

# Fusion oncoproteins and cooperating mutations define disease phenotypes in *NUP98*-rearranged leukemia

Masayuki Umeda<sup>1</sup>, Ryan Hiltenbrand<sup>1</sup>, Nicole L. Michmerhuizen<sup>1</sup>, Juan M. Barajas<sup>1</sup>, Melvin E. Thomas<sup>1</sup>, Bright Arthur<sup>1</sup>, Michael P Walsh<sup>1</sup>, Guangchun Song<sup>1</sup>, Jing Ma<sup>1</sup>, Tamara Westover<sup>1</sup>, Amit Kumar<sup>2</sup>, Petri Pölonen<sup>1</sup>, Cristina Mecucci<sup>3</sup>, Danika Di Giacomo<sup>1,4</sup>, Franco Locatelli<sup>5,6</sup>, Riccardo Masetti<sup>7,8</sup>, Salvatore N. Bertuccio<sup>9</sup>, Martina Pigazzi<sup>10</sup>, Shondra M. Pruett-Miller<sup>11,12</sup>, Stanley Pounds<sup>13</sup>, Jeffrey Rubnitz<sup>14</sup>, Hiroto Inaba<sup>14</sup>, Kyriakos P. Papadopoulos<sup>15</sup>, Michael J. Wick<sup>15</sup>, Ilaria Iacobucci<sup>1</sup>, Charles G. Mullighan<sup>1,2</sup>, Jeffery M. Kline<sup>1,2,\*</sup>

1 Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, US

2 Center of Excellence for Leukemia Studies (CELS), St. Jude Children's Research Hospital, Memphis, TN, US

3 Department of Medicine and Surgery, Hematology and Bone Marrow Transplantation Unit, University of Perugia, Perugia, Italy

4 Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy

5 Department of Onco-Haematology and Cell and Gene Therapy, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy

6 Department of Life Sciences and Public Health, Catholic University of the Sacred Heart, Rome, Italy

7 Pediatric Oncology and Hematology, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

8 Department of Medical and Surgical Sciences (DIMEC), University of Bologna, Bologna, Italy

9 IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

10 Department of Women's and Children's Health, Onco-hematology lab and clinic, University of Padova, Padova, Italy

11 Center for Advanced Genome Engineering, St. Jude Children's Research Hospital, Memphis, TN, US

12 Department of Cell and Molecular Biology, St. Jude Children's Research Hospital, Memphis, TN, US

13 Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, US

14 Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, US

15 XenoSTART/ The START Center for Cancer Research, San Antonio, TX, US

\*Corresponding Author

Jeffery M. Kline, MD, PhD

262 Danny Thomas Place, MS 342, D4047B, Memphis, TN 38105

Jeffery.Kline@STJUDE.ORG

Phone: 901-595-6807

Fax: 901-595-5947

## Abstract

Leukemias with *NUP98* rearrangements exhibit heterogeneous phenotypes correlated to fusion partners, whereas the mechanism responsible for this heterogeneity is poorly understood. Through genome-wide mutational and transcriptional analyses of 177 *NUP98*-rearranged leukemias, we show that cooperating alterations are associated with differentiation status even among leukemias sharing the same *NUP98* fusions, such as *NUP98::KDM5A* acute megakaryocytic leukemia with *RB1* loss or T-cell acute lymphoblastic leukemia with *NOTCH1* mutations. CUT&RUN profiling reveals that *NUP98* fusion oncoproteins directly regulate differentiation-related genes, with binding patterns also influenced by differentiation stage. Using *in vitro* models, we show *RB1* loss cooperates with *NUP98::KDM5A* by blocking terminal differentiation toward platelets and expanding megakaryocyte-like cells, whereas *WT1* frameshifts skew differentiation toward dormant lympho-myeloid primed progenitor cells and cycling granulocyte-monocyte progenitor cells. *NUP98::KDM5A* models with *RB1* or *WT1* alterations have different sensitivities to menin inhibition, suggesting cellular differentiation stage-specific resistant mechanism against menin inhibitors with clinical implications for *NUP98*-rearranged leukemia.

# Introduction

The prognosis of children with acute myeloid leukemia (AML) remains unsatisfactory with high rates of disease relapse<sup>1-3</sup>. This is partly attributed to pediatric AML subtypes associated with poor treatment response, such as *KMT2A* rearrangement (*KMT2Ar*)<sup>4,5</sup>, *UBTF* tandem duplications (*UBTF-TD*)<sup>6</sup>, *CBFA2T3::GLIS2*<sup>7,8</sup>, and *NUP98* rearrangement (*NUP98r*)<sup>5,9,10</sup>. Understanding the genetic background and disease mechanisms of these subtypes is needed to develop efficient treatment strategies based on biology and molecular mechanisms.

*NUP98r* with various fusion partners are known to be associated with specific hematologic malignancies<sup>11</sup>. For example, *NUP98::NSD1* is predominantly found in adolescents with AML, whereas *NUP98::KDM5A* is often found in acute megakaryocytic leukemia (AMKL) in young children<sup>5,8,9</sup>. Many other *NUP98r* with a variety of fusion partners have been reported in other hematological malignancies, such as T-cell acute lymphoblastic leukemia (T-ALL)<sup>12</sup>, myelodysplastic neoplasms (MDS)<sup>13</sup> or mixed phenotype acute leukemia (MPAL)<sup>14</sup>. Recent genomic studies focusing on *NUP98r* AML confirmed previously known associations of *NUP98::NSD1* with *FLT3* internal tandem duplications (ITD) and *WT1* mutations or *NUP98::KDM5A* with chromosome 13 loss involving the *RB1* locus<sup>8,10,15</sup>. However, more comprehensive and genome-wide studies on various *NUP98r* leukemias are needed to better understand the association of genomic features, including cooperating mutations, with disease phenotypes.

From a mechanistic standpoint, recent studies revealed that *NUP98* fusion oncoproteins (FOs) directly bind to gene loci expressed in *NUP98r* AML, such as *MEIS1*, *HOXA/B* clusters, and *CDK6*, with some similarities to AMLs with *NPM1* mutations or *KMT2A* rearrangements<sup>16-18</sup>. This binding is partly driven by intrinsically disordered regions of the N-terminal *NUP98*, which mediate transcriptional condensates<sup>19,20</sup>, and DNA/chromatin binding domains of the C-terminal fusion partners, leading to target gene activation. This knowledge has mostly been derived from mouse or non-hematopoietic human models due to a lack of appropriate human *NUP98r* leukemia models. An understanding of molecular mechanisms of leukemogenesis by *NUP98* FOs in human hematopoietic cells, including the association of genomic binding with histone modification and transcriptional profiles, is necessary for developing mechanism-based novel therapeutics for *NUP98r* leukemias.

In this study, we performed genome-wide genetic characterization of 177 pediatric and young adult *NUP98r* hematologic malignancies including AML and T-ALL, using RNA sequencing (RNAseq) and whole genome/exome sequencing (WGS/WES), to reveal phenotype-specific mutational profiles. These data were corroborated by single cell transcriptomics and epigenetic profiling by CUT&RUN<sup>21</sup>, revealing *NUP98*-FO binding profiles associated with fusion partners and disease phenotypes of patient samples. Functional characterization of recurrent gene alterations using cord blood CD34+ cell models showed the *NUP98* fusion- and phenotype-dependent impact on cell growth and transcriptional profiles associated with drug responses, which can be a basis for overcoming these refractory diseases.

## Results

### *Transcriptional heterogeneity of pediatric NUP98r leukemia*

We first established a pediatric and young adult *NUP98r* leukemia cohort with 185 samples from 177 patients (median age: 8.6, range: 0.4-28) by collecting new cases and integrating data across published cohorts<sup>3,6,8,10,13,22-30</sup> (Fig.S1A, Supplementary Table 1), including 8 patients with data at multiple time points (Fig.1A). AML was the dominant disease (n=162), followed by T-ALL (n=18), therapy-related myeloid neoplasms (t-MN, n=4), and MDS (n=1), reflecting the heterogeneity of *NUP98r* hematologic neoplasms. We obtained RNAseq data for all samples (n=185) and DNA data for 91 samples (WGS:43, WES:16, both:32, Supplementary Tables 2-9). Fusion genes were identified by CICERO<sup>31</sup> and manual inspection of sequencing data, confirming *NUP98* fusion transcripts in all cases.

Among fusion partners, *NSD1* was the most frequent (n=92, 51.7%), followed by *KDM5A* (n=45, 25.3%, Fig.1B, Fig.S1B, Supplementary Table 8). These fusions were enriched in specific age groups (Fig.1C). *RAP1GDS1* (n=11, 6.2%) was the third most common partner, with nine *NUP98::RAP1GDS1* were found in T-ALL cases. Most of the other partners fall into functional categories of epigenetic regulators (n=12, 6.7%) or homeobox proteins (n=9, 5.1%).

**A.** Details of *NUP98*-rearranged (*NUP98r*) leukemia samples (n=185, **left**) and analytical pipelines (**right**). **B.** Numbers and functional annotations of fusion partners in the study cohort. Colors indicate protein functional groups. **C.** Age distribution related to fusion partners and disease types. Colors indicate disease types. **D.** UMAP (Uniform Manifold Approximation and Projection) plots of transcriptional cohort (n=2,321) colored according to leukemia subtypes (**left**) and *NUP98* fusion partners (**mid**) and enrichment of fusion partners in transcriptional clusters (**right**). The shapes of dots indicate disease types (circles-acute myeloid leukemia: AML, triangles-acute lymphoblastic leukemia: ALL), and colors in the heatmap indicate enrichment of fusions in each cluster, asterisks indicate statistically significant adjusted *P*-values from two-sided Fisher's exact tests and the Benjamini-Hochberg adjustment (\* $<0.05$ , \*\* $<0.01$ , \*\*\* $<0.001$ , black: enriched, blue: exclusive). In **A** and **C**, lines of the box plots represent 25% quantile, median, and 75% quantile and the upper whisker represents the higher value of maxima or 1.5 x interquartile range (IQR), and the lower whisker represents the lower value of minima or 1.5 x IQR. Abbreviations. tMN: therapy-related myeloid neoplasm, MDS/MPN: myelodysplastic syndrome/ myeloproliferative neoplasms, AEL/AMKL: acute erythroid leukemia/ acute megakaryocytic leukemia, ETP: Early T-cell precursor ALL. Abbreviations in leukemia subtypes are found in Table S10.

The majority of the partners with epigenetic regulatory functions, including NSD1 and KDM5A, had plant homeodomain (PHD) fingers in the retained C-terminus, suggesting structural redundancy among fusion partners (Fig.S1C). Functions of the remaining 6 partners (8 patients) were broad, ranging from RNA-binding (DDX1/10), maintenance of DNA structure (HMGB3 and CEP295), or without known function (LNP1 and CCDC28A), but all these partners had domains with multiple  $\alpha$ -helices with linker sequences similar to homeobox domains or PHD. *NUP98* breakpoints were heterogeneous, ranging from exon 7-15, although fusion partners were strongly associated with specific exons (*RAP1GDS1*-exon 11, *NSD1*-exon 12, and *KDM5A*-exon 13, Fig.S1B).

Appreciating the known heterogeneity of the disease phenotypes and transcriptional profiles of *NUP98r* leukemia, we performed transcriptional analyses with other subtypes of AML<sup>5</sup> (n=816) and T-ALL<sup>12</sup> (n=1,320) as a background cohort (Fig.S2A, Supplementary Table 10). UMAP (Uniform Manifold Approximation and Projection) and unsupervised clustering revealed 10 clusters driven by driver alterations and differentiation status. *NUP98r* malignancies were distributed to 6 of these clusters (Fig.1D, Fig.S2B-D). The majority of *NUP98::NSD1* samples were clustered with AML with *NPM1* mutations or *UBTF*-TD showing high *HOXA/B* gene expression (the *HOXB* cluster). By contrast, samples with *NUP98::KDM5A* were predominantly found in a cluster with other AMKL subtypes (the AMKL cluster), although rare cases clustered with subtypes with *KMT2A*-rearranged and *KAT6A*-rearranged AML with *HOXA* expression (the *HOXA* cluster) or with mature ALL (the *TLX* cluster). *NUP98::RAP1GDS1* samples were mainly clustered with other early T-cell precursor-like ALL (the ETP-like cluster), while they were also found in the *HOXB*, *TLX*, or immature AML clusters. Although *HOXA-B* gene expression is recognized as a hallmark of *NUP98r* leukemia, we observed that *NUP98::NSD1* samples lack posterior *HOXB* (*HOXB6-9*) expression and immature/T-ALL cases lack *HOXB* or both *HOXA-B* expression (Fig.S2E). These data show that *NUP98r* leukemias are heterogeneous and even the same fusion partners can contribute to variable phenotypes of these hematopoietic malignancies.

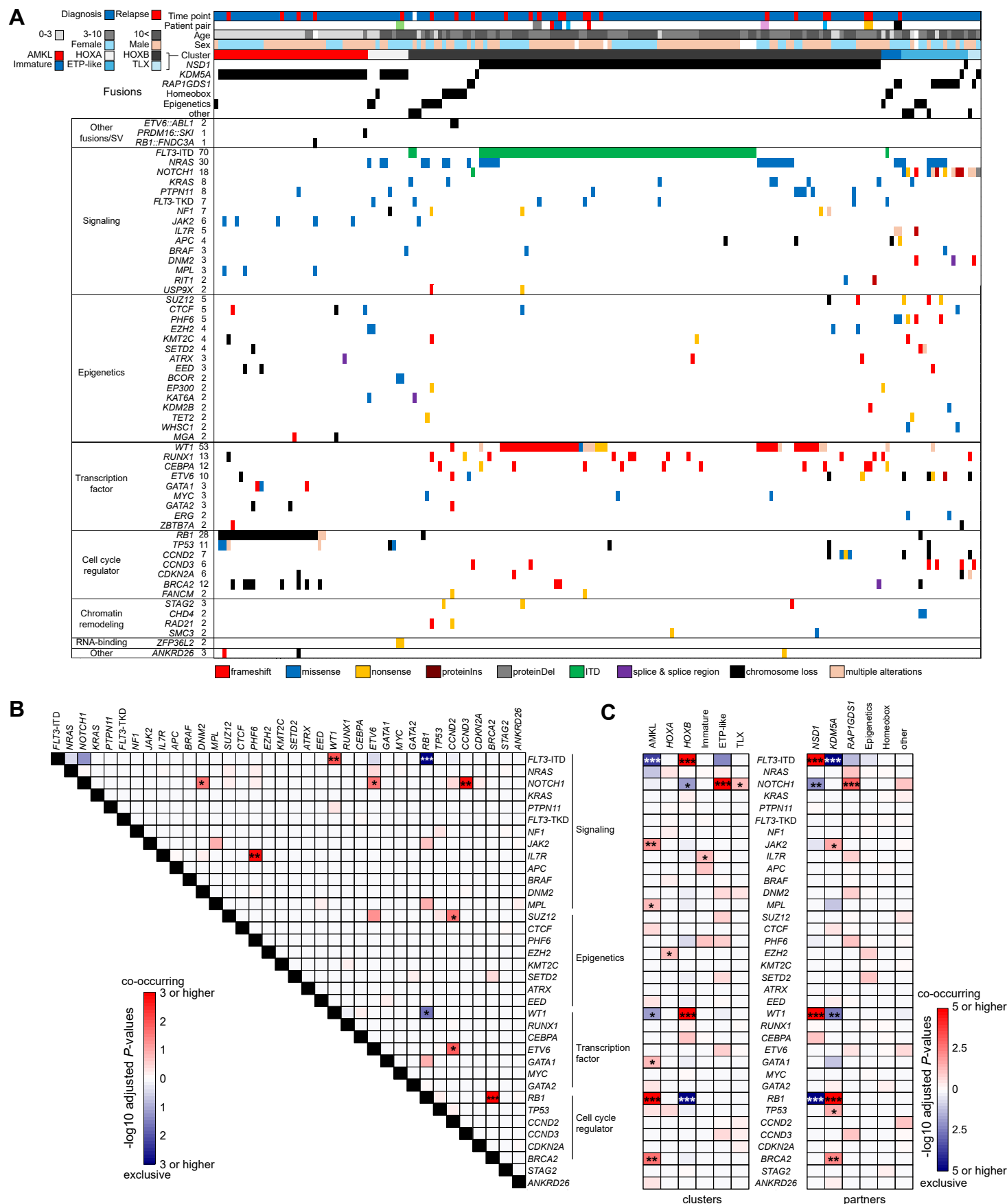
#### Mutational background of *NUP98r* leukemia associated with disease phenotypes

To assess genetic factors contributing to the transcriptional heterogeneity of *NUP98r*, we investigated the genomic landscape using DNA data and an RNAseq-based pipeline<sup>5,6</sup> (see Methods), including single nucleotide variants (SNV), insertions and deletions (indel), ITD, copy number variations (CNV), and fusion genes other than *NUP98* rearrangements (Supplementary Tables 5-9). Among 53 recurrently altered genes, the most frequent alteration was *FLT3*-ITD (n=70, 37.8%), followed by *WT1* (n=44, 23.8%) and *NRAS* mutations (n=30, 16.2%, Fig.2A-C). Chromosome 13 deletions involving *RB1*<sup>8,10</sup> were found predominantly in *NUP98::KDM5A* cases (n=23, 14.1%), with deleted regions extending in both 3' and 5' directions of *RB1*, often affecting the tumor suppressor *BRCA2* (Fig.S3A). We also identified *RB1* loss in cases with only RNAseq by assessing for allelic imbalance of single nucleotide polymorphisms (SNPs) within and around the *RB1* locus; however, *RB1* loss could still be underestimated as AMKL cases without detectable CNV also showed low *RB1* RNA expression (Fig.S3B-C). *RB1* loss was associated with high expression of stem cell markers (*GATA2*, *CD96*, and *BMI1*) among *NUP98::KDM5A* AMKL samples, and gene set enrichment analysis (GSEA) showed enrichment of genes involving AMKL phenotypes (Fig.S3D-F). GSEA also showed changes in genes related to cell growth and mitochondrial function, which are consistent with the known function of *RB1*/E2F<sup>32,33</sup> (Fig.S3F). Additionally, genome-wide profiling revealed *JAK2* (n=6, 3.2%) mutations enriched in *NUP98::KDM5A* and the AMKL cluster, while *GATA1* (n=3, 1.6%) and *MPL* (n=3) mutations were associated only with the AMKL cluster (Fig.2C), all of which were frequently found in non-Down syndrome AMKL<sup>8</sup>.

*FLT3*-ITD and *WT1* were highly co-occurring and enriched in *NUP98::NSD1* AMLs (Fig.2A-C). Variant allele frequencies (VAFs) of *WT1* and *FLT3*-ITD were generally high both at diagnosis and relapse (Fig.S4A-C). *FLT3*-ITD and *RAS*-related mutations were less likely to co-occur, especially among *NUP98::NSD1* samples, and VAFs of *RAS* mutations were more variable (Fig.2B, Fig.S4A-B), indicating that *RAS* mutations could exist in independent clones or subclonal to *FLT3*-ITD<sup>34</sup>. Differentially expressed gene (DEG) analysis showed high expression of immunity-related genes, such as *MRC1* (CD206) or *IL2RA* (CD25), in *FLT3*-ITD+ *NUP98::NSD1* AML, consistent with previous finding in *FLT3*-ITD+ AML, where these markers are associated with inferior outcomes<sup>35,36</sup>. GSEA



Fig.2



## Figure 2. Mutational background of *NUP98r* leukemia associated with disease phenotypes

**A.** Genetic profiles of *NUP98r* samples in the cohort. Colors indicate patient annotations (**top**) and types of gene alterations (**bottom**). **B.** Co-occurrence and mutual exclusivity among recurrent alterations ( $n \geq 3$ ). **C.** Enrichment of somatic alterations in transcriptional clusters (**left**) and fusion partners (**right**). In **B** and **C**, colors indicate adjusted *P*-values by two-sided Fisher's exact tests and the Benjamini-Hochberg adjustment (red: co-occurring, blue: mutually exclusive), and asterisks indicate statistically significant values (adjusted *P*-values \* $<0.05$ , \*\* $<0.01$ , \*\*\* $<0.001$ ). Annotations of genes in mutational heatmaps depend on known general functions.

confirmed known enrichment of inflammatory-related genes<sup>37</sup> (Fig.S4D-E). Samples with *WT1* mutations had significantly high expression of genes such as *DNTT* (TdT: Terminal deoxynucleotidyl transferase), which is a known marker for lymphoid-myeloid primed progenitors (LMPP)<sup>38,39</sup>, and gene sets related to mitochondrial function and translation by ribosomes. These data suggest that secondary alterations in *NUP98::NSD1* AMLs can influence expression profiles and differentiation status.

Clinically, all 12 *NUP98::RAP1GDS1* leukemias were either T-ALL and immature AML (FAB M0-1). Those with *NOTCH1* mutations ( $n=7$ , 58.3%) were predominantly found in the *TLX* and ETP-like clusters (6/7, 85.7%) and associated with T cell-related expression profiles, whereas *NRAS* mutations were mainly found in the ETP and immature AML clusters (Fig.S4F-H). Mutational profiling of samples in T-ALL/immature AML clusters showed *NOTCH1* mutations enriched in the ALL clusters, including those in *NUP98::KDM5A* ALL, whereas *IL7R* mutations were enriched in immature AML clusters (Fig.2C). Thus, among immature *NUP98r* leukemias, *NOTCH1* mutations likely contribute to the establishment of T-ALL phenotypes<sup>40</sup>. These data show the strong association among fusion partners, cooperating mutations, and disease phenotypes of *NUP98r* malignancies.

### Variety of cellular hierarchies in *NUP98r* leukemia associated with fusions and cooperating alterations

Recent research showed associations between differentiation status or cellular hierarchies with outcomes or drug responses in AML<sup>41-43</sup>. To investigate the differentiation status of *NUP98r* leukemia, we performed single cell RNAseq (scRNAseq) on 9 representative samples from the cohort and obtained publicly available data from one case with diagnosis-relapse paired (D-R) samples of *NUP98::NSD1* AML (PAXLWH)<sup>43</sup> (Fig.3A, Fig.S5A, Supplementary Table 11). UMAP analysis showed distinct clusters for each patient sample indicating unique expression profiles of each tumor, whereas T-cell or B-cell clusters consisted of cells from multiple patients reflecting normal lymphocytes, as reported previously<sup>43,44</sup> (Fig.3B-C, Fig.S5B). To infer the differentiation stages compared to normal hematopoiesis, we projected scRNAseq data onto reference normal bone marrow and thymocyte scRNAseq data<sup>44</sup>, which revealed unique distributions of tumor cells in each sample (Fig.3D-F, Fig.S5C). Three *NUP98::NSD1* samples (SJAML015363\_D5, SJAML016582\_D1, and SJAML061252\_R1) showed various cellular hierarchies ranging from LMPP-like to monocyte-like cells (Fig.3E-F). SJAML001441 D-R pair (*FLT3*-ITD+/WT1+) had enriched monocyte and granulocyte-macrophage progenitor (GMP)-like cells at diagnosis with increasing LMPP-like cells at relapse with an additional *WT1* mutation. PAXLWH D-R pair (*FLT3*-ITD+) showed homogenous LMPP-like cells at diagnosis but increased GMP or monocytic populations at relapse with an acquired *RUNX1* mutation, suggesting different patterns of relapse of these *NUP98::NSD1* cases. DEG analyses between cells from *NUP98::NSD1* patients and normal counterparts showed aberrantly high expression of *HOXA/B* genes along with low expression of marker genes (e.g., *S100A12* in monocyte and *ELANE* in GMP) in leukemic samples (Fig.S5D), confirming aberrant leukemia expression profiles. Hierarchies of bulk RNA samples inferred by deconvolution using CIBERSORT<sup>45</sup> revealed heterogeneous hierarchies among the entire cohort and variable patterns among both D-R pairs and somatic mutations (Fig.3G, Fig.S5E). Relapse samples commonly showed low monocyte and high LMPP signatures, indicating expansion of stem-like populations at relapse in a subset of *NUP98::NSD1* AML as reported in other subtypes<sup>43</sup>.

A *NUP98::KDM5A* AMKL case with *RB1* loss showed a cluster corresponding to megakaryocyte/erythrocyte progenitors (MEP) in the normal reference (Fig.3B,E-F). DEG analysis of the AMKL cluster with normal MEP revealed aberrant expressions of *HOXB* genes and immune response-related *IFI27* in AMKL cells, which were co-



### Figure 3. Variety of cellular hierarchies in *NUP98r* leukemia

**A.** Strategies for single cell RNAseq (scRNAseq) and deconvolution of bulk RNAseq data. **B.** UMAP plots of patient samples colored by sample source (**left**) and transcriptional clusters (**right**). **C.** Enrichment of cells in each cluster indicated by colors and sizes (**left**) and marker gene expression indicated by colors (averaged expression) and size (ratio of expressing cells: count>0). **D.** UMAP plot of reference bone marrow and thymocyte scRNAseq data, colored according to cell labels from the original reference data<sup>44</sup>. **E.** Distribution of patient sample scRNAseq on the reference data inferred by MapQuery function in Seurat package. Cells in normal hematopoietic cell clusters were excluded. Cells are colored according to the cell density on UMAP. Cooperating mutations found in bulk samples are also shown. **F.** Enrichment of cells with each cell label inferred by Seurat, indicated by colors and sizes. **G.** Cellular component of bulk RNA samples (n=185) inferred by CIBERSORT using signature matrix derived from reference scRNAseq data. Bars are colored by cell populations in each sample. Abbreviations. HSPC: hematopoietic stem and progenitor cell, LMPP: Lympho-myeloid primed progenitor, GMP: Granulocyte-monocyte progenitor, cDC: classic dendritic cell, pDC: plasmacytoid dendritic cell, CLP: common lymphoid progenitor, DP: CD4-CD8 double-positive T-cell, NK: natural killer T-cell, MEP: Megakaryocyte-erythroid progenitor

expressed at the single cell level (Fig.S6A-B). *NOTCH1*<sup>+</sup> *NUP98::KDM5A* ALL cells were enriched at Pro- and Pre-T cell stages with aberrant expression of hemoglobin or *HOXB* genes (Fig.S6A).

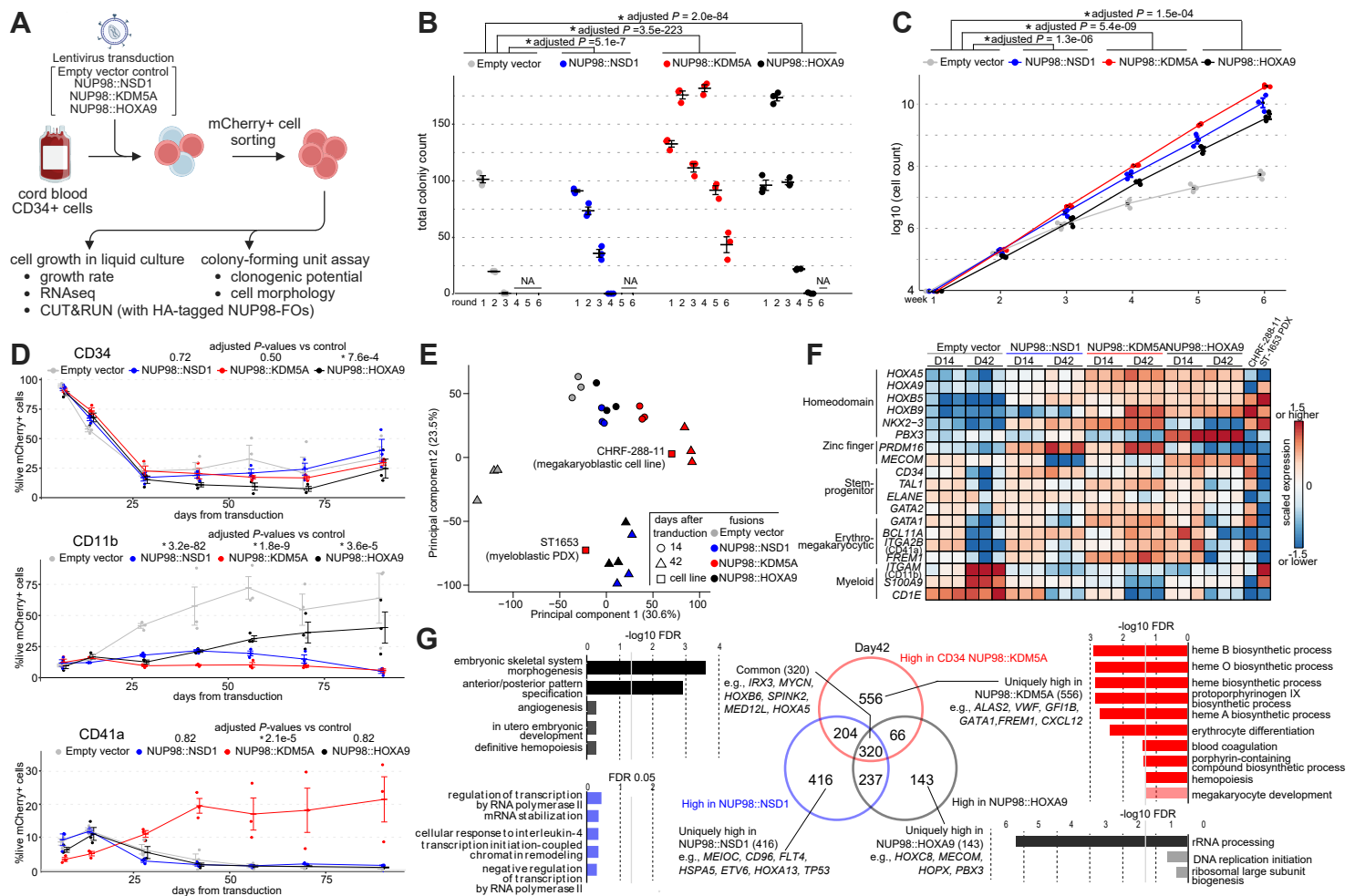
*NOTCH1*<sup>+</sup> *NUP98::RAP1GDS1* cells distributed broadly at the LMPP-ETP-Pro-T stages whereas *NUP98::RAP1GDS1* AML cells were enriched in the LMPP cluster (Fig.3E-F, Fig.S6C). Deconvoluted bulk *NUP98::RAP1GDS1* samples in immature clusters showed LMPP-CLP (common lymphoid progenitor) signatures, while those in the ALL clusters showed high Pro-Pre-T cell signatures, suggesting that the differentiation block of *NUP98::RAP1GDS1* occurs at various stages during T-cell development from hematopoietic stem cells (HSC), possibly affected by cooperating mutations<sup>40</sup> (Fig.S6C-E). Two *NUP98::KDM5A* ALL cases showed higher Pre-T signals than *NUP98::RAP1GDS1* cases (Fig.S6F). These observations indicate that *NUP98::KDM5A* could contribute to the development of more mature ALL than *NUP98::RAP1GDS1*, despite both cooperating with *NOTCH1* mutations.

### Cord blood CD34 models recapitulate phenotypes of *NUP98r* leukemia

There are limited human *NUP98r* leukemia models that recapitulate the disease phenotypes<sup>40,46-49</sup>. To fill this gap, we tested the transforming potential of various *NUP98* FOs by transducing cord blood CD34<sup>+</sup> cells (cbCD34) with lentiviral particles encoding major *NUP98*-FOs (pediatrics-*NUP98::NSD1* and *NUP98::KDM5A*, adult-*NUP98::HOXA9*). Expression of these *NUP98*-FOs increased clonogenic potential in methylcellulose and enhanced cell growth in liquid culture compared with an empty vector control (Fig.4A-C), despite a gradual decrease in CD34 expression. Flow cytometry analyses showed low CD11b expression in all conditions compared with the control, whereas only *NUP98::KDM5A* showed high CD41a expression, indicating preferential differentiation toward megakaryocytes (Fig.4D, Fig.S7A-B). We performed RNAseq from liquid cultures at multiple time points and observed dynamic changes of transcriptional status unique to each FO (Fig.4E-G, Fig.S7C-E). Consistent with patient samples, *NUP98*-FO-transduced conditions showed high *TAL1* and *MYCN* expression in addition to upregulation of *HOXA/B* genes, whereas control conditions showed reduced *TAL1* and increased *ITGAM* (CD11b) expression, suggesting myeloid differentiation (Fig.4F). We identified 320 common genes with high expression in conditions with *NUP98* FOs, which enriched homeobox genes (Fig.4G). Notably, we also identified uniquely high genes with each FO related to specific functions, such as *GFI1B* or *GATA1* related to erythromegakaryocytic differentiation in cells with *NUP98::KDM5A*, consistent with surface markers. Comparison with the patient cohort showed clustering of cbCD34 models with patient samples with similar fusions/phenotypes (Fig.S7E). Two cell lines with *NUP98::KDM5A*, CHRF-288-11<sup>46</sup> (megakaryocytic) and ST1653<sup>50</sup> (myelomonocytic, patient-derived xenograft: PDX) also showed unique patterns of expression (e.g., *ITGA2B* and *CD34* in CHRF-288-11, and *ITGAM* in ST1653), confirming the heterogeneity of leukemia with *NUP98::KDM5A* (Fig.4E-F, Fig.S7E).



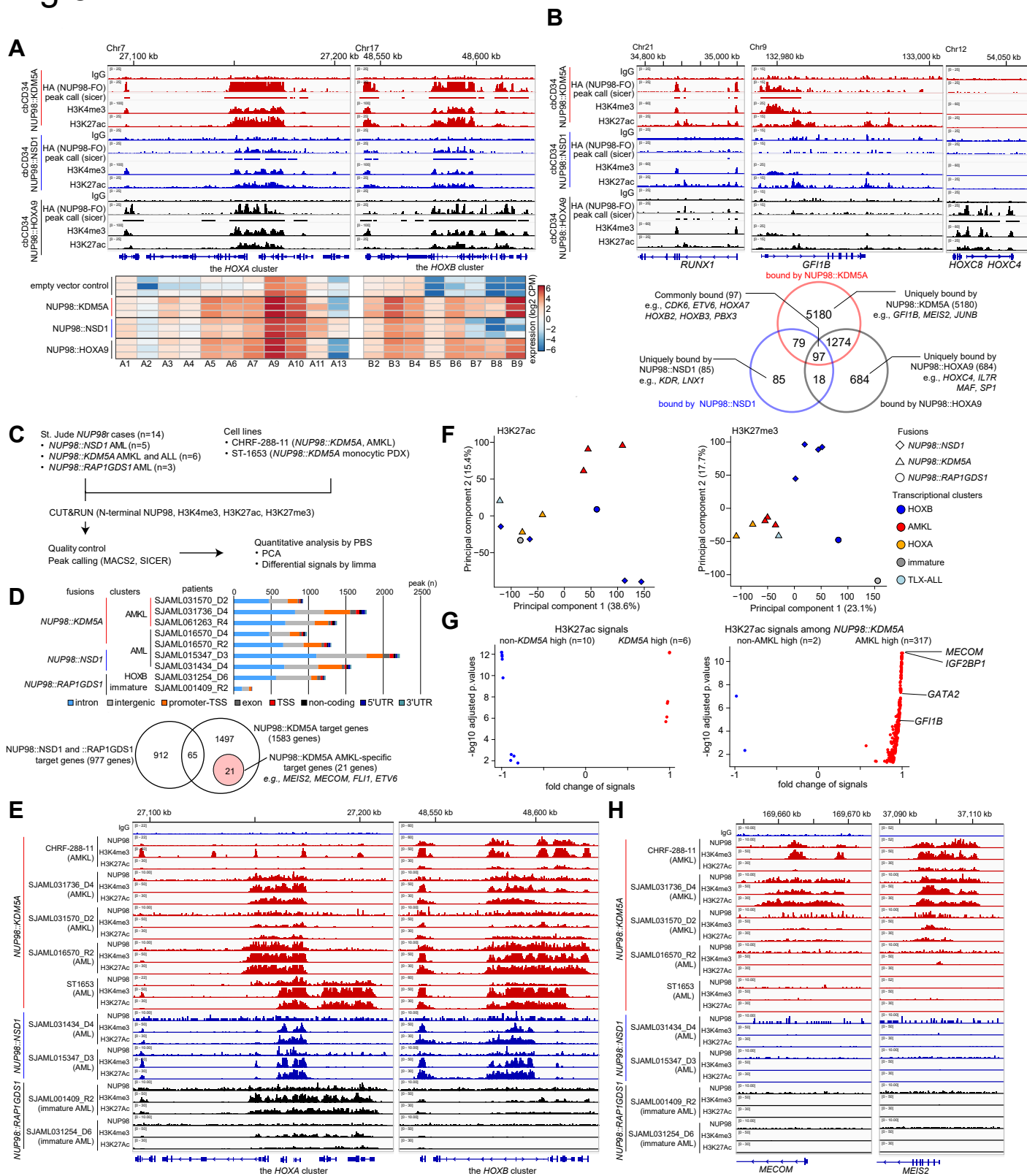
Fig.4



**Figure 4. Cord blood CD34 models recapitulate phenotypes of NUP98r leukemia**

**A.** Experimental schema using cord blood CD34+ cell models (cbCD34). **B.** Colony-forming unit assays of cbCD34 models with empty control vectors or NUP98::NSD1, NUP98::KDM5A, or NUP98::HOXA9-expressing vectors. **C.** Cell growth assays of cbCD34 models in liquid culture. **D.** Flow cytometric analysis of cbCD34 models in liquid culture (top: CD34+, mid: CD11b+, bottom: CD41a+ population ratio: % in mCherry+ live cells). **E.** Principal component analysis (PCA) of RNAseq data from liquid culture. Colors indicate NUP98 fusions, and shapes indicate days after transduction. **F.** Heatmap showing expression of representative genes related to stemness or differentiation of hematopoietic cells. Colors of cells indicate expression levels normalized among samples, and genes are annotated on the left. **G.** Comparison of differentially expressed genes (DEGs) in each cbCD34 model compared with empty vector controls at day 42. Venn diagram showing overlaps of highly expressed genes in each model (mid) and GO term analyses of shared or specific DEGs are shown (left, right). Data was obtained from three biological replicates (different lots of cord blood). In B-D, statistical tests were performed by generalized linear mixed effect model with Poisson (B) and Gaussian (C-D) distributions followed by comparison with empty vector control and the Benjamini-Hochberg adjustment, asterisks indicating adjusted  $P$ -values  $< 0.05$ . Error bars indicate mean  $\pm$  s.e.m. Abbreviations: FO: fusion oncoprotein, PDX: patient-derived xenograft, FDR: false-discovery rate.

Fig.5



## Figure 5. Differential gene regulation by NUP98 fusion oncoproteins

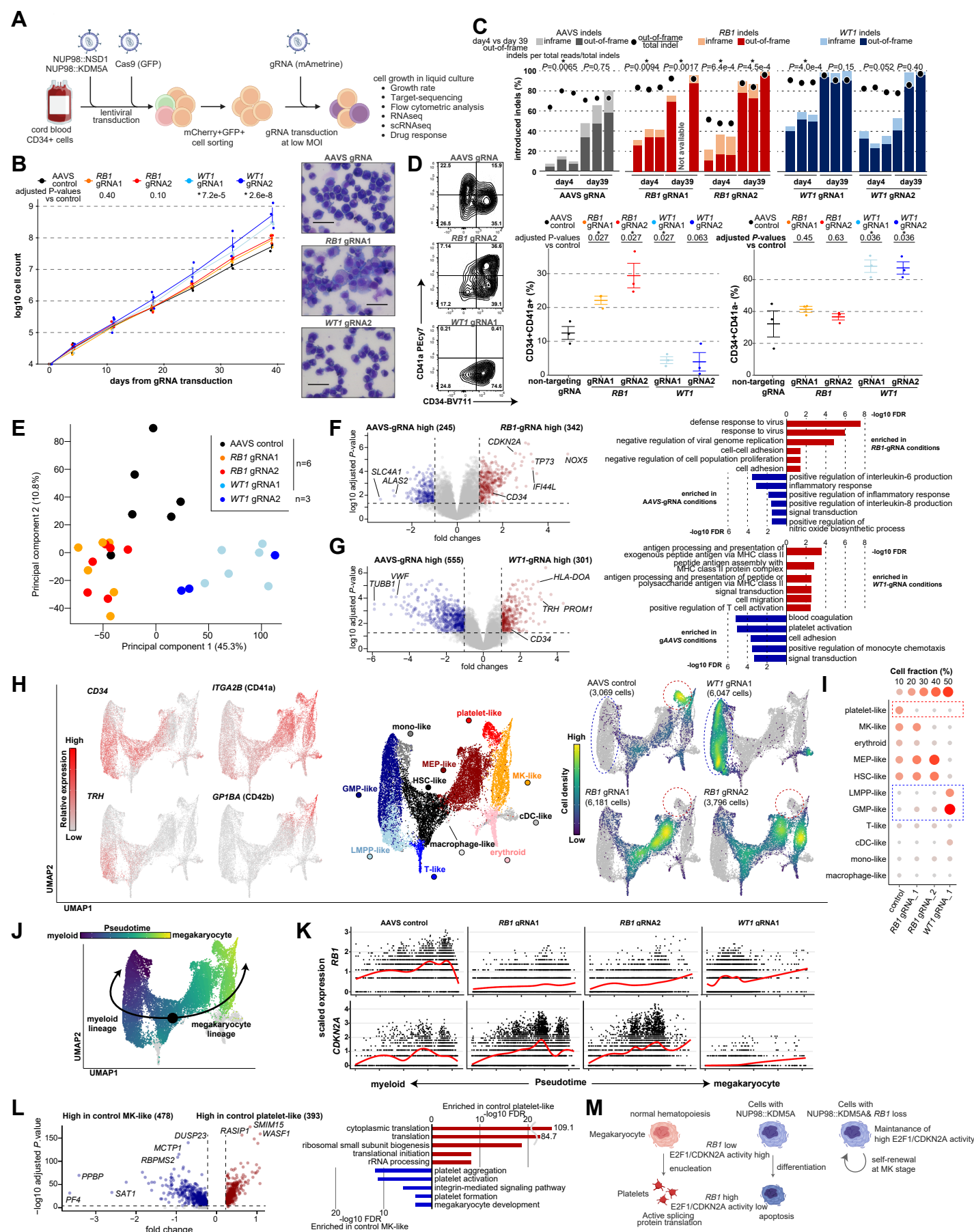
**A.** IGV tracks of the *HOXA-B* clusters from CUT&RUN using HA, H3K4me3, and H3K27ac antibodies in HA-tagged *NUP98r* cbCD34 models (**top**: HA-NUP98::KDM5A-red, HA-NUP98::NSD1-blue, HA-NUP98::HOXA9-black) and heatmap showing expression levels of *HOXA-B* genes (**bottom**). **B.** IGV tracks of differentiation-related gene loci (**top**: *RUNX1*, *GF1B*, and *MECOM*) and Venn diagram showing overlap of protein-coding genes with annotated peaks (**bottom**: FDR<0.00001). **C.** CUT&RUN strategy from primary patient samples or *NUP98::KDM5A* cell lines (CHRF-288-11 and ST1653). **D.** Counts of peaks from the N-terminus NUP98 antibody in primary samples (**top**) and overlaps of target genes among non-NUP98::KDM5A and NUP98::KDM5A. NUP98::KDM5A AMKL-specific 21 target genes are highlighted. Colors indicate peak annotations. **E.** IGV tracks of the *HOXA-B* cluster from CUT&RUN using N-terminal NUP98, H3K4me3, and H3K27ac antibodies in primary leukemia samples and *NUP98::KDM5A* cell lines. **F.** PCA of genome-wide PBS (probability of being signals) scores of H3K27ac (**left**) and H3K27me3 (**right**) from primary samples. Colors indicate expression clusters and shapes indicate fusion partners. **G.** Differential signal analysis using H3K27ac PBS scores between *NUP98::KDM5A* and other (*NUP98::NSD1* and *NUP98::RAP1GDS1*) samples (**left**) and *NUP98::KDM5A* AMKL and non-AMKL (**right**) calculated by limma followed by the Benjamini-Hochberg adjustment. Only regions with significant enrichment (adjusted *P*-values < 0.05) are shown. **H.** IGV tracks of the *MECOM* and *MEIS2* gene loci from CUT&RUN using N-terminal NUP98, H3K4me3, and H3K27ac antibodies in primary leukemia samples and *NUP98::KDM5A* cell lines.

### 1 Differential gene regulation by NUP98 fusion oncoproteins

2 To further study how NUP98-FOs drive leukemogenesis, we profiled the genomic binding of N-terminal HA-tagged  
3 NUP98-FOs (Fig.S8A), using anti-HA antibody in cbCD34 models along with histone modifications (H3K4me3 and  
4 H3K27ac) and menin using CUT&RUN<sup>21</sup> (Fig.5A-B, Fig.S8B-D). These studies revealed FO-binding peaks enriched  
5 in the promoter and intronic regions of protein-coding genes with high reproducibility among biological replicates  
6 (Fig.S8C). All three FOs showed binding to the *HOXA-B* cluster genes, whereas the binding patterns along the  
7 anterior-posterior axes were unique to each fusion, corresponding to RNA expression, active histone modifications,  
8 and menin binding as previously reported using different models<sup>16-19</sup> (Fig.5A, Fig.S8D). Comparisons of protein-  
9 coding genes with FO-binding peaks showed 97 common target genes for all three NUP98-FOs, including genes  
10 involved in leukemogenesis and hematopoiesis (e.g., *RUNX1*, *CDK6*). However, each FO also showed specific  
11 targets (NUP98::KDM5A: 5180 genes, NUP98::HOXA9: 684 genes, NUP98::NSD1: 85 genes, Fig.5B). Notably,  
12 NUP98::KDM5A uniquely showed peaks on *GF1B* and *MEIS2*, which are megakaryocyte/erythroid differentiation  
13 regulators<sup>51,52</sup>, indicating that direct regulation of differentiation-related genes could drive phenotypes of  
14 *NUP98::KDM5A* AMKL.

15 We next sought to profile endogenous NUP98-FO binding without epitope tags in patient samples or cell lines. We  
16 tested antibodies targeting the N-terminus of NUP98 in cbCD34, which showed similar binding patterns albeit with  
17 weaker signals yielding fewer peaks than HA-antibody, possibly underestimating weak FO-binding (Fig.S9A). N-  
18 terminus NUP98 antibody for CHRF-288-11 and ST1653 revealed NUP98 signals at genes highly expressed in  
19 these cell lines, such as *HOXA/B* genes and *MEIS1*, with menin co-localization (Fig.S9B). Contrarily, weak or low  
20 signals were observed in *RUNX1::RUNX1T1* AML (lacking *HOX* gene expression) or *UBTF*-TD AML (similar to  
21 *NUP98::NSD1* AML)<sup>6</sup>. We then applied CUT&RUN using antibodies for N-terminus NUP98 and histone  
22 modifications (H3K4me3, H3K27ac, H3K27me3) for 14 *NUP98r* patient samples covering *NUP98::NSD1*,  
23 *NUP98::KDM5A*, and *NUP98::RAP1GDS1* (Fig.5C, Fig.S9C, Supplementary Table 12). N-NUP98 antibodies  
24 revealed target genes for each FO, including *HOXA* or *HOXB* clusters with heterogeneous patterns even within the  
25 same NUP98-FO groups (Fig.5D-E). Also, N-NUP98 antibodies showed NUP98::KDM5A-specific targets (1,518  
26 genes), whereas 21 genes, including *MEIS2* or *MECOM*, were *NUP98::KDM5A* AMKL-specific targets. Genes with  
27 H3K4me3 peaks showed a broad overlap among samples, whereas 61 genes, including *IGF2BP3* associated with  
28 megakaryocyte development<sup>53</sup>, were unique to *NUP98::KDM5A* AMKL (Fig.S9D). For H3K27ac and H3K27me3  
29 signals with broad peaks that could span multiple genes, we applied the PBS<sup>54</sup> (probability of being signal) approach  
30 to quantify the genome-wide signals, which revealed that genome-wide H3K27ac status correlated with disease  
31 phenotypes, whereas H3K27me3 status was associated with fusion partners (Fig.5F). Comparisons of  
32 *NUP98::KDM5A* with the other fusions (*NUP98::NSD1* and *NUP98::RAP1GDS1*) showed limited regions with  
33 differential H3K27ac signals (Fig.5G). In contrast, comparison between *NUP98::KDM5A* AML and AMKL showed  
34 enrichment of *MECOM*, *IGF2BP1*, or *GATA2* regions in *NUP98::KDM5A* AMKL, indicating phenotype-specific  
35 epigenetic status (Fig.5G-H, Fig.S9E). These data suggest both the C-terminus of FOs and differentiation status  
36 can affect FO-binding patterns to establish heterogeneous epigenetic and expression profiles in *NUP98r* leukemia.

Fig.6





## Figure 6. Functional characterization of recurrent somatic alterations in NUP98r leukemia

**A.** Experimental schema of induction of cooperating alterations (*RB1*, *WT1*) in cbCD34/Cas9 models. **B.** Cell growth assays of cbCD34/Cas9 NUP98::KDM5A with gRNAs targeting the AAVS, *RB1*, or *WT1* loci (**left**), and cytospin of cells on day 35 (**right**). **C.** Induction rates of indel (insertions and deletions) at day 4 and 39 in each condition. Bars represent fractions of indel rates in all target sequence reads, and dots represent out-of-frame indel ratio among total indels. **D.** Flow gating (**left**), CD34+ CD41a+ positivity (**mid**), and CD34+ CD41a- (**right**) among mCherry+ GFP+ mAmetrine+ live cells. **E.** PCA of RNAseq of gRNA-transduced NUP98::KDM5A cbCD34 models at day 35. **F.** DEG analysis between AAVS controls and *RB1*-gRNA conditions (**left**) and gene ontology (GO) term analysis of DEGs (**right**). Colors indicate DEGs and GO terms (red: high in *RB1*-gRNA conditions, blue: low in *RB1*-gRNA conditions). **G.** DEG analysis between AAVS controls and *WT1*-gRNA conditions as shown in **F.** **H.** UMAP plots of scRNAseq data from gRNA-transduced NUP98::KDM5A cbCD34 models at day35, showing marker gene expression (**left**), annotated clusters (**mid**), and cell distributions among conditions (**right**). Colors in plots indicate relative expression levels, clusters, and cell density, respectively. **I.** Enrichment of cells with each cluster indicated by colors and sizes. **J.** Pseudotime along myeloid (HSC→GMP→monocytes) and platelet (HSC→MEP→MK→platelet) trajectories. Colors represent pseudotime scores of each single cell inferred by Slingshot. **K.** *RB1* (**top**) and *CDKN2A* (**bottom**) expression along the pseudotime axis in each condition with red curves show average expressions. **L.** DEG analysis between the platelet-like and MK-like clusters in the AAVS-control condition (**left**) and GO term analysis (**right**) of genes high in the platelet-like cluster (red) and the MK-like cluster (blue). **M.** Schematics illustrating platelet differentiation in normal hematopoiesis and NUP98::KDM5A models. Assay data was obtained in technical triplicates from an established NUP98::KDM5A/Cas9 line and independent experiments. One data point in **C** was not obtained due to technical errors. RNAseq data was obtained from six independent experiments. In **B-D**, statistical tests were performed by linear mixed effect model (**B**) or two-sided Student's t-test by comparing day 4 and day 39 (**C**) or gRNA conditions and AAVS controls (**D**), and limma (**F, G**) followed by the Benjamini-Hochberg adjustment when applicable. DEGs in scRNAseq (**L**) were identified using FindMarker function in Seurat package with default settings, which calculate adjusted *P*-values with limma implementation of the Wilcoxon rank-sum test followed by Bonferroni correction. Asterisks indicating *P*-values or adjusted *P*-values <0.05. Error bars indicate mean ± s.e.m.

### 1 Functional characterization of recurrent somatic alterations in NUP98r leukemia

2 Given the differential NUP98-FO binding patterns and cooperating mutations related to disease phenotypes, we  
3 investigated the impact of co-occurring alterations on NUP98r cbCD34 models. We first established cbCD34 models  
4 with constitutive Cas9 expression followed by lentiviral transduction of gRNA to mimic *WT1* frameshift mutations or  
5 *RB1* loss (Fig.6A, Fig.S10A-B). In NUP98::KDM5A cbCD34 models, *WT1* gRNAs showed enhanced cell growth  
6 compared with AAVS-targeting controls, whereas out-of-frame indels in *RB1* were enriched at later time points,  
7 indicating growth advantages of cells with *RB1* loss (Fig.6B-C). Morphologically, AAVS-controls showed cells with  
8 various degrees of myeloid and erythrocyte/megakaryocyte differentiation, whereas *RB1*-gRNA conditions yielded  
9 larger cells with features of maturing megakaryocytes, and *WT1*-gRNA enriched homogeneously small blast-like cells  
10 (Fig.6B). Flow cytometric analysis showed that *RB1*-gRNA conditions enriched CD34+CD41a+ populations,  
11 whereas *WT1*-gRNA conditions depleted the CD41a+ population with increased CD34+CD41a- cells (Fig.6D,  
12 Fig.S10C). NUP98::NSD1 cbCD34 models showed growth advantages with *WT1*-gRNA, whereas *RB1*-gRNAs did  
13 not show either growth advantage or consistent enrichment of out-of-frame indels, suggesting that *RB1* loss  
14 specifically synergizes with NUP98::KDM5A (Fig.S10D-E). RNAseq revealed global transcriptional changes in  
15 *RB1*- or *WT1*-gRNA conditions (Fig.6E). *RB1*-gRNA conditions showed upregulation of *CD34* and *CDKN2A*, which  
16 is downstream RB1/E2F and forms a feedback loop<sup>55,56</sup>, and changes in immunity genes compared with AAVS-  
17 controls (Fig.6F, Fig.S11A). *WT1*-gRNA also showed increased *CD34* expression, along with increased MHC-class  
18 II and decreased megakaryocyte/platelet-related gene expression (Fig.6G).

19 To study how co-alterations affected cellular hierarchies, we performed scRNAseq from these conditions, which  
20 collectively showed differentiation trajectories resembling normal hematopoiesis (Fig.6H-I, Fig.S11B). A control  
21 condition showed broad distributions with enrichment at the platelet-like populations adjacent to the MEP-like or  
22 megakaryocyte (MK)-like clusters. By contrast, *RB1*-gRNA conditions showed enrichment of cells at the MEP or  
23 MK-like clusters and lacked platelet-like populations, and *WT1*-gRNA showed strong bias at the GMP- and LMPP-  
24 like populations rarely found in control and *RB1*-gRNA conditions. Pseudotime analyses showed that *RB1* was  
25 upregulated through the platelet differentiation trajectory and highest in the platelet-like cluster, whereas *CDKN2A*  
26 expression was high in the MEP or MK clusters but low or undetectable in the platelet cluster (Fig.6J-K, Fig.S11C).  
27 Comparison of platelet-like and MK-clusters showed upregulation of translation-related genes in the platelet-like  
28 cluster (Fig.6L), recapitulating changes in translation in normal platelet production from MK or proplatelet<sup>57</sup>. We also  
29 observed negative correlation with *RB1* and *CDKN2A* among NUP98::KDM5A patient samples (Fig.S11D). These

suggest that *RB1* loss inhibits terminal platelet differentiation in the NUP98::KDM5A model, leading to the accumulation of cells at the MK-like or MEP-like stages, similarly to megakaryocytic differentiation block seen in E2F transgenic mice<sup>58</sup> (Fig.6M). The LMPP population enriched in the *WT1*-gRNA condition was characterized by high MHC-class II gene expression and low cell-division gene expression, and inference of cell-cycles from scRNAseq indicated enrichment of cells at G1-stage in the LMPP cluster and at S/G2M in the GMP cluster (Fig.S11E-G), suggesting unique cellular hierarchy of proliferating GMP-like cells and dormant LMPP-like cells in the *WT1*-gRNA condition.

### *Cooperating alterations and differentiation status affect sensitivity to menin inhibition*

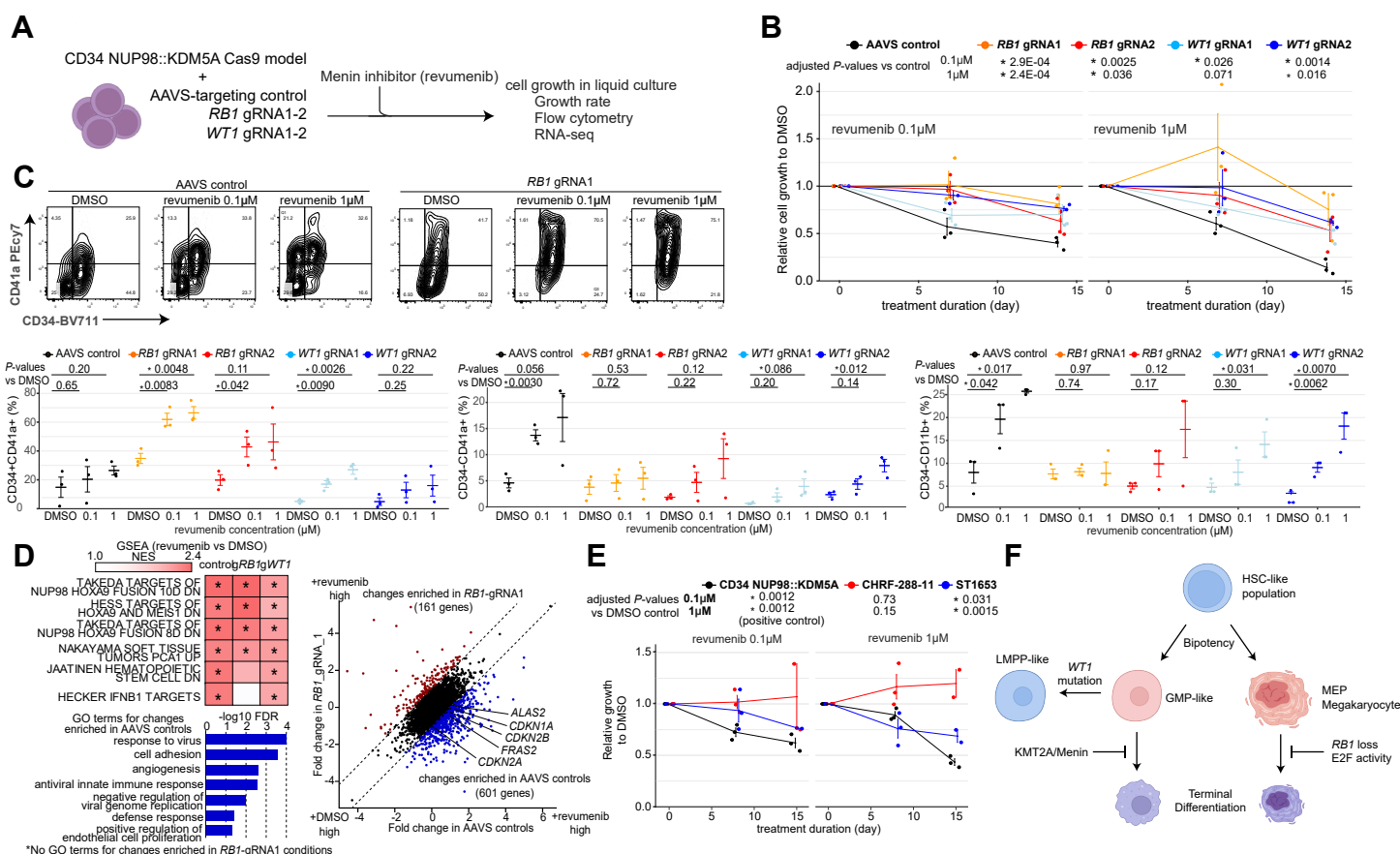
*NUP98r* leukemia has been reported to be dependent on KMT2A/menin complex<sup>18</sup>, and various menin inhibitors are currently being tested in clinical trials enrolling patients with AML including those with *NUP98r*<sup>59</sup>. Mouse models showed menin inhibition led to myeloid differentiation in *NPM1*-mutant and *NUP98r* leukemia models<sup>18,60-62</sup>, although whether menin is required for megakaryocytic lineages or AMKL is currently unclear. We tested a menin inhibitor (revumenib) in NUP98::KDM5A CD34 models with cooperating alterations to assess the sensitivity and impact on cellular populations (Fig.7A). Treatment of AAVS-control conditions showed a decreased cell growth ( $14.1 \pm 4.6\%$ ; mean  $\pm$  standard error, at day 14 of  $1\mu\text{M}$  revumenib treatment compared with DMSO), whereas *RB1*- or *WT1*-gRNA conditions were less sensitive (*RB1*-gRNA:  $64.3 \pm 10.3\%$ , *WT1*-gRNA,  $56.8 \pm 4.0\%$ , Fig.7B). Flow cytometric analysis at day 14 showed significant increases in the CD11b+ population in AAVS-control and *WT1*-gRNA conditions, indicating myeloid differentiation, contrasting to an increase in the CD34+CD41a+ population in *RB1*-gRNA conditions (Fig.7C), suggesting adaptive changes of cellular hierarchies. RNAseq analyses of samples on day 7 of treatment showed increased expression of canonical genes suppressed by NUP98-FOs in all conditions, whereas *CDKN2* family<sup>63</sup>, inflammation, and erythroid/megakaryocyte-related genes were enriched in AAVS compared with *RB1*-gRNA conditions, suggesting differentiation induced by menin inhibition (Fig.7D). *WT1*-gRNA showed immunity-related changes as in AAVS controls but more in activated monocyte/lymphocyte genes (e.g., *CD48*, *MAFB*, Fig.S11H), suggesting partial differentiation from dominant GMP or LMPP clusters. We also tested the sensitivity of ST1653-PDX (myelomonocytic) and CHRF-288-11 (megakaryocytic) cells to menin inhibition, which showed that ST1653 was sensitive to menin inhibition but CHRF-288-11 was not, further confirming the association of differentiation status with menin-dependency (Fig.7E). These collective data indicate that cellular differentiation and cooperating mutations can both influence the dependencies within *NUP98r* leukemias (Fig.7F).

### *Discussion*

Biology-informed treatment of many subtypes of pediatric leukemia is paving the way for improved clinical outcomes<sup>3,64</sup>; however, the biology driving the heterogeneous phenotypes of *NUP98r* leukemia, including somatic alterations or NUP98-FO binding, has not yet been fully characterized. Using genome-wide mutational and transcriptional analyses of 185 *NUP98r* leukemia samples, we expand on the known association of *NUP98r* with *WT1*, *FLT3*-ITD, and *RB1* loss to highlight *JAK2*, *GATA1*, and *MPL* alterations associated with AMKL and *NOTCH1* alterations with T-ALL phenotypes. Profiling with scRNAseq and deconvolution of bulk RNAseq data additionally show heterogeneous cellular hierarchies even with the same fusion partners or in the same patients, suggesting that both fusion partners and cooperating mutations can contribute to the expression profiles and cellular hierarchies<sup>41,43</sup>.

Further, using CUT&RUN in cbCD34 models, we showed that NUP98 FO-binding patterns are unique to each fusion partner, whereas profiling of patient samples and cell lines showed that the binding patterns are not static and could be dynamic according to the differentiation status, represented by AMKL-specific NUP98::KDM5A-binding to *MEIS2* or *MECOM*. Previous knowledge of NUP98-binding patterns is derived from mouse models<sup>16-18</sup> or in a human non-hematopoietic cell model<sup>19</sup>, whereas our data emphasize the need for studies using human hematopoietic cells to profile NUP98 FO-binding and related epigenetic status contributing to the pathogenesis of *NUP98r* leukemia.

## Fig.7



**Figure 7. Cooperating alterations and differentiation status affect sensitivity to menin inhibition**

**A.** Experimental schema showing revumenib treatment. **B.** Relative cell growth of cdCD34 NUP98::KDM5A models with gRNA treated with revumenib (0.1-1μM) compared with DMSO controls. **C.** Flow gating (top), CD34+CD41a+ population (left), CD34-CD41a+ population (mid), and CD34-CD11b+ populations (right) among mCherry+ GFP+ mAmetrine+ live population at day 15. **D.** GSEA analyses between DMSO and revumenib-treated conditions, colors showing NES (top-left), comparison of expression changes between AAVS and RB1-gRNA1 conditions (right), and GO term analysis of changes enriched (difference of fold changes >1) in AAVS conditions (bottom-left). **E.** Relative cell growth of unedited cdCD34 NUP98::KDM5A (control, black), CHRF-288-11 (red) and ST1653 PDX (blue) treated with revumenib (0.1-1μM) compared with DMSO controls. **F.** Schematics illustrating cellular hierarchy of NUP98::KDM5A models. Data was obtained from three technical replicates using gRNA-transduced cells in Fig.6. One data point at day 15 was excluded for technical errors. Statistical tests were performed by generalized linear mixed effect model with Gaussian distribution followed by the Benjamini-Hochberg adjustment (**B**, **E**) or Student's t-test by comparing gRNA conditions with AAVS controls (**C**). Asterisks indicating *P*-values or adjusted *P*-values \*<0.05. Error bars indicate mean ± s.e.m.

1 We also show that the introduction of cooperating alterations, namely *RB1* deletion and *WT1* frameshift mutations,  
2 could render cells with growth advantages and alter the cellular hierarchies of *NUP98r* models *in vitro* in a fusion  
3 partner-specific manner. Previously, transplant models of cbCD34 *NUP98::KDM5A* cells showed heterogeneous  
4 disease phenotypes<sup>49</sup>, and whether these cooperating mutations can affect phenotypes *in vivo* is to be investigated  
5 in future studies. Although acquired mutations in the *MEN1* gene have been clinically observed as mechanisms of  
6 resistance to menin inhibitors in other AML subtypes<sup>65,66</sup>, our experimental models also suggest that cellular  
7 differentiation-specific menin dependency can be alternative resistance mechanisms for *NUP98r* leukemia. It will  
8 be critical to investigate clinical outcomes of *NUP98r* leukemia treated with menin inhibitors according to fusion  
9 partners, cooperating mutations, or expression profiles to identify patients who will not benefit from menin inhibition  
10 and treat them with alternative approaches.

11 Mechanistically, Rb1 is a well-known repressor of E2F1 proteins, which in turn activate the expression of cell cycle-  
12 related genes including *CDKN2* families<sup>55,56</sup>. E2F1 activity inhibits normal platelet differentiation leading to the  
13 accumulation of abnormal megakaryocytes in E2F1 transgenic mice<sup>58</sup>, whereas *Cdkn2a*-deficient mice showed  
14 enhanced platelet production<sup>67</sup>. *NUP98-FO* binding data from cbCD34 models suggests that preferential  
15 differentiation toward MEP and platelets is *NUP98::KDM5A*-intrinsic, whereas *RB1* loss could lead to high E2F  
16 activity and high expression of *CDKN2* family at the MEP/MK stages, which lead to accumulation of megakaryocytic  
17 cells and ultimately development of AMKL. However, our cbCD34 models do not fully explain why *NUP98::KDM5A*  
18 AMKL is exclusively found in infants and *NUP98::KDM5A* myeloblastic AML is only rarely found in adolescence.  
19 Future studies using mice at the developmental stages or transplantation models could dissect the contribution of  
20 cell-of-origin (fetal or adult definitive hematopoiesis) or microenvironments (fetal liver or bone marrow) to the  
21 *NUP98r* leukemogenesis and provide mechanistic insights needed for better treatment for these refractory diseases.

22

23

## 24 **Methods**

### 25 **Subject cohorts and sample details.**

26 Tumor samples from patients with acute leukemia from the St. Jude Children's Research Hospital tissue  
27 biorepository were obtained with written informed consent in clinical trials<sup>68,69</sup> using a protocol approved by the St.  
28 Jude Children's Research Hospital institutional review board (IRB). Sequence data of tumor samples from the  
29 AIEOP (Associazione Italiana Ematologia/Oncologia Pediatrica) and Perugia University were also obtained with  
30 written informed consent in each institute. Studies were conducted in accordance with the International Ethical  
31 Guidelines for Biomedical Research Involving Human Subjects. No patient received compensation for the  
32 enrollment to this study. Patient names were de-identified and replaced with unique research IDs, which are  
33 maintained using an honest broker, in accordance with institutional protocols and ethical guidelines. This unique  
34 research ID cannot be linked to patient identification outside of the research team. Samples for RNA sequencing  
35 (RNAseq: n=10), whole genome sequencing (WGS: n=2), and whole exome sequencing (WES: n=2) were newly  
36 sequenced in this study, and the rest of the data were obtained from previous publications (Supplementary Table  
37 1-4)<sup>3,5,6,8,12,13,22-28</sup> or public databases (see details in **Data availability** and Table S1). Single cell RNA sequencing  
38 data of a *NUP98::NSD1* case was obtained from GEO<sup>43</sup> (GSM7494309, GSM7494310).

39

40 **Genotype fingerprints.** We performed a pairwise genotype concordance comparison using the estimated  
41 genotype from SNPs with  $\geq 20\times$  coverage in RNAseq BAM files to make sure that the study cohort cases represent  
42 unique individuals as we performed previously<sup>5</sup>. We set genotype concordance percentage cutoff  $\geq 90\%$  of SNPs  
43 shared between two individuals to identify potential duplicates.

44

45 **Sample processing, library preparation, and sequencing.** For newly sequenced samples with low tumor purity  
46 (below 60%), the leukemic cell population was enriched either by flow cytometric sorting or T cell depletion by  
47 magnetic beads (EasySep Human CD3 Positive Selection Kit II, StemCell Technologies, 17851). For flow cytometric  
48 sorting, CD45<sup>dim</sup>CD33<sup>dim</sup> positive population was sorted using anti-CD45 PerCP-Cyanine5.5 (eBioscience, cat#  
49 8045-9459-120, Clone:2D1, RRID: AB\_1907397, 1:20 dilution), anti-CD33 APC (eBioscience, cat# 17-0338-



42, clone: WM53, RRID: AB\_10667893, 1:20 dilution), and DAPI (BD Biosciences, cat# 564907, RRID: AB\_2869624) using FACS Aria III instrument and FACS Diva v9.0 (both BD Biosciences) as we have reported<sup>5</sup>. CD34 gating using anti-CD34 PE (Beckman cat# IM1459U, Clones: QBE10; Immu133; Immu409, RRID: AB\_131210, 1:5 dilution) was added depending on the positivity of each patient sample. Enrichment of the tumor population was confirmed by flow cytometric analysis of the post-sorting samples (generally > 90%). Libraries were constructed using the TruSeq Stranded Total RNA Kit, with RiboZero Gold (Illumina, 20020598) for RNAseq of patient samples and *NUP98* cbCD34 models in Fig.4, Ovation RNA sequencing system v2 (Tecan, 7102-32) for the first batch of *NUP98* CD34/Cas9 models in Fig.6-7, and Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero Plus (Illumina, 20040529) for the second batch. The TruSeq DNA PCR-Free Library Prep Kit (Illumina, 20015963) was used for WGS, and the TruSeq Exome Kit v1 (Illumina, 20020614) for WES according to the manufacturer's instructions. After library quality and quantity assessment, samples were sequenced on HiSeq2000 or 2500 (Illumina, RRID: SCR\_020132, RRID: SCR\_016383) instruments with paired-end (2 x 101 bp, 2 x 126 bp, or 2 x 151 bp) sequencing using TruSeq SBS Kit v3-HS (Illumina, FC-401-3001) or TruSeq Rapid SBS Kit (Illumina, FC-402-4023) and HiSeq Control Software with most recent version at the time of sequencing.

**Whole genome and whole exome sequencing data analysis.** The previous genomic lesion calls for the cases (WGS; n=73, WES; n=46) from published studies<sup>5,6,8,12,13,23-25,27,28</sup> were collected from their respective publications. For the unpublished cases with DNA data (WGS; n=2, WES; n=2), DNA reads were mapped using BWA<sup>70,71</sup> (WGS: v0.7.15-r1140 and v0.5.9-r26-dev; WES: v0.5.9-r26-dev and v0.5.9, RRID: SCR\_010910) to the GRCh38/hg38 human genome assembly. Aligned files were merged, sorted, and de-duplicated using Picard tools 1.65 (broadinstitute.github.io/picard/). SNVs and Indels were called using Bambino. For cases paired with matched germline controls, germline variants were filtered out if it's present in the matched germline sample. For unpaired cases, possible germline variants were filtered and classified. Structural variations (SV) were analyzed using CREST<sup>72</sup> (v1.0, RRID: SCR\_005257), and CNVs were analyzed using CONCERTING (1.0)<sup>73</sup> on the WGS data. CNVs were also called on cases with only WES DNA data using the following methods. Briefly, Samtools<sup>74</sup> (v1.16, RRID: SCR\_002105) mpileup command was used to generate a mpileup file from matched germline and tumor BAM files with duplicates removed. If a matched germline was not available, a high-quality normal sample was used to pair with the tumor sample. VarScan<sup>75</sup> (v2.3.5, RRID: SCR\_006849) was then used to take the mpileup file to call somatic CNVs after adjusting for normal/tumor sample read coverage depth and GC content. Circular Binary Segmentation algorithm<sup>76</sup> implemented in the DNACopy R package (v1.52.0, RRID: SCR\_012560) was used to identify the candidate CNVs for each sample. B-allele frequency info was also used to assess allelic imbalance.

**RNAseq mapping, fusion detection, and large-scale copy number variant calling.** RNA reads from newly sequenced samples and from publications were mapped to the GENCODE (RRID: SCR\_014966) human genome assembly release 38 gene annotation (GRCh38/hg38) using the StrongARM pipeline<sup>77</sup>. RNAseq data on hg19 from previous studies were re-mapped onto hg38 using STAR (v.2.7.9a, RRID: SCR\_004463). Chimeric fusion detection was carried out using CICERO<sup>31</sup> (v0.3.0) on hg19 BAM files and the results were lifted over to hg38 using LiftOver (not versioned, RRID: SCR\_018160) from UCSC genome browser (RRID: SCR\_005780). For the cases with only RNAseq data, RNAseqCNV<sup>78</sup> (v1.2.1) was used to call large-scale copy number variants (CNV) as previously performed<sup>5</sup>. For possible chromosomal loss involving *RB1*, we manually checked allele frequencies of single nucleotide polymorphisms (SNPs) within the *RB1* gene, around *RB1* genes, and broader regions on chromosome 13 to identify allelic imbalance indicative of loss of heterozygosity of the region (Fig.S3B). Those with allelic imbalance and low *RB1* expression (z-score in the transcriptome cohort < -0.5) were regarded as *RB1* loss (Table S6).

**Somatic mutation calling from RNAseq.** To detect single-nucleotide variants (SNV) and insertions and deletions (Indel) from RNAseq data, we applied our previous approach that showed high sensitivity and specificity on selected genes.<sup>5</sup> In this study, we focused on a panel of previously screened 87 genes and 25 genes with somatic alterations found in DNA data from this study cohort (Supplementary Table 5). Candidate SNVs/Indels were called by Bambino<sup>79</sup> (v1.07, RRID: SCR\_005649) or RNAIndel<sup>80,81</sup> (v3.0.4), annotated by VEP<sup>82</sup> (v95, RRID: SCR\_007931), filtered by excluding variants with gnomAD (v2.1.1, RRID: SCR\_014964)<sup>83</sup> population allele frequency > 0.1% as possible germline variants, and in turn, classified for putative pathogenicity with PeCanPie/MedalCeremony<sup>84</sup> (not versioned). Candidate variants with putative pathogenicity were considered germline or artifacts if present in > 5% of the cases. Candidate variants were further filtered if the number of supporting reads was ≤ 5 or if the variant allele fraction (VAF) was ≤ 5%.

1  
2 **Gene expression data summarization, dimension reduction, and clustering.** Reads from aligned BAM files  
3 were assigned to genes and counted using RSEM<sup>85</sup> (v1.3.0, RRID:SCR\_000262) with the GRCh38/hg38 GTF file.  
4 Expression data of other T-ALL or AML subtypes from previous studies were obtained in gene x sample matrices  
5<sup>5,12</sup> and combined with *NUP98* other leukemia data to establish RNAseq data of 2,321 samples. For a gene to be  
6 considered as expressed, we required that at least 4 samples should have  $\geq 20$  read counts per million (CPM) reads  
7 sequenced. The count data were transformed to log2CPM using Voom<sup>86</sup> available from R package limma<sup>87</sup> (v3.50.3,  
8 RRID: SCR\_010943). We did not perform batch correction between stranded and unstranded library or data from  
9 TARGET or not, as this information confounded with disease subtypes (AML vs ALL). The R package Seurat<sup>88-</sup>  
10<sup>91</sup>(v4.1.0, RRID:SCR\_016341) was used for dimension reduction and sample clustering. Briefly, the top 400 variable  
11 protein-coding genes were selected using the “vst” method. The expression data were then scaled and used for  
12 PCA (Principal Component Analysis), and the top 25 principal components were used for dimension reduction using  
13 UMAP<sup>92,93</sup> (Uniform Manifold Approximation and Projection, RRID:SCR\_018217) (n\_neighbors=15 and  
14 min\_dist=0.3). Samples were clustered using the top 25 principal components by first constructing a K nearest-  
15 neighbor graph and then iteratively optimizing the modularity using Louvain algorithm (resolution=0.4). Annotations  
16 of each cluster are based on enriched subtypes or known expression profiles. For cord blood CD34+ cell (cbCD34)  
17 models, data was processed in the same process, and principal component analysis (PCA) was performed using  
18 prcomp function. For combined analyses of patient data and cbCD34 models, shared genes in each data set were  
19 used to combine, followed by batch correction between patient and in vitro data using the ComBat  
20 (RRID:SCR\_010974) method available from R package SVA<sup>94</sup>(v3.42.0, RRID:SCR\_012836). UMAP plot was  
21 performed using the top 750 variable genes with the following parameters: dims=1:35, min.dist=0.25,  
22 n.neighbors=15L. For cbCD34 NUP98 Cas9 models, batch correction was performed between different library  
23 preparations.

24 Differential expressed gene (DEG) analyses of bulk samples were performed by limma<sup>87</sup>, and we set Log2 CPM =  
25 0 if it is < 0 based on the Log2 CPM data distribution. *P*-values were adjusted by the Benjamini-Hochberg method  
26 to calculate the adjusted *P*-values using R function p.adjust. Genes with absolute fold change > 1 and adjusted *P*-  
27 values < 0.05 were regarded as significantly differentially expressed. Gene Set Enrichment Analysis (GSEA)<sup>95</sup> was  
28 performed by GSEA (v4.2.3, RRID: SCR\_003199) using MSigDB gene sets c2.all and c5 (v7.2.1,  
29 RRID:SCR\_016863), comparing samples or conditions indicated in figure or figure legends. Permutations were  
30 done 1000 times among gene sets with sizes between 15 and 1500 genes. Normalized enrichment scores (NES)  
31 and FDR for top representative gene sets related to hematopoiesis, leukemia phenotype, biological processes were  
32 shown, and all data is available in source data. Gene Ontology (GO) term analyses of DEGs were performed with  
33 DAVID<sup>96</sup> (v6.8, RRID:SCR\_001881) using differentially expressed gene, and results for GO term, biological process  
34 (GOTERM\_BP\_DIRECT) were exported.

35  
36 **Single cell RNA sequencing (scRNAseq).** ScRNAseq was performed on 8 patients (9 samples) and 4 conditions  
37 of cbCD34/Cas9 models. Frozen patient samples were thawed followed by dead cell removal using EasySep Dead  
38 Cell Removal Kit (StemCell Technologies, 17899) and resuspended in DPBS (Dulbecco's Phosphate Buffered  
39 Saline, Gibco, 14190144)+0.04%BSA (Bovine Serum Albumin, Sigma-Aldrich, A7030-50G). Cultured cbCD34 cells  
40 are directly suspended in DPBS+0.04% BSA. Cells were prepared and quantified following the 10x Genomics cell  
41 thawing protocol for single cell assays (CG000447, rev B). Cells were isolated for single-cell barcoding and 3'  
42 (*NUP98::NSD1* diagnosis and relapse pair) and 5' (*NUP98::KDM5A* and *NUP98::RAP1GDS1*, CD34 models) GEX  
43 library preparation using the 10x Genomics Chromium Controller Genetic Analyzer (RRID:SCR\_019326) and Next-  
44 GEM Single Cell V(D)J Reagent Kit (PN-1000128, PN-1000167) following the manufacturer's protocol. Libraries  
45 were pooled and sequenced on an Illumina NovaSeq 6000 (RRID:SCR\_016387) using SP Reagent kit v1.5  
46 (Illumina, 20028401). The feature-barcode count data matrix for each sample was generated from raw sequence  
47 FASTQ files using Cell Ranger (v3.0, RRID:SCR\_017344) and the GRCh38/hg38 human genome assembly. Public  
48 scRNAseq data (GSM7494309, GSM7494310) was downloaded as processed count matrix, gene barcodes, and  
49 genes, and processed equally with other samples. The output data from Cell Ranger was further analyzed in R  
50 environment (v4.0.2, RRID: SCR\_001905) using R package Seurat (v4.1.0, RRID: SCR\_007322).

51 For patient data, count data matrices from each patient were combined using the merge function, followed by  
52 excluding cells with less than 200 detected genes or greater than 20% mitochondrial transcripts, which was adjusted  
53 for the variety of primary cells with low quality, and we further filtered genes to include those in reference bone  
54 marrow and thymocyte data set. Feature counts for each cell were normalized to  $1 \times 10^4$  counts/cell and natural log  
55 transformed using the Normalize Data function. The dimensional reduction was performed by principal component

analysis (PCA) using the top 1500 variable genes identified by vst analysis using the Find Variable Features function, followed by Jack Straw analysis and UMAP (uniform manifold approximation and projection, RRID: SCR\_018217) using the following parameters (dims=1:20, n.neighbors=15L, min.dist=0.5). Clusters were identified by the Find Neighbors and FindClusters functions (resolution = 0.1) and annotated by enrichment of leukemic samples or normal hematopoietic cells. Marker genes were identified by the FindAllMarkers function in Seurat (min.pct = 0.25) and DEGs were identified by FindMarker gene function in Seurat (v4.1.0), both of which calculate adjusted *P*-values with limma implementation of the Wilcoxon rank-sum test followed by Bonferroni correction.

Projection onto the normal bone marrow and thymocyte reference scRNAseq data was performed using MapQuery function in Seurat. Briefly, scaled data were processed through FindTransferAnchors, TransferData, and MapQuery functions, and the reference data and UMAP models were reproduced using following parameters (dims = 1:25, min.dist = 0.5) and used as reference input. Annotations of the cell types are imported from the original publication, and cell density on the original UMAP was calculated by get\_density function from ggplot2 library (v3.3.6, RRID:SCR\_014601).

Analyses of scRNAseq data of CD34 models were performed by excluding cells with less than 1500 detected genes or greater than 20% mitochondrial transcripts. Regression of cell cycle related gene was applied using CellCycleScoring functions and ScaleData function with the vars.to.regress parameter. The dimensional reduction was performed by PCA using the top 2000 variable genes and UMAP using the following parameters (dims=1:30, n.neighbors=40L, min.dist=0.2). Clusters were identified by the FindNeighbors and FindClusters functions (resolution = 0.35). Cell lineages were obtained by getLineages function in Slingshot (v1.6.1, RRID:SCR\_017012), specifying monocyte, macrophage, cDC, erythroid, T-like, and platelet-like clusters as terminal clusters. Pseudotime was calculated by getCurves function (approx\_points = 300, thresh = 0.01, stretch = 0.8, allow.breaks = FALSE, shrink = 0.99) by excluding macrophage, cDC, erythroid, and T-like population.

**Cellular deconvolution using CIBERSORT.** Inference of cellular hierarchy by CIBERSORT<sup>45</sup> (RRID:SCR\_016955) was performed by the web interface of CIBERSORTx<sup>97</sup>. We first established signature matrix using the bone marrow and thymocyte reference count data as reference, from which data was subsampled to include 500 cells or less for one cell type (total 49623 → 5500 cells) to accommodate with the data limit in the web application. Mature T-cell types and B-lineage cell types are excluded from the signature matrix, considering our NUP98r cohort including T-ALL and AML only, and parameters of signature matrix were set as follows: replicates: 5, sampling: 0.5, fraction: 0, k.max: 999, q.value: 0.01, G.min: 300, G.max: 500. Deconvolution of bulk RNAseq data was performed in absolute mode with S-mode batch correction with 100 permutations using normalized CPM values as input.

**CUT&RUN (cleavage under targets and release using nuclease).** CUT&RUN from cbCD34 models and primary leukemia samples were performed using CUTANA™ ChIC/CUT&RUN Kit (SKU: 14-1048) according to manufacturer's instructions. Briefly, 5.0x10<sup>5</sup> cells per condition were used as a maximum input, whereas the input cells could be as low as 5.0x10<sup>4</sup> depending on the availability of samples. For primary samples, frozen samples were thawed followed by dead cell removal using EasySep Dead Cell Removal Kit (StemCell Technologies, 17899), generally resulting in 85-100% of viable cells, and by additional nuclear extraction steps according to the protocol version 3, section 3.2, which was critical to detect endogenous NUP98-FO detection. Cells from cbCD34 models were processed without nuclear extraction. Extracted DNA was quantified using a Qubit fluorimeter (Thermo Fisher, RRID:SCR\_018095) and dsDNA Quantification Assay Kits, High Sensitivity (Thermo Fisher, Q32854). Antibodies against H3K27ac (Cell Signaling, 8173S, clone: D5E4, RRID:AB\_10949503), H3K4me3 (Epiccypher, 13-0041, Mixed Monoclonal, RRID:AB\_3076423), H3K27me3 (Cell Signaling, 9733, clone: C36B11, RRID:AB\_2616029), menin (Bethyl Cat# A300-105A, polyclonal, RRID:AB\_2143306), KMT2A (Epiccypher, 13-2004, unknown clone, RRID:AB\_3076423), N-terminus NUP98 (Abcam Cat# ab50610, clone:2H10, RRID:AB\_881769), N-terminus NUP98 (Cell Signaling, 2288, clone: L205, RRID:AB\_561204, for preliminary experiment only), HA (Cell Signaling, 3724, clone: C29F4, RRID:AB\_1549585), and control IgG antibody (Epiccypher Cat# 13-0042, RRID:AB\_2923178) were used at 1:50 dilution in all assays.

**CUT&RUN library preparation and sequencing.** DNA from CUT&RUN experiments were subjected for library preparation using NEBNext Ultra™ II DNA Library Prep Kit for Illumina® (E7645) according to a modified protocol based on protocol.io (dx.doi.org/10.17504/protocols.io.bagaibse). Specifically, DNA from CUT&RUN was used as



input (6ng or lower) and adaptor concentrations were proportionally to DNA concentrations (e.g. 3.75 pmol for 6 ng DNA, 0.8 pmol for 1 ng DNA) to avoid extra adaptor dimers. DNA purification with AMPure XP beads were performed as one step (after adapter ligation: 80 µL as 1.75x, after PCR: 36 µL as 1.2x) to enrich short fragments. PCR cycles were set at 12 cycles for 6 ng input and increased up to 15 cycles when input DNA concentration was too low and undetectable. Library amount was quantified using Qubit™ dsDNA Quantification Assay Kits (High Sensitivity) and analyzed by Agilent Bioanalyzer (RRID:SCR\_018043) using High Sensitivity DNA chip (Agilent, 5067-4626) before sequencing. Libraries were sequenced on NovaSeq 6000 with paired-end (2x100bp) sequencing using SP Reagent kit v1.5 (Illumina, 20028401) targeting 20 million reads.

## CUT&RUN analytic pipeline.

To obtain quality reads, raw reads were processed and trimmed with Trim Galore (v0.4.4, RRID:SCR\_011847) from cutadapt (RRID:SCR\_011841) and FASTQC analysis. A default quality score cutoff of Q20 is used. Quality reads were mapped to the GRCh38/hg38 human genome assembly using bwa (v0.7.12-r1039, RRID:SCR\_010910) and converted to a bam file by samtools (v1.2, RRID:SCR\_002105). Duplicate reads were marked with biobambam2 (v2.0.87, RRID:SCR\_003308). Uniquely mapped and properly paired reads were then extracted with samtools (v1.2, "-F 1804 -q 1") and sorted by read name using biobambam2 (v2.0.87, RRID:SCR\_003308). bedtools<sup>98</sup> (v2.17.0, RRID:SCR\_006646) were used to convert bam file to bedpe file as fragments. Fragments with size shorter than 2000bp were extracted and center 80bp of each fragment were used to generate bigwig track files by UCSC tools (v4, RRID:SCR\_005780) and visualized using Integrative Genomics Viewer (IGV, v2.17.4, RRID:SCR\_011793). Peak calling was performed using Macs2 (v2.1.1, RRID:SCR\_013291) targeting narrow peaks and SICER (v1.1, RRID:SCR\_010843) targeting broad peaks with following command lines.

```
macs2 callpeak -t $ [treatmentBed] -c $ [controlBed] -f BEDPE -g hg38 --keep-dup all -n ["output file"] -q [0.05]
```

```
sicer -t $ [treatmentBed] -c $ [controlBed] -s hg38 -rt 1 -w 200 -f 200 -egf 0.8-6 -g 600 -fdr 0.00001
```

Peaks called by macs2 or SICER were filtered according to peak scores (>50, corresponding FDR<1.0e-5) and annotated using HOMER (v4.10, RRID:SCR\_010881). Low-quality data with high background signal from IgG controls or low assay signals were excluded according to peak call count and manual inspection of bigwig files on IGV, which are listed in Fig.S9. Overlaps of coding genes with peaks were assessed by hypergeometric test (one-sided) using gmp library (v0.6-5), with total number of coding genes in hg38 (19,385 genes) as a denominator. Quantitative analysis of H3K27ac and H3K27me with broad signals from various primary samples, we utilized PBS<sup>54</sup> (probability of being signal, v02072024) which is a bin-based strategy of quantification of epigenetic data. PBS was performed by default setting, whose output included normalized signals (0 to 1) on 574,152 genomic regions binned at 5kb length, further filtered for regions with non-zero average signals among samples. PCA was performed by prcomp function, and differential analysis was performed by limma followed by the Benjamini-Hochberg adjustment to calculate adjusted *P*-values. PBS 5kb-binned regions are also annotated using HOMER.

**Construction of lentivirus vectors and virus Production.** *NUP98::NSD1* and *NUP98::HOXA9* cDNA were cloned from patient sample into pDonor221 Gateway entry vector (Thermo Fisher, 12536017) and then transferred to Gateway-compatible vectors (CL20-MSCV-ires-GFP and MND-mPGK-mCherry). *NUP98::KDM5A* expression vectors<sup>49</sup> were kindly provided by Dr. Tanja A. Gruber and cloned into MND-mPGK-mCherry. N-terminal HA tags (YPYDVPDYA) were introduced by amplification during cloning. cDNA encoding Cas9 expressing lentivirus vector (Addgene, FUCas9GFP, Plasmid #85555) and used without cloning. gRNA expressing vector were established by Center for Advanced Genome Engineering (CAGE) in St. Jude Children's Research Hospital using validated gRNAs and LentiGuide-Ametrine backbone expressing mAmetrine under mPGK promoter as described previously<sup>99</sup>. These lentivirus vectors were co-transfected with packaging vectors (pHDM-G, pCAGGHIVgpcr, and pCAG4-RTR2, provided from the St. Jude Vector Laboratory) into 50%–60% confluent HEK293T cells (CVCL\_0063, ATCC #CRL-3216, obtained from ATCC in 2016) using FuGene HD Transfection Reagent (Promega, E2311). Supernatant containing lentiviral particles was harvested at 48 hours after transfection and concentrated for cord blood CD34+ transduction using Amicon Ultra-0.5 Centrifugal Filter Units (MilliporeSigma, UFC905096). HEK293T cells used for virus production were maintained in DMEM (Gibco, #11965) supplemented with 10% Fetal Bovine Serum (FBS, R&D Systems, S11550H) and 100 U/mL penicillin–streptomycin (Gibco, 15140122), which were validated by short tandem repeat (STR) analysis and were generally used within 15 passages after thawing. Mycoplasma testing was intermittently performed using MycoAlert Mycoplasma Detection Kit (Lonza, #LT08–118).



1 **Virus transduction of NUP98-FOs into cord blood CD34+ cells.** Commercially available cord blood CD34+ cells  
2 were purchased from Lonza (catalog no. 2C-101) or the Carolinas Cord Blood Bank/Duke University. After thawing,  
3 cells were cultured for 24 hours in StemSpan SFEM II media (STEMCELL Technologies, #09655) supplemented  
4 with penicillin–streptomycin, l-glutamine, and recombinant human SCF, FLT-3, TPO, and IL6 (all 50 ng/mL,  
5 PeproTech, HHSC6), SR-1 (1 µmol/L, STEMCELL Technologies, 72344), and UM171 (35 nmol/L, STEMCELL  
6 Technologies, 72914, for Fig.4) or UM729 (10 nmol/L, STEMCELL Technologies, 72332, for Fig.6-7). Cells were  
7 transduced with MND-PGK-mCherry lentivirus expressing untagged NUP98-FO (NUP98::KDM5A, NUP98::NSD1,  
8 and NUP98::HOXA9) or empty vector controls at an MOI (multiplicity of infection) of ≥50 by adding 100µL of  
9 concentrated virus to 2.0x10<sup>4</sup> cells in 100µL media (1:1 ratio) supplemented with Polybrene (1µg/ml, Millipore #TR-  
10 1003-G) in a 96-well dish (Corning, 3799). Transduced mCherry+ cells were enriched by flow cytometric sorting on  
11 day 7 of transduction and plated for colony-forming unit (CFU) assay in MethoCult H4435 (STEMCELL, 04435) at  
12 1.0x10<sup>3</sup> cells/dish, followed by weekly serial replating. Cell growth assay was performed by weekly passaging cells  
13 at the density of 1.0x10<sup>4</sup> cells in 100 µL/well in 96-well dishes. HA-tagged NUP98-FOs were transduced similarly  
14 for CUT&RUN experiments. Experiments were performed in biological triplicates using different lots of cord blood.

15

16 **Virus transduction of Cas9 and gRNA into cbCD34 models.** Established cbCD34 models initially generated by  
17 transducing NUP98-FO were maintained in supplemented with penicillin–streptomycin, l-glutamine, and  
18 recombinant human SCF, FLT-3, TPO, and IL6, SR-1 and UM729 as described above. Cells were then transduced  
19 with MND-PGK-GFP lentivirus expressing Cas9 at an MOI of ≥50 by adding 100 µL of concentrated virus to 5.0x10<sup>4</sup>  
20 cells in 100 µL media (1:1 ratio). Transduced mCherry+GFP+ double positive cells were enriched by flow cytometric  
21 sorting on day 7 or later and expanded. Transduction of gRNA expressing vectors (AAVS gRNA controls or gRNAs  
22 for *WT1* and *RB1*) were performed around day 40-60 from transduction of HA-tagged NUP98-FO, by adding 25µL  
23 of concentrated virus to 2.0x10<sup>5</sup> cells in 100µL media (1:4 ratio) to achieve a low MOI around 0.5. Cell growth assay  
24 was performed by weekly passaging cells at the density of 2.0x10<sup>4</sup> cells/100 µL/well in 96-well dishes (Corning).  
25 Confirmation of introduction of indels was performed by target sequencing using genomic DNA from cells at day 39  
26 (cbCD34 NUP98::KDM5A) or day 32 (cbCD34 NUP98::NSD1). Drug sensitivity tests were performed using HA-  
27 tagged NUP98::KDM5A/Cas9/gRNA cbCD34 models at day 35 or later from gRNA transduction in 96-well dishes  
28 using revumenib (or SNDX-5613, MedChemExpress, cat# HY-136175). Experiments were performed in technical  
29 triplicates using established Cas9+ cell lines.

30

31 **Cytospin.** For assessment of cell morphology, 100,000 cells were washed with 1X PBS and spun onto Superfrost  
32 Plus Microscope slides (Fisher Scientific, 12-660-16) at 800 rpm for 5 min and then subjected to standard Wright-  
33 Geimsa staining as previously. Pictures were taken using OLYMPUS BX60 and INFINITY ANALYZE 7 software  
34 (TELEDYNE), and TIFF files were resized by Windows Paint.

35

36 **Cell line culture.** CHRF-288-11 cell line<sup>46</sup> (CVCL\_A280), which is acute megakaryocytic leukemia cell line with  
37 *NUP98::KDM5A*, was kindly provided by Dr. Tanja A. Gruber and maintained in RPMI 1640 (Gibco, 11875093)  
38 supplemented with 10% FBS and 100 U/mL penicillin–streptomycin, which were validated by short tandem repeat  
39 (STR) analysis. ST1653<sup>50</sup>, which is a patient-derived xenograft (PDX) cell line, was provided by XenoSTART/ The  
40 START Center for Cancer Research with a written agreement for use for research purposes and maintained in  
41 RPMI 1640 supplemented with 10% FBS and 100 U/mL penicillin–streptomycin. For drug sensitivity test, cells were  
42 plated at the density of 2.5x10<sup>5</sup> cells/ 330 µl in a 24-well dish (Corning, 3524). Mycoplasma testing was intermittently  
43 performed using MycoAlert Mycoplasma Detection Kit (Lonza, #LT08–118).

44

45 **Flow cytometric analysis and cell sorting.** Cells were washed with FACS buffer (DPBS supplemented 2%FBS  
46 and penicillin–streptomycin) and stained with the following combination of antibodies and dyes along with fluorescent  
47 proteins (mCherry for Fig.4 and Fig.S7, and mCherry/GFP/mAmetrine for Fig.6-7)

48 Fig.4 and Fig.S7; LIVE/DEAD Fixable Aqua Dead Cell Stain Kit, for 405 nm excitation (Thermo Fisher, L34966),  
49 anti-CD34 AlexaFluoro700 (BioLegend, 343526, RRID:AB\_2561495, clone: 581, 1:50 dilution), anti-CD11b APC-  
50 Cy7 (BioLegend, 301342, clone:ICRF44, RRID:AB\_2563395, 1:50 dilution), and anti-CD41a BV605 (BD Horizon™,  
51 563250, clone: HIP8, RRID:AB\_2738096, 1:50 dilution). Cells were also stained with anti-CD45 PerCP-Cy5.5 (BD  
52 Pharmingen™, 564105, clone: HI30, RRID:AB\_2744405, 1:50 dilution), anti-CD33 APC (Invitrogen, 17-0338-42,  
53 clone:WM-53, RRID:AB\_10667893, 1:50 dilution), anti-CD235a BV421 (BD Horizon™, 562938, clone: GA-R2 ,

1 RRID:AB\_2721016, 1:50 dilution), and anti-CD71 BV711 (BD Horizon™, 563767, clone: M-A712,  
2 RRID:AB\_2738413, 1:50 dilution) whereas we did not include these signals in the gating strategy in this manuscript.  
3 Fig.6-7 and Fig.S10; anti-CD34 BV711 (BD, 745543, clone:8G12, RRID:AB\_2743070, 1:40 dilution), anti-CD11b  
4 APC (BD, 561015, clone:ICRF44, RRID:AB\_10561676, 1:20 dilution), and anti-CD41 PEcy7 (Biolegend, 303718,  
5 clone:HIP8, RRID:AB\_10899413, 1:100 dilution).

6 Cells are incubated at room temperature or on ice for 15-30 min, washed by FACS buffer, and resuspended in FACS  
7 buffer with DAPI Staining Solution (Miltenty Biotech., 130-111-570, 1:1000 dilution, for Fig.6-7 only), and analyzed by  
8 BD LSRFortessa X-20 Cell Analyzer (RRID:SCR\_025285, Fig.4) or BD FACSymphony A5 Cell Analyzer  
9 (RRID:SCR\_022538, Fig.6-7). Output FCS files were analyzed FlowJo (v10.10.0, RRID:SCR\_008520).

10 Cell sorting of fluorescent+ (mCherry for NUP98 FOs, and GFP for Cas9) was performed by BD FACSAria III Cell  
11 Sorter (RRID:SCR\_016695) or Sony MA900 Multi-Application Cell Sorter (no RRID) without staining of cell surface  
12 markers.

13  
14 **Western blotting.** HA-tagged NUP98::KDM5A/Cas9/gRNA cbCD34 models at day 35 or later time point were  
15 subjected to Western blotting for RB1 and WT1 proteins. 25 micrograms of purified proteins were separated by  
16 SDS-PAGE on a 4-20% gradient gel (Bio-Rad, #4561033) and transferred onto nitrocellulose membranes (0.2  
17 umol/L, Bio-Rad, 1620252). Membranes were incubated overnight at 4°C with mouse anti-Rb (Cell signaling, 9309T,  
18 clone: 4H1, RRID:AB\_823629 1:1000 dilution), rabbit anti-WT1 (Cell signaling, 83535, clone: D8I7F,  
19 RRID:AB\_2800020, 1:1000 dilution) or anti-GAPDH (Cell signaling, 2118S, clone 14C10, RRID:AB\_561053, 1:1000  
20 dilution). Amersham ECL Rabbit IgG, HRP-linked F(ab')<sub>2</sub> fragment (from donkey)(Cytiva, NA9340, 1:2,000 dilution,  
21 RRID:AB\_772191) and Sheep Anti-Mouse IgG-Horseradish Peroxidase (Cytiva, NA931, 1:2,000 dilution,  
22 RRID:AB\_772210) were used as secondary antibodies. Imaging was performed on the BIO-RAD Chemidoc Touch  
23 imaging system and processed using Image Lab (v6.1).

24

25 **Statistics & Reproducibility.** No sample size, power calculation, or randomization of patients was performed in  
26 this study utilizing retrospective profiling of patients with available materials or sequence data. No analysis  
27 depending on patient background was performed in this study. No blinding was performed in the enrollment of  
28 patients or data collection of public data, and blinding in group allocation was not possible as the grouping is based  
29 on the molecular characteristics of individual patients. For discrete values of the molecular category and the  
30 mutation frequency in cohorts, statistical significance and mutual exclusivity were assessed by two-sided Fisher's  
31 exact test using fisher.test function. Comparison of variant allele frequencies (VAFs) or CIBERSORT scores were  
32 performed by the two-sided Wilcoxon rank-sum test using wilcox.test function. Adjustment of multiple testing was  
33 performed by the Benjamini-Hochberg method using p.adjust function on R when appropriate. The effect size for  
34 Pearson's correlation and statistical significance were assessed by cor and cor.test functions on R.

35 Experiments using cord blood cells are performed in biological triplicates using cord blood cells from different donors  
36 (Lonza, cat# 2C-101A., Donor#24690, Donor#25195, Donor#24987). Experiments of cooperating mutations using  
37 CRISPR/Cas9 system were performed using cbCD34 model lines with Cas9, one for each NUP98::KDM5A and  
38 NUP98::NSD1, and in technical triplicates from three independent experiments at different passages and one week  
39 apart. Sample size (n=3) was determined according to preliminary experiment. Statistical significance was assessed  
40 using generalized linear mixed effect model for growth assays and colony assays using glmer (lmer) function (lme4  
41 package, v1.1-35.1) and emmeans function (emmeans package, v1.5.0), and unpaired two-sided Student's t-test  
42 with equal variances (t.test function) assuming normality according to preliminary experiments, followed by  
43 adjustment of multiple testing was performed by the Benjamini-Hochberg method using p.adjust function on R when  
44 appropriate. Missing data was treated as null in statistical analyses and indicated in figures or legends. Definition  
45 of box plots or error bars are included in figure legends of each figure, and individual data points were shown when  
46 possible. Data used for figures are included in source data spreadsheets.

47

48 **Visualization.** Mutational heatmaps, pie charts, and bar plots were visualized using Windows Excel. Schemas for  
49 fusion partners were made by ProteinPaint (<https://proteinpaint.stjude.org/>). Heatmaps of expression data and  
50 mutual-exclusivity matrix were made by the pheatmap library (v1.0.12, RRID:SCR\_016418). Other data  
51 visualizations were performed by the ggplot function of the ggplot2 library (v3.3.6, RRID:SCR\_014601), the  
52 ggbeeswarm library (v0.7.2), Seurat package (v4.0.2, RRID:SCR\_016341) and the base plot function in the R

statistical environment (v4.0.2). Schemas were created with BioRender.com (RRID:SCR\_018361). Figures are incorporated and edited using Adobe Illustrator (2021, RRID:SCR\_010279).

**Data availability.** Genomic analyses in this study were based on the GENCODE GRCh38/hg38, and gnomAD (v2.1.1, RRID: SCR\_014964) was used for classification for germline and somatic mutations. The genomic data and expression data newly generated in this study (RNAseq: n=10, WGS: n=2, WES: n=2) and undeposited data from previous studies (RNAseq: n=3) have been deposited in the European Genome-Phenome Archive (EGA, RRID:SCR\_004944), which is hosted by the European Bioinformatics Institute (EBI), under accession [EGAS00001005760](#). For the remaining RNAseq data for 175 samples, 56 samples are sequenced in St. Jude and available either on EGA or St. Jude Cloud <sup>5,6,8,13,28</sup> or on the original publication <sup>26</sup>. For 9 published cases <sup>27</sup>, we downloaded the BAM files from EGA ([EGAS00001004701](#)). Of the remaining WGS data for 73 samples, 34 are sequenced at St. Jude samples, data for all 34 samples from the original publications<sup>5,6,13,28</sup> are available on either EGA or St. Jude Cloud, and for other 7 published cases <sup>27</sup>, we downloaded the BAM files from EGA ([EGAS00001004701](#)). For the remaining WES data for 46 samples, 21 are sequenced at St. Jude and all data from the original publications <sup>5,6,8,13,24,28</sup> are available either on St. Jude Cloud or EGA, and for the other 9 published cases <sup>27</sup>, we downloaded the BAM files from EGA ([EGAS00001004701](#)).

The data generated by the TARGET-AML initiative <sup>23,25</sup> (RNA:n=18, WGS:n=18, WES:n=2) is available under accession [phs000218](#) (TARGET-AML) and [phs000465](#) (TARGET sub-study, data is available as a part of phs000218), managed by the NCI. Data from the AAML1031 study lead by COG is also available at GDC data portal under [phs000218](#) (TARGET-AML). Public data from the COG AALL0434 trial (RNA:n=14, WGS:n=14, WES:n=14) is available on dbGaP under [phs002276.v2.p1](#) (phs000218 and phs000464 for T-ALL TARGET samples) and the Kids First data portal (<https://portal.kidsfirstdrc.org/dashboard>). Information about TARGET can be found at <http://ocg.cancer.gov/programs/target>. Background RNAseq data of other T-ALL<sup>12</sup> was also obtained from SYNAPSE under project ID of [syn54032669](#), and data of other AML subtypes were from publication<sup>5</sup>. All public sequencing data and relevant clinical information were obtained with de-identified sample/patient IDs.

Output files from Cell Ranger for scRNAseq data of 9 samples from 8 *NUP98* patient and 4 samples from cbCD34 models, bulk RNAseq data (log2CPM count data) from cbCD34 models, CUT&RUN data (.bigwig files) from primary patient samples and CD34 models have been deposited to GEO and will be available upon publication. Public patient scRNAdata from the COG<sup>43</sup> (GSM7494309, 7494310) is available at GEO under accession number GSE235063, and the patient data was obtained from supplemental tables from the publication. Other data generated in this study are available in Supplemental tables or Source data files or upon request to the corresponding author. We did not use any custom code in this study.

## Acknowledgments

We thank all the patients and their families at St. Jude Children's Research Hospital (SJCRH) for their contribution to the biological specimens used in this study. We also thank the Biorepository, the Flow Cytometry and Cell Sorting Core, the Center for Applied Bioinformatics (CAB), the Center for Advanced Genome Engineering (CAGE), and the Hartwell Center for Biotechnology at SJCRH for their essential services, which are supported by the NIH (P30 CA021765, Cancer Center Support Grant). This work was funded by the American Lebanese and Syrian Associated Charities of St. Jude Children's Research Hospital and grants from the NIH (U54 CA243124, Fusion Oncoproteins in Childhood Cancers (FusOnC2) Consortium granted for C.G. Mullighan and J.M. Kilo, R35CA197695 for C.G. Mullighan, and R01 CA276079 to J.M. Kilo). The content, however, does not necessarily represent the official views of the NIH and is solely the responsibility of the authors. J.M. Kilo holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a previous recipient of the V Foundation Scholar Award (Pediatric). D.D.G received financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union-NextGenerationEU, Project Title "Experimental modeling to identify molecular mechanisms and therapeutic vulnerabilities in NUP98-rearranged leukemia", CUP:J53D23017510001-Grant Assignment Decree No.1384 adopted on 01/09/2023 by the Italian Ministry of Ministry of University and Research (MUR).

## References

1. Rubnitz, J.E. & Kaspers, G.J.L. How I treat pediatric acute myeloid leukemia. *Blood* **138**, 1009-1018 (2021).
2. Rubnitz, J.E. *et al.* Clofarabine Can Replace Anthracyclines and Etoposide in Remission Induction Therapy for Childhood Acute Myeloid Leukemia: The AML08 Multicenter, Randomized Phase III Trial. *J Clin Oncol* **37**, 2072-2081 (2019).
3. Pollard, J.A. *et al.* Sorafenib in Combination With Standard Chemotherapy for Children With High Allelic Ratio FLT3/ITD+ Acute Myeloid Leukemia: A Report From the Children's Oncology Group Protocol AAML1031. *J Clin Oncol* **40**, 2023-2035 (2022).
4. Balgobind, B.V. *et al.* Novel prognostic subgroups in childhood 11q23/MLL-rearranged acute myeloid leukemia: results of an international retrospective study. *Blood* **114**, 2489-96 (2009).
5. Umeda, M. *et al.* A new genomic framework to categorize pediatric acute myeloid leukemia. *Nat Genet* (2024).
6. Umeda, M. *et al.* Integrated Genomic Analysis Identifies UBTF Tandem Duplications as a Recurrent Lesion in Pediatric Acute Myeloid Leukemia. *Blood Cancer Discov* **3**, 194-207 (2022).
7. Gruber, T.A. *et al.* An Inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia. *Cancer Cell* **22**, 683-97 (2012).
8. de Rooij, J.D. *et al.* Pediatric non-Down syndrome acute megakaryoblastic leukemia is characterized by distinct genomic subsets with varying outcomes. *Nat Genet* **49**, 451-456 (2017).
9. Hollink, I.H. *et al.* NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. *Blood* **118**, 3645-56 (2011).
10. Bertrams, E.J.M. *et al.* Comprehensive molecular and clinical characterization of NUP98 fusions in pediatric acute myeloid leukemia. *Haematologica* (2023).
11. Michmerhuizen, N.L., Klco, J.M. & Mullighan, C.G. Mechanistic insights and potential therapeutic approaches for NUP98-rearranged hematologic malignancies. *Blood* **136**, 2275-2289 (2020).
12. Polonen, P., Mullighan, C.G. & Teachey, D.T. Classification and risk stratification in T-lineage acute lymphoblastic leukemia. *Blood* (2024).
13. Schwartz, J.R. *et al.* The genomic landscape of pediatric myelodysplastic syndromes. *Nat Commun* **8**, 1557 (2017).
14. Takahashi, K. *et al.* Integrative genomic analysis of adult mixed phenotype acute leukemia delineates lineage associated molecular subtypes. *Nat Commun* **9**, 2670 (2018).
15. Tian, J. *et al.* The landscape of NUP98 rearrangements clinical characteristics and treatment response from 1491 acute leukemia patients. *Blood Cancer J* **14**, 81 (2024).
16. Xu, H. *et al.* NUP98 Fusion Proteins Interact with the NSL and MLL1 Complexes to Drive Leukemogenesis. *Cancer Cell* **30**, 863-878 (2016).
17. Oka, M. *et al.* Chromatin-prebound Crm1 recruits Nup98-HoxA9 fusion to induce aberrant expression of Hox cluster genes. *Elife* **5**, e09540 (2016).
18. Heikamp, E.B. *et al.* The menin-MLL1 interaction is a molecular dependency in NUP98-rearranged AML. *Blood* **139**, 894-906 (2022).
19. Ahn, J.H. *et al.* Phase separation drives aberrant chromatin looping and cancer development. *Nature* **595**, 591-595 (2021).
20. Chandra, B. *et al.* Phase Separation Mediates NUP98 Fusion Oncoprotein Leukemic Transformation. *Cancer Discov* **12**, 1152-1169 (2022).
21. Skene, P.J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**(2017).
22. Togni, M. *et al.* Identification of the NUP98-PHF23 fusion gene in pediatric cytogenetically normal acute myeloid leukemia by whole-transcriptome sequencing. *J Hematol Oncol* **8**, 69 (2015).
23. Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat Med* **24**, 103-112 (2018).
24. Iacobucci, I. *et al.* Genomic subtyping and therapeutic targeting of acute erythroleukemia. *Nat Genet* **51**, 694-704 (2019).
25. McNeer, N.A. *et al.* Genetic mechanisms of primary chemotherapy resistance in pediatric acute myeloid leukemia. *Leukemia* **33**, 1934-1943 (2019).
26. Buelow, D.R. *et al.* Uncovering the Genomic Landscape in Newly Diagnosed and Relapsed Pediatric Cytogenetically Normal FLT3-ITD AML. *Clin Transl Sci* **12**, 641-647 (2019).
27. Fornerod, M. *et al.* Integrative Genomic Analysis of Pediatric Myeloid-Related Acute Leukemias Identifies Novel Subtypes and Prognostic Indicators. *Blood Cancer Discov* **2**, 586-599 (2021).



28. Schwartz, J.R. *et al.* The acquisition of molecular drivers in pediatric therapy-related myeloid neoplasms. *Nat Commun* **12**, 985 (2021).
29. Umeda, M. *et al.* A new genomic framework to categorize pediatric acute myeloid leukemia. *Nat Genet* **56**, 281-293 (2024).
30. Polonen, P. *et al.* The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature* **632**, 1082-1091 (2024).
31. Tian, L. *et al.* CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol* **21**, 126 (2020).
32. Chellappan, S.P., Hiebert, S., Mudryj, M., Horowitz, J.M. & Nevins, J.R. The E2F transcription factor is a cellular target for the RB protein. *Cell* **65**, 1053-61 (1991).
33. Benevolenskaya, E.V. & Frolov, M.V. Emerging links between E2F control and mitochondrial function. *Cancer Res* **75**, 619-23 (2015).
34. McMahon, C.M. *et al.* Clonal Selection with RAS Pathway Activation Mediates Secondary Clinical Resistance to Selective FLT3 Inhibition in Acute Myeloid Leukemia. *Cancer Discov* **9**, 1050-1063 (2019).
35. Gonen, M. *et al.* CD25 expression status improves prognostic risk classification in AML independent of established biomarkers: ECOG phase 3 trial, E1900. *Blood* **120**, 2297-306 (2012).
36. Xu, Z.J. *et al.* The M2 macrophage marker CD206: a novel prognostic indicator for acute myeloid leukemia. *Oncoimmunology* **9**, 1683347 (2020).
37. Lasry, A. *et al.* An inflammatory state remodels the immune microenvironment and improves risk stratification in acute myeloid leukemia. *Nat Cancer* **4**, 27-42 (2023).
38. de Pooter, R.F. *et al.* Cutting Edge: Lymphomyeloid-Primed Progenitor Cell Fates Are Controlled by the Transcription Factor Tal1. *J Immunol* **202**, 2837-2842 (2019).
39. Kim, Y. *et al.* Terminal deoxynucleotidyl transferase and CD84 identify human multi-potent lymphoid progenitors. *Nat Commun* **15**, 5910 (2024).
40. Crescenzi, B. *et al.* NUP98/11p15 translocations affect CD34+ cells in myeloid and T lymphoid leukemias. *Leuk Res* **39**, 769-72 (2015).
41. Zeng, A.G.X. *et al.* A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. *Nat Med* **28**, 1212-1223 (2022).
42. Bottomly, D. *et al.* Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer Cell* **40**, 850-864 e9 (2022).
43. Lambo, S. *et al.* A longitudinal single-cell atlas of treatment response in pediatric AML. *Cancer Cell* **41**, 2117-2135 e12 (2023).
44. Xu, J. *et al.* A multiomic atlas identifies a treatment-resistant, bone marrow progenitor-like cell population in T cell acute lymphoblastic leukemia. *Nat Cancer* (2024).
45. Newman, A.M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782 (2019).
46. Fugman, D.A., Witte, D.P., Jones, C.L., Aronow, B.J. & Lieberman, M.A. In vitro establishment and characterization of a human megakaryoblastic cell line. *Blood* **75**, 1252-61 (1990).
47. Takeda, A., Goolsby, C. & Yaseen, N.R. NUP98-HOXA9 induces long-term proliferation and blocks differentiation of primary human CD34+ hematopoietic cells. *Cancer Res* **66**, 6628-37 (2006).
48. Chung, K.Y. *et al.* Enforced expression of NUP98-HOXA9 in human CD34(+) cells enhances stem cell proliferation. *Cancer Res* **66**, 11781-91 (2006).
49. Cardin, S. *et al.* Human models of NUP98-KDM5A megakaryocytic leukemia in mice contribute to uncovering new biomarkers and therapeutic vulnerabilities. *Blood Adv* **3**, 3307-3321 (2019).
50. Diaz, A., 3rd *et al.* Abstract C099: Establishment and characterization of an NUP98-KDM5A-driven AML XPD model. *Molecular Cancer Therapeutics* **22**, C099-C099 (2023).
51. Saleque, S., Cameron, S. & Orkin, S.H. The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. *Genes Dev* **16**, 301-6 (2002).
52. Cai, M. *et al.* Dual actions of Meis1 inhibit erythroid progenitor development and sustain general hematopoietic cell proliferation. *Blood* **120**, 335-46 (2012).
53. Elagib, K.E. *et al.* Neonatal expression of RNA-binding protein IGF2BP3 regulates the human fetal-adult megakaryocyte transition. *J Clin Invest* **127**, 2365-2377 (2017).
54. Hecht, V. *et al.* Analyzing histone ChIP-seq data with a bin-based probability of being signal. *PLoS Comput Biol* **19**, e1011568 (2023).
55. Foulkes, W.D., Flanders, T.Y., Pollock, P.M. & Hayward, N.K. The CDKN2A (p16) gene and human cancer. *Mol Med* **3**, 5-20 (1997).
56. Stott, F.J. *et al.* The alternative product from the human CDKN2A locus, p14(ARF), participates in a regulatory feedback loop with p53 and MDM2. *EMBO J* **17**, 5001-14 (1998).

57. Mills, E.W., Green, R. & Ingolia, N.T. Slowed decay of mRNAs enhances platelet specific translation. *Blood* **129**, e38-e48 (2017).
58. Guy, C.T., Zhou, W., Kaufman, S. & Robinson, M.O. E2F-1 blocks terminal differentiation and causes proliferation in transgenic megakaryocytes. *Mol Cell Biol* **16**, 685-93 (1996).
59. Swaminathan, M., Bourgeois, W., Armstrong, S.A. & Wang, E.S. Menin Inhibitors in Acute Myeloid Leukemia-What Does the Future Hold? *Cancer J* **28**, 62-66 (2022).
60. Uckelmann, H.J. *et al.* Therapeutic targeting of preleukemia cells in a mouse model of NPM1 mutant acute myeloid leukemia. *Science* **367**, 586-590 (2020).
61. Uckelmann, H.J. *et al.* Mutant NPM1 Directly Regulates Oncogenic Transcription in Acute Myeloid Leukemia. *Cancer Discov* **13**, 746-765 (2023).
62. Wang, X.Q.D. *et al.* Mutant NPM1 Hijacks Transcriptional Hubs to Maintain Pathogenic Gene Programs in Acute Myeloid Leukemia. *Cancer Discov* **13**, 724-745 (2023).
63. Soto-Feliciano, Y.M. *et al.* A Molecular Switch between Mammalian MLL Complexes Dictates Response to Menin-MLL Inhibition. *Cancer Discov* **13**, 146-169 (2023).
64. Issa, G.C. *et al.* The menin inhibitor revumenib in KMT2A-rearranged or NPM1-mutant leukaemia. *Nature* **615**, 920-924 (2023).
65. Perner, F. *et al.* MEN1 mutations mediate clinical resistance to menin inhibition. *Nature* **615**, 913-919 (2023).
66. Tiong, I.S., Ritchie, D.S. & Blombery, P. Response and Resistance to Menin Inhibitor in UBTF-Tandem Duplication AML. *N Engl J Med* **390**, 2323-2325 (2024).
67. Wang, W. *et al.* Enhanced Megakaryopoiesis and Platelet Activity in Hypercholesterolemic, B6-Ldlr<sup>-/-</sup>, Cdkn2a-Deficient Mice. *Circ Cardiovasc Genet* **9**, 213-22 (2016).
68. Inaba, H. *et al.* Phase I pharmacokinetic and pharmacodynamic study of the multikinase inhibitor sorafenib in combination with clofarabine and cytarabine in pediatric relapsed/refractory leukemia. *J Clin Oncol* **29**, 3293-300 (2011).
69. Alexander, T.B. *et al.* Phase I Study of Selinexor, a Selective Inhibitor of Nuclear Export, in Combination With Fludarabine and Cytarabine, in Pediatric Relapsed or Refractory Acute Leukemia. *J Clin Oncol* **34**, 4094-4101 (2016).
70. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
71. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
72. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8**, 652-4 (2011).
73. Chen, X. *et al.* CONCERTING: integrating copy-number analysis with structural-variation detection. *Nat Methods* **12**, 527-30 (2015).
74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
75. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-76 (2012).
76. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
77. Wu, G. *et al.* The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat Genet* **46**, 444-450 (2014).
78. Jakubek, Y.A. *et al.* Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer. *Nat Biotechnol* **38**, 90-96 (2020).
79. Edmonson, M.N. *et al.* Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* **27**, 865-6 (2011).
80. Hagiwara, K. *et al.* RNAIndel: discovering somatic coding indels from tumor RNA-Seq data. *Bioinformatics* **36**, 1382-1390 (2020).
81. Hagiwara, K., Edmonson, M.N., Wheeler, D.A. & Zhang, J. indelPost: harmonizing ambiguities in simple and complex indel alignments. *Bioinformatics* (2021).
82. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
83. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
84. Edmonson, M.N. *et al.* Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline variants. *Genome Res* **29**, 1555-1565 (2019).
85. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

86. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
87. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
88. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
89. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
90. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e21 (2019).
91. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e29 (2021).
92. McInnes L, H.J., & Melville J (2018). . UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Retrieved from <http://arxiv.org/abs/1802.03426>.
93. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2018).
94. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-3 (2012).
95. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
96. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
97. Steen, C.B., Liu, C.L., Alizadeh, A.A. & Newman, A.M. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods Mol Biol* **2117**, 135-157 (2020).
98. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
99. Wei, J. *et al.* Targeting REGNASE-1 programs long-lived effector T cells for cancer therapy. *Nature* **576**, 471-476 (2019).