were log-transformed with those less than 0.01% (detection limit) assigned a value of 0.005% before log-transformation. Leukemic blast proportions from deconvolution of bulk RNA-seq data were scaled to have zero mean and unit variance before regression. Multiple linear regression was performed adjusting for age at diagnosis, white blood cell counts at diagnosis, and treatment protocol. Regression of T-ALL data also included a term adjusting for ETP/non-ETP status. P-values were adjusted for multiple testing using the Benjamini-Hochberg method.
LC50 values taken from[14] were log-transformed and normalized using the following equation LC50_normalized = (LC50_measured − LC50_min)/(LC50_max − LC50_min). Multiple linear regression was performed between normalized LC50 values and blast proportions from bulk RNA-Seq deconvolution described above, and p-values were adjusted for multiple testing using the Benjamini-Hochberg method.

***In silico* drug screening**

To identify targetable genes and eliminate the off-target effect on the normal HSPCs, we computed two sets of DEGs and obtained 672 genes that were up-regulated in at least one comparison in the DEG Set1, removing the genes that are up-regulated in the normal HSPCs (DEG Set2). This included 6 comparisons of DEG analysis. The genes were given a score of 1 if one comparison fulfilled the following cutoffs: FDR < 0.05, and log2FC(HSPC-like/Lineage-like) or log2FC(Normal HSPC/HSPC-like) > 0.25. Thus, the maximum of DEG scores for a gene would be 6.

We then employed 3 public drug databases to identify the genes with available drugs: DGIdb (https://www.dgidb.org/), Open Targets (https://www.opentargets.org/), and Therapeutic Target Database (https://db.idrblab.net/ttd/). The genes were given a score of 1 if the drug can be found in each of the 3 drug databases. Thus, the maximum of drug database score for a gene would be 3.

To search for drugs that could specifically modify the HSPC-like transcriptome state, we inputted top HSPC-like DEGs and top lineage-like DEGs of three leukemia subtypes into the LINCS1000 database (https://clue.io/query#l1000) using default parameters. Top DEGs were defined by Log2FC >= 0.25 and FDR < 0.001. Perturbation results were filtered using a custom R script which filtered compound-mediated perturbations for compounds with defined targets, statistical significance (log10(FDR) > 1), effect size (normalized connectivity score > 0.8),

specificity to HSPC-like state (raw connectivity score > 0), and activity in 2 or more leukemia cell lines. Specificity to HSPC-like state was defined by positive connectivity score in the case of gene knockdown, and negative connectivity score in the case of gene overexpression. DEGs that up-regulated in HSPC-like blasts targeted by top compound perturbations and/or genes within gene overexpression/KD were given a score of 1. Thus, the maximum of LINCS1000 database score for a gene would be 3.

Finally, to predict HSPC-like DEGs that could mediate cytotoxicity when targeted, we downloaded cell line dependency data from the cancer dependency map (DEPMAP) portal and searched for HSPC-like DEGs which showed increased dependency in leukemia cell lines (n=59) compared to non-leukemia/lymphoma cell lines (n=1,052). Genes with negative dependency scores in leukemia cell lines (mean dependency score < -0.2), dependency fold-change > 2, and >20% expression in HSPC-like blasts were assigned a score of 1.

To prioritize genes, we summed DE evidence (ranging from 0-6) and drug database evidence (ranging from 0-3), and functional genomics database evidence (ranging from 0-4) was taken to rank HSPC-like DEGs for follow-up experimental studies.

Construction of Transcriptional Regulatory Network (TRN)

TRN was constructed using the following three steps. In Step one, two strategies were employed to determine the transcription factors (TFs) that regulate HSPC-like leukemic blasts: 1) upregulated TFs as determined by differential gene expression analysis using scRNA-Seq data; 2) increased TF activity in HSPC-like blasts as determined using scATAC-Seq. For TF activity, we performed differential TF motif enrichment analysis between HSPC-like and lineage-like blasts for each leukemic subtype. We first used the chromVAR algorithm[15] to identify TF motifs that had enriched genome-wide chromatin accessibility in each cell population. We then tested differential enrichment between the two populations 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% of cells in the group. In Step two, we identify TF-target pairs by regression analysis. We first co-embedded scRNA-Seq and scATAC-Seq data using Seurat (V4.0.5) and called metacells sample-by-sample using the WGCNA package with k=20 nearest-neighbors[16,17]. We then predicted enhancer-promoter (E-P) interactions using a regression-based method described previously[18]. A linear regression was conducted for each gene, with the gene expression in each metacell as the dependent variable and the normalized accessibility of chromatin peaks (enhancers) within +/- 500kb of the gene promoter as the independent variables. If the regression coefficient of an enhancer (chromatin peak) is greater than 0.1 and the Benjamini-Hochberg adjusted P-value is smaller than 0.05, the E-P pair was called significant. Cell-type-specific E-P pairs were called if the enhancer peak had higher chromatin accessibility ($log_2FC > 1$) in the given cell type than the compared cell type(s). In Step three, the TRN was constructed using differentially expressed genes and enriched TFs in HSPC-like blasts compared to lineage-like blasts. 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.

Primary-relapse RNA-seq analysis

Paired diagnosis and relapse samples from the AALL0434 dataset were analyzed by computing the HSPC-like TRN signature score. AUCell v1.26.0 was used with the top 50% of expressed genes (aucMaxRank parameter).

***In silico* transcription factor perturbation analysis**

A global enhancer-driven transcriptional regulatory network was re-constructed using SCENIC+ v1.0a1[19]. The scRNA-seq and scATAC-seq data were co-embedded and metacells were called as described above to construct pseudo-multiome data. Then, the metacells were normalized and

projected sample-by-sample to the healthy reference using the FindTransferAnchors and MapQuery functions with the top 30 dimensions and k.weight=40. Then SCENIC+ pipeline was run largely in accordance with the suggested protocol for multiome data (https://scenicplus.readthedocs.io/en/latest/tutorials.html). Each subtype (T-ALL, B-ALL, AML) was analyzed separately. Briefly, topic modeling was performed and 80, 70, and 35 topics were selected for T-ALL, B-ALL, and AML respectively. The topics were binarized using the otsu method with ntop=3000. Differentially accessible chromatin regions were selected with the find_diff_features function with adjusted p-value <0.05 and log2FC >1.5 based on the projected cell type annotations.

A custom cistarget database was constructed for each leukemia subtype using the consensus peaks. FASTA files were prepared with 1000 bp padding and the created_cistarget_motif_databases function was used with the v10nr_clust_public motif database. The SCENIC+ pipeline was run with a search space defined as 0-500 kb. Regulons were filtered with the following parameters: rho_threshold=0.03, min_regions_per_gene=0, and min_target_genes=10. All other parameters were maintained as the defaults. Then, *in silico* perturbation was conducted by using the train_gene_expression_models with all genes followed by the simulate_perturbation function in which each of the transcription factors in the HSPC-like TRN were set to 0 and parameter n_iter=10.This produced a modified count matrix that was subsequently loaded into Seurat, cells with projected identity of HSPC were subset, and the TRN signature was scored as above.

SUPPLEMENTAL TABLES

Table S1. Summary of patient cohorts and datasets generated and used in this study**.**

Table S2. Summary of patient clinical data and sample information used in this study.

Table S3. List of antibodies used in this study.

Table S4. List of PCR primers for Genotyping of Transcriptomes (GoT) assay.

Table S5. List of differentially expressed genes (abs(log2FC) > 0.25 & FDR < 0.05) between HSPC-like and lineage-like blasts. DEGs were organized into clusters identified by k-means clustering. Up-regulated genes in HSPC-like blasts were listed in the first sheet, down-regulated genes in HSPC-like blasts were listed in the second sheet.

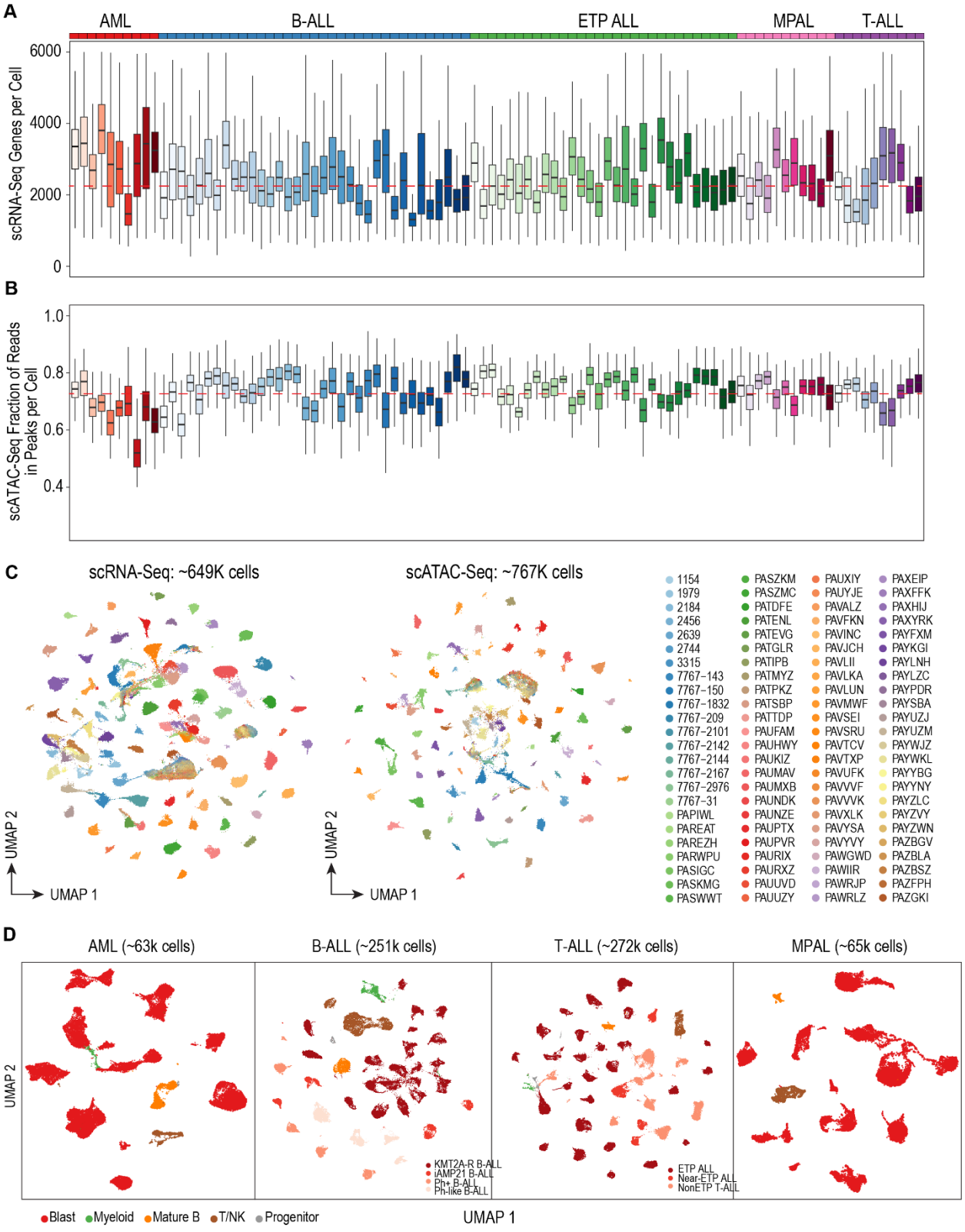
Table S6. List of all 125 transcription factors that are activated in at least one leukemia type based on both scRNA-Seq and scATAC-Seq data.

Table S7. Shared TRN targets among three major types of leukemias.

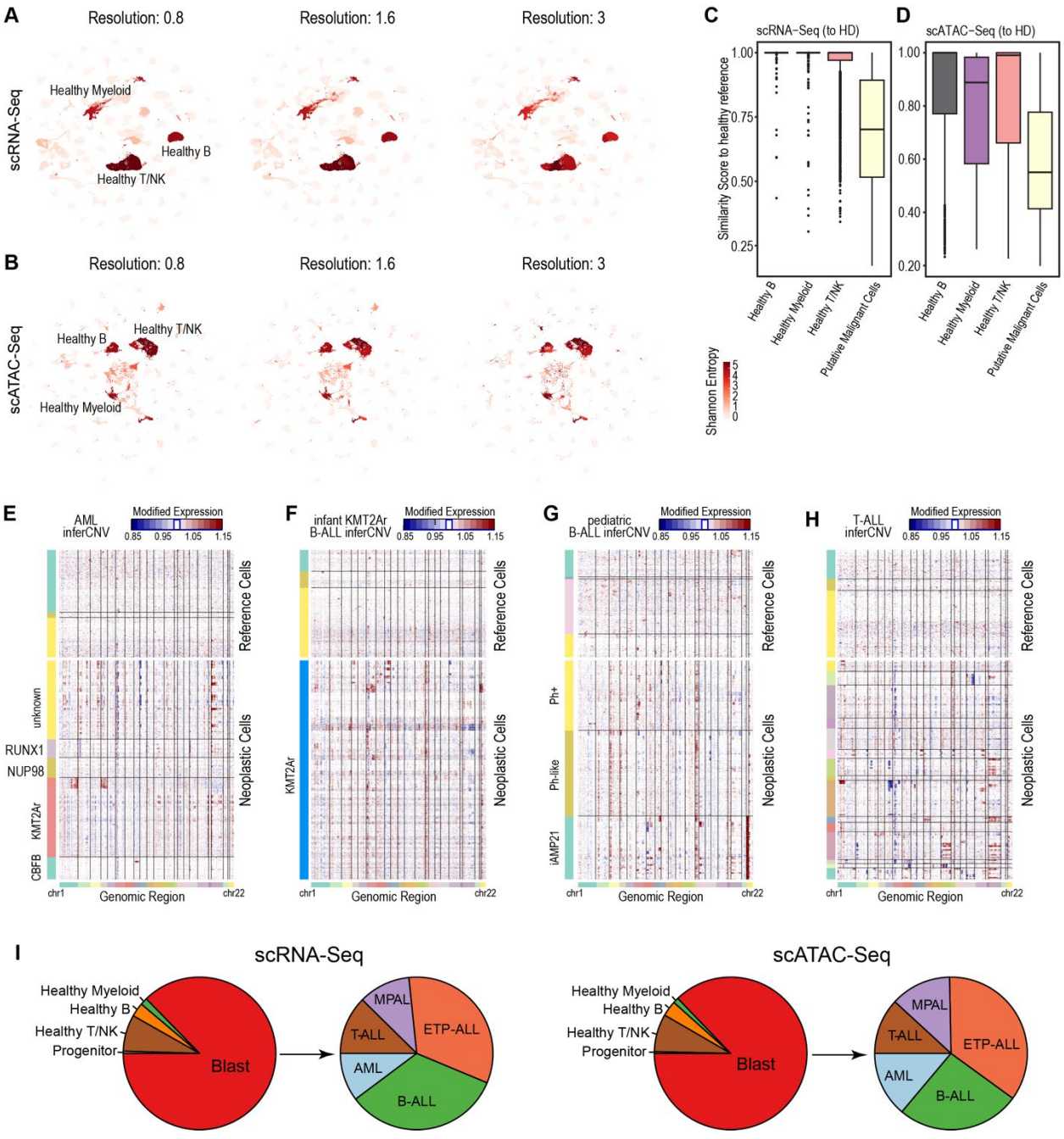
Table S8. List of candidate drug targets against HSPC-like blasts based on in silico screening.

Table S9. List of 38 drugs used in *ex vivo* drug screening.

SUPPLEMENTAL FIGURES

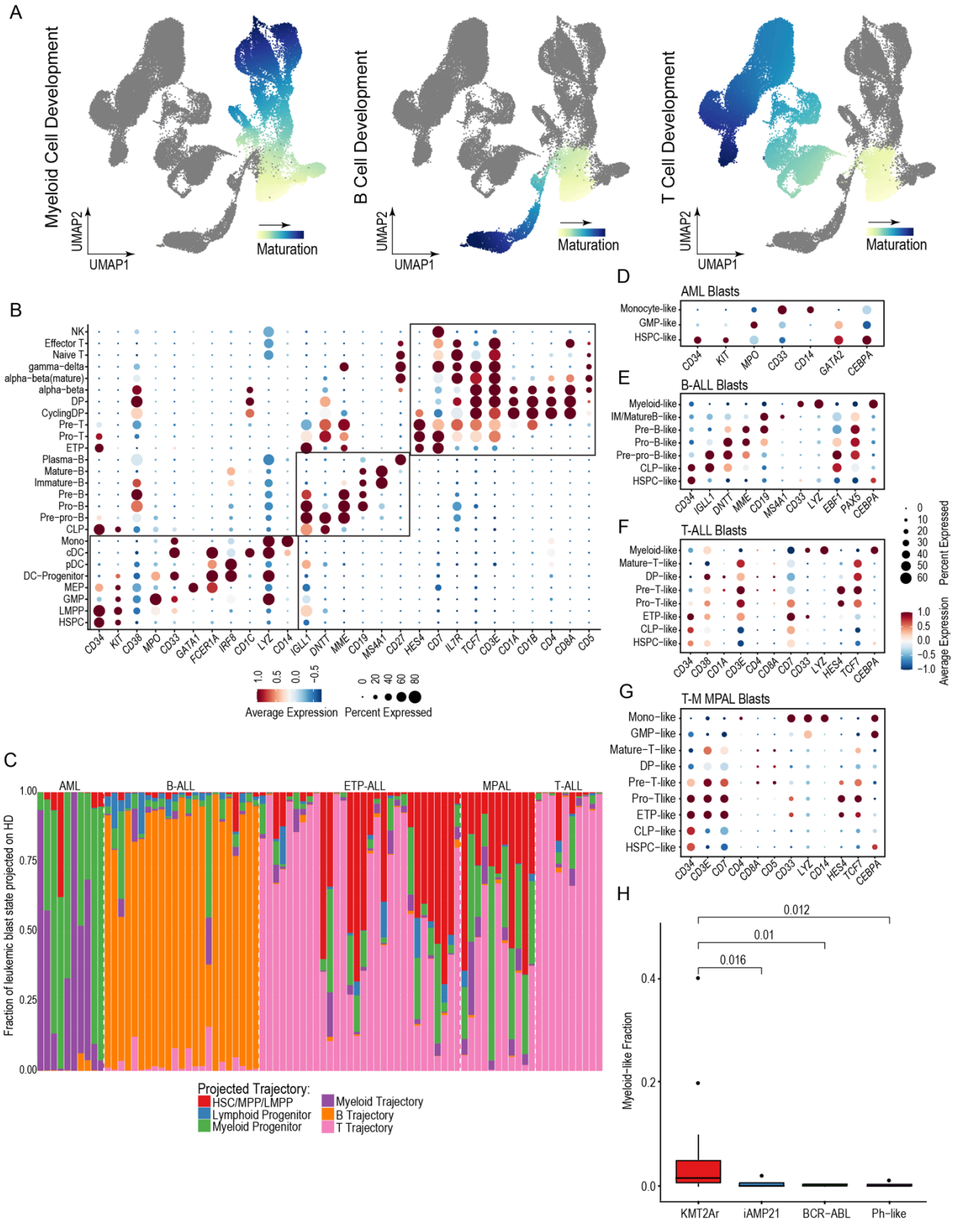


Supplemental Figure 1. Quality assessment of single-cell multi-omics data (related to Figure 1). Dataset used in this figure: scRNA-Seq and scATAC-Seq in this study. A, Boxplot showing the number of genes identified per cell in scRNA-Seq data after QC and removal of low-quality cells. **B,** Boxplot showing the fraction of accessible fragment reads located in peaks (FRIP) in scATAC-Seq data after QC and removal of low-quality cells. Red horizontal line indicates average value across all samples. **C,** Overall UMAPs of all scRNA-Seq (left panel) and scATAC-Seq (middle panel) cells of 96 leukemia samples, colored by samples. Right panel, color legend for samples. Total numbers of sequenced cells are indicated. **D,** UMAPs for each leukemia subtype studied, colored by annotated cell populations. Different shades are used to further distinguish leukemic blasts of different sub-subtypes within each major leukemia subtype. Total numbers of sequenced cells are indicated.

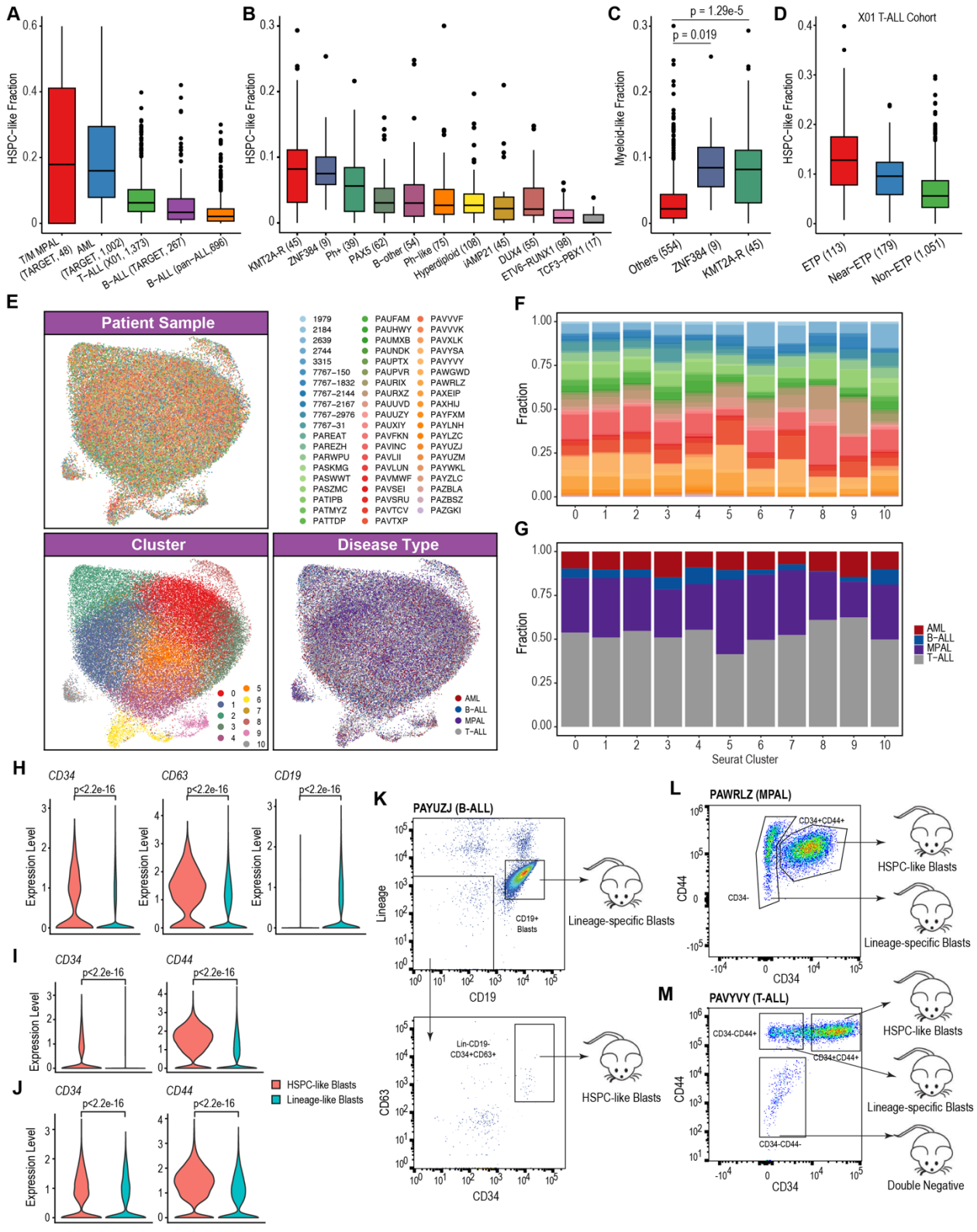


Supplemental Figure 2. Annotation of malignant leukemic blasts and non-malignant cell populations (related to Figure 1). Dataset used in this figure: scRNA-Seq and scATAC-Seq in this study. A-B, Shannon entropy scores of scRNA-Seq **(A)** and scATAC-Seq **(B)** clusters at k=0.8, k=1.6, and k=3 clustering resolutions. Healthy populations are indicated on the left UMAP with clustering resolutions k=0.8. **C-D,** Similarity scores of annotated cell types in patient samples to healthy BM and thymus samples based on scRNA-Seq **(C)** and scATAC-Seq (**D**) data. Patient cells were mapped to cell types in healthy samples using the rPCA method of Seurat 4.0 using k=30 nearest neighbors. The average similarity score to 30 nearest healthy sample neighbors in principal component space is shown for each cell. **E-H,** InferCNV analysis of candidate malignant blasts by subtypes: AML (**E**), infant B-ALL (**F**), pediatric B-ALL (**G**), and T-ALL (**H**). Heatmap showing the normalized expression values across all samples. Top panel shows result using the reference healthy cells, including healthy B, T/NK, and myeloid cells from the patient samples.

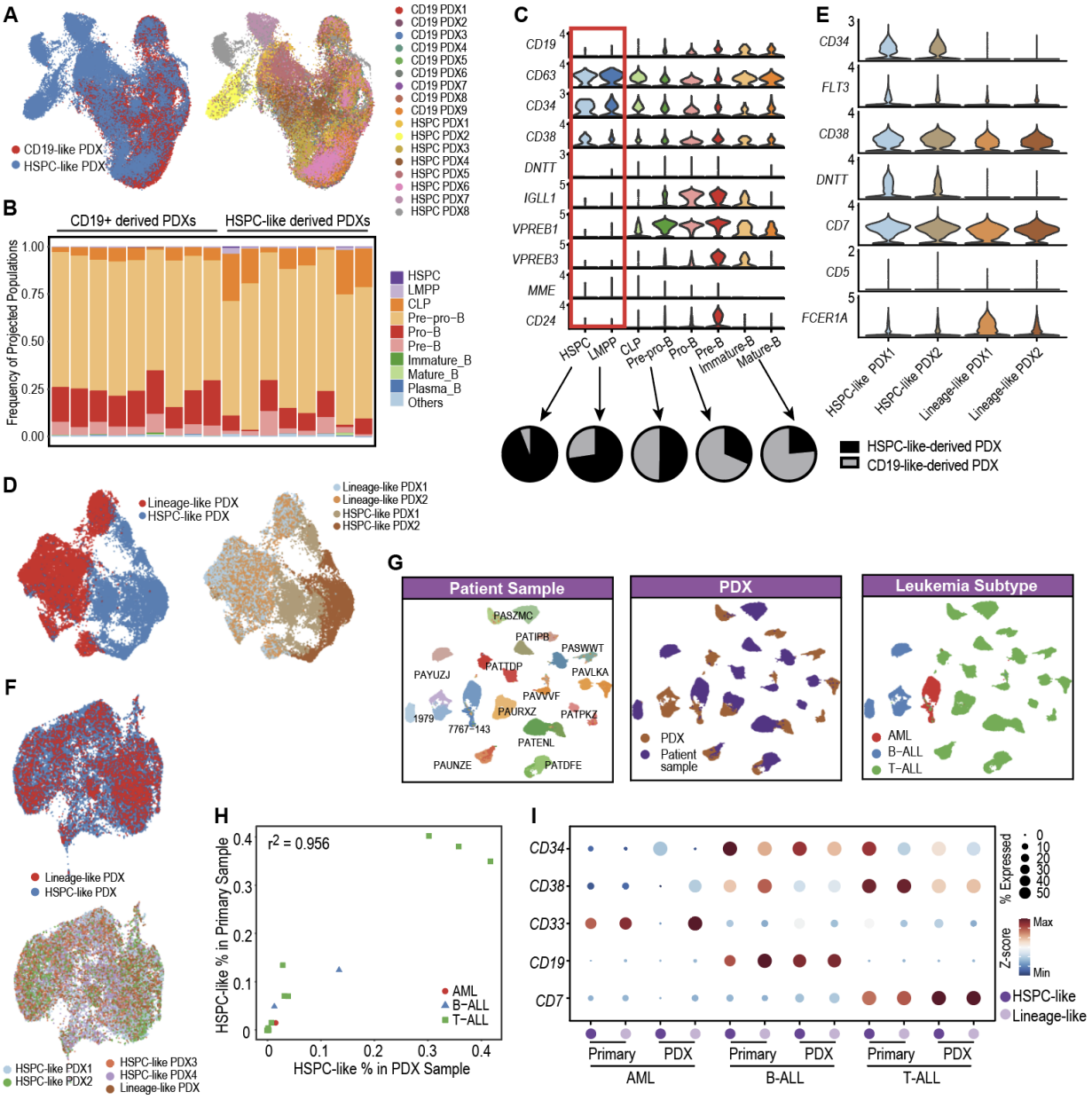
Bottom panel shows result using candidate leukemic blasts, clustered by genetic lesions. CNVs were called using InferCNV with default parameters. **I,** Pie charts showing the cell type compositions of scRNA-Seq (left panel) and scATAC-Seq (right panel) data. The proportions of leukemic blasts for each subtype are indicated in the right pie charts. In total, 56,661 AML, 187,092 B-ALL, 253,051 T-ALL, and 60,293 MPAL leukemic blasts were annotated.


A
Myeloid Cell Development
B Cell Development
T Cell Development
UMAP1
UMAP2
Maturation
B
NK
Effector T
Naive T
gamma−delta
alpha−beta(mature)
alpha−beta
DP
CyclingDP
Pre−T
Pro−T
ETP
Plasma−B
Mature−B
Immature−B
Pre−B
Pro−B
Pre−pro−B
CLP
Mono
cDC
pDC
DC−Progenitor
MEP
GMP
LMPP
HSPC
CD34 KIT CD38 MPO CD33 GATA1 FCER1A IRF8 CD1C LYZ CD14 IGLL1 DNTT MME CD19 MS4A1 CD27 HES4 CD7 IL7R TCF7 CD3E CD1A CD1B CD4 CD8A CD5
Average Expression
Percent Expressed
C
AML
B-ALL
ETP-ALL
MPAL
T-ALL
Fraction of leukemic blast state projected on HD
Projected Trajectory:
HSC/MPP/LMPP
Lymphoid Progenitor
Myeloid Progenitor
Myeloid Trajectory
B Trajectory
T Trajectory
D
AML Blasts
Monocyte-like
GMP-like
HSPC-like
CD34 KIT MPO CD33 CD14 GATA2 CEBPA
E
B-ALL Blasts
Myeloid-like
IM/MatureB−like
Pre−B−like
Pro−B−like
Pre−pro−B−like
CLP−like
HSPC−like
CD34 IGLL1 DNTT MME CD19 MS4A1 CD33 LYZ EBF1 PAX5 CEBPA
F
T-ALL Blasts
Myeloid-like
Mature−T−like
DP−like
Pre−T−like
Pro−T−like
ETP−like
CLP−like
HSPC−like
CD34 CD38 CD1A CD3E CD4 CD8A CD7 CD33 LYZ HES4 TCF7 CEBPA
Percent Expressed
Average Expression
G
T-M MPAL Blasts
Mono−like
GMP−like
Mature−T−like
DP−like
Pre−T−like
Pro−Tlike
ETP−like
CLP−like
HSPC−like
CD34 CD3E CD7 CD4 CD8A CD5 CD33 LYZ CD14 HES4 TCF7 CEBPA
H
0.012
0.01
0.016
Myeloid−like Fraction
KMT2Ar
iAMP21
BCR−ABL
Ph−like

Supplemental Figure 3. Healthy reference trajectories and patient sample projection (related to Figure 1). Dataset used in this figure: scRNA-Seq and scATAC-Seq in this study. A, UMAPs showing three major lineages, colored by pseudotime values computed by Slingshot[20] with default parameters. Left, Myeloid developmental trajectory; middle, B-cell developmental trajectory; right, T-cell developmental trajectory. **B,** Dot plot of manually curated markers for normal hematopoietic development in assigned populations. Color represents scRNA-Seq Z-score. Dot size represents the percentage of expression in particular populations. Boxes highlight three major lineages, Myeloid, B cell, and T cell lineages from bottom to top. **C,** Bar plots showing proportions of patient cells projected onto stem/progenitor, T-lineage, B-lineage, and Myeloid-lineage in all samples from this study using scATAC-Seq. **D-G**, Dot plot of manually curated markers for projected populations in different leukemias, AML (**D**), B-ALL (**E**), T-ALL (**F**), and T-M MPAL (**G**). **H,** Fraction of leukemic blasts projected as myeloid progenitor-like cells based on scRNA-Seq in four B-ALL subtypes. Case number: *KMT2A*-R n=23; iAMP21 n=4; BCR-ABL n=4; Ph-like n=4. P-values were computed using Student's t-test.

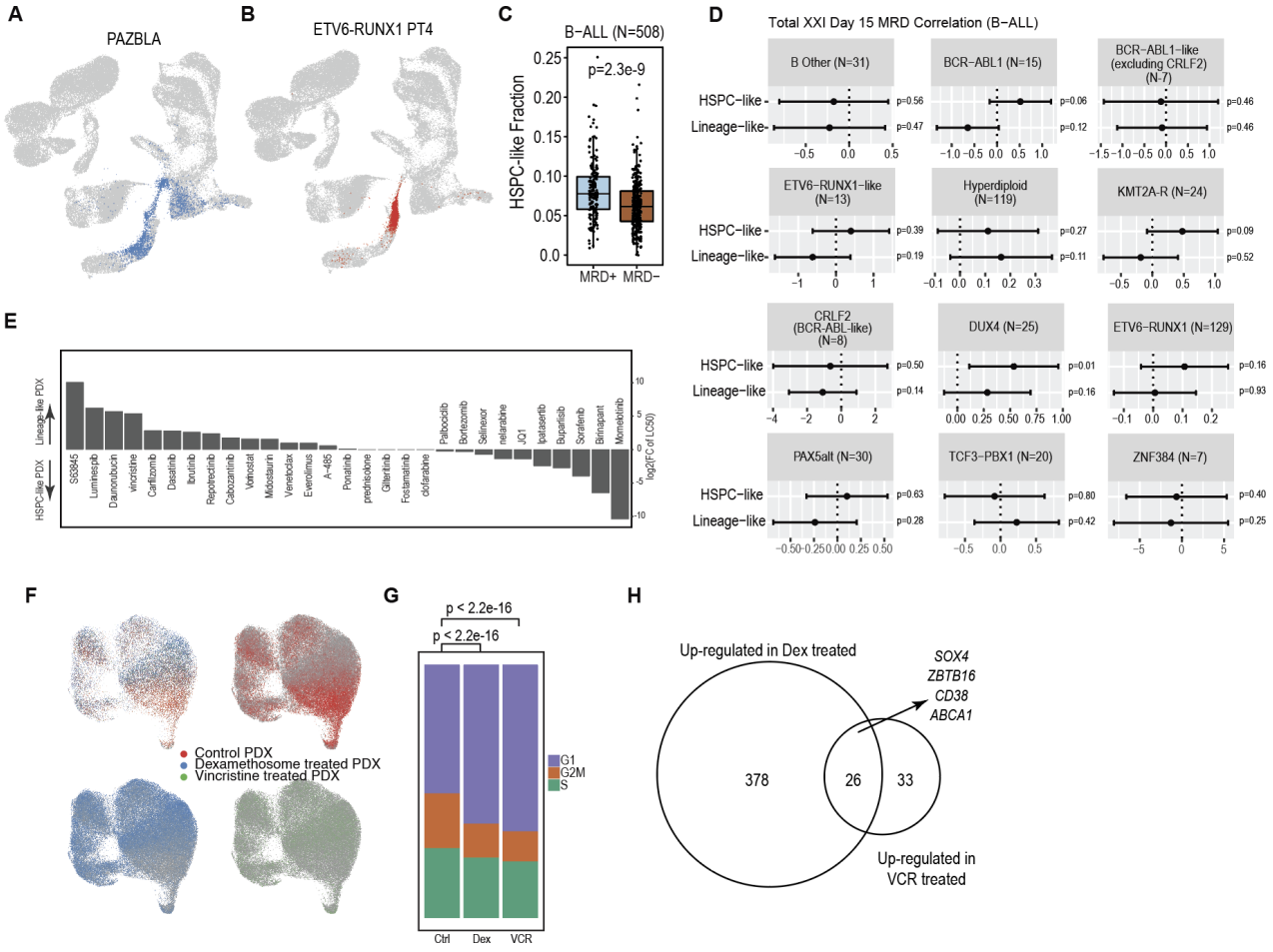

A
HSPC−like Fraction
T/M MPAL (TARGET, 48)
AML (TARGET, 1,002)
T−ALL (X01, 1,373)
B−ALL (TARGET, 267)
B−ALL (pan−ALL,696)
B
HSPC−like Fraction
KMT2A−R (45)
ZNF384 (9)
Ph+ (39)
PAX5 (62)
B−other (54)
Ph−like (75)
Hyperdiploid (108)
iAMP21 (45)
DUX4 (55)
ETV6−RUNX1 (98)
TCF3−PBX1 (17)
C
Myeloid−like Fraction
p = 1.29e−5
p = 0.019
Others (554)
ZNF384 (9)
KMT2A−R (45)
D
X01 T−ALL Cohort
HSPC−like Fraction
ETP (113)
Near−ETP (179)
Non−ETP (1,051)
E
Patient Sample
Cluster
Disease Type
AML
B−ALL
MPAL
T−ALL
F
Fraction
G
Fraction
Seurat Cluster
AML
B−ALL
MPAL
T−ALL
H
CD34
CD63
CD19
p<2.2e−16
Expression Level
I
CD34
CD44
p<2.2e−16
Expression Level
J
CD34
CD44
p<2.2e−16
Expression Level
HSPC−like Blasts
Lineage−like Blasts
K
PAYUZJ (B−ALL)
Lineage
CD19
CD19+ Blasts
Lineage−specific Blasts
CD63
CD34
Lin−CD19−
CD34+CD63+
HSPC−like Blasts
L
PAWRLZ (MPAL)
CD34+CD44+
CD34−
CD44
CD34
HSPC−like Blasts
Lineage−specific Blasts
M
PAVYVY (T−ALL)
CD34−CD44+
CD34+CD44+
CD34−CD44−
CD44
CD34
HSPC−like Blasts
Lineage−specific Blasts
Double Negative

Supplemental Figure 4. HSPC-like leukemic blasts are commonly observed across high-risk acute leukemias and can give rise to leukemia in mice (related to Figure 2). Dataset used in this figure: A-D: public bulk RNA-Seq; E-J: scRNA-Seq in this study. A, Frequencies of HSPC-like blasts in public bulk RNA-Seq cohorts based on CIBERSORTx deconvolution (Supplementary Methods), ranked by the median of frequencies in each leukemia subtype. Five datasets are used in this plot: TARGET T/M MPAL (n = 48), TARGET AML (n = 1,002), X01 T-ALL (n = 1,373), TARGET B-ALL (n = 267), and pan ALL B-ALL (n=696). **B,** Frequencies of HSPC-like blasts in B-ALL from the pan ALL cohort, grouped by major B-ALL subtypes, ranked by the median of frequencies in each subtype. PAX5 cohort includes patients with both PAX5 alt and PAX5 P80R genotypes. Dataset: pan ALL bulk RNA-Seq[21], n=696. **C,** Frequency of Myeloid-like blasts in the pan ALL B-ALL cohort, grouped by B-ALL subtypes. Dataset: pan ALL bulk RNA-Seq[21], n=696. P-values were computed using Student's t-test. **D,** Frequencies of HSPC-like blasts in T-ALL from bulk RNA-Seq X01 cohort[22], grouped by T-ALL phenotypes, ranked by the median of frequencies in each subtype. The patient number of each subtype was indicated (ETP-ALL, n=113; near-ETP, n=179; non-ETP, n=1051). **E,** Integrated UMAP for all HSPC-like blasts identified in this study. Upper panel, colored by patient samples; bottom left panel, colored by cluster; bottom right panel, colored by disease type. **F,** HSPC-like blast fraction of each patient in different clusters; **G,** HSPC-like blast fraction of each disease in different clusters. **H-J,** Expression levels of surface marker used for flow sorting in both HSPC-like and lineage-like populations in B-ALL (H), T-ALL (I), and MPAL (J) scRNA-Seq data. P-values were computed using two-sided Wilcoxon test. **K,** Representative plots showing FACS sorting strategy of HSPC-like blasts and lineage-like blasts from KMT2A-R B-ALL patients. Sorted cells were injected into NSG mice. **L-M,** Representative plots showing FACS sorting strategy of HSPC-like blasts from T-ALL (**L**) and T-myeloid MPAL patients (**M**). Sorted cells were injected into NSG mice for T-ALL and NSGS mice for MPAL.

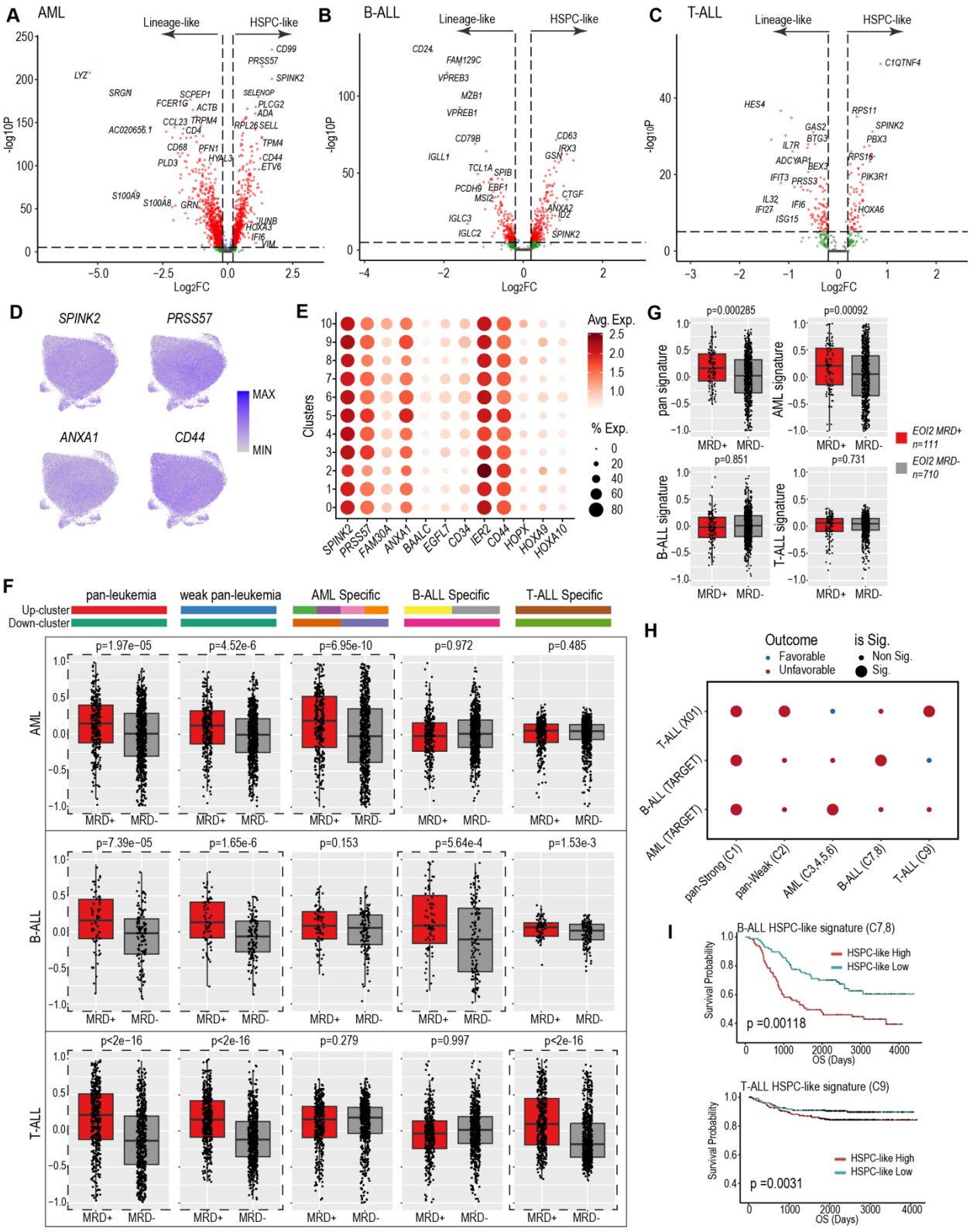


Supplemental Figure 5. HSPC-like leukemic blasts give rise to distinct phenotypes of leukemia in B-ALL and T-ALL patients, but not in MPAL (related to Figure 2). Dataset used in this figure: scRNA-Seq in this study. A, Overall UMAPs of scRNA-Seq data using sorted PDX cells from an infant KMT2A-R B-ALL sample. Left UMAP is colored by source of PDXs, blue dots represent cells derived from HSPC-like (CD34+CD63+) blasts, red dots represent cells derived from CD19+ lineage-like blasts. Right UMAP is colored by different PDXs. **B,** Bar plots showing the proportions of PDX cells projected onto B-cell developmental trajectory. **C,** Upper panel: violin plots showing the canonical surface markers for B-cell development. Red box highlights PDX cells projected onto HSC/LMPP population. Lower panel: pie charts showing the fractions of sources from which PDX cells were derived at each projected developmental stage. **D,** Overall UMAPs of scRNA-Seq data using sorted PDX cells from an ETP ALL sample. Left UMAP is colored by source of PDXs, blue dots represent cells derived from HSPC-like (CD34+CD44+) blasts, red dots represent cells derived from CD34- lineage-like blasts. Right UMAP is colored by different PDXs. **E,** Violin plots showing the canonical surface markers for T-cell development in different PDX models. **F,**

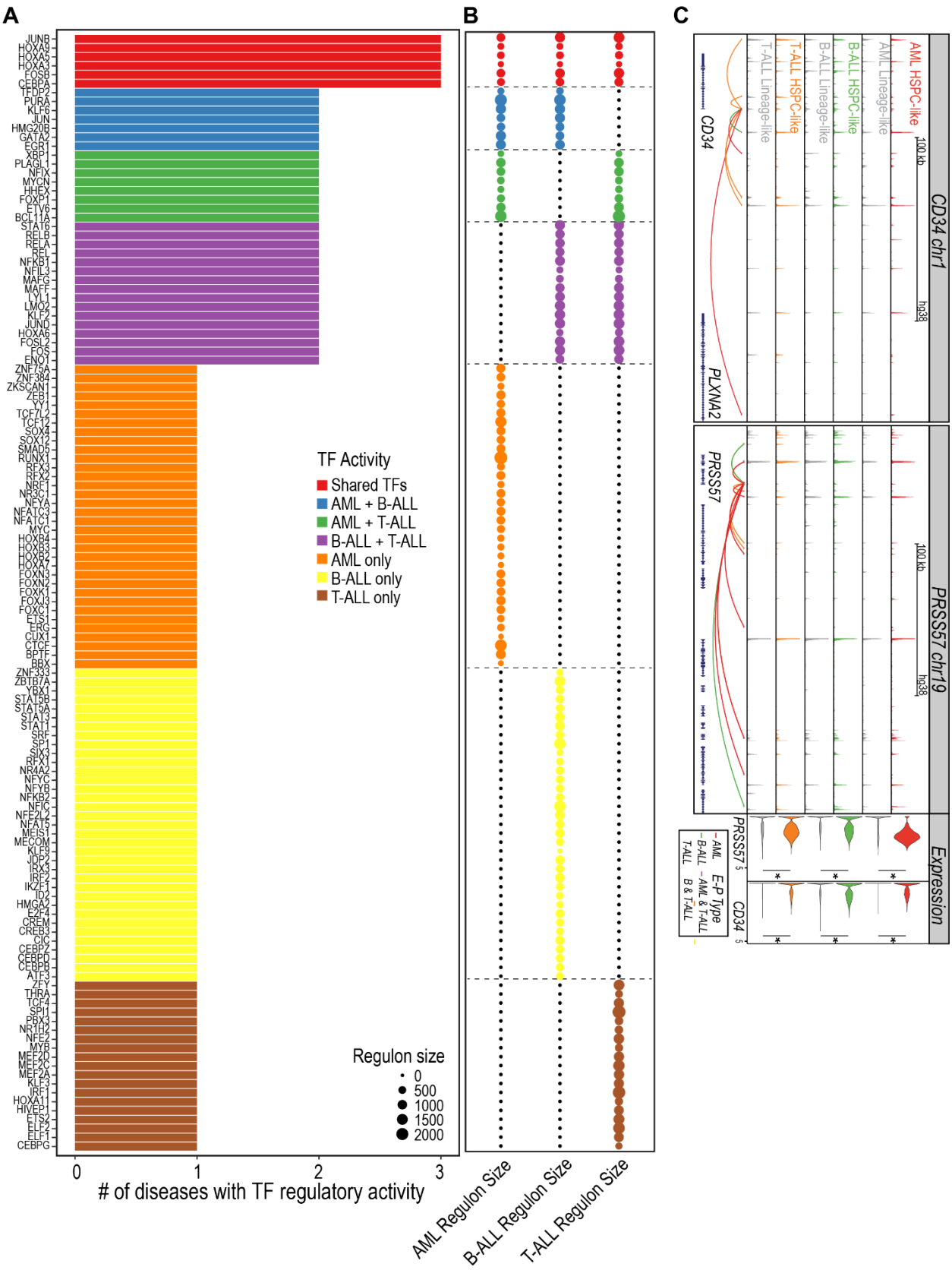
Overall UMAPs of scRNA-Seq data using sorted PDX cells from one MPAL sample. Left UMAP is colored by source of PDXs, blue dots represent cells derived from HSPC-like (CD34+CD44+) blasts, red dots represent cells derived from CD34- lineage-like blasts. Right UMAP is colored by different PDXs. **G,** UMAPs of scRNA-Seq data of primary patient sample and PDX pairs. **H,** Correlation of fraction of projected HSPC-like blasts between paired primary and PDX cells. **I,** Dot plots showing expression of manually curated markers for HSPC/lineage-like blasts in primary and PDX samples. Color represents scRNA-Seq Z-score. Dot size represents percentage of expression in particular populations.



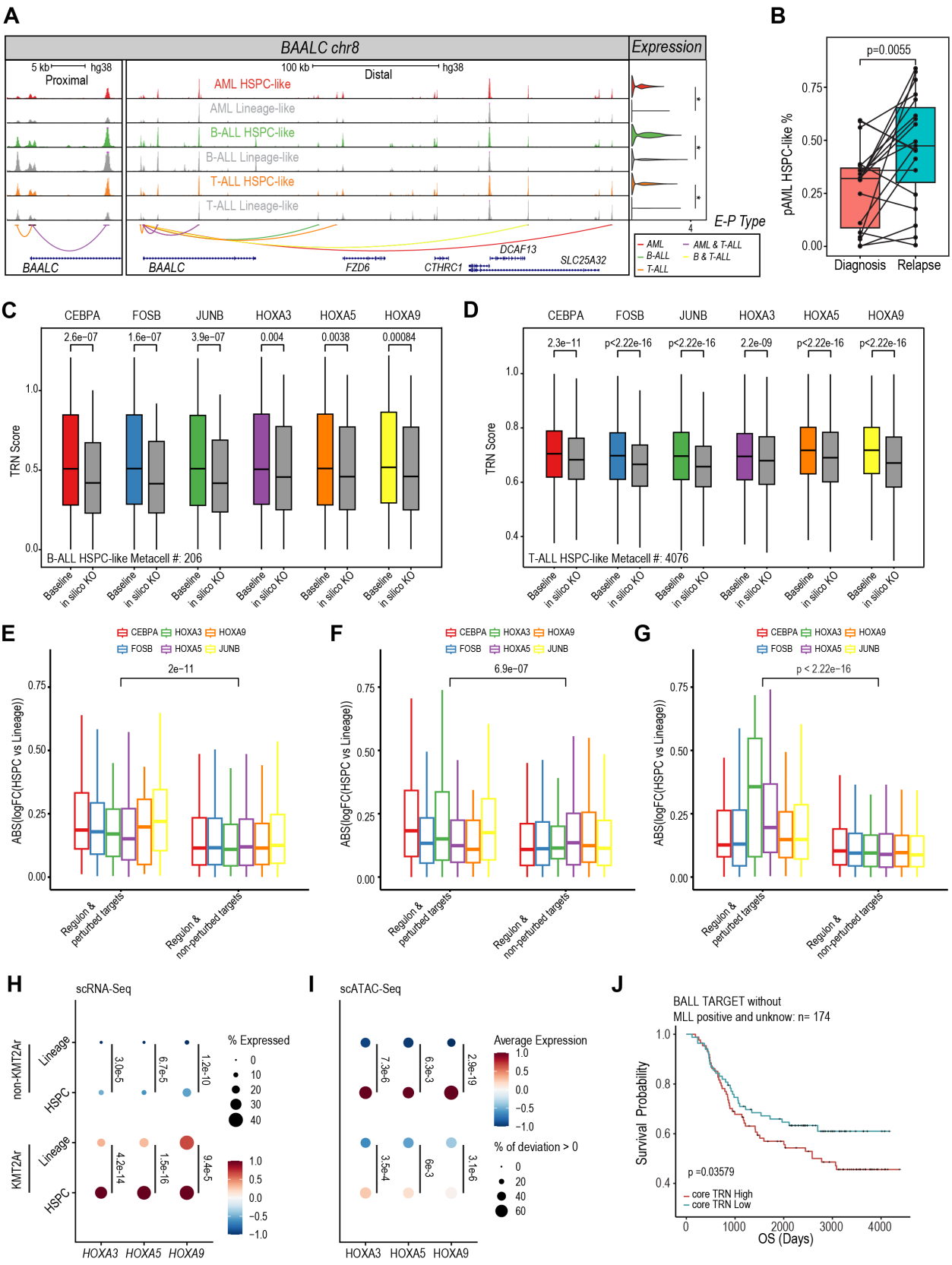
Supplemental Figure 6. HSPC-like blasts are resistant to chemotherapy across acute leukemias (related to Figure 3). Dataset used in this figure: A-B: scRNA-Seq in this study and public scRNA-Seq; C-D: bulk RNA-Seq from Total XXI cohort. E-H: scRNA-Seq of PDX models in this study. A-B, UMAPs showing projected blasts from two pediatric B-ALL patients: high-risk subtype harboring KMT2A-R **(A)**; and low-risk subtype harboring the ETV6-RUNX1 translocation[23] **(B)**. **C,** Boxplots showing the frequencies of HSPC-like blasts in MRD+ vs MRD- pediatric B-ALL patients based on deconvoluted bulk RNA-Seq data. P-values was computed using Student's t-test. **D,** Association between frequency of HSPC-like/lineage-like blasts and MRD values after induction chemotherapy in different genetic subtypes of B-ALL based on multiple linear regression (Supplementary Methods). The error bars show the 95% confidence interval of regression coefficient. The frequencies of HSPC-like and lineage-like blasts were computed using CIBERSORTx and bulk RNA-Seq data from the TotalXVI dataset. **E,** Relative drug sensitivity between HSPC-like-derived PDX cells and lineage-like-derived PDX cells. **F,** Overall UMAPs of scRNA-Seq data using PDX cells from dexamethosome (Dex) treated, vincristine (VCR) treated, and control mice. Red dots, cells from control PDXs; blue dots, cells from Dex-treated PDXs; green dots, cells from VCR-treated PDXs. **G,** Distribution of cell cycle stages in control, Dex treated, and VCR treated PDX cells. P-values were computed using the proportion test. **H,** Venn diagram showing the overlapped DEGs that were up-regulated in both Dex and VCR treated PDX cells. Key DEGs are labeled.


A
AML
Lineage-like
HSPC-like
-log10P
Log2FC
B
B-ALL
C
T-ALL
D
SPINK2
PRSS57
ANXA1
CD44
MAX
MIN
E
Clusters
Avg. Exp.
% Exp.
SPINK2
PRSS57
FAM30A
ANXA1
BAALC
EGFL7
CD34
IER2
CD44
HOPX
HOXA9
HOXA10
F
pan-leukemia
weak pan-leukemia
AML Specific
B-ALL Specific
T-ALL Specific
Up-cluster
Down-cluster
AML
B-ALL
T-ALL
MRD+
MRD-
G
pan signature
AML signature
B-ALL signature
T-ALL signature
p=0.000285
p=0.00092
p=0.851
p=0.731
EOI2 MRD+ n=111
EOI2 MRD- n=710
H
Outcome
Favorable
Unfavorable
is Sig.
Non Sig.
Sig.
T-ALL (X01)
B-ALL (TARGET)
AML (TARGET)
pan-Strong (C1)
pan-Weak (C2)
AML (C3,4,5,6)
B-ALL (C7,8)
T-ALL (C9)
I
B-ALL HSPC-like signature (C7,8)
HSPC-like High
HSPC-like Low
p =0.00118
Survival Probability
OS (Days)
T-ALL HSPC-like signature (C9)
p =0.0031

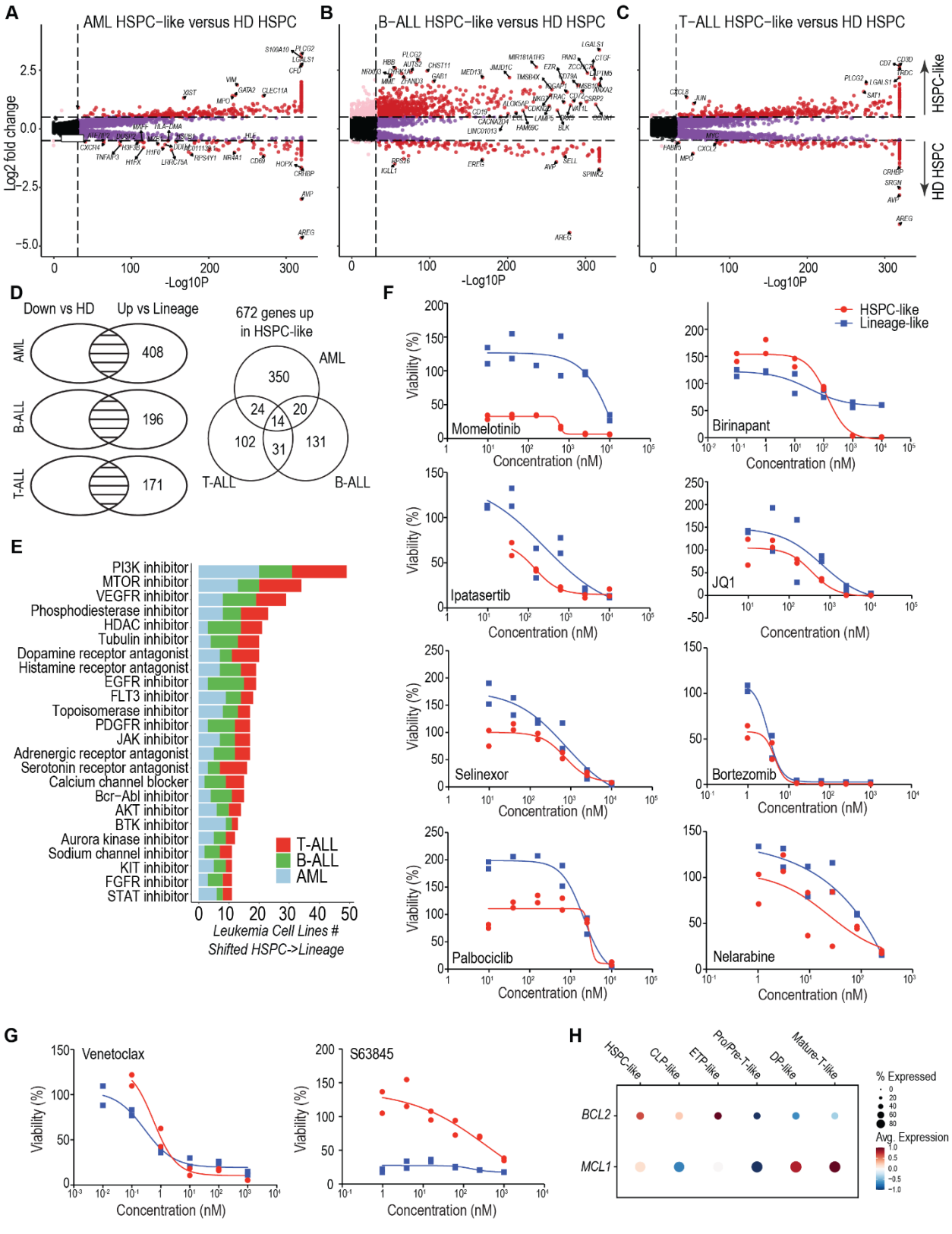
Supplemental Figure 7. HSPC-like blasts transcriptomic signatures predict drug response and clinical outcome (related to Figure 4). Dataset used in this figure: A-E: scRNA-Seq in this study; F-I: bulk RNA-Seq from TARGET for AML and B-ALL, X01 for T-ALL[22]. A-C, Volcano plots showing the DEGs comparing HSPC-like blasts and lineage-like blasts in three leukemia subtypes. **D-E,** expression level of representative commonly upregulated genes in HSPC-like blasts identified in **Figure 4A** in different clusters. **F,** HSPC-like signatures can stratify Day 29 end of induction 1 MRD level in bulk RNA-Seq cohort. Rows, leukemia subtypes; columns, single-cell derived signatures. Each patient cohort was divided as MRD+ (>= 0.1% for AML, >= 0.01% for B-ALL and T-ALL) and MRD- based on their Day 29 MRD level at the end of induction. Source of AML and B-ALL cohorts: TARGET; T-ALL cohort: X01. Dashed boxes highlight the comparison that related to each signature. P-values were computed using the two-sided Student's t-test. **G,** HSPC-like signatures can stratify end of induction 2 MRD level in AML TARGET bulk RNA-Seq cohort. Cohort was divided as MRD+ (>= 0.1%) and MRD- based on their MRD level at the end of induction 2. Sample numbers for each group were indicated. P-values were computed using the two-sided Student's t-test. **H,** Survival analysis of all HSPC-like signature scores identified in Fig. 4A. Color represents favorable or unfavorable outcome determined by the Cox proportional hazards test. Size of dot represents the significance of each score in each dataset. The p-values were calculated for log-likelihood statistic of Cox proportional hazards test with CNS status and WBC as co-variates. "Sig.", p-value < 0.05; "Non Sig.", p-value >= 0.05. **I,** ALL-specific HSPC-like signature stratifies long-term outcomes of B-ALL (upper) and T-ALL (bottom) patients. Source of patient cohort data and corresponding HSPC-like signatures are indicated at the top of each plot. The p-values were calculated for log-likelihood statistic of Cox proportional hazards test with CNS status and WBC as co-variates.


A
B
C
# of diseases with TF regulatory activity
TF Activity
Shared TFs
AML + B-ALL
AML + T-ALL
B-ALL + T-ALL
AML only
B-ALL only
T-ALL only
Regulon size
0
500
1000
1500
2000
AML Regulon Size
B-ALL Regulon Size
T-ALL Regulon Size
CD34 chr1
PRSS57 chr19
Expression
AML HSPC-like
AML Lineage-like
B-ALL HSPC-like
B-ALL Lineage-like
T-ALL HSPC-like
T-ALL Lineage-like
CD34
PLXNA2
PRSS57
100 kb
hg38
E-P Type
AML
B-ALL
T-ALL
AML & T-ALL
B & T-ALL

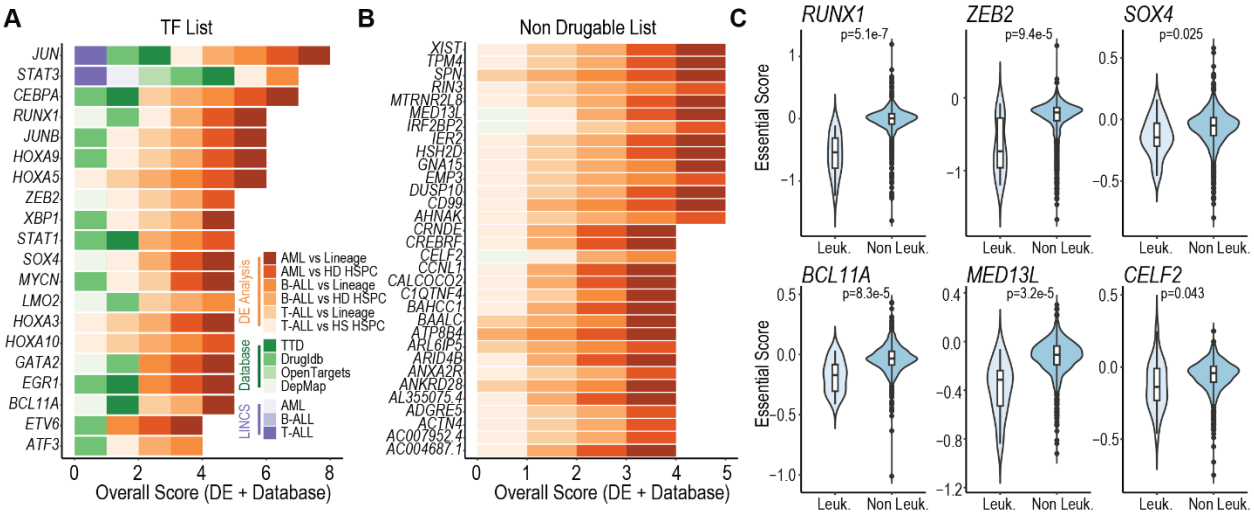
Supplemental Figure 8. Transcriptional regulatory circuitries of HSPC-like blasts (related to Figure 5). Dataset used in this figure: scRNA-Seq and scATAC-Seq in this study. A, Distribution of 125 active TF regulons across three leukemia subtypes organized in 7 groups: Shared by three leukemia subtypes; Shared by AML and B-ALL; Shared by AML and T-ALL; Shared by B-ALL and T-ALL; and subtype-specific. **B,** Bubble plot showing TF regulon sizes. **C,** Aggregated scATAC-Seq signals of HSPC-like and lineage-like blasts in each leukemia for CD34 (upper) and PRSS57 (lower) genomic loci, two representative genes that are upregulated in HSPC-like blasts and regulated by the shared TFs. Predicted enhancer-promoter interactions are shown as arcs, colored by leukemia subtypes. Bottom panel, violin plots of CD34 and PRSS57 expression for HSPC-like and lineage-like blasts in different leukemic subtypes. Adjusted p-values are indicated at the top of each violin plot. * P-value < 0.01.


A
BAALC chr8
Expression
5 kb
hg38
Proximal
100 kb
Distal
hg38
AML HSPC-like
AML Lineage-like
B-ALL HSPC-like
B-ALL Lineage-like
T-ALL HSPC-like
T-ALL Lineage-like
E-P Type
AML
B-ALL
T-ALL
AML & T-ALL
B & T-ALL
BAALC
BAALC
FZD6
CTHRC1
DCAF13
SLC25A32
B
p=0.0055
pAML HSPC-like %
Diagnosis
Relapse
C
CEBPA
FOSB
JUNB
HOXA3
HOXA5
HOXA9
2.6e−07
1.6e−07
3.9e−07
0.004
0.0038
0.00084
TRN Score
B-ALL HSPC-like Metacell #: 206
Baseline
in silico KO
D
CEBPA
FOSB
JUNB
HOXA3
HOXA5
HOXA9
2.3e−11
p<2.22e-16
p<2.22e-16
2.2e-09
p<2.22e-16
p<2.22e-16
TRN Score
T-ALL HSPC-like Metacell #: 4076
Baseline
in silico KO
E
CEBPA
HOXA3
HOXA9
FOSB
HOXA5
JUNB
2e−11
ABS(logFC(HSPC vs Lineage))
Regulon &
perturbed targets
Regulon &
non-perturbed targets
F
CEBPA
HOXA3
HOXA9
FOSB
HOXA5
JUNB
6.9e−07
ABS(logFC(HSPC vs Lineage))
Regulon &
perturbed targets
Regulon &
non-perturbed targets
G
CEBPA
HOXA3
HOXA9
FOSB
HOXA5
JUNB
p < 2.22e−16
ABS(logFC(HSPC vs Lineage))
Regulon &
perturbed targets
Regulon &
non-perturbed targets
H
scRNA-Seq
non-KMT2Ar
KMT2Ar
Lineage
HSPC
3.0e-5
6.7e-5
1.2e-10
4.2e-14
1.5e-16
9.4e-5
% Expressed
HOXA3
HOXA5
HOXA9
I
scATAC-Seq
7.3e-6
6.3e-3
2.9e-19
3.5e-4
6e-3
3.1e-6
Average Expression
% of deviation > 0
HOXA3
HOXA5
HOXA9
J
BALL TARGET without
MLL positive and unknow: n= 174
Survival Probability
p =0.03579
core TRN High
core TRN Low
OS (Days)

Supplemental Figure 9. Transcriptional regulatory circuitries of HSPC-like blasts (related to Figure 5). Dataset used in this figure: A, C-I: scRNA-Seq and scATAC-Seq in this study; B: scRNA-Seq from public scRNA-Seq AML cohort[24]; J: bulk RNA-Seq TARGET cohort. A, Aggregated scATAC-Seq signals of HSPC-like and lineage-like blasts in each leukemia for BAALC genomic loci, that are upregulated in HSPC-like blasts and regulated by the shared TFs. Predicted enhancer-promoter interactions are shown as arcs, colored by leukemia subtypes. Bottom panel, violin plots of BAALC expression for HSPC-like and lineage-like blasts in different leukemic subtypes. Adjusted p-values are indicated at the top of each violin plot. * P-value < 0.01. **B,** Frequencies of HSPC-like blasts from diagnostic-relapsed longitudinal pediatric AML scRNA-Seq data[24]. P-values were computed using paired t-test. **C-D**, Core TRN signature in baseline and *in silico* TF KO expression matrices in B-ALL HSPC-like blasts (C) and T-ALL HSPC-like blasts (D). Color boxes represent baseline signatures, grey boxes represent *in silico* TF KO signatures. P-values were computed using Student's t-test. **E-G,** Boxplots showing absolute value of log2 expression fold change of core TF targets between HSPC-like and lineage-like blasts: **(E)** AML HSPC-like blasts, **(F)** B-ALL HSPC-like blasts, and **(G)** T-ALL HSPC-like blasts. Left boxplots, core TRN TF targets with expression changes due to *in silico* KO of the TF; Right boxplots, core TRN TF targets with no expression changes due to i*n silico* KO of the TF. Different color boxes represent different TFs in the core TRN. **H,** Dot plot of three *HOXA* family gene expression levels in both HSPC-like and lineage-like blasts in two patient groups: *KMT2A*-R and non-*KMT2A*-R. Color represents scRNA-Seq Z-score. Dot size represents the percentage of expression in particular populations. P-value for each comparison were added to the side of dots. P-values were computed using the two-sided Wilcoxon test. **I,** Dot plot of three HOXA family TF activity (deviation score) in both HSPC-like and lineage-like blasts in two patient groups: *KMT2A*-R and non-*KMT2A*-R. Color represents deviation Z-score. Dot size represents the percentage of cells with deviation score > 0. P-value for each comparison were added to the side of dots. P-values were computed using the two-sided Wilcoxon test. **J,** Core TRN signature stratifies long-term outcomes of B-ALL without *KMT2A*-R cases (n=174). Source of patient cohort data and sample number are indicated at the top. The p-values were calculated for log-likelihood statistic of Cox proportional hazards test with CNS status and WBC as co-variates.


A
AML HSPC−like versus HD HSPC
B
B−ALL HSPC−like versus HD HSPC
C
T−ALL HSPC−like versus HD HSPC
Log2 fold change
-Log10P
HSPC-like
HD HSPC
D
Down vs HD
Up vs Lineage
AML
B-ALL
T-ALL
408
196
171
672 genes up
in HSPC-like
350
24
20
14
102
131
31
E
PI3K inhibitor
MTOR inhibitor
VEGFR inhibitor
Phosphodiesterase inhibitor
HDAC inhibitor
Tubulin inhibitor
Dopamine receptor antagonist
Histamine receptor antagonist
EGFR inhibitor
FLT3 inhibitor
Topoisomerase inhibitor
PDGFR inhibitor
JAK inhibitor
Adrenergic receptor antagonist
Serotonin receptor antagonist
Calcium channel blocker
Bcr−Abl inhibitor
AKT inhibitor
BTK inhibitor
Aurora kinase inhibitor
Sodium channel inhibitor
KIT inhibitor
FGFR inhibitor
STAT inhibitor
T-ALL
B-ALL
AML
Leukemia Cell Lines #
Shifted HSPC->Lineage
F
HSPC-like
Lineage-like
Viability (%)
Concentration (nM)
Momelotinib
Birinapant
Ipatasertib
JQ1
Selinexor
Bortezomib
Palbociclib
Nelarabine
G
Venetoclax
S63845
H
HSPC-like
CLP-like
ETP-like
Pro/Pre-T-like
DP-like
Mature-T-like
BCL2
MCL1
% Expressed
Avg. Expression

Supplemental Figure 10. In silico and ex vivo screening of drug targets against HSPC-like blasts (related to Figure 6). Dataset used in this figure: scRNA-Seq in this study. A-C, Volcano plots showing the DEGs comparing HSPC-like blasts and normal HSPCs in three leukemia subtypes. **D,** Venn diagrams showing the removal of highly expressed genes (shadow) in normal HSPCs from up-regulated DEGs from HSPC-like vs lineage-like blasts comparisons. **E,** Top predicted drugs from LINCS1000 database. DEGs up-regulated in HSPC-like blasts and lineage-like blasts of different leukemias were input into LINCS1000 database. Drug treated leukemia cell lines were filtered for statistical significance (FDR < 0.1) and connectivity score (NCS > 0.8). Drugs were ranked by the number of leukemia cell lines with favorable transcriptomic shift after treatment (downregulation of HSPC-like DEGs, upregulation of lineage-like DEGs). **F,** Drug dose response curves showing the different response to JAKi (Momelotinib), PI3Ki (Ipatasertib), IAPi (Birinapant), BETi (JQ1), Nuclear Export inhibitor (Selinexor), NFKBi (Bortezomib), CDKi (Palbociclib), and DNA inhibitor (Nelarabine) between HSPC-like-derived and lineage-like-derived PDX cells. **G,** Drug response curves showing the different response to BCL2i (Venetoclax) and MCL1i (S63845) between HSPC-like-derived and lineage-like-derived PDX cells. **H,** Dot plot of BCL2 and MCL1 expression in assigned populations on T-cell development trajectory. Color represents scRNA-Seq Z-score. Dot size represents the percentage of expression in particular populations.



Supplemental Figure 11. *In silico* screening of de novo and TF targets against HSPC-like blasts (related to Discussion). Dataset used in this figure: scRNA-Seq in this study. A, Overall scores of top 20 TFs based on the *in silico* screening. **B,** Overall scores of top 30 genes without known drugs based on the *in silico* screening. **C,** DepMap essentiality scores for selected TFs and novel targets, including RUNX1, ZEB2, SOX4, BCL11A, MED12L, and CELF2 in leukemia cell lines and non-leukemia cell lines, respectively. P-values were computed using Student's t-test.

References

1. Chen C, Yu W, Alikarami F, et al. Single-cell multiomics reveals increased plasticity, resistant populations and stem-cell-like blasts in KMT2A-rearranged leukemia. *Blood*. 2021;
2. Nam AS, Kim K-T, Chaligne R, et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*. 2019;571(7765):355–360.
3. Hu J, Jarusiewicz J, Du G, et al. Preclinical evaluation of proteolytic targeting of LCK as a therapeutic approach in T cell acute lymphoblastic leukemia. *Sci. Transl. Med.* 2022;14(659):eabo5228.
4. Rowland L, Smart B, Brown A, et al. Ex vivo drug sensitivity imaging-based platform for primary acute lymphoblastic leukemia cells. *Bio Protoc.* 2023;13(15):e4731.
5. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29.
6. 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.
7. Zhu Q, Conrad DN, Gartner ZJ. deMULTIplex2: robust sample demultiplexing for scRNA-seq. *bioRxivorg*. 2023;
8. 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.
9. Xu J, Chen C, Sussman JH, et al. A multiomic atlas identifies a treatment-resistant, bone marrow progenitor-like cell population in T cell acute lymphoblastic leukemia. *Nat. Cancer*. 2024;
10. Tickle, T. I., Georgescu, C., Brown, M. & Haas, B. inferCNV of the Trinity CTAT Project (2019). *https://github.com/broadinstitute/inferCNV*. .
11. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* 2024;42(2):293–304.
12. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 2019;37(7):773–782.
13. Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling cell type abundance and expression in bulk tissues with CIBERSORTx. *Methods Mol. Biol.* 2020;2117:135–157.
14. Lee SHR, Yang W, Gocho Y, et al. Pharmacotypes across the genomic landscape of pediatric acute lymphoblastic leukemia and impact on treatment response. *Nat. Med.* 2023;29(1):170–179.
15. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*. 2017;14(10):975–978.
16. Morabito S, Miyoshi E, Michael N, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* 2021;53(8):1143–1155.
17. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Rep. Methods*. 2023;3(6):100498.
18. Zhu Q, Gao P, Tober J, et al. Developmental trajectory of pre-hematopoietic stem cell formation from endothelium. *Blood*. 2020;
19. Bravo González-Blas C, De Winter S, Hulselmans G, et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods*. 2023;20(9):1355–1367.
20. Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*. 2018;19(1):.
21. Brady SW, Roberts KG, Gu Z, et al. The genomic landscape of pediatric acute lymphoblastic leukemia. *Nat. Genet.* 2022;54(9):1376–1389.
22. Pölönen P, Di Giacomo D, Seffernick AE, et al. The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature*. 2024;632(8027):1082–1091.

23. Mehtonen J, Teppo S, Lahnalampi M, et al. Single cell characterization of B-lymphoid differentiation and leukemic cell states during chemotherapy in ETV6-RUNX1-positive pediatric leukemia identifies drug-targetable transcription factor activities. *Genome Med.* 2020;12(1):99.
24. Lambo S, Trinh DL, Ries RE, et al. A longitudinal single-cell atlas of treatment response in pediatric AML. *Cancer Cell*. 2023;41(12):2117-2135.e12.