GTAC enables parallel genotyping of multiple genomic loci with chromatin accessibility profiling in single cells

Sven Turkalj[1,2†], Niels Asger Jakobsen[1,2,3†], Angus Groom[1,2], Marlen Metzner[1,2], Simone G. Riva[1,4], E. Ravza Gür[1,4], Batchimeg Usukhbayar[1,2], Mirian Angulo Salazar[1,2], Lance D. Hentges[1,4], Gerda Mickute[1,2], Kevin Clark[5], Paul Sopp[5], James O.J. Davies[1,2,3], Jim R. Hughes[1,4], Paresh Vyas[1,2,3].

[1]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK [2]Oxford Centre for Haematology, NIHR Oxford Biomedical Research Centre, Oxford, UK. [3]Department of Haematology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [4]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [5]Flow Cytometry Facility, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK

[†]These authors contributed equally

Corresponding and lead contact authors: Paresh Vyas (paresh.vyas@imm.ox.ac.uk).

Summary

Understanding clonal evolution and cancer development requires experimental approaches for characterizing the consequences of somatic mutations on gene regulation. However, no methods currently exist that efficiently link high-content chromatin accessibility with high-confidence genotyping in single cells. To address this, we developed Genotyping with the Assay for Transposase-Accessible Chromatin (GTAC), enabling accurate mutation detection at multiple amplified loci, coupled with robust chromatin accessibility readout. We applied GTAC to primary acute myeloid leukemia, obtaining high-quality chromatin accessibility profiles and clonal identities for multiple mutations in 88% of cells. We traced chromatin variation throughout clonal evolution, showing the restriction of different clones to distinct differentiation stages. Furthermore, we identified switches in transcription factor motif accessibility associated with a specific combination of driver mutations, which biased transformed progenitors towards a leukemia stem cell-like chromatin state. GTAC is a powerful tool to study clonal heterogeneity across a wide spectrum of pre-malignant and neoplastic conditions.

Introduction

Somatic genetic variation leads to clonal mosaicism[1], pre-cancerous clonal outgrowth[2,3] and neoplastic transformation[4,5]. Cancerous clonal evolution driven by serial and/or branched acquisition of somatic mutations is a cause of intra-tumoral heterogeneity[6], therapy resistance[7], and disease relapse[8,9]. Hence, knowledge of how somatic mutations corrupt cellular phenotype is a prerequisite for understanding the mechanisms underlying clonal dynamics and for the development of personalized therapeutic approaches.

Chromatin is the medium through which transcription factors (TFs), epigenetic regulators and signalling pathways control gene expression and cell phenotype[10,11]. Epigenetic dysregulation is a common feature in a wide range of human cancers, contributing to disease progression through activation of oncogenes, silencing of tumor suppressor genes, or disruption of differentiation programs[12]. Furthermore, many human cancers and pre-cancerous clonal outgrowths harbour somatic mutations in genes encoding epigenetic regulators, which have a direct effect on gene regulation, clonal expansion, and differentiation bias[2,12-15]. An example of such cancers are hematopoietic neoplasms such as acute myeloid leukemia (AML)[16]. Methods that enable the epigenetic consequences of somatic variation to be characterized in primary patient samples are therefore critical to understand the mechanisms underlying cancer development.

Recent strategies combining single-cell RNA-sequencing (scRNA-seq) and targeted genotyping allow linkage of genetic lineage history with transcriptomic heterogeneity within primary human samples, in which mutant and wild-type (WT) cells are often admixed without any known surface markers allowing for their purification[14,17-20]. Albeit powerful, these approaches fail to capture epigenetic variation and specifically chromatin accessibility, limiting our understanding of how somatic variation is linked to epigenetic state and TF activity. The assay for transposase-accessible chromatin with sequencing (ATAC-seq)[10,21,22] and single-cell ATAC-seq (scATAC-seq)[23-25] have enabled the study of chromatin accessibility variation across heterogeneous cell types whilst methods coupling scATAC-seq with transcriptome sequencing[26-29] and surface[30] or intra-cellular protein profiling[31] capture multiple layers of molecular and phenotypic information. Finally, integration of mitochondrial genome sequencing with chromatin accessibility linked clonal dynamics with epigenetic state[32]. However, no method links chromatin with driver somatic events at high efficiency in primary tissues, representing a major unmet need in our analyses of somatic disease.

Here we present a technique that addresses this methodological gap, termed Genotyping with the Assay for Transposase-Accessible Chromatin (GTAC), which couples single-cell genotyping of multiple genomic loci with robust scATAC-seq in FACS-sorted nuclei. We first

benchmarked GTAC in cell lines, showing accurate mutation calling enabled by parallel multi-allelic capture of several genomic targets, whilst maintaining a high quality scATAC library. Next, by applying GTAC to a primary AML sample, we captured clone-specific chromatin accessibility at defined stages of differentiation, providing mechanistic insights into the epigenetic consequences of specific driver mutations within a complex clonal hierarchy.

Results

GTAC workflow

Coding mutations are not consistently covered in scATAC libraries because most known driver mutations occur within exons, whereas fragments from scATAC libraries map mostly to regulatory genomic loci. We therefore designed GTAC, a method for simultaneous genotyping of multiple driver mutations and high-quality chromatin accessibility profiling from the same cell (Figure 1). Cell lysis and Tn5 tagmentation are performed on a cell population, followed by FACS sorting of single nuclei into plates[33,34]. Genotyping targets are captured by the addition of target-specific primers to the initial PCR, which amplifies genomic targets in parallel with fragments of accessible chromatin. Careful optimization of primer design (STAR Methods) is crucial for minimal interference with scATAC library generation. We optimised both the PCR protocol and the indexing strategy, to improve scATAC library complexity and increase cell throughput. The incorporation of 384 ‘i7’ cell barcodes and ‘i5’ plate barcodes during the initial PCR step enables scATAC libraries from many plates to be sequenced together without additional PCR reactions. After the PCR, 80% of the material is pooled and purified to yield scATAC libraries. The remaining 20% of the initial library material is transferred to another plate and used for further target amplification via two additional PCR reactions (PCR1 and PCR2), which fully barcode the genotyping libraries. scATAC and genotyping libraries are then sequenced and pre-processed separately. After single-cell mutation calling for each amplified locus, genotype information for every cell is linked to its accessibility profile via cell barcodes (STAR Methods).

GTAC allows for sensitive mutation detection at multiple genomic loci in cell lines

To assess efficient capture of genotyping targets, we first performed GTAC on K562 cells, amplifying previously reported mutant or variant loci (Figure 2A and Table S1). We compared coverage across the five loci in K562 cells between libraries prepared with standard scATAC without genotyping primers and GTAC (Figure 2B). The median number of reads covering the variant nucleotide was 0 for scATAC libraries, whereas it was 7433 reads for GTAC libraries across all loci, indicating that genotyping primers are needed for efficient locus capture. Target amplification was equally efficient in the absence of Tn5 (Figure 2C), showing that locus

capture is unrelated to Tn5 integration. This indicates that genotyping locus amplification does not depend on its chromatin accessibility.

We next benchmarked the accuracy and sensitivity of genotyping, targeting previously reported mutant or variant loci in several cell lines, using THP1 cells as negative wild type (WT) control (Figure 2D). For all five mutant or variant loci in K562, the single-cell variant allele frequency (scVAF) of mutant reads was <1% in THP1 cells (Figures 2E and S1A-S1E), allowing us to set thresholds for mutation calling. scVAF distributions observed in K562 were consistent with the reported zygosity[35] (Figure 2E; note that K562 is a triploid cell line, in which a heterozygous mutation is present on either 1 out of 3, or 2 out of 3, alleles). In addition, we targeted *RUNX1* and *NRAS* loci in JURKAT and OCI-AML3 cells, respectively, and observed a marked difference in scVAF distributions comparing the two (Figures 2F), consistent with the *RUNX1* mutation being heterozygous in JURKAT (average scVAF was 0.47) and the *NRAS* mutation being homozygous in OCI-AML3 cells (average scVAF was 0.99)[35], providing evidence that our genotyping data can discriminate between a heterozygous and a homozygous mutant population.

Next, we estimated allelic drop-out (ADO) rates for the profiled heterozygous loci across cell lines (Figure 2G). ADO, caused by failed amplification of one allele during PCR, can result in misassignment of heterozygous mutant loci as WT, compromising the accuracy of downstream analysis. For the nine heterozygous loci, we considered that ADO occurred if a heterozygous mutant cell line was wrongly assigned as homozygous mutant or homozygous WT. By using scVAF values from THP1 cells to set thresholds for genotype calling (Figures S1A-S1J), we observed variable but consistently low ADO rates across the profiled loci, ranging from 1.3% for *NOTCH1* up to 11.9% for *HOXA9* (Figure 2G).

Finally, given that accurate genotyping of multiple loci is crucial for high-confidence reconstruction of clonal hierarchies, we tested whether GTAC could reliably detect multiple mutations in the same single cells by simultaneously amplifying 5 mutant loci in K562 cells. All 5 mutations were detected in 79% of cells, at least 4 mutations were detected in 94%, while only 6% of cells failed genotyping at 2 or more loci (Figure 2H). Given that failed genotyping of an early mutation can still lead to correct clonal assignment if later mutations are efficiently detected in the same cells (STAR Methods), this suggests that GTAC can be used for correct clonal assignment of most single cells.

We thus demonstrate that GTAC enables high-confidence, accurate genotyping of multiple genomic loci within single cells.

Targeting multiple genotyping loci does not disrupt scATAC library quality

We next assessed whether GTAC consistently yielded high-quality scATAC libraries with increasing numbers of genotyped loci. We performed GTAC on K562 cells targeting 4, 6 or 8 loci and compared scATAC library metrics with those of an scATAC only control. All genotyping amplicons were detected in over 90% of cells in all GTAC conditions (98.9%, 92.6% and 93.7%, respectively), indicating no interference between genotyping amplicons when increasing numbers of loci are genotyped (Figure S1K).

Library complexity (Figure 3A) and signal-to-noise ratio (fraction of reads in peaks; FRIP) (Figure 3B) were comparable across GTAC conditions and not significantly different from the scATAC control. GTAC showed improved library complexity and FRIP compared to 10X Chromium droplet-based technology[36]. Median mitochondrial read contamination was <1% in all GTAC and our scATAC conditions, in contrast to >5% with the published plate-based method (Figure 3C). Additionally, transcription start site (TSS) enrichment (Figure 3D) and aggregated fragment size distribution (Figure 3E) were unaltered between GTAC and scATAC. While most cells displayed TSS enrichment values above QC thresholds, subtly lower median values in our conditions compared to prior data (Figure 3D) are likely related to more efficient capture of di-nucleosomal fragments (Figure S1L; eukaryotic promoters are largely nucleosome-free[37]). Comparison of peak calls showed strong correlation between the three GTAC and the scATAC-seq conditions ($r = 0.96$ for all comparisons) and with two public K562 datasets produced via plate-based scATAC[33] (Figure 3F). In addition, visual inspection of aggregated data from single cells suggested no qualitative differences between these conditions (Figure 3G). Importantly, visual inspection of scATAC genomic coverage surrounding the profiled mutant loci in K562 cells showed no difference between conditions, suggesting no disruption of scATAC fragment amplification in proximity of genotyped loci (Figure S1M). We observed >90% of cells passing scATAC quality control (QC) regardless of the number of genotyped loci, comparable to standard plate-based scATAC[33] (Figure 3H). Likewise, the percentage of cells passing scATAC QC in which more than 30 sequencing reads were obtained for all genotyped loci was >90% in all conditions (Figure 3I). Crucially, the percentage of cells passing scATAC QC in which all profiled heterozygous mutations were correctly detected was 80% or higher across conditions (Figure 3J), leading to high numbers of cells useful for analysis.

To test integration of scATAC and genotyping data, we performed a cell line mixing experiment with K562, JURKAT and OCI-AML3 cells, targeting 3 genomic loci in each cell line (1 mutant and 2 WT loci per cell line). Clustering based on scATAC data alone revealed three distinct clusters (Figure 3K). Integrating chromatin accessibility data with genotyping information revealed correct mutational calls for 97.2% of cells (Figures 3L-3N). Taken together, we

demonstrated that GTAC generates high quality scATAC data unperturbed by targeted genotyping of multiple loci, and that scATAC and genotyping data can be integrated.

GTAC links clonal hierarchy to accessibility landscape and identifies restriction of clones to distinct differentiation stages in human acute myeloid leukemia

To investigate how clonal structure correlates with chromatin accessibility in a primary malignant tissue, we applied GTAC to 3520 AML cells from a patient with expanded granulocyte-monocyte progenitor (GMP) and lymphoid-primed multipotent progenitor (LMPP) immunophenotypic compartments (Figures S2A and S2B). We genotyped five mutant loci previously identified by bulk sequencing (Figure 4A), as well as two heterozygous germline single nucleotide polymorphisms (SNPs) to estimate ADO in primary sample (Figure S2C). Distribution of scVAFs for two germline SNPs in *SH2B3* and *ASXL2* indicated very low WT (2.0% and 3.4%) and variant (2.2% and 3.1%) ADO rates (Figure S2D). This allowed bi-allelic detection in 95.8% of cells for the *SH2B3* locus and in 93.5% of cells for the *ASXL2* locus, which was consistent across plates (Figure S2E).

To reconstruct clonal hierarchy, we performed mutation calling on the five mutant loci. High genotyping success rates (from 94% to 98% per amplicon) resulted in 3409 cells (96.8%) being assigned to a clone (Figure 4B). Moreover, 3113 cells (88.44%) passed stringent scATAC QC (STAR Methods) and were successfully genotyped (Figure 4C). Mutant cell scVAF distributions revealed that *TET2* and *ASXL1* mutations were heterozygous, while *RUNX1* and the two X-linked *BCOR* mutations were homozygous and hemizygous, respectively (Figure 4D). Of note, the *RUNX1* R201G missense mutation disrupts DNA binding[38-40], suggesting that *RUNX1* R201G homozygous mutant cells in this sample harbour a complete *RUNX1* loss-of-function (LoF). By integrating genotype information across loci (Figure 4E), we reconstructed the most likely clonal hierarchy in which *TET2*, *ASXL1*, and *RUNX1* mutations were acquired in linear fashion (T, TA and TAR clones), followed by branched acquisition of two distinct *BCOR* mutant clones (TARB1 and TARB2) (Figure 4F). Notably, mutant cell fractions obtained with GTAC correlated well with clone size estimates from bulk sequencing (Figure S2F).

scATAC data analysis indicated high library quality (Figures S2G-S2K), with improved metrics as compared to those obtained for leukemia samples processed with commercial droplet-based technology[41] (Figures S2G-S2I). We first used chromatin accessibility data to perform dimensionality reduction and clustering agnostic to genotyping information. Of note, we did not observe plate-related batch effects (Figure S2L). Cluster identities (Figure 4G) were assigned using gene activity (i.e. accessibility) scores of key marker genes, binding motif accessibility for key transcription factors, as well as differential enrichment of marker peaks

(Figure S3A-S3D, STAR Methods). Projection of our dataset onto a healthy human hematopoiesis scATAC-seq reference[41] (Figure S3E, STAR Methods) was broadly consistent with our cluster assignment, though there were differences, which is unsurprising given that leukemogenic mutations disrupt differentiation trajectories. Readout of chromatin openness in regulatory and/or coding regions confirmed higher accessibility of the quiescence-related *PBX1* gene[42] in the hematopoietic stem cells/multipotent progenitors (HSC/MPP) and megakaryocytic/erythroid progenitors (Mk/EP), of *GATA1* in Mk/EP cells, of *CD3E/D/G* genes in T/NK cells, and of *CD79B* in the B progenitor-like cluster (Figure S4A). In contrast, the *SETBP1* gene, preferentially expressed in leukemic stem cells (LSCs)[43], as well as *FLT3*, were more accessible in the clusters resembling lympho-myeloid progenitors (Figure S4B).

Next, by projecting AML driver mutation clonal structure onto chromatin accessibility data, we noticed a striking segregation of genotypes into separate epigenomic entities (Figure 4H). 97.5% of T/NK cells were WT with the remaining cells being *TET2* mutant. 87.1% of HSC/MPP and 82.5% of Mk/EP clusters were composed of *TET2* single mutant cells intermingled with a minority of *TET2/ASXL1* double-mutant cells, suggestive that these populations were preleukemic[3,44,45]. In stark contrast, *TET2*/*ASXL1* double-mutant clones that acquired either a *RUNX1* mutation, or *RUNX1* and *BCOR* mutations, were distributed almost exclusively within the lympho-myeloid, myeloid and B-cell progenitor-like clusters, contributing to more than 99% of these populations. We and others have previously shown that the AML LSC populations reside in these populations[46-48]. This suggests that in this patient, the *RUNX1* mutation acts as a key transforming event, leading from preleukemia to leukemia. Notably, *RUNX1* mutant clones displayed increased accessibility of several genes reported as upregulated in *RUNX1* mutant compared to *RUNX1* WT AML[49,50], some of which are lymphoid markers, *DNTT*, *BLNK*, *IRF8* and *MS4A1,* as well as *MN1,* associated with worse AML prognosis (Figure S4C)[51]. Although *BCOR* mutant clones did not form separate progenitor clusters within the leukemic progenitor populations, they were significantly enriched within the most primitive LMPP-like differentiation stage (Figure 4I).

Unsurprisingly, transcription factor motif openness (Figure 4J) reflected the differentiation stage of the clones. WT cells were enriched for DNA-binding motifs for T cell-associated TFs EOMES[52,53], LEF1[54], and RUNX[55]. The preleukemic T/TA clones were enriched for TF motifs important in HSC/MPP (HOX, GATA2) and in megakaryocytic and erythroid cells (GATA1, the heterodimeric bZIP TF NFE2/MAFF, and members of the nuclear factor family NFIA/NFIC). Finally, in the leukemic TAR/TARB1/TARB2 clones, enriched TF motifs included those important for lympho-myeloid progenitor differentiation[56,57] (SPI1, CEBP, and members of the IRF, AP-1, MEF2 and TCF families). MEF2C/D factors, whose motifs were particularly enriched in *BCOR* mutants, have been implicated in AML chemotherapy resistance[58-60], and

MEF2C is upregulated in LSCs compared to leukemic progenitor cells[43]. Footprinting analyses confirmed higher DNA binding of the EOMES in T/NK cells, HOXB4 in preleukemic clones, and of SPI1 in leukemic clones (Figures S4D).

To understand whether the segregation of leukemic and preleukemic clones into different chromatin states was mainly driven by the *RUNX1* mutation or by a difference in differentiation stage, we integrated genotyping information with the healthy human hematopoiesis reference projection[41] (Figure S4E). This showed that 97/595 genetically preleukemic (*TET2* (*T*) and *TET2/ASXL1* (*TA*)) cells, as well as most genetically leukemic (TAR/TARB1/TARB2) LMPP-like cells, mapped to the normal CMP/LMPP population (Figures S3E, S4E, S4F). In contrast, only 1/2141 genetically leukemic (TARB2) cell mapped to healthy HSC. This suggests two possibilities: (1) the preleukemic cell-of-origin of leukemic clones transited through the CMP/LMPP like stage when *RUNX1* mutation was acquired, or (2) a preleukemic HSC acquired the *RUNX1* mutation, but this caused clonal dominance of its progeny at the LMPP stage. Importantly, both scenarios suggest profound chromatin reprogramming at the LMPP stage upon *RUNX1* mutation acquisition on the background of preleukemic mutations and are concordant with our prior finding that the LMPP population is the closest normal counterpart of LMPP-like LSCs[46]. Finally, to understand chromatin changes associated specifically with transformation, we compared the accessibility profiles of genetic preleukemic and leukemic cells that mapped to the normal CMP/LMPP compartment. Interestingly, while *RUNX1* gene activity was greater in leukemic CMP/LMPP cells (Figure S4G), its binding motif was less accessible (Figure S4H), consistent with disrupted DNA binding. Leukemic cells mapping to CMP/LMPP also showed higher accessibility of WT1 and MEIS1 motifs (Figure S4H), linked to AML maintenance[61,62]. Consistent with higher MYC motif accessibility in leukemic cells mapping to CMP/LMPP (Figure S4H), accessible genes were enriched for pathways associated with proliferation (MYC targets and cell cycle regulators), MAPK and PI3K signalling (Figure S4I). These data, which take differentiation stage into account, indicate that leukemic transformation at the CMP/LMPP stage correlates with acquisition of greater binding by TFs known to support LSC function and proliferation.

Taken together, these results demonstrate that in a primary AML sample, GTAC captures the stages of clonal evolution from WT cells through to preleukemia and into the fully transformed AML leukemic stage, showing how distinct clones correlate with different chromatin accessibility landscapes.

GTAC reveals a differentiation delay of *BCOR* mutant clones within the leukemic hierarchy, which biases cells towards an LSC-like chromatin state

Further selection of two independent *BCOR* mutant clones suggests the acquisition of *BCOR* mutations conferred additional fitness advantage in this somatic background. BCOR is a key component of the repressive non-canonical PRC1.1 complex. Both *BCOR* mutations are likely LoF (both occur in exon 4; one results in a premature stop codon, while the other is a frameshift mutation), suggesting these may lead to loss of PRC1.1 function in these clones. Compared to other leukemic stages, we noticed higher gene activity of *BCOR* and other PRC1.1 components within the LMPP-like cluster (Figure S5A), where *BCOR* mutant cells accumulate (Figure 4I), suggesting PRC1.1 may play a role for either entry into, or exit from the LMPP-like compartment. Interestingly, the LMPP-like cluster displayed the highest LSC score[43] amongst all leukemic stages, as well as higher accessibility of several motifs bound by TFs functionally shown to promote self-renewal and AML maintenance[61,63-67] (Figures 5A, 5B). Furthermore, LMPP-like and Mono/Neut progenitor-like cells exhibited higher accessibility of the *TIM-3* gene that discriminates between LSCs and HSCs[68], and LMPP-like cells had lower accessibility of *CD38* that is downregulated in LSCs in some patient samples[43] (Figure S5B). Consistently, open chromatin within LMPP-like cells was enriched for HOXA and FOXO target genes, compared to other leukemic clusters (Figure S5C). Compared to TAR cells, *BCOR* mutant cells were enriched for genes upregulated in HSCs and LSCs (Figure 5C), consistent with *BCOR* mutant cells being enriched for an LSC-like chromatin state.

To characterize dynamic chromatin changes associated with differentiation of specific leukemic subclones, we computed three putative differentiation trajectories undertaken by leukemic cells, from the LMPP-like compartment along the: pDC/NKP pathway; Mono/Neut pathway; or the B cell pathway (Figures 5D and S5D-S5F). We observed accumulation of cells from both *BCOR* mutant clones earlier in pseudotime along all trajectories (Figures 5D and S5G), but the distribution of cells across pseudotime was different between cells from the two *BCOR* mutant clones (Figure 5D). The larger *BCOR* mutant clone (TARB2) showed a more pronounced delay in the Mono/Neut and pDC/NKP trajectories compared to TARB1, which was not the case across the leukemic B-cell trajectory (Figure S5G). Concordantly, *BCOR* mutant cells showed delayed accessibility of motifs for lineage-affiliated TFs (Figures 5E and S5H). Notably, the magnitude of CEBPA motif accessibility through Mono/Neut pseudotime was lower in *BCOR* mutant cells, and this was particularly pronounced in the TARB2 clone (Figure 5E). Consistent with aberrant myeloid differentiation, there was delayed accessibility of regulatory elements co-accessible with the *TLR4* gene and of chromatin peaks mapping to the *CEBPD* promoter and gene body in TARB2 compared to TAR cells (Figure 5F). Moreover, TARB2 Mono/Neut progenitor-like cells displayed lower accessibility of peaks linked to genes associated with neutrophilic and monocytic differentiation, compared to TAR cells (Figure S5I). Conversely, motifs for ETV6, ERG, PBX3, FLI1, YY1, and NF-Y family members, functionally

implicated in stem/early progenitor cell self-renewal and/or AML maintenance[63,65-67,69-72], remained accessible for longer in *BCOR* mutant cells along the pDC/NKP trajectory (Figure 5G). In contrast, there was delayed accessibility of motifs for AP-1 complex members in *BCOR* mutant cells along the same trajectory (Figure 5H). Interestingly, at the early stages of the Mono/Neut pathway, *BCOR* WT cells gained AP-1 family motif accessibility faster compared to *BCOR* mutant cells (Figure S5J). The AP-1 family is involved in cell cycle regulation and myelomonocytic differentiation[73,74]; in addition, JUNB exerts a negative effect on leukemic self-renewal capacity[75] and, when overexpressed, leads to loss of HSCs in mice[76]. Indeed, within LMPP-like cells, motif accessibility of JUNB was negatively correlated to that of FOXO1 and ETV6, mediators of quiescence and self-renewal, and positively correlated to CEBPA motif accessibility, essential for myeloid differentiation (Figure 5I). Altogether, these data suggest that *BCOR* mutations delay differentiation, biasing leukemic cells towards stem-like chromatin states.

GTAC reveals switches in chromatin accessibility within the LMPP-like differentiation stage upon subclonal evolution of leukemia

Next, we wanted to study chromatin changes related to *BCOR* mutations within the LMPP-like cluster, to provide putative molecular mechanisms for the accumulation of *BCOR* mutant cells at this LSC-like stage. We first compared gene activity between *TET2/ASXL1/RUNX1* mutant *BCOR* WT and *TET2/ASXL1/RUNX1/BCOR* mutant LMPP-like cells (Figures 6A and 6B and Table S2). Notably, *BCOR* mutant LMPP-like cells were enriched in genes upregulated in LSC and HSC (Figure 6C). Both *BCOR* mutant clones showed more activity of the *INPP4B*, *SOCS2* and *NR4A2* genes (Figures 6A and 6B). *INPP4B* and *SOCS2,* putative direct PRC1.1 targets in primary human AML[77]*,* have been functionally shown to promote HSC and LSC maintenance[78-80]. Increased activity of the *NR4A2* gene, which promotes HSC quiescence[81], suggests a more quiescent state of *BCOR* mutant LMPP-like cells. Additionally, the larger *BCOR* mutant clone showed higher activity of *PBX3*, *BTG3,* and *JMJD1C* genes, putative PRC1.1 targets related to quiescence and/or promoting leukemogenesis[47,66,67,82,83]. Of note, both INPP4B, an enzyme which dephosphorylates inositol 3,4-biphosphate, and BTG3, which interacts with AKT[82], are negative regulators of the PI3K/AKT pathway. Conversely, the promoter region and several distal elements linked to *CDK20,* a positive cell cycle regulator*,* were less accessible in the TARB2 compared to TAR cells (Figure 6D).

We leveraged chromatin information to generate mechanistic hypotheses about aberrant gene regulation associated with *BCOR* LoF. Prior evidence suggests that PRC1.1 binding is strongly biased towards promoter regions[77,84]. We found that peaks more accessible in either TARB2 or TARB1, compared to TAR LMPP-like cells, were enriched for promoters over distal elements, consistent with de-repression of promoters upon loss of PRC1.1 activity (Figure

S6A). Consistently, both *BCOR* mutant clones showed increased accessibility of the *INPP4B* promoter region compared to TAR cells (Figures 6E and S6B). Of note, higher *INPP4B* gene activity persisted in *BCOR* mutant cells throughout all leukemic trajectories and was particularly pronounced in the TARB2 clone (Figure 6F). Interestingly, *INPP4B* is amongst the genes likely to be positively regulated by RUNX1 in mixed-lineage acute leukemia[41]. Consistent with this, *RUNX1* mutant leukemic cells displayed lower *INPP4B* activity compared to pre-leukemic cells (Figure S6C). It is tempting to speculate that *BCOR* mutations increase fitness advantage in this context by de-repressing the *INPP4B* promoter, compensating for lower *INPP4B* activity in *RUNX1* mutant cells (Figure S6C). In addition, TSS-proximal regions of *NR4A2*, *MLLT10*, *BTG3*, and *SOCS2* genes were also more accessible in *BCOR* mutant cells (Figures 6G and S6D-S6F), and their activity was increased through differentiation in one or both *BCOR* mutant clones (Figures 6H and S6G-S6I).

To better understand how *BCOR* mutations correlated with altered regulatory landscape, we compared TF motif accessibility between *TET2/ASXL1/RUNX1* mutant *BCOR* WT and *TET2/ASXL1/RUNX1/BCOR* mutant LMPP-like cells (Figures 6I and 6J and Table S3). Both *BCOR* mutant clones displayed higher accessibility of HOX family and SMAD motifs, associated with LSC and cancer stem cell function, consistent with previous evidence that *BCOR* mutations are associated with increased expression of *HOX* cluster genes[85,86]. However, there were also unexpected differences between the two *BCOR* mutant clones. TARB1 cells were enriched for ERG, ETV6, WT1, as well as for MEF and FOXO family members, whereas motifs for NFKB, VENTX, and NANOG were enriched in TARB2 cells. Interestingly, NF kB signalling was shown to promote HSC self-renewal and survival[87-89]. Accordingly, TF footprinting analysis confirmed more RelA binding in TARB2 LMPP-like cells (Figure S6J). In addition, several other motifs more accessible in one or both *BCOR* mutant clones were bound by TFs previously linked to stemness[65,70,90,91]. Conversely, AP-1, TCF7, IRF4, BCL11A/B, and CEBPA/D motifs, linked to differentiation and/or cell cycle, were more open in *BCOR* WT LMPP-like cells (Figure 6J).

Finally, to corroborate the unexpected differences between TARB1 and TARB2 clones, we directly compared their gene activity scores across leukemic populations. This revealed increased activity of genes involved in HSC and/or LSC function, including *HOXA9, HOXA10, KIT, BMI1*, *NR4A2* and *INPP4B,* in the larger TARB2 clone (Figure 6K). Conversely, *CDK20* and *CDC20*, which promote cell cycle, were more accessible in the TARB1 clone. Moreover, binding motifs of IRF4, BCL11A/B and SPI1 TFs were also more accessible in the TARB1 clone (Figure S6K). This raises the intriguing possibility that the TARB2 clone had a fitness advantage over the TARB1 clone, occupying a more stem-like chromatin state within the leukemic hierarchy.

Collectively, the above data suggest that *BCOR* mutations in this somatic context alter the balance between chromatin elements regulating quiescence, cell cycle and differentiation (Figure S6L). Differential gene and motif openness indicate aberrant reactivation of molecular actors promoting quiescence and maintenance of LSCs, at the expense of differentiation. GTAC thus provides specific insights of how mutational cooperativity correlates with chromatin heterogeneity, within defined leukemic differentiation stages.

Discussion

We present GTAC, a single-cell multi-modal approach mapping high quality chromatin accessibility profiles to complex clonal structures in malignant tissues. While our scATAC-seq library construction protocol simplifies and increases cell throughput compared to previous plate-based methods[33,34], targeted genotyping is easily implementable in any laboratory. Moreover, our analyses adopt broadly utilized computational pipelines, making GTAC a widely accessible method.

Functional assays that rely on either targeted mutagenesis or overexpression in model systems may miss the full breadth of clonal complexity underlying human cancer. Hence, primary samples provide an important benchmark to establish the cellular, epigenetic and molecular landscape in human pre-malignant and fully transformed clonal populations. While existing approaches demonstrated the correlation of altered genotypes with transcriptomic[17,18] or surface protein heterogeneity[92], linking somatic mutations to chromatin accessibility leads to several unique advantages. Firstly, chromatin is a more accurate indicator of cell identity compared to RNA[10], allowing for more granular analysis of cell populations. Secondly, chromatin accessibility enables study of dynamic alterations in TF networks downstream of mutations. Crucially, GTAC reveals proximal or distal regulatory elements with clone-specific activity, enabling mechanistic hypotheses to be generated about the impact of mutations on gene regulation.

We demonstrate the utility of GTAC in human AML, where linking epigenomic profiles to clonal progression identified a *RUNX1* mutation as a key driver event associated with leukemic transformation. Furthermore, more subtle chromatin changes downstream of *BCOR* mutations suggest a de-regulated epigenetic balance between self-renewal and differentiation in *BCOR* mutant progenitors, which gradually expand the pool of LSC-like cells in this patient. Finally, the unexpected difference between two *BCOR* mutant clones suggests that even cells with apparently similar somatic genotypes can display notable chromatin heterogeneity, resulting either from additional unidentified genetic/epigenetic driver(s), differences in the cells-of-origin, or due to distinct functional properties of the two mutations. Crucially, GTAC enabled comparison of different clones within specific scATAC-defined populations, a level of

granularity not achievable by conventional scATAC approaches. Going forward, GTAC could also be applied to longitudinal patient samples to elucidate clone-specific changes in chromatin profiles throughout treatment, providing insights into epigenetic mechanisms of therapy resistance.

GTAC has important advantages compared to droplet-based technologies[93]. First, high genotyping efficiency and an average ADO of only ~5% per locus make it possible to confidently assign the most likely phylogenetic tree across a range of putative clonal hierarchies, making GTAC uniquely suitable for the study of somatically complex disease. Low ADO rates are crucial to minimise error in downstream analyses as clone-specific differences in chromatin accessibility would progressively be diluted with increasing ADO rates. Second, locus capture does not depend on chromatin accessibility proximal to the amplified region, allowing for uniform genotyping across different cell states. Third, at parity of Tn5 concentration and cell type, GTAC yields substantially more complex scATAC libraries compared to other technologies[23-25,93,94], providing greater granularity, which is important when studying subtle differences arising from intratumoral heterogeneity. This is currently achieved with a cost per cell of approximately $3 (excluding sequencing). Finally, we speculate that our scATAC library complexity may be further improved by integrating the recently published s3-ATAC strategy[95].

In conclusion, we provide a robust and highly sensitive single-cell tool capable of resolving chromatin differences between somatic clones within the same tissue. Given that scATAC performs well on a vast spectrum of human cancers[96], GTAC will be widely applicable across the field of somatic mosaicism and tumor evolution.

Limitations of the study

The main limitation of GTAC is lower throughput compared to droplet-based strategies. In terms of cell numbers, one tagmentation reaction on 50,000 cells usually yields up to ten 384-well plates. This number can be increased depending on starting cell numbers and populations of interest. Timewise, the method requires around five days to process ten 384-well plates, but the time required progressively increases with the number of genotyped loci. Altogether, data from 10,000-12,000 cells can be generated in a few weeks. Lower throughput can be limiting when very small clones are studied: in such cases, GTAC is likely underpowered and may miss the full breadth of clone-specific chromatin accessibility. It should also be considered that GTAC relies on the analysis of mutations previously detected in the patient sample by either targeted panel or whole exome sequencing. Finally, for genetically complex tumors, in which hundreds of mutations are present within the same sample, it would be necessary to limit GTAC analysis to a selected subset of mutations.

Acknowledgements. P.V. acknowledges funding from the Medical Research Council Molecular Haematology Unit Programme Grant (MC_UU_00029/8), Blood Cancer UK Programme Continuity Grant 13008, NIHR Senior Fellowship and the Oxford BRC Haematology Theme. S.T. is supported by a Scatcherd European Scholarship in partnership with The Medical Research Council/Radcliffe Department of Medicine and The Clarendon Fund. N.A.J. is supported by an Medical Research Council and Leukaemia UK Clinical Research Training Fellowship (MR/R002258/1). A.G. is supported by a DPhil studentship funding from the Lady Tata Memorial Trust. S.G.R., E.R.G., and J.R.H. acknowledge funding from the Medical Research Council (MC_UU_00029/3) and a Wellcome Trust Strategic Award (106130/Z/14/Z). E.R.G. is supported by "The Ministry of National Education Selection and Placement of Candidates Sent Abroad for Postgraduate Education (YLSY) scholarship", Republic of Türkiye, Ministry of National Education. L.D.H. was supported by the National Institutes of Health (USA) (R24DK106766). G.M. is supported by a Cancer Research UK DPhil Studentship (C2195/A31310). J.O.J.D. is supported by an Medical Research Council Clinician Scientist Fellowship (MR/R008108) and the Medical Research Council Molecular Haematology Unit Programme Grant (MC_UU_00029/4) and the Lister Institute Fellowship. The authors thank Prof. Ketan Patel, Prof. Thomas Milne, Dr. Sally-Ann Clark, Dr. Bilyana Stoilova, Dr. David Cruz Hernandez, and Mr. Felix Radtke, for insightful comments and discussions. The authors also acknowledge the MRC WIMM Flow Cytometry Facility and Single Cell Facility. Some of the Figures in this manuscript were created using BioRender.

Author Contributions Conceptualization, S.T., N.A.J., A.G., P.V.; Methodology, S.T., N.A.J., A.G.; Investigation, S.T., N.A.J., A.G., M.M., B.U., M.A.S., G.M., K.C., P.S.; Formal Analysis, S.T., N.A.J., S.G.R., E.R.G., L.D.H.; Funding Acquisition, P.V.; Supervision, J.O.J.D., J.R.H., P.V.; Writing – Original Draft, S.T., N.A.J. and P.V.; Writing – Review & Editing, all authors.

Inclusion and Diversity We support inclusive, diverse, and equitable conduct of research.

Figure Legends

Figure 1. GTAC workflow.

Schematic overview of GTAC. Primer legend is shown on the left. Details regarding experimental procedures are described in STAR Methods.

Figure 2. GTAC enables sensitive mutation detection from multiple amplified genomic loci.

(A) Workflow for validation of genotyping amplicon capture in K562 cells, using scATAC alone (red) or the GTAC assay with (blue) or without (purple) the Tn5 enzyme.

(B) Number of sequencing reads in single K562 cells across the variant nucleotide in the indicated loci, comparing scATAC only (red) and GTAC libraries (blue). Median coverage in scATAC and GTAC libraries, respectively, for each locus: *TET2*: 0 and 4823; *NOTCH1*: 0 and 7993; *CBL*: 0 and 7461; *HOXA9*: 0 and 6961; *ASXL1*: 0 and 7980. Number of cells tested by GTAC is 240 for *ASXL1*, *CBL*, *HOXA9*, *NOTCH1*; 80 for *TET2*. Number of cells tested by scATAC is 95 for each locus.

(C) Percentage of K562 cells showing ≥30 sequencing reads across the amplified loci, comparing conditions with (blue) or without (purple) Tn5. Error bars show the standard error. Fisher's exact test was used (ns – not significant).

(D) Single-cell genotype calling validation workflow. Top, table shows cell lines tested, ploidy at interrogated loci, mutations/variants expected, and genomic loci amplified. Bottom, cells were tested with GTAC, and mutation calling was performed.

(E) Single-cell VAF (scVAF) values for variant loci in K562, in cells with ≥30 reads across the variant nucleotide. Median scVAF in K562 and THP1 cells (WT control), respectively, are: *ASXL1:* 0.345 and 0.001; *CBL*: 0.352 and 0.002; *NOTCH1*: 0.342 and 0.001; *HOXA9*: 0.327 and 0.002; *TET2*: 0.642 and 0.001. Median values are shown.

(F) scVAF for the *RUNX1* locus in JURKAT cells and for the *NRAS* locus in OCI-AML3 cells. Each dot is a single cell (n=95, each). Median scVAF for *RUNX1* and *NRAS* is 0.471 and 0.996, respectively.

(G) Percentage of cells in which both mutant and WT (green), only the mutant (purple), or only the WT (orange) allele was detected, across 9 heterozygous loci. Cell lines are indicated above. n = 236 (*ASXL1*), 238 (*CBL*), 235 (*HOXA9*), 232 (*NOTCH1*), 80 (*TET2*), 93 (*RUNX1*), 149 (*NPM1*), 151 (*DNMT3A*), 144 (*FLT3*). Mutant and WT allelic drop-out (ADO) rates for each locus are indicated below.

(H) Frequency of detection of heterozygous mutations in K562 cells when 5 mutant loci were amplified in parallel. The percentage of cells in which all 5 (5 detected; 79%), or at least 4 (≥4 detected; 94%), 3 (≥3 detected; 96%), 2 (≥2 detected; 98%) or 1 (≥1 detected; 98%) mutant loci were detected is shown. 2% of cells failed genotyping at all loci (Failed).

Figure 3. scATAC library quality is not altered with increasing numbers of amplified genomic loci.

In panels (A) to (H), 80 K562 cells were assayed in 4 different conditions: GTAC (blue) where either 4, 6 or 8 genotyping loci were co-amplified, or scATAC alone (red). In panels (A) to (D), (F), and (H), metrics were compared to published plate-based K562 scATAC-seq data[33] (green, Chen plate-based); in panels (A), (B) and (D), metrics were compared with published

K562 data obtained with a 10X Genomics droplet-based method[36] (purple, 10X Chromium). In panels (A) to (D), ordinary one-way ANOVA with Dunnett's multiple comparisons was used (ns – non-significant, adj. $P > 0.05$).

(A) Log_{10} of the number of unique nuclear ATAC fragments. Median values are: GTAC with 4 loci: 5.172; GTAC with 6 loci: 5.231; GTAC with 8 loci: 5.178; scATAC alone: 5.174; Chen plate-based: 4.589; 10X Chromium: 4.001. Median values are shown within boxplots. Whiskers extend from the 2.5^{th} to the 97.5^{th} percentile.

(B) Percentage of ATAC reads in called peaks. Median values are: GTAC with 4 loci: 34.2%; GTAC with 6 loci: 35.1%; GTAC with 8 loci: 35.0%; scATAC alone: 35.0%; Chen plate-based: 39.4%; 10X Chromium: 27.5%. The rest is as in (A).

(C) Percentage of reads mapping to the mitochondrial genome. Median values are: GTAC with 4 loci: 0.53%; GTAC with 6 loci: 0.56%; GTAC with 8 loci: 0.47%; scATAC alone: 0.48%; Chen plate-based: 6.05%. The rest is as in (A).

(D) Transcription start site (TSS) enrichment. Median values are: GTAC with 4 loci: 10.65; GTAC with 6 loci: 10.93; GTAC with 8 loci: 10.85; scATAC alone: 10.99; Chen plate-based: 11.95; 10X Chromium: 12.33.

(E) Aggregated fragment size distribution across our conditions.

(F) Correlation matrix of Pearson's coefficients comparing GTAC conditions and scATAC alone with published K562 (Chen K562 replicate 1 and Chen K562 replicate 2) and human skin fibroblast scATAC-seq datasets.

(G) Aggregated reads from genome positions chr19:35620000-35790000, across our conditions and plate-based K562 data. Positions of labelled genes are shown.

(H) Percentage of K562 cells passing scATAC QC across our conditions and for the published K562 data (STAR Methods).

(I) Percentage of K562 cells passing both scATAC QC and achieving successful detection of all amplified genotyping loci (STAR Methods).

(J) Percentage of K562 cells passing both scATAC QC and achieving successful detection of all four heterozygous mutations (STAR Methods).

In panels (K) to (N), 95 cells from each of the three indicated cell lines were subject to GTAC with amplification of 3 genomic loci, *ASXL1*, *RUNX1* and *NRAS*.

(K) UMAP based on scATAC data from 285 profiled single cells.

(L)-(N) The UMAP from (K) overlaid with genotyping data from three genomic loci. Legend on the left indicates whether a mutation was detected (red), only the WT sequence was detected

(blue), or if there were insufficient reads across the locus (grey). In each panel, the percentage of cells that are false negative (i.e. a mutant cell is called WT) and of cells displaying insufficient reads (Detection failure rate) is shown. Overall, 97% of cells were assigned the correct genotype.

Figure 4. GTAC analysis of primary human acute myeloid leukemia maps clonal hierarchy onto accessibility profiles and reveals the expansion of specific clones at distinct stages of differentiation.

(A) Application of GTAC to 3520 FACS-sorted single nuclei (85% $CD34^+$ and 15% $Lineage^+$ cells) from a primary human AML sample. Five loci previously identified as mutant by bulk sequencing were targeted.

(B) Percentage of cells assigned to a clone (dark blue, 3407 of 3520 cells), based on successful genotyping (STAR Methods).

(C) Percentage of cells passing scATAC QC and assigned to a clone (dark blue, 3113 of 3520 cells, STAR Methods).

(D) scVAF distributions in cells containing the mutation in the indicated genotyped loci. 'BCOR-1' and 'BCOR-2' refer to the two distinct *BCOR* mutations. Median scVAFs for each locus are: *TET2*: 0.507; *ASXL1*: 0.493; *RUNX1*: 0.997; *BCOR*-1: 0.987; *BCOR*-2: 0.996. Median and quartile values are shown for *TET2* and *ASXL1*.

(E) Top, single-cell genotyping results. Each row represents a locus (indicated on the left). Each column is a single cell. Colours indicate whether a cell was called WT (grey), mutant (cyan), or undetected (white), for each locus. Bottom, clonal assignment for each cell (colour legend on the right). WT, wild-type. T, *TET2* mutant. TA, *TET2*/*ASXL1* mutant. TAR, *TET2*/*ASXL1*/*RUNX1* mutant. TARB1, *TET2*/*ASXL1*/*RUNX1*/*BCOR-1* mutant. TARB2, *TET2*/*ASXL1*/*RUNX1*/*BCOR-2* mutant. Only cells assigned to a clone are displayed.

(F) Most likely clonal structure based on mutation calls across the five loci. Numbers of cells assigned to a clone are indicated. Abbreviations are as in (E).

(G) UMAP based on scATAC data of 3113 single cells from the AML sample. Cluster differentiation stage is defined by chromatin accessibility profiles. HSC/MPP: hematopoietic stem cells/multipotent progenitors; Mk/EP: megakaryocytic/erythroid progenitors; LMPP-like: lymphoid-primed multipotent progenitor-like; Mono/Neut progenitor-like: monocyte/neutrophil progenitor-like; LM progenitor-like: lympho-myeloid progenitor-like; pDC/NKP-like: plasmacytoid dendritic cell/NK progenitor-like.

(H) The UMAP from (G) overlaid with single-cell genotyping data. Clones are labelled below. Abbreviations are as in (F). Cell numbers: WT: 377; T: 576; TA: 19; TAR: 861; TARB2: 834; TARB1: 446.

(I) Percentage of cells from each clone within scATAC clusters. Fisher's exact test was used to assess the frequency of *BCOR* mutant clones between the LMPP-like and other progenitor clusters (significant P-values reported above the graph).

(J) Heatmap displaying enrichment of TF motifs (rows) in marker peaks for each clone. Each column is a clone. Colours indicate the normalized motif enrichment ($-\log_{10}$(adjusted P-value)) on a scale of 0-100. The maximum motif $-\log_{10}$(adjusted P-value) is indicated in brackets. Marker peaks were selected with FDR < 0.05 and $\log_2$FC > 0.5. Only motifs for transcription factors expressed in the hematopoietic system[97,98] are shown.

Figure 5. GTAC identifies clone-specific chromatin dynamics throughout the leukemic hierarchy, highlighting a differentiation delay of *BCOR* mutant clones at an LSC-like state

(A) LSC module score deviation from the average of all cells, shown for leukemic populations. Whiskers extend from the minimum to maximum value. Kruskall-Wallis test with Dunnett's multiple comparison was used and P-values are reported.

(B) Left: UMAP from Figure 4G overlaid with the LSC module score. Right: leukemic cells overlaid with motif accessibility for ERG, ETV6, PBX3, and WT1.

(C) GSEA of genes upregulated in HSCs and LSCs compared to committed progenitors[47]. All *BCOR* mutant cells were compared to all TAR cells. Normalized enrichment score (NES) and FDR values are indicated.

(**D**) Above: UMAP from Figure 4G depicting three differentiation trajectories within leukemia (STAR Methods). Cells are colour-coded by pseudotime score. Below: for each trajectory, clonal density is plotted as a function of pseudotime. The Kolmogorov-Smirnov test was used to compare distributions. P-values for each comparison are indicated.

(**E**) TCF4, CEBPA and EBF1 motif accessibility as a function of pseudotime, when each trajectory was split by clone (STAR Methods), indicated on the left.

(**F**) Normalized scATAC read coverage around the *TLR4* (left) and *CEBPD* (right) genes. Clone-specific Mono/Neut trajectories were split into five quantiles from low to high pseudotime (Q1-Q5). Peaks and co-accessibility are plotted below.

(G)-(H) Heatmaps of motif accessibility across clone-specific pDC/NKP pseudotime (x-axis), for a selected subset of most variable motifs (each row represents a motif). In each panel, the

upper heatmap relates to TAR and the lower heatmap to TARB2. Motifs bound by TFs discussed in the Results section are labelled on the right.

(**I**) Correlation between JUNB and ETV6 (left), FOXO1 (middle) and CEBPA (right) motif accessibility in LMPP-like cells (dots). Lines represent the linear fit and shadowed areas represent 95% confidence intervals. Spearman correlation coefficients and P-values are reported.

Figure 6. GTAC reveals specific switches in chromatin accessibility within the LMPP-like differentiation stage upon subclonal acquisition of *BCOR* mutations

(A)-(B) Volcano plots of differential gene activity between TARB1 (n=130) and TAR (n=167) clones (A), and TARB2 (n=293) and TAR clones (B), within the LMPP-like cluster. Differentially accessible genes display FDR < 0.05 and absolute $log_2FC > 0.25$. Lines indicate log_2FC and FDR thresholds. Differentially accessible putative PRC1.1 targets in AML[77] are labelled in green.

(C) GSEA of genes upregulated in HSCs and LSCs compared to committed progenitors[47]. *BCOR* mutant and WT LMPP-like cells were compared. NES and FDR values are indicated.

(D) Normalized scATAC read coverage around the *CDK20* gene, comparing TARB2 and TAR LMPP-like cells. Peaks and co-accessibility are plotted below. The arrow indicates gene orientation.

(E) Overlaid normalized scATAC read coverage around the *INPP4B* promoter region, comparing TAR and TARB2 LMPP-like cells. Peaks and gene annotation are indicated below. The arrow indicates gene orientation.

(F) *INPP4B* gene activity as a function of pseudotime for the three trajectories. Clones are colour coded (legend on the right). The loess method was used.

(G) Same as (E) but around the *NR4A2* gene.

(H) Same as (F) but relative to *NR4A2* gene activity.

(I)-(J) Volcano plots of differential TF motif accessibility between TARB1 and TAR clones (I), and TARB2 and TAR clones (J), within LMPP-like cells. Differentially accessible motifs display FDR < 0.05 and absolute $\Delta z > 0.25$ and are labelled in red. Lines indicate Δz and FDR thresholds.

(K) Volcano plot of differential gene activity between TARB1 (n=446 cells) and TARB2 (n=834 cells) clones across all leukemic cells. Differentially accessible genes display FDR < 0.05 and absolute $log_2FC > 0.20$ and are labelled in red. Lines indicate log_2FC and FDR thresholds.

STAR METHODS

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Paresh Vyas (paresh.vyas@imm.ox.ac.uk).

Materials Availability

The list of all oligo sequences designed in this study and used for single-cell genotyping can be found in Table S4. These include both target-specific oligos used in the first PCR, and nested barcoded target-specific oligos used in genotyping PCR1.

Data and Code Availability

Raw and processed scATAC-seq (GEO: GSE214812) data have been deposited at GEO and are publicly available as of the date of publication. Raw targeted single-cell genotyping data produced with GTAC (BioProject: PRJNA887752) have been deposited at NCBI's SRA data repository and are publicly available as of the date of publication. Python and Perl scripts used for scATAC or single-cell genotyping data pre-processing were published previously[17,33] and are available at the respective Github pages. Custom scripts used for downstream analysis are accessible at https://github.com/sventurkalj/GTAC as of the date of publication. Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture

K562, MOLM-13, THP1 and JURKAT human cell lines were cultured in RPMI-1640 medium (Cat# 21875034, Gibco) supplemented with 10% FBS and 1% V/V Pen-Strep (Cat# 15140122, Gibco). OCI-AML3 cells were cultured in Alpha-MEM medium with nucleosides (Cat# 12571063, Gibco) supplemented with 20% FBS and 1% V/V Pen-Strep. All cell lines were regularly screened for Mycoplasma contamination using the MycoAlert Mycoplasma Detection Kit (Cat# LT07-218). Cells were passaged every 2-3 days and seeded at approximately 500 000 cells/mL. All cell lines were kept in a CO_2 incubator at 37 $^{\circ}$C. Prior to GTAC procedures, viability was evaluated to assure >90% viability.

Patient samples

Patient BM samples were collected with informed consent under ethically approved protocols (MREC 06/Q1606/110 or NHS REC 17/YH/0382). Written informed consent was obtained in accordance with the Declaration of Helsinki. Cytogenetic analyses were performed in clinical

laboratories. Mononuclear cells (MNCs) were isolated by Ficoll density gradient centrifugation and viably frozen in 90% FBS/ 10% DMSO. Genomic DNA was extracted using a DNeasy Blood & Tissue Kit (Cat# 69506, Qiagen). The diagnostic human AML BM sample used for GTAC experiments was derived from a male AML patient aged 81 at the time of sample collection. Wild-type BM used for GTAC experiments was derived from a female orthopedic patient undergoing hip replacement surgery, aged 65 at the time of sample collection.

METHOD DETAILS

Targeted DNA sequencing

Library preparation and sequencing: Targeted DNA sequencing was performed on bone marrow MNCs. Pre-capture DNA libraries were prepared using the KAPA HyperPlus protocol (Cat# KK8514, Roche). 100ng of genomic DNA was fragmented by enzymatic fragmentation. Following end repair and A-tailing, adapter ligation was performed using KAPA dual-indexed adapters (Cat# KK8722, Roche). Library cleanup and double-sided size selection was performed using Agencourt AMPure XP beads (Cat# A63881, Beckman Coulter) to obtain fragments of ~320 bp. Libraries were amplified by ligation-mediated PCR for 6 cycles using a KAPA HiFi HotStart high-fidelity DNA polymerase (Cat# 7958927001, Roche) and purified using AMPure XP beads.

Targeted capture was performed using a custom pool of biotinylated capture probes (SeqCap EZ Prime Choice, Roche) targeting 97 genes recurrently mutated in myeloid malignancies spanning 347 kb[2,16,99-102]. Amplified DNA libraries were hybridized to the capture probes in pools of 10-12 samples according to the manufacturer's instructions. The captured DNA libraries were amplified by 14 cycles of PCR using a KAPA HiFi HotStart DNA polymerase and purified using AMPure XP Beads.

Post-capture amplified DNA libraries were quantified by Qubit (Life Technologies) and size distribution and quality analyzed using a Bioanalyser chip (Agilent Technologies). Libraries were pooled in equimolar concentrations and were sequenced on an Illumina NextSeq 500 using paired-end reads.

Variant calling in targeted panel sequencing: Sequencing data were analyzed in a custom pipeline based on the GATK Best Practices (GATK v4.1.2.0 and Picard v2.21.3). Raw sequencing reads were converted to an unmapped BAM file and adapter sequences soft-clipped using Picard *MarkIlluminaAdapters*. Following conversion back to a FASTQ file, reads were mapped to the hg38 human reference genome assembly using the Burrows–Wheeler aligner v0.7.17[103] with alternate contig-aware alignment. Mapped, and unmapped, BAM files were merged using *MergeBamAlignment* and reads from different sequencing lanes were combined. Duplicate reads were marked using Picard *MarkDuplicates* and base quality scores

recalibrated with GATK BaseRecalibrator and ApplyBQSR. Somatic variant calling was performed on the pre-processed BAM files using VarDictJava v1.7.0[104] and Mutect2[105] in tumour-only mode. For VarDict, variants were called with a minimum variant allele frequency of 0.01, minimum base quality score of 25 and minimum supporting reads of 2, with indel realignment and removal of adapter sequences. For Mutect2, a minimum tumor LOD of 2 was used and variants were filtered for sequence context-dependent artefacts using *FilterMutectCalls* and *FilterByOrientationBias*. Indels were left-aligned and normalized using bcftools norm (v1.9). Variants were annotated using Annovar[106]. Target enrichment metrics and coverage was calculated using Picard *CollectHsMetrics* and custom scripts.

VarDict and Mutect2 variant calls were analyzed separately to identify a consensus list of high-confidence variants. The following post-processing filters were applied to VarDict calls to exclude likely sequencing artefacts: (1) Minimum of 5 variant reads for SNVs (with at least 2 reads in forward and reverse directions), or minimum of 10 variant reads for indels (with at least 4 reads in forward and reverse directions). (2) Minimum base quality score 30. (3) Minimum mapping quality score 40, except for variants in *U2AF1;U2AF1L5*, where the mapping quality was ignored. This because, there is a duplication of the *U2AF1* gene on chromosome 21 called U2AF1L5, in hg38, which results in reads being flagged as multi-mapped. (4) Maximum strand bias Fisher p-value of 0.0001. (5) No position bias towards beginning or end of reads. The following post-processing filters were applied to Mutect2 calls: (1) Passed all default Mutect2 filters or only failing the *clustered_events* filter. (2) Minimum of 5 variant reads for SNVs (with at least 2 reads in forward and reverse directions), or minimum of 10 variant reads for indels (with at least 4 reads in forward and reverse directions).

Variants were flagged as likely germline, or sequencing artefacts, if any of the following applied: (1) Variant allele with a population allele frequency >1 in 1,000 according to any of three large polymorphism databases (Gnomad, 1000 Genomes Project, ESP6500) that is not a hotspot driver mutation with a COSMIC (v88) occurrence count of >100 cases. (2) Present in a panel of normal cord blood samples. (3) Recurrent in the cohort unless present at least 5 times in COSMIC.

Primary bone marrow cell sorting

Thaw media was prepared with IMDM buffer (Cat# 21056023, Gibco) supplemented with 20% FBS and 11% DNAse I (110 µg/mL). Cryopreserved bone marrow samples were thawed at 37 °C in a water bath, 1 mL warm FBS was added and the suspension then diluted by dropwise addition of 8 mL thaw media. Cells were centrifuged at 350 g for 10 min at room temperature and then resuspended in FACS buffer (IMDM with 10% FBS and 10 µg/mL DNAse I), filtered, and placed on ice. Cells were stained for flow cytometry with antibodies listed in Key

Resources Table for 20-40 min on ice. After washing, cells were centrifuged at 350 g for 5 min at 8 °C, resuspended in 200-500 µL FACS buffer containing DAPI at 1 in 10^4 dilution for dead cell exclusion. Cells were sorted on a Sony MA900 instrument. Single stains were used to set initial voltages. Fluorescence minus one controls (FMOs) and positive and negative cell populations were used to set gates. Single, live $CD34^+$ cells were sorted into IMDM with 10% FBS, in tubes that were pre-coated with IMDM containing 10% FBS.

GTAC cell lysis and tagmentation

Tagmentation of cell populations was performed using the Omni-ATAC protocol[22,34]. Briefly, 50,000-90,000 FACS-sorted cells were centrifuged, washed with 500 µL of cold PBS containing 0.5% BSA, and centrifuged again at 500 g in a pre-chilled (4 °C) fixed-angle centrifuge for 5 min. The supernatant was carefully removed. Cell pellets were then resuspended in 50 µL ATAC resuspension-digitonin buffer (ATAC-RSB-DTN buffer; 0.01% digitonin, 0.1% NP-40, 0.1% Tween-20, 10 mM Tris-HCl pH 8.0, 10 mM NaCl, and 3 mM MgCl2) and left on ice for 3 min. Following lysis, 1 mL chilled RSB-Wash buffer (0.1% Tween-20, 10 mM Tris-HCl pH 8.0, 10 mM NaCl, and 3 mM MgCl2) was added and the tube was inverted gently a few times. Nuclei were centrifuged at 1000 g for 3-5 min at 4 °C and the supernatant removed leaving 2-3 µL of buffer. Nuclei were resuspended in chilled tagmentation buffer (16% DMF, 66mM potassium acetate, 10mM magnesium acetate, 33mM Tris-HCl pH 8.0, 0.1% digitonin) containing 5 µL of Tn5 (Cat# 20034197, Illumina). Tubes were incubated in a pre-heated thermo-mixer for 30 min at 37 °C with 800 rpm shaking. After incubation, 50 µL tagmentation stop buffer containing 10 mM EDTA and 0.5% BSA was added, the suspension was pipette-mixed gently 10 times and left on ice for a minimum of 10 min.

GTAC lysis plate preparation, single-nucleus sorting and preamplification

All cell line experiments were done in the 96-well plate format, while primary samples were processed using a 384-well plate format. Here we describe the procedure for 384-well plates. All the volumes were doubled for the 96-well format, as recommended previously[34]. Note that all the experiments prior to the first PCR (lysis buffer preparation, plate pre-processing after sort) were carried under the hood in a designated pre-PCR clean room. The full list of the 384 barcoded i7 indexes have been published[34]. Targeted oligo sequences used for genotyping and barcoded nested target-specific primers used for PCR1 are listed in Table S4.

1.1X lysis buffer was prepared as detailed previously[34]. 27 µL of 1.1X lysis buffer was aliquoted into each well of a 384-well plate (Cat# 4ti-0384/B, FrameStar) and 3 µL of 100 µM i7 indexed primers (containing well barcodes; IDT) were added to each well to obtain a 10µM i7 index stock plate with unique barcodes in each well. Sorting plates were prepared by transferring 1 µL from the 10µM i7 index stock plate into a new 384-well plate and adding 1 µL of 10 µM i5

indexed primer (containing a plate barcode; IDT) in 1X lysis buffer, using a unique i5 index for each plate. Sorting plates were stored at -80˚C and thawed prior to nuclei sorting.

Nuclei were sorted on a Sony MA900 instrument. 500-1500 µL of nucleus dilution buffer (0.5% BSA in 0.5X PBS and nuclease-free water) was added to the tagmented nuclei suspension and 1:1000 1 mg/mL DAPI was added for a final DAPI dilution $1:10^5$ (an unstained sample was used to set the DAPI-negative gate). Single nuclei were identified using a combination of FSC, SSC and DAPI fluorescence and sorted into each well of the 384-well plates, excluding clumps and doublets with two consecutive singlet gates. Plates were stored at -80 ˚C.

Next, plates were incubated at 65 ˚C for 15 min followed by addition of 1 µL Tween-20 10% (Cat# 11332465001, Sigma). PCR preamplification was performed by adding 5 µL PCR mix containing 2x NEB High Fidelity Q5 master-mix (Cat# M0515, New England Biolabs) and target-specific genotyping primers (IDT) at a final concentration of 100 nM per primer. Plates were incubated on a thermo-cycler with the following PCR conditions: 72 ˚C for 10 min, 98 ˚C for 1 min, then 15-18 cycles of 98 ˚C 10 s, 60 ˚C 30 s and 72 ˚C for 30 s, then a final elongation at 72 ˚C for 5 min. 18 cycles of amplification were used for cell lines and 16 cycles of amplification for primary human cells.

After the preamplification PCR, 20% of the PCR product was transferred to a new plate containing 4 µL nuclease-free water (Cat# AM9932, Invitrogen) per well to make a stock plate for single-cell genotyping which was then stored at -80 ˚C. To purify scATAC libraries, the remaining 80% of the volume from each 384-well plate was pooled into a Falcon tube. Each pool was mixed with 5 volumes of Qiagen PB buffer (Cat# 19066, Qiagen) and 10-30 µL of sodium acetate pH 5.5 (Cat# AM9740, ThermoFisher) was added. The pool was purified using a QIAquick spin column (Cat# 28104, Qiagen) and eluted into a total of 42 µL of nuclease-free water (two elutions in 21 µL were done). The pooled libraries were further purified twice with AMPure XP beads first with a 0.5:1 ratio of beads to sample, then with a 1.2:1 ratio of beads to sample. Libraries were quantified with Qubit dsDNA HS Assay Kit (Cat# Q32854, Invitrogen) and fragment size was analysed with Agilent Bioanalyzer (High Sensitivity DNA Kit, Agilent, Cat# 5067-4626) or Agilent Tapestation (High Sensitivity D5000 ScreenTape and Reagents, Agilent, Cat# 5067-5592 and Cat# 5067-5593). Libraries were pooled in equimolar concentrations and diluted to 4 nM for sequencing.

GTAC single-nucleus genotyping library preparation

To generate Illumina-compatible libraries for single-cell genotyping, two PCR steps are required following the initial preamplification PCR. In the first PCR step (genotyping PCR1), nested target-specific primers containing universal CS1 (forward primer) or CS2 (reverse primer) adapters are used to amplify the target loci of interest. Incorporation of a barcode

sequence specific to each plate into these primers enables libraries from different plates to be pooled later (Figure 1). Genotyping stock plates with the preamplification PCR product were thawed and spun at 3000 g at room temperature. PCR1 reactions were set up using 3.25 µL of KAPA 2G Robust HS Ready Mix (Cat# KK5702, Sigma-Aldrich), 1.5 µL of preamplification PCR product and 300 nM target-specific primers containing plate barcodes (IDT) in a 6.5 µL reaction. PCR1 was performed with the following PCR conditions: 95 °C for 3 min, 22-25 cycles of 95 °C 15 s, 60 °C 20 s and 72 °C for 1 min, then a final elongation at 72 °C for 5 min. For cell lines, 22 cycles of PCR were used and for primary human cells, 24 cycles were used.

In the second PCR step (genotyping PCR2), Illumina-compatible adapters containing a 10 bp cell barcode are attached to the genotyping PCR1 product by binding to the CS1/CS2 adapters. PCR2 reactions were performed using FastStart High Fidelity polymerase (Sigma-Aldrich) with 1.0 µL of PCR1 product and 1.2 µL of each barcode primer mix (Access Array Barcode Library for Illumina Sequencers- 384, Single Direction, Cat# 100-4876, Fluidigm) in a 6.2 µL reaction. PCR2 was performed with the following program: 95 °C for 10 min, then 10 cycles of 95 °C 15 s, 60 °C 30 s and 72 °C for 1 min, then a final elongation at 72 °C for 5 min.

After PCR2, 1 µL from each well was pooled using Mosquito nanolitre liquid handling platform (TTP Labtech) and purified with AMPure XP beads with a 1.2:1 ratio of beads to PCR product. Libraries were quantified with Qubit dsDNA HS Assay Kit and fragment size was analysed with Agilent Tapestation (High Sensitivity D1000 ScreenTape and Reagents, Agilent, Cat# 5067-5583 and Cat# 5067-5584). Next, libraries were pooled at equimolar ratios taking into consideration average library size of the library, total number of amplicons per pool (number of amplicons per cell * number of cells) and Qubit measurements. The pool underwent a final AMPure XP bead purification with a double-size selection, using first a 0.6:1 ratio of beads to library, followed by a 1.2:1 ratio of beads to library to remove traces of scATAC library and preserve only genotyping amplicons. The final library was diluted to 4 nM prior to sequencing.

GTAC sequencing

For scATAC library sequencing, the number of reads per cell depends on the library complexity of a given sample. From our experience, at least 200k reads per cell are recommended. K562 cells used for scATAC library comparison (Figure 3) were sequenced at average depth of 435080 reads/cell. The AML sample was sequenced at average depth of 256064 reads/cell. For scATAC library sequencing, we either used Illumina NextSeq 500/550 Mid-Output kits or NextSeq 500/550 High-Output kits. All sequencing runs were performed using 150-cycle kits (75 bp paired-end).

For genotyping libraries, we aimed for a minimum of 2000 reads per cell per amplicon. We used either Illumina MiSeq v3 or NextSeq Mid Output kits (75 bp paired-end). Three custom

primers were used for sequencing: two primers anneal to the CS1 and the CS2 adaptors and sequence the plate barcode and the specific region of interest (500 nM CS1 primer and 500 nM CS2 primer), while the CS2rc primer anneals to the reverse complement of the CS2 sequence, reading the 10-bp Fluidigm well barcode (500 nM CS2rc primer) [107]. All sequencing primers were diluted in 700 mL of HT1 buffer.

GTAC primer design and validation

Target-specific pre-amplification primer pairs were designed using Primer-BLAST [108], using the Homo sapiens database, using *Thermodynamic Oligo Alignment* option. Ideally, amplicons should be 250 bp or shorter in length, given that longer regions might fail amplification due to Tn5 tagmentation (if the region is in open chromatin). Primers were checked with the IDT Oligo Analyzer tool to exclude annealing to i7 and i5 ATAC primers, avoiding those with $\Delta G < -9$kcal/mol. Nested target-specific primers used in genotyping PCR1 were also designed with Primer-BLAST and additionally contain universal CS adaptors (CS1 adaptor for forward nested primer and CS2 adaptor for the reverse nested primers) followed by a 6 bp-long custom plate barcode. We note that when genotyping libraries are sequenced with 75 bp paired-end sequencing, at least one of the nested primers should not be more than 69 bp from the locus of interest. Primer pairs were first tested using extracted genomic DNA and appropriate enzymes. Pairs displaying clear bands of expected sizes and low or no unspecific amplification were tested in single-cell reactions. For that purpose, single OCI-AML3 or primary CD34+ bone marrow cells were sorted into lysis buffer containing generic i5 and i7 oligos and GTAC library preparation was carried out as described in the previous sections. Each primer pair was tested on 16 single cells. After 18 cycles of PCR, an aliquot of the sample was taken for genotyping testing, in which the PCR1 reaction was performed for each single cell with 32 cycles of amplification, followed by amplicon evaluation on a 2% agarose gel. To confirm correct scATAC library formation, the remaining material from each set of 16 cells was pooled and purified as described above, followed by fragment size analysis on an Agilent Bioanalyzer. Primer pairs forming primer dimers or interfering with correct library formation were excluded and re-designed.

Single-cell genotyping data pre-processing and mutation calling

Single-cell genotyping data was pre-processed using the published custom pipeline used previously for TARGET-seq (https://github.com/albarmeira/TARGET-seq)[107]. Reads were first demultiplexed using the 384 well barcodes introduced via the genotyping PCR2 reaction, followed by demultiplexing based on plate barcodes introduced during genotyping PCR1. This generated separate fastq files for each single cell. Reads were aligned to hg38 using STAR version 2.7.3a with default settings. Variant calling was performed with *mpileup* (samtools

version 1.1, options: *--minBQ 30*, *--count-orphans*, *--ignore overlaps*) and results were summarized using the custom pipeline. This generates summary files for each variant, providing the number of A, G, T, C or indel reads for the specified coordinates for each cell. We note that in case of indels, it might be necessary to provide the coordinates mapping 1-2 bp downstream of the first affected base, in order to avoid the indel being missed by the variant call. For FLT3-ITD genotype pre-processing (MOLM-13 cell line), getITD[109] pipeline was used instead, providing the primer and CS1/2 adaptor sequences, and using *-min_read_length 60*.

Mutational calling on single cells was then performed with custom R scripts, separately for each variant. Coverage for each cell was calculated as the sum of all reads across the variant locus for that cell. Empty wells routinely displayed zero or very few reads (usually up to 2), indicating no cross-well contamination. Cells were filtered out if the coverage was less than 30 reads across the variant locus. The scVAF for each cell was calculated with the following formula:

scVAF = variant reads / coverage

PCR amplification and next-generation sequencing have an inherent error rate, which needs to be accounted for when calling cell genotypes. Control cells that were WT for the locus were used to set scVAF thresholds for mutation calling. For mutational analysis in cell lines, we used cell line(s) that were WT for the locus as a control. For mutational analysis of the human AML sample, we profiled 300 CD34+ bone marrow cells from a healthy individual WT for the loci of interest. The scVAF threshold for calling a cell mutant was set using the following formula:

VAF threshold = max(scVAF in WT control) + 2 x SD of scVAF in WT control

Furthermore, we set a threshold for the minimum number of mutant reads for a cell to be called mutant, calculated with the following formula:

Minimum number mutant reads = Maximum number mutant reads in WT sample + 1

Thus, cells were called mutant for a locus if both the scVAF threshold and the minimum number of mutant reads were passed.

Inference of clonal hierarchies

For the AML sample, each single cell was assessed for detection or non-detection of the five mutations present in the sample. The pattern of mutational co-occurrence was used to determine the most likely clonal structure and assign a clonal identity to each cell using principles based on those in our previous study[8]. The occurrence of ADO means that in some cells, a mutation that is present will not be detected. For most mutations in this sample, the sequence of acquisition was clear. However, we observed a small number of cells belonging

to TA and TR clones, suggesting either the *ASXL1* or the *RUNX1* mutation was acquired after *TET2*. We observed 21 cells with the TA genotype compared to 11 cells with the TR genotype, suggesting that T → TA → TAR was the most likely sequence of acquisition.

While our highly sensitive single-cell genotyping resulted in straightforward clone assignment for most cells, a fraction of cells failed detection or displayed ADO for one or more loci. In some of these cases, we were still able to assign a cell to a clone. For example, a small proportion of cells were mutant for *TET2* and one of the *BCOR* loci but the *ASXL1* or *RUNX1* mutations were not detected. In these cases, we inferred that ADO had occurred, and the cell was assigned to the appropriate daughter clone. We applied the same reasoning in cells that were called WT for *TET2* but mutant for *ASXL1*, *RUNX1* and/or one of the *BCOR* loci, in which we inferred ADO of the mutant *TET2* allele had occurred. In all cases, this was a rare occurrence consistent with our estimates of the ADO rate.

We could not apply the same ADO correction for loci harbouring later mutations (*BCOR* mutations), given the absence of a later 'control' event. In a fraction of cells, mutations were called on *TET2*, *ASXL1* and *RUNX1*, while one *BCOR* locus was called WT and the other failed detection (less than 30 reads covering the variant allele). In those cases, we assigned the cell to the TAR clone. If mutations were called on *TET2*, *ASXL1* and *RUNX1* but both *BCOR* loci failed detection, we removed these cells from further analysis. We also noticed 18 cells where both *BCOR* mutations were detected, and scVAFs were consistent with heterozygosity, suggesting they were doublets. Consistent with this, the majority of these cells had a high number of unique scATAC fragments. We therefore excluded these from downstream analysis.

To confirm that our phylogenetic tree for the AML sample represented the statistically most likely course of somatic events, we used infSCITE[110] with default parameters and '-r 200 -l 10000 -fd 0.01 -ad 0.02 -e 0.2 -p 1000'. As input, we used the matrix containing the mutational status for each locus, for each analysed cell. The estimated allelic drop-out rate determined by infSCITE was 2%.

Allelic drop-out (ADO) estimation

To estimate ADO rates for a range of loci, we analysed variants known to be present in a heterozygous state in all cells and calculated the frequency with which they were called homozygous (either reference or alternative) when they should be heterozygous.

For the cell line analysis, we profiled 5 heterozygous loci in K562 cells (*HOXA9*, *CBL*, *NOTCH1*, *ASXL1, TET2*), 2 in OCI-AML3 (NPM1, DNMT3A), 1 in MOLM-13 (FLT3) and 1 in JURKAT (*RUNX1*). In the primary AML sample, we profiled two germline heterozygous SNPs (in the *SH2B3* and *ASXL2* loci) (Figures S2C-S2E), which were previously identified via

targeted DNA sequencing of BM MNCs. For each locus, detection of the variant allele was called by setting a scVAF threshold based on WT control cells as described previously:

scVAF threshold for detection of variant allele = max scVAF in WT + 2 x SD of the scVAF mean in WT

To determine detection of the reference (or WT) allele, we defined a scVAF threshold using the following formula:

scVAF threshold for detection of reference allele = 1 – scVAF threshold for detection of variant allele

For each allele, we then calculated the percentage of cells in which ADO occurred:

ADO percentage = (Cells below scVAF threshold) / (Total number of cells) * 100

Finally, total ADO rate was calculated as:

Total ADO percentage = Variant ADO percentage + WT ADO percentage

We therefore estimate that the percentage of cells in which bi-allelic detection occurred as:

Bi-allelic detection percentage = 100 – Total ADO percentage

We report variant ADO, WT ADO and bi-allelic detection percentages in Results. We note that, when profiling heterozygous mutations at single-cell level, WT ADO is not usually problematic because the cell can still be called mutant. The same is not true for mutant ADO, which leads to a mutant cell being wrongly assigned as WT.

Coverage calculation from scATAC data

To calculate the number of reads mapping to target loci of interest within scATAC libraries, we used *bedtools coverage* on non-deduplicated BAM files for each scATAClibrary, providing a BED file containing variant coordinates. We then plotted the coverage as total reads mapping to the locus derived from each cell.

scATAC data pre-processing

Raw bcl files were first demultiplexed using *bcl2fastq* (version 2.20) and specifying *- - no-lane-splitting*. Files were demultiplexed using unique i7/i5 combinations present in each well. Raw data was pre-processed with a published pipeline written in snakemake [33], with minor modifications. In brief, trimming was performed with cutadapt42 to remove the Nextera sequence at the 3′-insert end. Next, reads were mapped to hg38 with Hisat2. Reads with mapping quality less than 30 were removed by samtools (*-q 30* flag). Next, each bam file was de-duplicated using the Picard *MarkDuplicates* function and indexed using *samtools index*. Files with aggregate size distribution for each cell were generated using samtools after removal of the reads mapping to chrM. De-duplicated reads from single cells were merged using samtools (*ulimit -n 16384* was specified to avoid errors if more than 1024 cells were analysed). Average fragment size was computed again on the merged file excluding chrM.

The merged BAM file was converted to BED format via *bam2bed*. The BED file was used for peak calling with MACS2 and specifying *--nomodel --shift -100 --extsize 200*. The resulting BDG files were used to generate a genome coverage file in BW format using *bdg2bw,* which was used as input for UCSC genome browser visualization, as well as for LanceOtron peak caller[111] used to generate peak sets for correlation analyses (Figure 3F, detailed below). A count matrix over the union of peaks was generated by counting the number of reads from individual cells that overlap the union peaks using *coverageBed* from bedTools. Additional per-cell quality metrics were generated from individual BAM files using MACS2-produced .narrowPeak files for peak coordinates.

Correlation of peak calls

Merged BAM files were indexed using *samtools index* function. Then, coverage files in BW format were produced using *bamCoverage - - extendReads -bs 1 - - normalizeUsing RPKM*. Coverage files were loaded into the LanceOtron peak caller in order to extract bona-fide peaks of open chromatin and filter out noise peaks to avoid inflating correlation (model used: wide-and-deep-jan-2021). Output files were filtered for true peaks via peak scores – a bona-fide peak of open chromatin had the peak score higher than 0.5. Next, these files were merged to create a union of peak coordinates sorted by chromosome and start coordinates. The final union file was generated with *bedtools merge,* containing all bona-fide peaks called across all samples. *multiBigwigSummary* quantified sample correlation across peaks present in the union bed file, taking as input original coverage files, excluding chrM peaks. To plot correlation results, *plotCorrelation* was used with the following parameters: *- - corData results.npz - - corMethod pearson - - whatToPlot heatmap - - removeOutliers - - skipZeros - - plotNumbers*.

Merging scATAC and genotyping data

To link clonal information with ATAC-seq data for each single cell, genotyping data was merged with metadata produced by the pre-processing pipeline using shared cell IDs. ArchR[112] was used for most downstream scATAC analyses (details in the next section). To generate Arrow Files and the ArchR project, the fragment file produced by the pre-processing pipeline was used as input. Arrow Files were generated using *maxFrags = 120000*, *filterTSS = 8* and *filterFrags = 10000* and *addGeneScoreMat = TRUE* to add gene score information used for later analyses. After ArchR project generation, the relevant metadata was extracted using *cellColData*. Next, this table was merged with the pre-processing metadata via the cellNames column to link clonal information with the QC metrics assigned during Arrow File generation, maintaining the row order from the ArchR metadata, and subtracting cells filtered out during Arrow File generation. The column containing clonal information was then extracted

in vector format and added as new column to the ArchR project metadata. Finally, cells which could not be assigned to a clone were removed from the ArchR project.

Doublet exclusion

We added doublet scores for each cell using the *addDoubletScores* function in ArchR, with *k = 10*, *knnMethod = 'UMAP'* and *LSIMethod = 1*. Briefly, this step generates synthetic doublets by using the provided data and by mixing sequencing reads derived from thousands of single-cell combinations. Synthetic doublets are then projected on a UMAP embedding and their closest neighbour is identified. The procedure is iterated thousands of times to identify most likely doublets within the data. Identified doublets were removed using *filterDoublets* with *cutEnrich = 1* and *filterRatio = 1*.

Dimensionality reduction and clustering

We performed Latent Semantic Indexing (LSI) dimensionality reduction using the *addIterativeLSI* function in ArchR, with *iterations = 2*, *varFeatures = 30000*, *dimsToUse = 1:30*, and *clusterParams* including *resolution = c(0.2)* and *sampleCells = 4000*. Clustering was performed using the *addClusters* function with the LSI dimensions as input using *method = 'Seurat'*, *resolution = 0.55*, and visualised by computing a UMAP using the *addUMAP* function with *nNeighbours = 30*, *minDist = 0.5*.

We used several orthogonal strategies to annotate clusters. First, we obtained marker genes for each cluster based on gene activity scores [112], a measure of accessibility within and around gene body. Marker genes were identified using the *getMarkerFeatures* function with *useMatrix = 'GeneScoreMatrix', testMethod = 'wilcoxon', bias = c('TSSEnrichment', 'log10(nFrags)')*. This uses the Wilcoxon rank-sum test to identify marker genes while controlling for the TSS enrichment and $\log_{10}$(unique fragments) of cells when sampling the background set of cells. Marker genes with $\log_2 FC \geq 0.25$ and $FDR \leq 0.05$ were used to test for enrichment of published human haematopoietic gene signatures[113-118] using g:Profiler[119] (Figure S3B), using 'Only annotated genes' as statistical domain scope. Second, we computed motif accessibility deviation scores of key transcription factors (Figure S3C). Third, we identified differentially accessible marker peaks for each cluster using the *getMarkerFeatures* function in ArchR (as described below) and identified the overlap of those marker peaks with curated peak sets from previously published bulk ATAC-seq data using the *peakAnnoEnrichment* function in ArchR (Figure S3D). Finally, we confirmed cluster annotations by projecting our data onto a healthy human bone marrow dataset[41] (details below, Figure S3E).

Gene activity and module score

We calculated gene activity scores in ArchR as previously described [112]. To plot gene activity scores on the UMAP embedding, MAGIC was used for signal smoothing across nearby cells to account for data sparsity, via *plotEmbedding* coloured by the GeneScoreMatrix and using *imputeWeights = getImputeWeights*.

In order to compute the LSC score, we used the *addModuleScore* function in ArchR, providing as input the previously published LSC-specific signature[43], and using gene activity scores from the GeneScoreMatrix.

Pseudo-bulk replicate generation, peak calling and marker identification

We generated pseudo-bulk replicates using clonal information – each clone constituted a pseudo-bulk replicate. We used *addGroupCoverages* with *groupBy = 'clone'* and called peaks with MACS2 using *addReproduciblePeakSet* with *groupBy = 'clone'*, *shift = -75, extsize = 150*, *additionalParams = ' - - nomodel - - nolambda', peaksPerCell = 1000*, performing an iterative overlap procedure for merging peaks specific to ArchR. We called 175563 peaks in the AML sample. We added the peak matrix via *addPeakMatrix*.

Differential accessibility analysis and motif enrichment

We used several strategies to estimate chromatin accessibility differences between clones and clusters. We first defined marker peaks for clones and clusters using *getMarkerFeatures* with *useMatrix = 'PeakMatrix', bias = 'c('TSSEnrichment', 'log10(nFrags)'), testMethod = 'wilcoxon*'. Next, we extracted motif enrichment using *peakAnnoEnrichment* defining the *seMarker* object, with *peakAnnotation = 'Motif'*. Unless otherwise specified, we used the following cutoffs to select enriched motifs: FDR ≤ 0.05 and Log2FC ≥ 1. Heatmaps were generated using *plotEnrichHeatmap* and *ComplexHeatmap* functions. For motifs of TF belonging to TF families, we selected the TFs that were expressed in the cells where the motif was identified and where prior published evidence indicated a functional role. To obtain custom annotations for motif enrichment, we first added the annotation via *addArchRAnnotations*, specifying the *collection* parameter, and then added that through *peakAnnotation* within *peakAnnoEnrichment*.

For pairwise comparisons of motif enrichments between clones or clusters, we used deviations scores computed with chromVAR[120] within ArchR. We first added a background peak set via *addBgdPeaks* and then computed the deviation matrix using *addDeviationMatrix* with *peakAnnotation = 'Motif'*. For pairwise comparisons, we used *getMarkerFeatures* specifying *useGroups, bgdGroups, useMatrix = 'MotifMatrix', useSeqnames = 'z'* to compute differences in z-scores for motifs. To compare two different clones, we specified *groupBy = 'clone'* and clone names in *useGroups* and *bgdGroups*. We then extracted motifs with the respective FDR and 'MeanDiff' values to generate volcano plots. For pairwise comparison of peak

accessibility, we used *getMarkerFeatures* specifying *useMatrix = 'PeakMatrix',* and utilized the information about the closest gene computed by ArchR. We similarly performed pairwise analyses between clones using gene activity scores; in this case, we used *getMarkerFeatures* specifying *useMatrix = 'GeneScoreMatrix'* and the groups to compare. For pathway enrichment analysis, we used the fgsea (v1.20.0) package[121]. We first used *getMarkerFeatures* to compare groups of interest followed by *getMarkers* with *cutOff = 'FDR <= 1'*. We then ranked the differentially accessible genes by multiplying the $-\log_{10}$(FDR) by the sign of the Log2FC (can be either 1 or -1). The ranked list was used within the *fgsea::fgseaMultilevel* function with *minSize = 5*, *maxSize = 1000*, and *nPermSimple = 10000*. We tested the ranked list against Hallmark, CGP, TFT, and GO human collections from MSigDB. The result was filtered only for pathways showing FDR <0.05.

Co-accessibility

We added loops indicative of peak co-accessibility using the *addCoAccessibility* function. Co-accessibility was generated together with the coverage tracks using the *getCoAccessibility* function, using *resolution = 1* and *corCutOff = 0.5*.

Projection analysis

To project the cell type label from healthy human hematopoiesis reference scATAC-seq data to our query AML dataset produced with GTAC, we firstly removed the unknown clusters, obtaining 33819 single-cells from the reference dataset[41]. We then converted our original fragment file to the hg19 annotation and repeated peak calling as described above. Then, peaks from both datasets were intersected, obtaining 162505 shared peaks by employing *GetIntersectingFeatures* in Signac (v1.9.0) package[122]. To work on the same genomic regions, peaks in the AML dataset were replaced with peak counts corresponding to the intersecting genomic regions from the healthy human hematopoiesis dataset. For this task, we used Signac's *CreateFragmentObject* and *FeatureMatrix* functions. Consequently, we performed the LSI on this set of peaks by employing Signac functions in the following order: *RunTFIDF* (with default parameters), *FindTopFeatures* (using 10% as the threshold for feature to be encompassed), and *RunSVD* (with *n = 30*). Next, we used *FindTransferAnchors* function from the Seurat (v4.3.0) package[123] to find the anchor set between the reference and query dataset on shared peaks, using the following parameters: *reduction = 'lsiproject'*, *reference.reduction = 'lsi'*, and *dim = 2:30*, finding 3265 anchors. Finally, we ran *MapQuery* Seurat wrapper for transferring reference cell types onto the query cells, integrating both datasets, and projecting the query into the reference UMAP providing anchors and using *reference.reduction='lsi'* and *reduction.model='umap'* (loaded using *uwot* (v0.1.14)).

Trajectory analysis

To compute differentiation trajectories within the leukemic populations, we used ArchR's supervised trajectory tools. We first reasoned that three broadly different lineages were present within the leukemic populations. Next, we reasoned that the starting point for each differentiation trajectory should be the leukemic cluster with the highest LSC score and lowest CD38 gene accessibility, namely the LMPP-like cluster. This was further justified by our projection of healthy hematopoiesis reference. We generated each of the three trajectories with the *addTrajectory* function, which computes a pseudotime vector as described in the ArchR pipeline. To plot gene scores across trajectories, we used the *plotTrajectory* function with the *colorBy = 'GeneScoreMatrix'* option and *name = 'gene of interest'*. To plot clonal presence across pseudotime, the *geom_density* function from the ggplot2 (v3.4.0) package was used. To plot clone-specific gene activity across a common trajectory, *geom_smooth* with loess method was used.

We also computed the three trajectories in clone-specific manner, using genotype information, by retaining cells belonging only to one specific clone. We did this in order to understand how clone-specific dynamics of differentiation along the same lineage were reflected in chromatin accessibility of key TF motifs. To plot variation in motif accessibility of single motifs across trajectories, we used *plotTrajectory* with *colorBy = 'MotifMatrix'* and *name = 'motif of interest'*. To plot global TF motif deviation changes along these clone-specific trajectories, we used the *plotTrajectoryHeatmap* function, using the default smoothing window generating a smoothened pseudotime x motif matrix. This was used to generate clone-specific heatmaps of motif deviation change across pseudotime, which were then compared between different clones. To visualize how genomic elements changed accessibility across clone-specific trajectories, we split each clone-specific trajectory into five quantiles and used each quantile as pseudobulk for track plotting.

Motif correlation analysis

To correlate accessibility of TF motifs within the LMPP-like differentiation stage, we first extracted the per-cell z-scores computed by chromVAR from the peak matrix using *getMatrixFromProject*. We then subtracted these to contain only LMPP-like cells and plotted correlations of z-score for specific motifs. We used Spearman correlation to define whether there was a statistically significant positive or negative correlation.

Footprinting analysis

To perform footprinting analyses, we first added pseudobulk replicates via *addGroupCoverages*. We next obtained footprints for motifs of interest via *getFootprints* specifying the positions via *getPositions*. We used *plotFootprints* to generate the graphs. We substracted the Tn5 bias from the footprint to perform normalization.

Enrichment of genomic elements in differentially accessible peaks

To understand whether differentially accessible peaks of open chromatin were enriched in promoters versus distal elements or vice-versa, we joined differentially accessible peaks with peak metadata generated via *addReproduciblePeakSet.* We then filtered this list to retain only peaks mapping to promoter regions or to distal elements. Those peaks found in genomic regions ranging from 2kb upstream to 100bp downstream of annotated transcription start sites (default settings) were defined as promoter peaks. Peaks mapping to distal elements are all peaks that map neither to promoter regions nor to gene bodies (exons or introns). Fisher's exact test was used to estimate whether there was a significant overrepresentation of one peak type compared to the other, between sets of differentially accessible peaks.

Identification of putative PRC1.1 direct targets in AML from public data

We used previously published ChIP-seq data[77] to compile a list of putative direct PRC1.1 target genes in AML, defined as those genes bound by KDM2B (the DNA-binding component of PRC1.1.) in at least 5 out of 6 previously analysed AML patients. We do note that ChIP-seq results may vary between patients, depending on cellular context and specific populations expanded (PRC1.1 binding may be dependent on cellular stage of differentiation); hence, this list of targets likely does not represent all possible targets, but rather those genes most frequently bound by PRC1.1 in AML.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical details (type of statistical tests, exact sample size, dispersion and precision measures) are indicated in each figure legend. *P values < 0.05* were considered statistically significant and are reported in the figures. Data were analysed using GraphPad Prism 9 and R version 4.1.3. Detailed statistics related to Figures 4,5 and 6 (and related Supplemental Figures) are reported in Method Details.

Supplemental Item Titles

Table S1. Related to Figures 1 and 4. Variants and mutations screened in cell lines and primary AML sample.

Table S2. Related to Figure 6. Gene accessibility comparison between *BCOR* mutant and *BCOR* WT clones within the LMPP-like differentiation stage.

Table S3. Related to Figure 6. Motif accessibility comparison between *BCOR* mutant and *BCOR* WT clones within the LMPP-like differentiation stage.

Table S4. Related to Figures 2-6. Target-specific and nested barcoded target-specific oligos used for single-cell genotyping library preparation.

Declaration of Interests J.O.J.D and J.R.H are co-founders of Nucleome Therapeutics and provide consultancy to the company. The remaining authors declare no competing interests.

References

1. Forsberg, L.A., Gisselsson, D., and Dumanski, J.P. (2017). Mosaicism in health and disease - clones picking up speed. Nat Rev Genet *18*, 128-142. https://doi.org/10.1038/nrg.2016.145.
2. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med *371*, 2488-2498. https://doi.org/10.1056/NEJMoa1408617.
3. Bowman, R.L., Busque, L., and Levine, R.L. (2018). Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell Stem Cell *22*, 157-170. https://doi.org/10.1016/j.stem.2018.01.011.
4. Sperling, A.S., Gibson, C.J., and Ebert, B.L. (2017). The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. Nat Rev Cancer *17*, 5-19. https://doi.org/10.1038/nrc.2016.112.
5. Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483-1489. https://doi.org/doi:10.1126/science.aab4082.
6. McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell *168*, 613-628. https://doi.org/10.1016/j.cell.2017.01.018.
7. Woyach, J.A., Furman, R.R., Liu, T.M., Ozer, H.G., Zapatka, M., Ruppert, A.S., Xue, L., Li, D.H., Steggerda, S.M., Versele, M., et al. (2014). Resistance mechanisms for the Bruton's tyrosine kinase inhibitor ibrutinib. N Engl J Med *370*, 2286-2294. https://doi.org/10.1056/NEJMoa1400029.
8. Quek, L., David, M.D., Kennedy, A., Metzner, M., Amatangelo, M., Shih, A., Stoilova, B., Quivoron, C., Heiblig, M., Willekens, C., et al. (2018). Clonal heterogeneity of acute myeloid leukemia treated with the IDH2 inhibitor enasidenib. Nat Med *24*, 1167-1177. https://doi.org/10.1038/s41591-018-0115-6.
9. Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S.A., Yan, Y., et al. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nature Communications *11*, 5327. https://doi.org/10.1038/s41467-020-19119-8.
10. Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet *48*, 1193-1203. https://doi.org/10.1038/ng.3646.
11. Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat Rev Genet *20*, 207-220. https://doi.org/10.1038/s41576-018-0089-8.
12. Ushijima, T., Clark, S.J., and Tan, P. (2021). Mapping genomic and epigenomic evolution in cancer ecosystems. Science *373*, 1474-1479. https://doi.org/10.1126/science.abh1645.
13. Izzo, F., Lee, S.C., Poran, A., Chaligne, R., Gaiti, F., Gross, B., Murali, R.R., Deochand, S.D., Ang, C., Jones, P.W., et al. (2020). DNA methylation disruption reshapes the hematopoietic differentiation landscape. Nat Genet *52*, 378-387. https://doi.org/10.1038/s41588-020-0595-4.
14. Nam, A.S., Dusaj, N., Izzo, F., Murali, R., Myers, R.M., Mouhieddine, T., Sotelo, J., Benbarche, S., Waarts, M., Gaiti, F., et al. (2022). Single-cell multi-omics of human

clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. bioRxiv, 2022.2001.2014.476225. https://doi.org/10.1101/2022.01.14.476225.

15. Consortium, T.I.T.P.-C.A.o.W.G. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93. https://doi.org/10.1038/s41586-020-1969-6.

16. Cancer Genome Atlas Research, N., Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A., Hoadley, K., Triche, T.J., Jr., Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med *368*, 2059-2074. https://doi.org/10.1056/NEJMoa1301689.

17. Rodriguez-Meira, A., Buck, G., Clark, S.A., Povinelli, B.J., Alcolea, V., Louka, E., McGowan, S., Hamblin, A., Sousos, N., Barkas, N., et al. (2019). Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. Mol Cell *73*, 1292-1305 e1298. https://doi.org/10.1016/j.molcel.2019.01.009.

18. Nam, A.S., Kim, K.T., Chaligne, R., Izzo, F., Ang, C., Taylor, J., Myers, R.M., Abu-Zeinah, G., Brand, R., Omans, N.D., et al. (2019). Somatic mutations and cell identity linked by Genotyping of Transcriptomes. Nature *571*, 355-360. https://doi.org/10.1038/s41586-019-1367-0.

19. Velten, L., Story, B.A., Hernandez-Malmierca, P., Raffel, S., Leonce, D.R., Milbank, J., Paulsen, M., Demir, A., Szu-Tu, C., Fromel, R., et al. (2021). Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. Nat Commun *12*, 1366. https://doi.org/10.1038/s41467-021-21650-1.

20. Rodriguez-Meira, A., Norfo, R., Wen, W.X., Chédeville, A.L., Rahman, H., O'Sullivan, J., Wang, G., Louka, E., Kretzschmar, W.W., Paterson, A., et al. (2022). Deciphering TP53 mutant Cancer Evolution with Single-Cell Multi-Omics. bioRxiv, 2022.2003.2028.485984. https://doi.org/10.1101/2022.03.28.485984.

21. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213-1218. https://doi.org/10.1038/nmeth.2688.

22. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods *14*, 959-962. https://doi.org/10.1038/nmeth.4396.

23. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486-490. https://doi.org/10.1038/nature14590.

24. Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology *37*, 925-936. https://doi.org/10.1038/s41587-019-0206-z.

25. Lareau, C.A., Duarte, F.M., Chew, J.G., Kartha, V.K., Burkett, Z.D., Kohlway, A.S., Pokholok, D., Aryee, M.J., Steemers, F.J., Lebofsky, R., and Buenrostro, J.D. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology *37*, 916-924. https://doi.org/10.1038/s41587-019-0147-6.

26. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin Potential Identified by Shared Single-

Cell Profiling of RNA and Chromatin. Cell *183*, 1103-1116 e1120. https://doi.org/10.1016/j.cell.2020.09.056.
27. Xu, W., Yang, W., Zhang, Y., Chen, Y., Zhang, Q., Wang, X., Song, K., Jin, W., and Chen, X. (2022). ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. bioRxiv, 2022.2001.2016.476488. https://doi.org/10.1101/2022.01.16.476488.
28. Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science *361*, 1380-1385. https://doi.org/doi:10.1126/science.aau0730.
29. Cheng, H., Pui, H.-p., Lentini, A., Kolbeinsdóttir, S., Andrews, N., Pei, Y., Reinius, B., Deng, Q., and Enge, M. (2021). Smart3-ATAC: a highly sensitive method for joint accessibility and full-length transcriptome analysis in single cells. bioRxiv, 2021.2012.2002.470912. https://doi.org/10.1101/2021.12.02.470912.
30. Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.S., Yeung, B.Z., Papalexi, E., et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol *39*, 1246-1258. https://doi.org/10.1038/s41587-021-00927-2.
31. Chen, A.F., Parks, B., Kathiria, A.S., Ober-Reynolds, B., Goronzy, J.J., and Greenleaf, W.J. (2022). NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. Nature Methods *19*, 547-553. https://doi.org/10.1038/s41592-022-01461-y.
32. Lareau, C.A., Ludwig, L.S., Muus, C., Gohil, S.H., Zhao, T., Chiang, Z., Pelka, K., Verboon, J.M., Luo, W., Christian, E., et al. (2021). Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. Nature Biotechnology *39*, 451-461. https://doi.org/10.1038/s41587-020-0645-6.
33. Chen, X., Miragaia, R.J., Natarajan, K.N., and Teichmann, S.A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. Nat Commun *9*, 5345. https://doi.org/10.1038/s41467-018-07771-0.
34. Xu, W., Wen, Y., Liang, Y., Xu, Q., Wang, X., Jin, W., and Chen, X. (2021). A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility. Nat Protoc *16*, 4084-4107. https://doi.org/10.1038/s41596-021-00583-5.
35. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2018). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research *47*, D941-D947. https://doi.org/10.1093/nar/gky1015.
36. Pierce, S.E., Granja, J.M., and Greenleaf, W.J. (2021). High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. Nat Commun *12*, 2969. https://doi.org/10.1038/s41467-021-23213-w.
37. Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nature Reviews Genetics *10*, 161-172. https://doi.org/10.1038/nrg2522.
38. Bravo, J., Li, Z., Speck, N.A., and Warren, A.J. (2001). The leukemia-associated AML1 (Runx1)–CBFβ complex functions as a DNA-induced molecular clamp. Nature Structural Biology *8*, 371-378. https://doi.org/10.1038/86264.
39. Ullah, H., Zhang, B., Sharma, N.K., McCrea, P.D., and Srivastava, Y. (2022). In-silico probing of AML related RUNX1 cancer-associated missense mutations: Predicted

relationships to DNA binding and drug interactions. Frontiers in Molecular Biosciences *9*. https://doi.org/10.3389/fmolb.2022.981020.

40. Song, W.-J., Sullivan, M.G., Legare, R.D., Hutchings, S., Tan, X., Kufrin, D., Ratajczak, J., Resende, I.C., Haworth, C., Hock, R., et al. (1999). Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. Nature Genetics *23*, 166-175. https://doi.org/10.1038/13793.

41. Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nature Biotechnology *37*, 1458-1465. https://doi.org/10.1038/s41587-019-0332-7.

42. Ficara, F., Murphy, M.J., Lin, M., and Cleary, M.L. (2008). Pbx1 regulates self-renewal of long-term hematopoietic stem cells by maintaining their quiescence. Cell Stem Cell *2*, 484-496. https://doi.org/10.1016/j.stem.2008.03.004.

43. Gentles, A.J., Plevritis, S.K., Majeti, R., and Alizadeh, A.A. (2010). Association of a Leukemic Stem Cell Gene Expression Signature With Clinical Outcomes in Acute Myeloid Leukemia. JAMA *304*, 2706-2715. https://doi.org/10.1001/jama.2010.1862.

44. Corces-Zimmerman, M.R., Hong, W.-J., Weissman, I.L., Medeiros, B.C., and Majeti, R. (2014). Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. Proceedings of the National Academy of Sciences *111*, 2548-2553. https://doi.org/doi:10.1073/pnas.1324297111.

45. Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L., Quake, S.R., and Majeti, R. (2012). Clonal Evolution of Preleukemic Hematopoietic Stem Cells Precedes Human Acute Myeloid Leukemia. Science Translational Medicine *4*, 149ra118-149ra118. https://doi.org/doi:10.1126/scitranslmed.3004315.

46. Goardon, N., Marchi, E., Atzberger, A., Quek, L., Schuh, A., Soneji, S., Woll, P., Mead, A., Alford, K.A., Rout, R., et al. (2011). Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. Cancer Cell *19*, 138-152. https://doi.org/10.1016/j.ccr.2010.12.012.

47. Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J., et al. (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. Nat Med *17*, 1086-1093. https://doi.org/10.1038/nm.2415.

48. Shlush, L.I., Mitchell, A., Heisler, L., Abelson, S., Ng, S.W.K., Trotman-Grant, A., Medeiros, J.J.F., Rao-Bhatia, A., Jaciw-Zurakowsky, I., Marke, R., et al. (2017). Tracing the origins of relapse in acute myeloid leukaemia to stem cells. Nature *547*, 104-108. https://doi.org/10.1038/nature22993.

49. Mendler, J.H., Maharry, K., Radmacher, M.D., Mrozek, K., Becker, H., Metzeler, K.H., Schwind, S., Whitman, S.P., Khalife, J., Kohlschmidt, J., et al. (2012). RUNX1 mutations are associated with poor outcome in younger and older patients with cytogenetically normal acute myeloid leukemia and with distinct gene and MicroRNA expression signatures. J Clin Oncol *30*, 3109-3118. https://doi.org/10.1200/JCO.2011.40.6652.

50. Greif, P.A., Konstandin, N.P., Metzeler, K.H., Herold, T., Pasalic, Z., Ksienzyk, B., Dufour, A., Schneider, F., Schneider, S., Kakadia, P.M., et al. (2012). RUNX1 mutations in cytogenetically normal acute myeloid leukemia are associated with a poor prognosis and up-regulation of lymphoid genes. Haematologica *97*, 1909-1915. https://doi.org/10.3324/haematol.2012.064667.

51. Heuser, M., Beutel, G., Krauter, J., Dohner, K., von Neuhoff, N., Schlegelberger, B., and Ganser, A. (2006). High meningioma 1 (MN1) expression as a predictor for poor

outcome in acute myeloid leukemia with normal cytogenetics. Blood *108*, 3898-3905. https://doi.org/10.1182/blood-2006-04-014845.

52. Intlekofer, A.M., Takemoto, N., Wherry, E.J., Longworth, S.A., Northrup, J.T., Palanivel, V.R., Mullen, A.C., Gasink, C.R., Kaech, S.M., Miller, J.D., et al. (2005). Effector and memory CD8+ T cell fate coupled by T-bet and eomesodermin. Nat Immunol *6*, 1236-1244. https://doi.org/10.1038/ni1268.

53. Paley, M.A., Kroy, D.C., Odorizzi, P.M., Johnnidis, J.B., Dolfi, D.V., Barnett, B.E., Bikoff, E.K., Robertson, E.J., Lauer, G.M., Reiner, S.L., and Wherry, E.J. (2012). Progenitor and terminal subsets of CD8+ T cells cooperate to contain chronic viral infection. Science *338*, 1220-1225. https://doi.org/10.1126/science.1229620.

54. Okamura, R.M., Sigvardsson, M., Galceran, J., Verbeek, S., Clevers, H., and Grosschedl, R. (1998). Redundant regulation of T cell differentiation and TCRalpha gene expression by the transcription factors LEF-1 and TCF-1. Immunity *8*, 11-20. https://doi.org/10.1016/s1074-7613(00)80454-9.

55. Wong, W.F., Kohu, K., Chiba, T., Sato, T., and Satake, M. (2011). Interplay of transcription factors in T-cell differentiation and function: the role of Runx. Immunology *132*, 157-164. https://doi.org/10.1111/j.1365-2567.2010.03381.x.

56. Rosenbauer, F., and Tenen, D.G. (2007). Transcription factors in myeloid development: balancing differentiation with transformation. Nature Reviews Immunology *7*, 105-117. https://doi.org/10.1038/nri2024.

57. Yu, Y., Wang, J., Khaled, W., Burke, S., Li, P., Chen, X., Yang, W., Jenkins, N.A., Copeland, N.G., Zhang, S., and Liu, P. (2012). Bcl11a is essential for lymphoid development and negatively regulates p53. Journal of Experimental Medicine *209*, 2467-2483. https://doi.org/10.1084/jem.20121846.

58. Brown, F.C., Still, E., Koche, R.P., Yim, C.Y., Takao, S., Cifani, P., Reed, C., Gunasekera, S., Ficarro, S.B., Romanienko, P., et al. (2018). MEF2C Phosphorylation Is Required for Chemotherapy Resistance in Acute Myeloid Leukemia. Cancer Discov *8*, 478-497. https://doi.org/10.1158/2159-8290.CD-17-1271.

59. Schwieger, M., Schuler, A., Forster, M., Engelmann, A., Arnold, M.A., Delwel, R., Valk, P.J., Lohler, J., Slany, R.K., Olson, E.N., and Stocking, C. (2009). Homing and invasiveness of MLL/ENL leukemic cells is regulated by MEF2C. Blood *114*, 2476-2488. https://doi.org/10.1182/blood-2008-05-158196.

60. Zhao, L., Zhang, P., Galbo, P.M., Zhou, X., Aryal, S., Qiu, S., Zhang, H., Zhou, Y., Li, C., Zheng, D., et al. (2021). Transcription factor MEF2D is required for the maintenance of MLL-rearranged acute myeloid leukemia. Blood Adv *5*, 4727-4740. https://doi.org/10.1182/bloodadvances.2021004469.

61. Raajit, R., and Maria, E.F. (2016). Wilms tumor 1 mutations in the pathogenesis of acute myeloid leukemia. Haematologica *101*, 672-679. https://doi.org/10.3324/haematol.2015.141796.

62. Wong, P., Iwasaki, M., Somervaille, T.C.P., So, C.W.E., and Cleary, M.L. (2007). Meis1 is an essential and rate-limiting regulator of MLL leukemia stem cell potential. Genes & Development *21*, 2762-2774. https://doi.org/10.1101/gad.1602107.

63. Hock, H., Meade, E., Medeiros, S., Schindler, J.W., Valk, P.J., Fujiwara, Y., and Orkin, S.H. (2004). Tel/Etv6 is an essential and selective regulator of adult hematopoietic stem cell survival. Genes Dev *18*, 2336-2341. https://doi.org/10.1101/gad.1239604.

64. Guo, H., Chu, Y., Wang, L., Chen, X., Chen, Y., Cheng, H., Zhang, L., Zhou, Y., Yang, F.-c., Cheng, T., et al. (2017). PBX3 is essential for leukemia stem cell maintenance in MLL-rearranged leukemia. International Journal of Cancer *141*, 324-335. https://doi.org/10.1002/ijc.30739.

65. Knudsen, K.J., Rehn, M., Hasemann, M.S., Rapin, N., Bagger, F.O., Ohlsson, E., Willer, A., Frank, A.-K., Søndergaard, E., Jendholm, J., et al. (2015). ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation. Genes & Development *29*, 1915-1929. https://doi.org/10.1101/gad.268409.115.
66. Li, Z., Chen, P., Su, R., Hu, C., Li, Y., Elkahloun, A.G., Zuo, Z., Gurbuxani, S., Arnovitz, S., Weng, H., et al. (2016). PBX3 and MEIS1 Cooperate in Hematopoietic Cells to Drive Acute Myeloid Leukemias Characterized by a Core Transcriptome of the MLL-Rearranged Disease. Cancer Research *76*, 619-629. https://doi.org/10.1158/0008-5472.Can-15-1566.
67. Li, Z., Zhang, Z., Li, Y., Arnovitz, S., Chen, P., Huang, H., Jiang, X., Hong, G.-M., Kunjamma, R.B., Ren, H., et al. (2013). PBX3 is an important cofactor of HOXA9 in leukemogenesis. Blood *121*, 1422-1431. https://doi.org/10.1182/blood-2012-07-442004.
68. Jan, M., Chao, M.P., Cha, A.C., Alizadeh, A.A., Gentles, A.J., Weissman, I.L., and Majeti, R. (2011). Prospective separation of normal and leukemic stem cells based on differential expression of TIM3, a human acute myeloid leukemia stem cell marker. Proceedings of the National Academy of Sciences *108*, 5009-5014. https://doi.org/doi:10.1073/pnas.1100551108.
69. He, X., Xu, Y., Huang, D., Yu, Z., Yu, J., Xie, L., Liu, L., Yu, Y., Chen, C., Wan, J., et al. (2023). P2X1 enhances leukemogenesis through PBX3-BCAT1 pathways. Leukemia *37*, 265-275. https://doi.org/10.1038/s41375-022-01759-y.
70. Kruse, E.A., Loughran, S.J., Baldwin, T.M., Josefsson, E.C., Ellis, S., Watson, D.K., Nurden, P., Metcalf, D., Hilton, D.J., Alexander, W.S., and Kile, B.T. (2009). Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. Proceedings of the National Academy of Sciences *106*, 13814-13819. https://doi.org/doi:10.1073/pnas.0906556106.
71. Lu, Z., Hong, C.C., Kong, G., Assumpção, A.L.F.V., Ong, I.M., Bresnick, E.H., Zhang, J., and Pan, X. (2018). Polycomb Group Protein YY1 Is an Essential Regulator of Hematopoietic Stem Cell Quiescence. Cell Reports *22*, 1545-1559. https://doi.org/10.1016/j.celrep.2018.01.026.
72. Bungartz, G., Land, H., Scadden, D.T., and Emerson, S.G. (2012). NF-Y is necessary for hematopoietic stem cell proliferation and survival. Blood *119*, 1380-1389. https://doi.org/10.1182/blood-2011-06-359406.
73. Behre, G., Whitmarsh, A.J., Coghlan, M.P., Hoang, T., Carpenter, C.L., Zhang, D.-E., Davis, R.J., and Tenen, D.G. (1999). c-Jun Is a JNK-independent Coactivator of the PU.1 Transcription Factor *. Journal of Biological Chemistry *274*, 4939-4946. https://doi.org/10.1074/jbc.274.8.4939.
74. Shafarenko, M., Amanullah, A., Gregory, B., Liebermann, D.A., and Hoffman, B. (2004). Fos modulates myeloid cell survival and differentiation and partially abrogates the c-Myc block in terminal myeloid differentiation. Blood *103*, 4259-4267. https://doi.org/10.1182/blood-2002-09-2704.
75. Steidl, U., Rosenbauer, F., Verhaak, R.G.W., Gu, X., Ebralidze, A., Otu, H.H., Klippel, S., Steidl, C., Bruns, I., Costa, D.B., et al. (2006). Essential role of Jun family transcription factors in PU.1 knockdown–induced leukemic stem cells. Nature Genetics *38*, 1269-1277. https://doi.org/10.1038/ng1898.
76. Passegué, E., Wagner, E.F., and Weissman, I.L. (2004). JunB Deficiency Leads to a Myeloproliferative Disorder Arising from Hematopoietic Stem Cells. Cell *119*, 431-443. https://doi.org/10.1016/j.cell.2004.10.010.
77. van den Boom, V., Maat, H., Geugien, M., Rodríguez López, A., Sotoca, A.M., Jaques, J., Brouwers-Vos, A.Z., Fusetti, F., Groen, R.W.J., Yuan, H., et al. (2016). Non-

canonical PRC1.1 Targets Active Genes Independent of H3K27me3 and Is Essential for Leukemogenesis. Cell Reports *14*, 332-346. https://doi.org/10.1016/j.celrep.2015.12.034.

78. Woolley, J.F., Chen, K., Genc, G.E., Lee, D.K.C., Dzneladze, I., He, R., Gabra, M.M., Saffi, G.T., Son, M.H., Schoof, E.M., et al. (2021). INPP4B promotes leukemia by restricting leukemic stem cell differentiation through regulation of lysosomal functions. bioRxiv, 2021.2003.2025.437029. https://doi.org/10.1101/2021.03.25.437029.

79. Nguyen, C.H., Glüxam, T., Schlerka, A., Bauer, K., Grandits, A.M., Hackl, H., Dovey, O., Zöchbauer-Müller, S., Cooper, J.L., Vassiliou, G.S., et al. (2019). SOCS2 is part of a highly prognostic 4-gene signature in AML and promotes disease aggressiveness. Scientific Reports *9*, 9139. https://doi.org/10.1038/s41598-019-45579-0.

80. Vitali, C., Bassani, C., Chiodoni, C., Fellini, E., Guarnotta, C., Miotti, S., Sangaletti, S., Fuligni, F., De Cecco, L., Piccaluga, P.P., et al. (2015). SOCS2 Controls Proliferation and Stemness of Hematopoietic Cells under Stress Conditions and Its Deregulation Marks Unfavorable Acute Leukemias. Cancer Research *75*, 2387-2399. https://doi.org/10.1158/0008-5472.Can-14-3625.

81. Sirin, O., Lukov, G.L., Mao, R., Conneely, O.M., and Goodell, M.A. (2010). The orphan nuclear receptor Nurr1 restricts the proliferation of haematopoietic stem cells. Nature Cell Biology *12*, 1213-1219. https://doi.org/10.1038/ncb2125.

82. Cheng, Y.C., Chen, P.H., Chiang, H.Y., Suen, C.S., Hwang, M.J., Lin, T.Y., Yang, H.C., Lin, W.C., Lai, P.L., and Shieh, S.Y. (2015). Candidate tumor suppressor B-cell translocation gene 3 impedes neoplastic progression by suppression of AKT. Cell Death & Disease *6*, e1584-e1584. https://doi.org/10.1038/cddis.2014.550.

83. Chen, M., Zhu, N., Liu, X., Laurent, B., Tang, Z., Eng, R., Shi, Y., Armstrong, S.A., and Roeder, R.G. (2015). JMJD1C is required for the survival of acute myeloid leukemia by functioning as a coactivator for key transcription factors. Genes & Development *29*, 2123-2139. https://doi.org/10.1101/gad.267278.115.

84. Wang, Z., Gearhart, M.D., Lee, Y.-W., Kumar, I., Ramazanov, B., Zhang, Y., Hernandez, C., Lu, A.Y., Neuenkirchen, N., Deng, J., et al. (2018). A Non-canonical BCOR-PRC1.1 Complex Represses Differentiation Programs in Human ESCs. Cell Stem Cell *22*, 235-251.e239. https://doi.org/10.1016/j.stem.2017.12.002.

85. Cao, Q., Gearhart, M.D., Gery, S., Shojaee, S., Yang, H., Sun, H., Lin, D.c., Bai, J.w., Mead, M., Zhao, Z., et al. (2016). BCOR regulates myeloid cell proliferation and differentiation. Leukemia *30*, 1155-1165. https://doi.org/10.1038/leu.2016.2.

86. Kelly, M.J., So, J., Rogers, A.J., Gregory, G., Li, J., Zethoven, M., Gearhart, M.D., Bardwell, V.J., Johnstone, R.W., Vervoort, S.J., and Kats, L.M. (2019). Bcor loss perturbs myeloid differentiation and promotes leukaemogenesis. Nature Communications *10*, 1347. https://doi.org/10.1038/s41467-019-09250-6.

87. Stein, S.J., and Baldwin, A.S. (2013). Deletion of the NF-κB subunit p65/RelA in the hematopoietic compartment leads to defects in hematopoietic stem cell function. Blood *121*, 5015-5024. https://doi.org/10.1182/blood-2013-02-486142.

88. Zhao, C., Xiu, Y., Ashton, J., Xing, L., Morita, Y., Jordan, C.T., and Boyce, B.F. (2012). Noncanonical NF-κB Signaling Regulates Hematopoietic Stem Cell Self-Renewal and Microenvironment Interactions. Stem Cells *30*, 709-718. https://doi.org/10.1002/stem.1050.

89. Yamashita, M., and Passegué, E. (2019). TNF-α Coordinates Hematopoietic Stem Cell Survival and Myeloid Regeneration. Cell Stem Cell *25*, 357-372.e357. https://doi.org/10.1016/j.stem.2019.05.019.

90. Tothova, Z., Kollipara, R., Huntly, B.J., Lee, B.H., Castrillon, D.H., Cullen, D.E., McDowell, E.P., Lazo-Kallanian, S., Williams, I.R., Sears, C., et al. (2007). FoxOs Are

Critical Mediators of Hematopoietic Stem Cell Resistance to Physiologic Oxidative Stress. Cell *128*, 325-339. https://doi.org/10.1016/j.cell.2007.01.003.

91. Alharbi, R.A., Pettengell, R., Pandha, H.S., and Morgan, R. (2013). The role of HOX genes in normal hematopoiesis and acute leukemia. Leukemia *27*, 1000-1008. https://doi.org/10.1038/leu.2012.356.

92. Miles, L.A., Bowman, R.L., Merlinsky, T.R., Csete, I.S., Ooi, A.T., Durruthy-Durruthy, R., Bowman, M., Famulare, C., Patel, M.A., Mendez, P., et al. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature *587*, 477-482. https://doi.org/10.1038/s41586-020-2864-x.

93. Myers, R.M., Izzo, F., Kottapalli, S., Prieto, T., Dunbar, A., Bowman, R.L., Mimitou, E.P., Stahl, M., El Ghaity-Beckley, S., Arandela, J., et al. (2022). Integrated Single-Cell Genotyping and Chromatin Accessibility Charts JAK2 V617F Human Hematopoietic Differentiation. bioRxiv, 2022.2005.2011.491515. https://doi.org/10.1101/2022.05.11.491515.

94. Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y., and Greenleaf, W.J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell *173*, 1535-1548.e1516. https://doi.org/10.1016/j.cell.2018.03.074.

95. Mulqueen, R.M., Pokholok, D., O'Connell, B.L., Thornton, C.A., Zhang, F., O'Roak, B.J., Link, J., Yardımcı, G.G., Sears, R.C., Steemers, F.J., and Adey, A.C. (2021). High-content single-cell combinatorial indexing. Nature Biotechnology *39*, 1574-1580. https://doi.org/10.1038/s41587-021-00962-z.

96. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*. https://doi.org/10.1126/science.aav1898.

97. Karlsson, M., Zhang, C., Mear, L., Zhong, W., Digre, A., Katona, B., Sjostedt, E., Butler, L., Odeberg, J., Dusart, P., et al. (2021). A single-cell type transcriptomics map of human tissues. Sci Adv *7*. https://doi.org/10.1126/sciadv.abh2169.

98. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/doi:10.1126/science.1260419.

99. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. New England Journal of Medicine *374*, 2209-2221. https://doi.org/10.1056/NEJMoa1516192.

100. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. New England Journal of Medicine *371*, 2477-2487. https://doi.org/10.1056/NEJMoa1409405.

101. Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nature Medicine *20*, 1472-1478. https://doi.org/10.1038/nm.3733.

102. Acuna-Hidalgo, R., Sengul, H., Steehouwer, M., van de Vorst, M., Vermeulen, S.H., Kiemeney, L.A.L.M., Veltman, J.A., Gilissen, C., and Hoischen, A. (2017). Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. The American Journal of Human Genetics *101*, 50-64. https://doi.org/10.1016/j.ajhg.2017.05.013.

103. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352.
104. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., and Dry, J.R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Research *44*, e108-e108. https://doi.org/10.1093/nar/gkw227.
105. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology *31*, 213-219. https://doi.org/10.1038/nbt.2514.
106. Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nature Protocols *10*, 1556-1566. https://doi.org/10.1038/nprot.2015.105.
107. Rodriguez-Meira, A., O'Sullivan, J., Rahman, H., and Mead, A.J. (2020). TARGET-Seq: A Protocol for High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. STAR Protoc *1*, 100125. https://doi.org/10.1016/j.xpro.2020.100125.
108. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134. https://doi.org/10.1186/1471-2105-13-134.
109. Blätte, T.J., Schmalbrock, L.K., Skambraks, S., Lux, S., Cocciardi, S., Dolnik, A., Döhner, H., Döhner, K., and Bullinger, L. (2019). getITD for FLT3-ITD-based MRD monitoring in AML. Leukemia *33*, 2535-2539. https://doi.org/10.1038/s41375-019-0483-z.
110. Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. Genome Biology *17*, 86. https://doi.org/10.1186/s13059-016-0936-x.
111. Hentges, L.D., Sergeant, M.J., Cole, C.B., Downes, D.J., Hughes, J.R., and Taylor, S. (2022). LanceOtron: a deep learning peak caller for genome sequencing experiments. Bioinformatics. https://doi.org/10.1093/bioinformatics/btac525.
112. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet *53*, 403-411. https://doi.org/10.1038/s41588-021-00790-6.
113. Popescu, D.-M., Botting, R.A., Stephenson, E., Green, K., Webb, S., Jardine, L., Calderbank, E.F., Polanski, K., Goh, I., Efremova, M., et al. (2019). Decoding human fetal liver haematopoiesis. Nature *574*, 365-371. https://doi.org/10.1038/s41586-019-1652-y.
114. Chen, L., Kostadima, M., Martens, J.H.A., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033. https://doi.org/doi:10.1126/science.1251033.
115. Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J., et al. (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. Nat Immunol *19*, 85-97. https://doi.org/10.1038/s41590-017-0001-2.
116. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nature Cell Biology *19*, 271-281. https://doi.org/10.1038/ncb3493.

117. Drissen, R., Thongjuea, S., Theilgaard-Mönch, K., and Nerlov, C. (2019). Identification of two distinct pathways of human myelopoiesis. Science Immunology *4*, eaau7148. https://doi.org/doi:10.1126/sciimmunol.aau7148.
118. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell *144*, 296-309. https://doi.org/10.1016/j.cell.2011.01.004.
119. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020). gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]. F1000Research *9*. https://doi.org/10.12688/f1000research.24956.2.
120. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature Methods *14*, 975-978. https://doi.org/10.1038/nmeth.4401.
121. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. bioRxiv, 060012. https://doi.org/10.1101/060012.
122. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat Methods *18*, 1333-1341. https://doi.org/10.1038/s41592-021-01282-5.
123. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e3529. https://doi.org/10.1016/j.cell.2021.04.048.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-human CD45RA (clone HI100)	BD	Cat#564552; RRID:AB_2738841
Mouse anti-human CD90 (clone 5E10)	BioLegend	Cat# 328124; RRID:AB_2561693
Mouse anti-human CD123 (clone 6H6)	BioLegend	Cat# 306006; RRID:AB_314580
Rat anti-human CD49f (clone GoH3)	BioLegend	Cat# 313626; RRID:AB_2616782
Mouse anti-human CD38 (clone HIT2)	BioLegend	Cat# 303526; RRID:AB_10983072
Mouse anti-human CD10 (clone HI10a)	BioLegend	Cat# 312222; RRID:AB_2562157
Mouse anti-human CD34 (clone 581)	BioLegend	Cat# 343510; RRID:AB_1877153
Mouse anti-human CD117 (clone 104D2)	BioLegend	Cat# 313238; RRID:AB_2629837
Mouse anti-human CD2 (clone RPA-2.10)	BioLegend	Cat# 300210; RRID:AB_314034
Mouse anti-human CD3 (clone HIT3a)	BioLegend	Cat# 300310; RRID:AB_314046
Mouse anti-human CD4 (clone RPA-T4)	BioLegend	Cat# 300510; RRID:AB_314078
Mouse anti-human CD8a (clone RPA-T8)	BioLegend	Cat# 301010; RRID:AB_314128
Mouse anti-human CD11b (clone ICRF44)	BioLegend	Cat# 301308; RRID:AB_314159
Mouse anti-human CD14 (clone 61D3)	eBioscience	Cat#15-0149-42; RRID:AB_2573058
Mouse anti-human CD19 (clone HIB19)	BioLegend	Cat# 302210; RRID:AB_314240
Mouse anti-human CD20 (clone 2H7)	BioLegend	Cat# 302308; RRID:AB_314256
Mouse anti-human CD56 (clone MEM188)	BioLegend	Cat# 304608; RRID:AB_314450

Mouse anti-human CD235ab (clone HIR2)	BioLegend	Cat# 306606; RRID:AB_314623
Chemicals, peptides, and recombinant proteins		
RPMI-1640	Gibco	Cat#21875034
Fetal Bovine Serum	StemCell Technologies	Cat#06950
Penicillin-Streptomycin (10,000 U/mL)	Gibco	Cat#15140122
MEM α, nucleosides	Gibco	Cat#12571063
IMDM	Gibco	Cat# 21056023
DPBS	ThermoFisher	Cat# 14190169
30% (wt/vol) BSA	Sigma-Aldrich	Cat#A8577
Digitonin	Promega	Cat# G9441
IGEPAL CA-630	Sigma-Aldrich	Cat# I8896
Tween-20	Sigma-Aldrich	Cat# 11332465001
Tris-HCl 1M pH 8.0	ThermoFisher	Cat# 15568025
NaCl 5M	Sigma-Aldrich	Cat# S6546
MgCl2 1M	ThermoFisher	Cat# AM9530G
N,N-Dimethylformamide	Sigma-Aldrich	Cat# D4551-250ML
Magnesium acetate solution 1M	Sigma-Aldrich	Cat# 63052-100ML
Potassium acetate solution 5M	Sigma-Aldrich	Cat# 95843-100ML-F
Tn5	Illumina	Cat# 20034197
EDTA 0.5M pH 8.0	ThermoFisher Scientific	Cat# AM9260G
SDS 10%	Sigma-Aldrich	Cat# L4509
DAPI	ThermoFisher Scientific	Cat# 62248
UltraPure Agarose	Invitrogen	Cat# 16500-500
Ethidium Bromide Solution	Invitrogen	Cat# 15585-011
2x NEB High Fidelity Q5 master-mix	New England Biolabs	Cat# M0515
Nuclease-free water	Invitrogen	Cat# AM9932

PB buffer	Qiagen	Cat# 19066
Sodium acetate	ThermoFisher	Cat# AM9740
KAPA 2G Robust HS Ready Mix	Sigma-Aldrich	Cat# KK5702
FastStart High Fidelity polymerase	Sigma-Aldrich	Cat#4738292001
Access Array Barcode Library for Illumina Sequencers-384, Single Direction	Fluidigm	Cat# 100-4876
KAPA dual-indexed adapters	Roche	Cat# KK8722
KAPA HiFi HotStart high-fidelity DNA polymerase	Roche	Cat#7958927001
SeqCap EZ Prime Choice (custom probes)	Roche	N/A
Critical commercial assays		
MycoAlert Mycoplasma Detection Kit	Lonza	Cat#LT07-218
DNeasy Blood & Tissue Kit	Qiagen	Cat#69506
KAPA HyperPlus Kit (96)	Roche	Cat#KK8514
Qiaquick spin column	Qiagen	Cat# 28104
Qubit dsDNA HS Assay Kit	Invitrogen	Cat# Q32854
Agilent Tapestation HS D5000 ScreenTape	Agilent	Cat#5067-5592
Agilent Tapestation HS D5000 Reagents	Agilent	Cat#5067-5593
Agilent Tapestation HS D1000 ScreenTape	Agilent	Cat# 5067-5583
Agilent Tapestation HS D1000 Reagents	Agilent	Cat#5067-5584
Agilent High Sensitivity DNA Kit	Agilent	Cat#5067-4626
Deposited data		
Raw and processed scATAC-seq data generated with GTAC (cell lines and primary samples)	This study	GEO: GSE214812
Targeted single cell genotyping sequencing data generated with GTAC (cell lines and primary samples)	This study	SRA: PRJNA887752
Experimental models: Cell lines		
K562	Vyas lab	N/A
JURKAT	Vyas lab	N/A
OCI-AML3	Milne lab	N/A
THP1	Vyas lab	N/A

MOLM-13	Vyas lab	N/A
Oligonucleotides		
See Table S5 for target-specific genotyping oligos	IDT (design: this paper)	N/A
See Table S5 for target-specific nested barcoded genotyping oligos	IDT (design: this paper)	N/A
384 indexed i7 primers	IDT (design: Xu et al., 2021)	https://www.nature.com/articles/s41596-021-00583-5
See Table S5 for Custom barcoded i5 primers	Illumina	N/A
CS1 sequencing primer (HPLC purified): A+CA+CTG+ACGACATGGTTCTACA	Rodriguez-Meira et al., 2020	https://www.sciencedirect.com/science/article/pii/S266616672030112X
CS2 sequencing primer (HPLC purified): T+AC+GGT+AGCAGAGACTTGGTCT	Rodriguez-Meira et al., 2020	https://www.sciencedirect.com/science/article/pii/S266616672030112X
CS2rc sequencing primer (HPLC purified): A+GAC+CA+AGTCTCTGCTACCGTA	Rodriguez-Meira et al., 2020	https://www.sciencedirect.com/science/article/pii/S266616672030112X
Software and algorithms		
Burrows-Wheeler aligner 0.7.17	Li et al., 2009	https://academic.oup.com/bioinformatics/article/25/14/1754/225615
VarDictJava 1.7.0	Lai et al., 2016	https://academic.oup.com/nar/article/44/11/e108/2468301
Annovar	Yang et al., 2015	https://www.nature.com/articles/nprot.2015.105
Primer-BLAST	Ye et al., 2012	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-134
TARGET-seq genotyping pipeline	Rodriguez-Meira et al., 2020	https://github.com/albarmeira/TARGET-seq

Bcl2fastq	Illumina	https://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq2-v2-20-software-guide-15051736-03.pdf
Snakemake scATAC-seq pre-processing pipeline	Chen et al., 2018	https://github.com/dbrg77/scATAC_snakemake
LanceOtron peak caller	Hentges et al., 2022	https://academic.oup.com/bioinformatics/article/38/18/4255/6648462
R 4.1.3	R-Project	
ArchR 1.0.2	Granja et al., 2021	https://www.nature.com/articles/s41588-021-00790-6
chromVAR 1.16.0	Schep et al., 2017	https://www.nature.com/articles/nmeth.4401
GraphPad Prism 9	GraphPad	https://www.graphpad.com
FlowJo	FlowJo	https://www.bdbiosciences.com/en-gb/products/software/flowjo-v10-software
g:Profiler	Kolberg et al., 2020	https://f1000research.com/articles/9-709/v2
fgsea	Korotkevich et al.,2021	https://www.biorxiv.org/content/10.1101/060012v3
infSCITE	Jahn et al. 2016	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0936-x
BioRender	BioRender	https://biorender.com

Mutect2	Cibulkis et al., 2013	https://www.nature.com/articles/nbt.2514
Other		
12.5 mL GRIPTIP, Sterile, Filter	INTEGRA Biosciences	Cat#6455
High Volume MANTIS Chip	FORMULATRIX	Cat# MCHVSMR6
Agencourt AMPure XP beads	Beckman Coulter	Cat# A63881

Figure 1

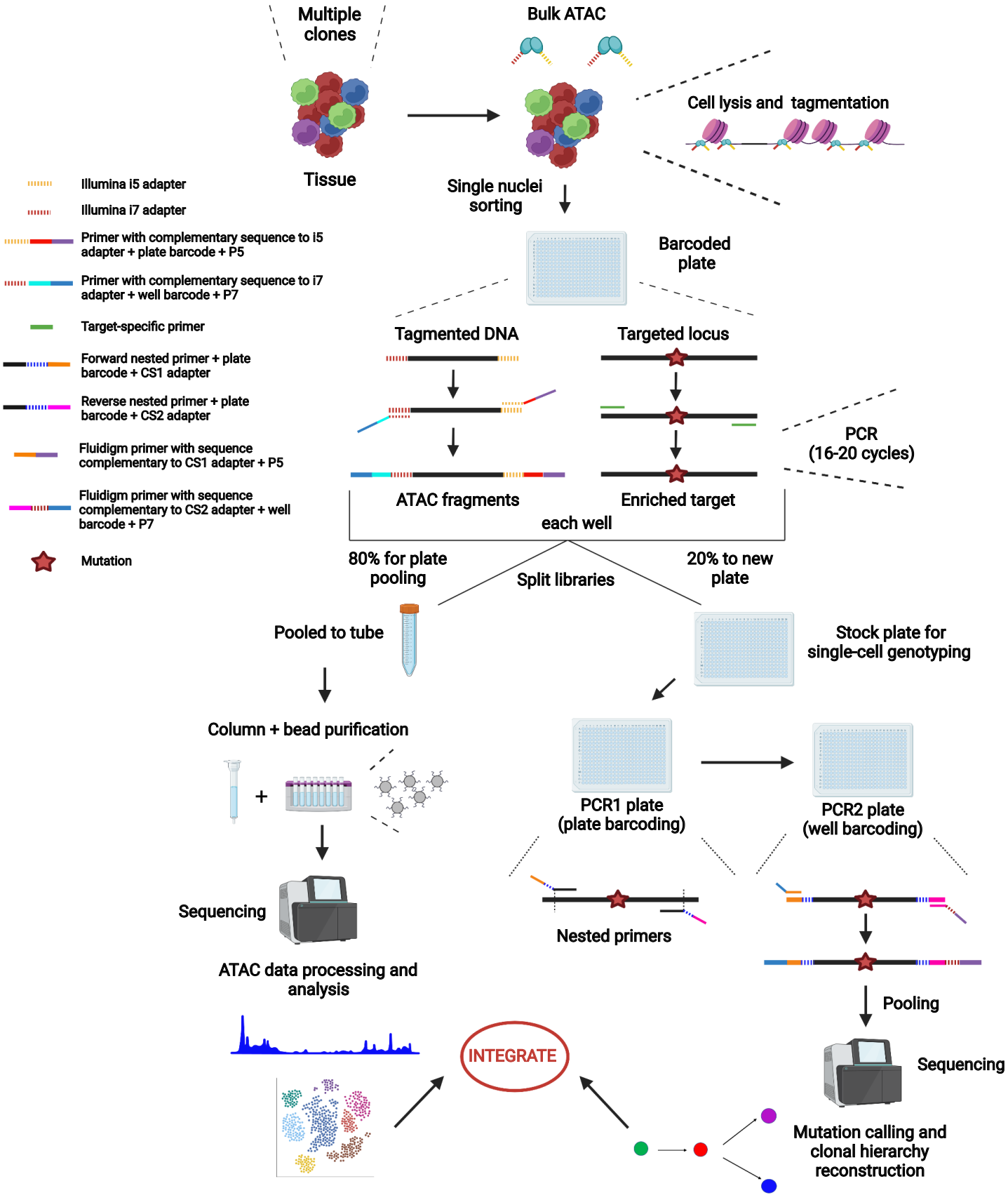

Multiple clones
Tissue
Bulk ATAC
Cell lysis and tagmentation
Single nuclei sorting
Barcoded plate
Illumina i5 adapter
Illumina i7 adapter
Primer with complementary sequence to i5 adapter + plate barcode + P5
Primer with complementary sequence to i7 adapter + well barcode + P7
Target-specific primer
Forward nested primer + plate barcode + CS1 adapter
Reverse nested primer + plate barcode + CS2 adapter
Fluidigm primer with sequence complementary to CS1 adapter + P5
Fluidigm primer with sequence complementary to CS2 adapter + well barcode + P7
Mutation
Tagmented DNA
Targeted locus
PCR (16-20 cycles)
ATAC fragments
Enriched target
each well
80% for plate pooling
Split libraries
20% to new plate
Pooled to tube
Stock plate for single-cell genotyping
Column + bead purification
PCR1 plate (plate barcoding)
PCR2 plate (well barcoding)
Sequencing
Nested primers
ATAC data processing and analysis
Pooling
INTEGRATE
Sequencing
Mutation calling and clonal hierarchy reconstruction

Figure 2

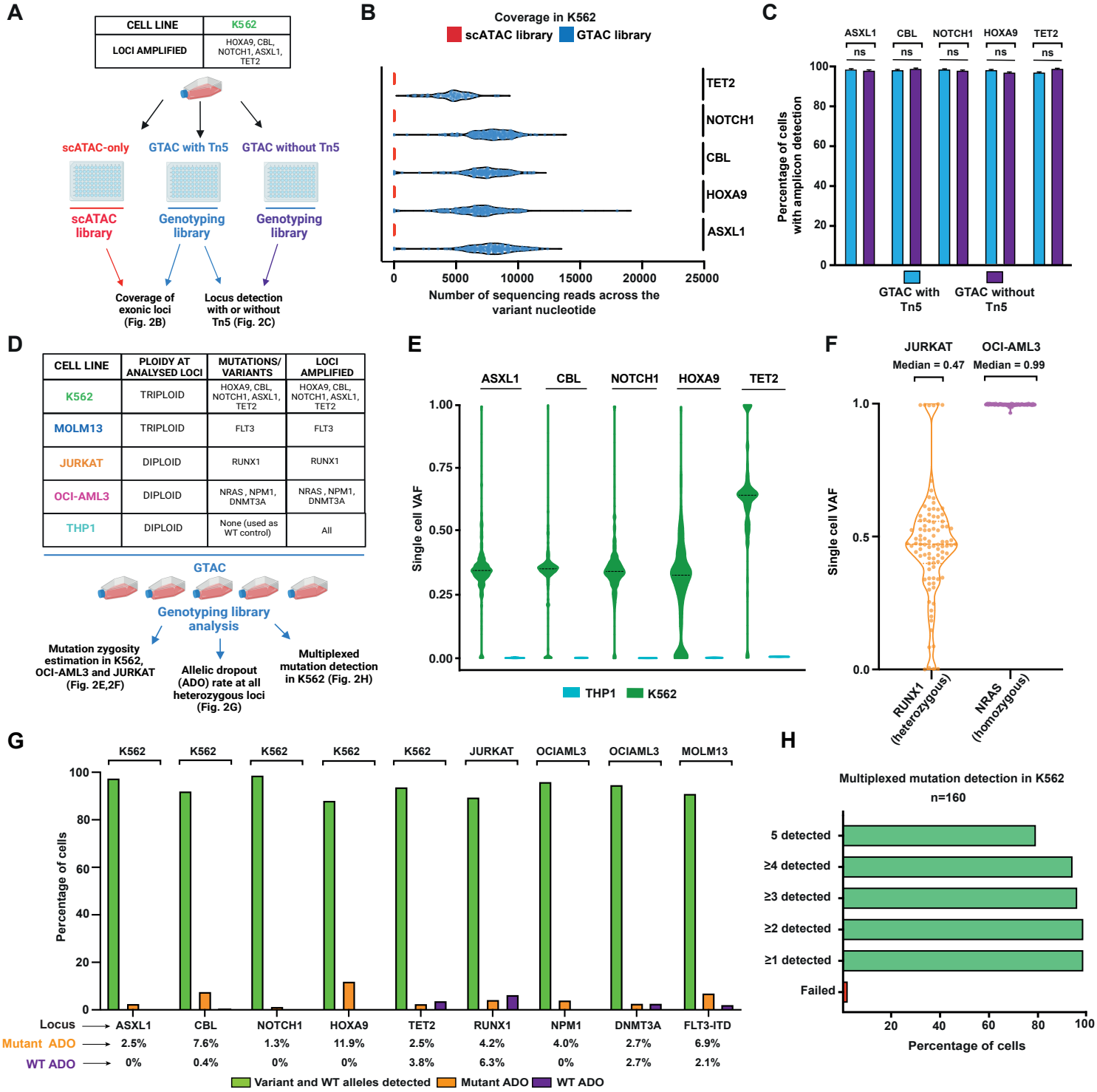

A
CELL LINE
K562
LOCI AMPLIFIED
HOXA9, CBL, NOTCH1, ASXL1, TET2
scATAC-only
GTAC with Tn5
GTAC without Tn5
scATAC library
Genotyping library
Genotyping library
Coverage of exonic loci (Fig. 2B)
Locus detection with or without Tn5 (Fig. 2C)
B
Coverage in K562
scATAC library
GTAC library
TET2
NOTCH1
CBL
HOXA9
ASXL1
0
5000
10000
15000
20000
25000
Number of sequencing reads across the variant nucleotide
C
ASXL1
CBL
NOTCH1
HOXA9
TET2
ns
Percentage of cells with amplicon detection
GTAC with Tn5
GTAC without Tn5
D
CELL LINE
PLOIDY AT ANALYSED LOCI
MUTATIONS/ VARIANTS
LOCI AMPLIFIED
K562
TRIPLOID
HOXA9, CBL, NOTCH1, ASXL1, TET2
HOXA9, CBL, NOTCH1, ASXL1, TET2
MOLM13
TRIPLOID
FLT3
FLT3
JURKAT
DIPLOID
RUNX1
RUNX1
OCI-AML3
DIPLOID
NRAS , NPM1, DNMT3A
NRAS , NPM1, DNMT3A
THP1
DIPLOID
None (used as WT control)
All
GTAC
Genotyping library analysis
Mutation zygosity estimation in K562, OCI-AML3 and JURKAT (Fig. 2E,2F)
Allelic dropout (ADO) rate at all heterozygous loci (Fig. 2G)
Multiplexed mutation detection in K562 (Fig. 2H)
E
ASXL1
CBL
NOTCH1
HOXA9
TET2
Single cell VAF
THP1
K562
F
JURKAT
OCI-AML3
Median = 0.47
Median = 0.99
Single cell VAF
RUNX1 (heterozygous)
NRAS (homozygous)
G
K562
JURKAT
OCIAML3
MOLM13
Percentage of cells
Locus
ASXL1
CBL
NOTCH1
HOXA9
TET2
RUNX1
NPM1
DNMT3A
FLT3-ITD
Mutant ADO
2.5%
7.6%
1.3%
11.9%
2.5%
4.2%
4.0%
2.7%
6.9%
WT ADO
0%
0.4%
0%
0%
3.8%
6.3%
0%
2.7%
2.1%
Variant and WT alleles detected
Mutant ADO
WT ADO
H
Multiplexed mutation detection in K562
n=160
5 detected
≥4 detected
≥3 detected
≥2 detected
≥1 detected
Failed
Percentage of cells

Figure 3

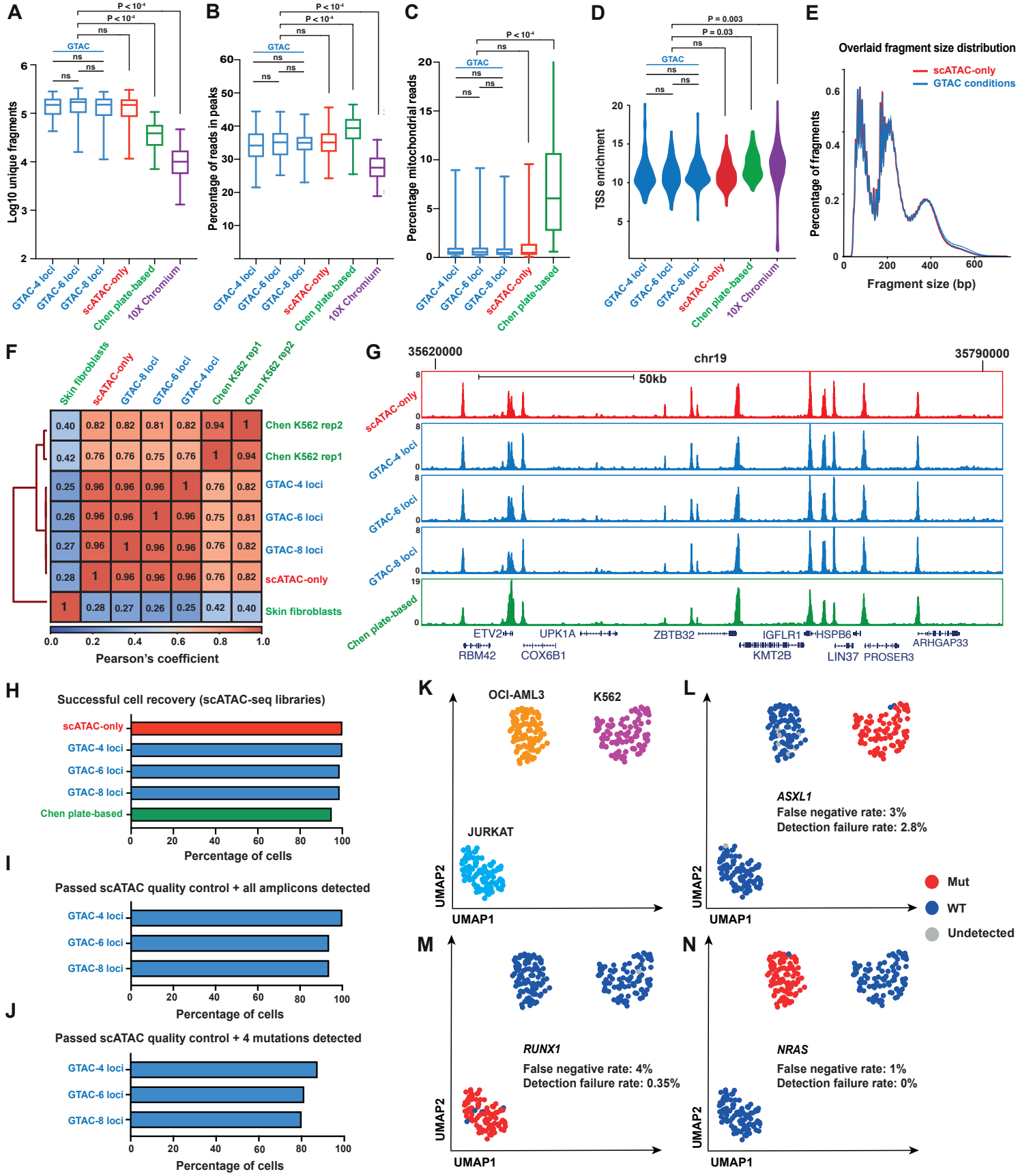

A
B
C
D
E
Log10 unique fragments
Percentage of reads in peaks
Percentage mitochondrial reads
TSS enrichment
Overlaid fragment size distribution
scATAC-only
GTAC conditions
Percentage of fragments
Fragment size (bp)
GTAC
ns
P < 10⁻⁴
P = 0.003
P = 0.03
GTAC-4 loci
GTAC-6 loci
GTAC-8 loci
scATAC-only
Chen plate-based
10X Chromium
F
Skin fibroblasts
scATAC-only
GTAC-8 loci
GTAC-6 loci
GTAC-4 loci
Chen K562 rep1
Chen K562 rep2
Pearson's coefficient
G
35620000
chr19
35790000
50kb
ETV2
UPK1A
ZBTB32
IGFLR1
HSPB6
ARHGAP33
RBM42
COX6B1
KMT2B
LIN37
PROSER3
H
Successful cell recovery (scATAC-seq libraries)
Percentage of cells
I
Passed scATAC quality control + all amplicons detected
J
Passed scATAC quality control + 4 mutations detected
K
OCI-AML3
K562
JURKAT
UMAP1
UMAP2
L
ASXL1
False negative rate: 3%
Detection failure rate: 2.8%
Mut
WT
Undetected
M
RUNX1
False negative rate: 4%
Detection failure rate: 0.35%
N
NRAS
False negative rate: 1%
Detection failure rate: 0%

Figure 4

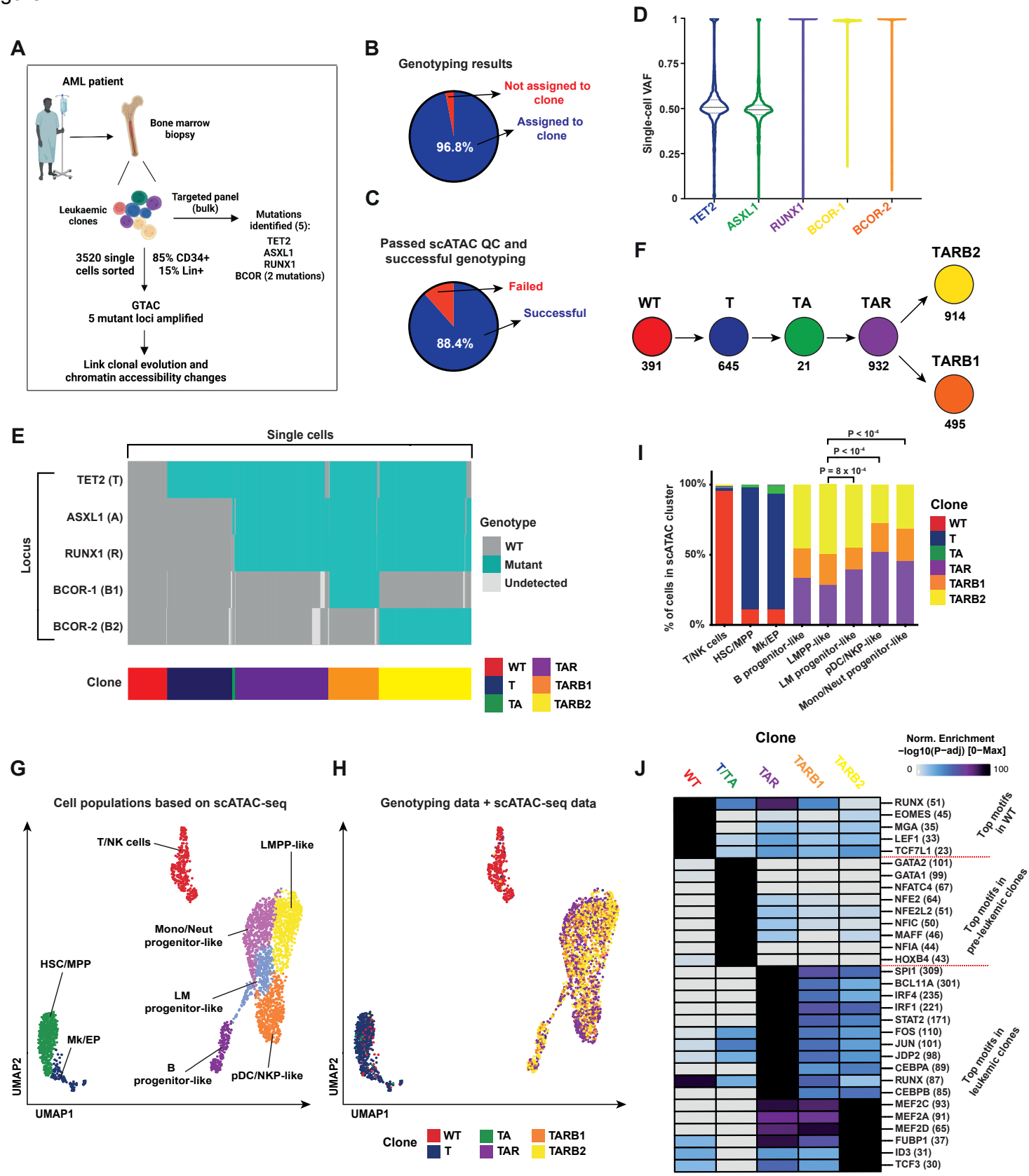

A
AML patient
Bone marrow biopsy
Leukaemic clones
Targeted panel (bulk)
Mutations identified (5):
TET2
ASXL1
RUNX1
BCOR (2 mutations)
3520 single cells sorted
85% CD34+ 15% Lin+
GTAC
5 mutant loci amplified
Link clonal evolution and chromatin accessibility changes
B
Genotyping results
Not assigned to clone
Assigned to clone
96.8%
C
Passed scATAC QC and successful genotyping
Failed
Successful
88.4%
D
Single-cell VAF
1
0.75
0.50
0.25
0
TET2
ASXL1
RUNX1
BCOR-1
BCOR-2
F
WT
T
TA
TAR
TARB2
TARB1
391
645
21
932
914
495
E
Single cells
Locus
TET2 (T)
ASXL1 (A)
RUNX1 (R)
BCOR-1 (B1)
BCOR-2 (B2)
Genotype
WT
Mutant
Undetected
Clone
WT
T
TA
TAR
TARB1
TARB2
I
P < 10-4
P < 10-4
P = 8 x 10-4
100%
50%
0%
% of cells in scATAC cluster
T/NK cells
HSC/MPP
Mk/EP
B progenitor-like
LMPP-like
LM progenitor-like
pDC/NKP-like
Mono/Neut progenitor-like
Clone
WT
T
TA
TAR
TARB1
TARB2
G
Cell populations based on scATAC-seq
T/NK cells
LMPP-like
Mono/Neut progenitor-like
HSC/MPP
LM progenitor-like
Mk/EP
B progenitor-like
pDC/NKP-like
UMAP2
UMAP1
H
Genotyping data + scATAC-seq data
UMAP2
UMAP1
Clone
WT
T
TA
TAR
TARB1
TARB2
J
Clone
WT
T/TA
TAR
TARB1
TARB2
Norm. Enrichment −log10(P−adj) [0−Max]
0
100
RUNX (51)
EOMES (45)
MGA (35)
LEF1 (33)
TCF7L1 (23)
GATA2 (101)
GATA1 (99)
NFATC4 (67)
NFE2 (64)
NFE2L2 (51)
NFIC (50)
MAFF (46)
NFIA (44)
HOXB4 (43)
SPI1 (309)
BCL11A (301)
IRF4 (235)
IRF1 (221)
STAT2 (171)
FOS (110)
JUN (101)
JDP2 (98)
CEBPA (89)
RUNX (87)
CEBPB (85)
MEF2C (93)
MEF2A (91)
MEF2D (65)
FUBP1 (37)
ID3 (31)
TCF3 (30)
Top motifs in WT
Top motifs in pre-leukemic clones
Top motifs in leukemic clones

Figure 5

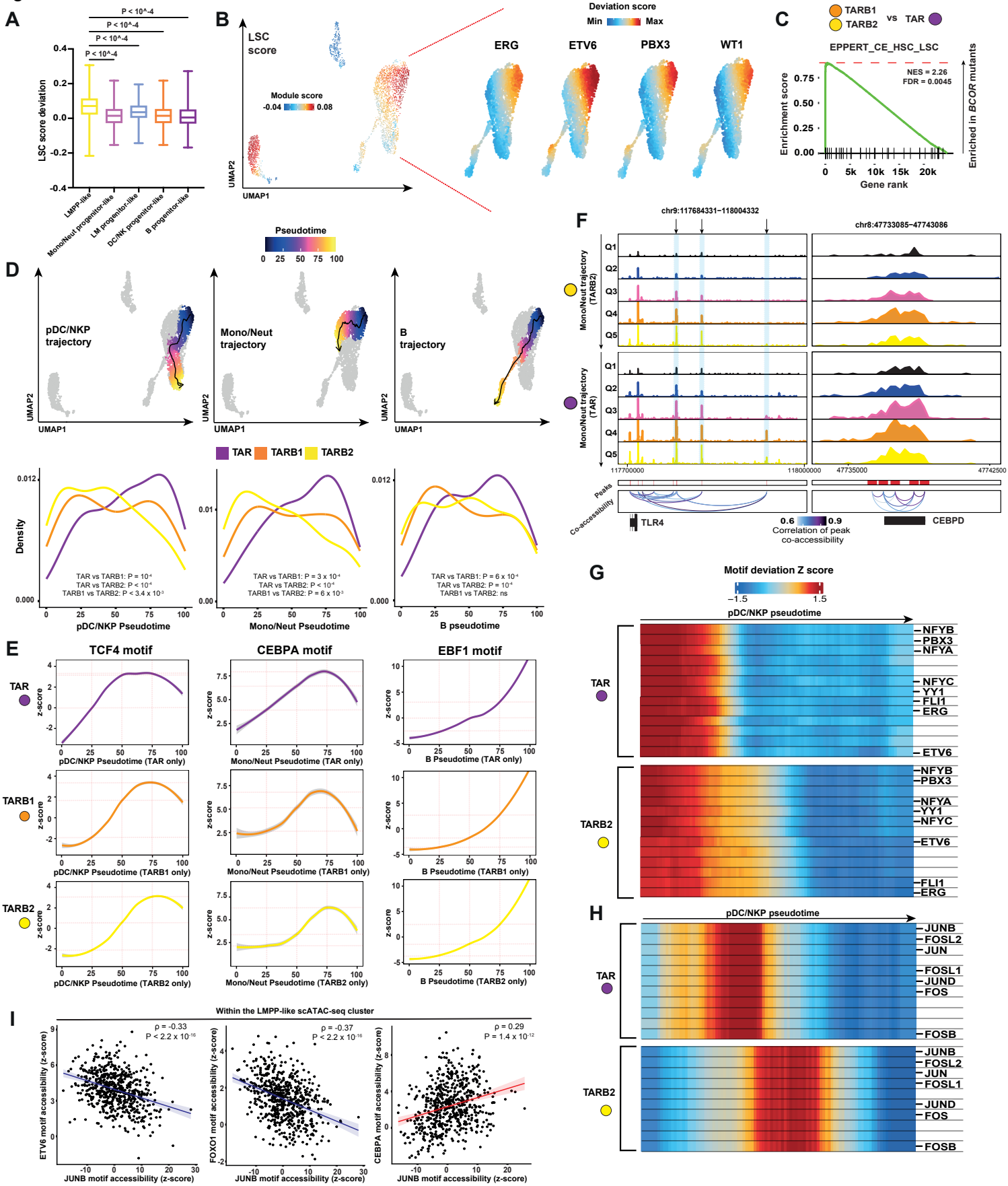
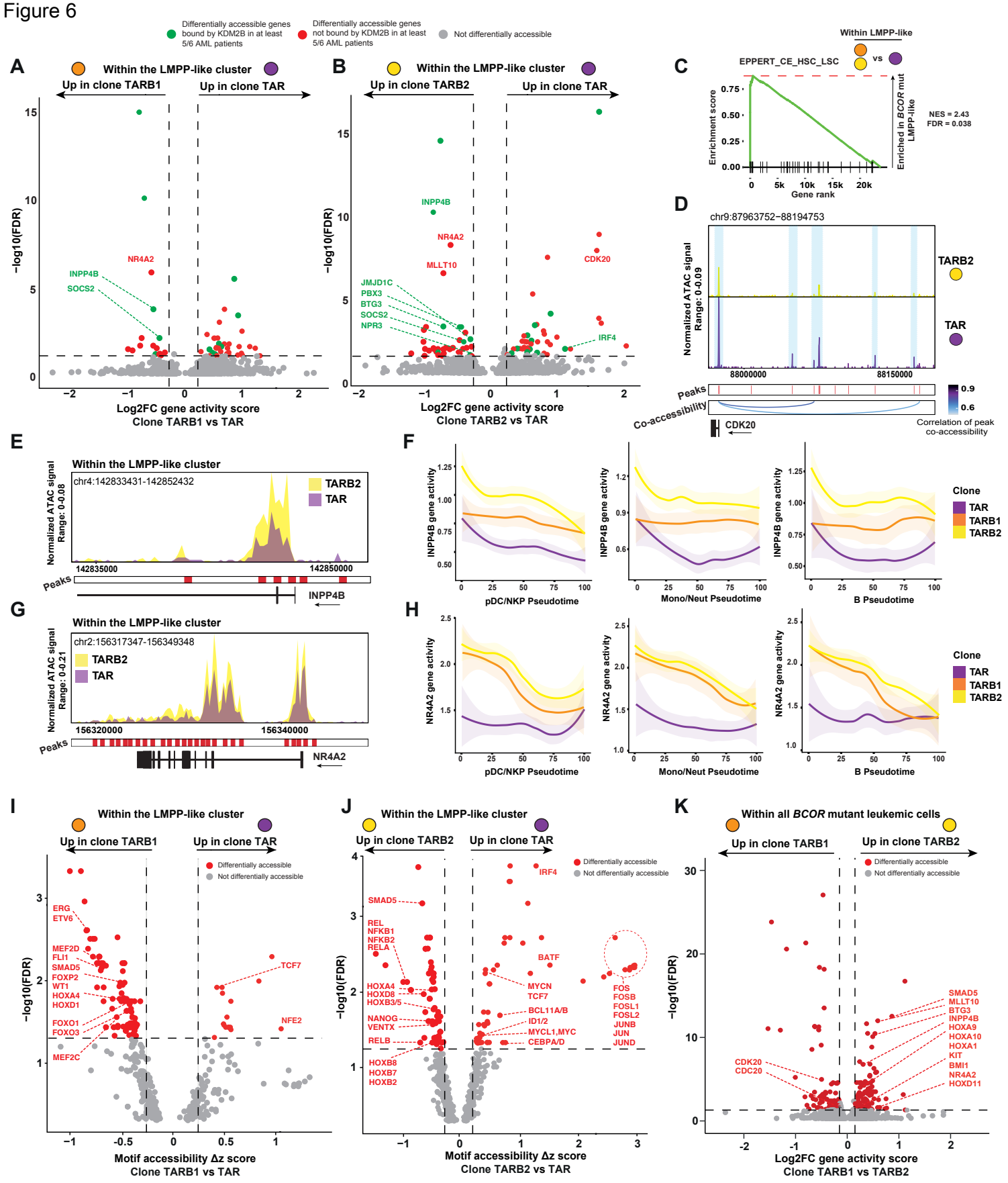
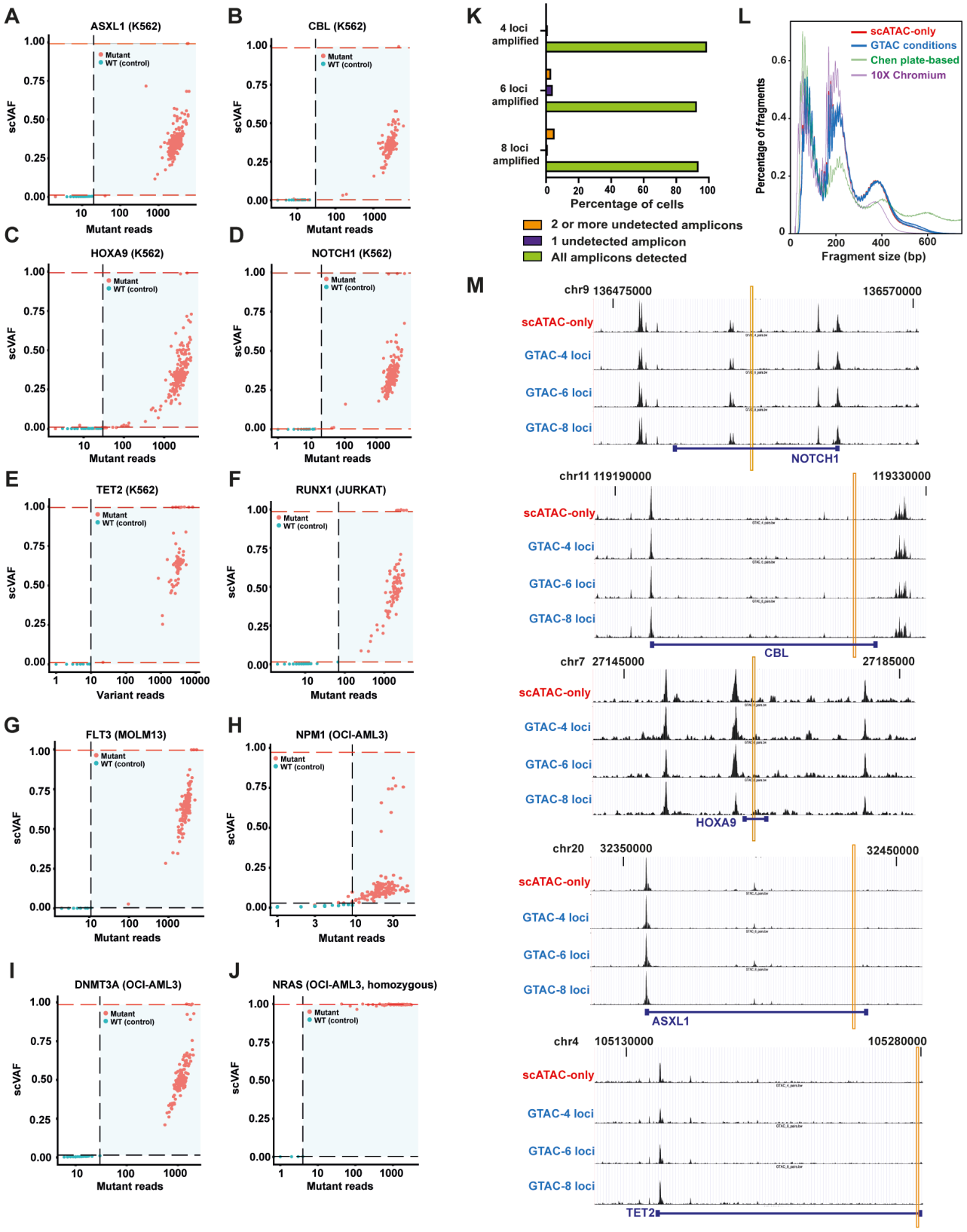

A
P < 10^-4
LSC score deviation
LMPP-like
Mono/Neut progenitor-like
LM progenitor-like
DC/NK progenitor-like
B progenitor-like
B
LSC score
Module score
-0.04
0.08
UMAP1
UMAP2
Deviation score
Min
Max
ERG
ETV6
PBX3
WT1
Pseudotime
C
TARB1
TARB2
vs
TAR
EPPERT_CE_HSC_LSC
NES = 2.26
FDR = 0.0045
Enrichment score
Enriched in BCOR mutants
Gene rank
D
pDC/NKP trajectory
Mono/Neut trajectory
B trajectory
TAR
TARB1
TARB2
Density
TAR vs TARB1: P = 10^-4
TAR vs TARB2: P < 10^-4
TARB1 vs TARB2: P < 3.4 x 10^-3
pDC/NKP Pseudotime
TAR vs TARB1: P = 3 x 10^-4
TAR vs TARB2: P < 10^-4
TARB1 vs TARB2: P = 6 x 10^-3
Mono/Neut Pseudotime
TAR vs TARB1: P = 6 x 10^-4
TAR vs TARB2: P = 10^-4
TARB1 vs TARB2: ns
B pseudotime
E
TCF4 motif
CEBPA motif
EBF1 motif
z-score
pDC/NKP Pseudotime (TAR only)
Mono/Neut Pseudotime (TAR only)
B Pseudotime (TAR only)
pDC/NKP Pseudotime (TARB1 only)
Mono/Neut Pseudotime (TARB1 only)
B Pseudotime (TARB1 only)
pDC/NKP Pseudotime (TARB2 only)
Mono/Neut Pseudotime (TARB2 only)
B Pseudotime (TARB2 only)
F
chr9:117684331–118004332
chr8:47733085–47743086
Mono/Neut trajectory (TARB2)
Mono/Neut trajectory (TAR)
Q1
Q2
Q3
Q4
Q5
Peaks
Co-accessibility
TLR4
CEBPD
Correlation of peak co-accessibility
G
Motif deviation Z score
pDC/NKP pseudotime
NFYB
PBX3
NFYA
NFYC
YY1
FLI1
ERG
ETV6
H
JUNB
FOSL2
JUN
FOSL1
JUND
FOS
FOSB
I
Within the LMPP-like scATAC-seq cluster
ρ = -0.33
P < 2.2 x 10^-16
ETV6 motif accessibility (z-score)
JUNB motif accessibility (z-score)
ρ = -0.37
P < 2.2 x 10^-16
FOXO1 motif accessibility (z-score)
ρ = 0.29
P = 1.4 x 10^-12
CEBPA motif accessibility (z-score)

Figure 6


Differentially accessible genes bound by KDM2B in at least 5/6 AML patients
Differentially accessible genes not bound by KDM2B in at least 5/6 AML patients
Not differentially accessible
A
Within the LMPP-like cluster
Up in clone TARB1
Up in clone TAR
−log10(FDR)
NR4A2
INPP4B
SOCS2
Log2FC gene activity score
Clone TARB1 vs TAR
B
Within the LMPP-like cluster
Up in clone TARB2
Up in clone TAR
INPP4B
NR4A2
MLLT10
CDK20
JMJD1C
PBX3
BTG3
SOCS2
NPR3
IRF4
Log2FC gene activity score
Clone TARB2 vs TAR
C
Within LMPP-like
EPPERT_CE_HSC_LSC
vs
Enrichment score
Enriched in BCOR mut LMPP-like
NES = 2.43
FDR = 0.038
Gene rank
D
chr9:87963752−88194753
Normalized ATAC signal Range: 0-0.09
TARB2
TAR
Peaks
Co-accessibility
CDK20
Correlation of peak co-accessibility
E
Within the LMPP-like cluster
chr4:142833431-142852432
Normalized ATAC signal Range: 0-0.08
TARB2
TAR
Peaks
INPP4B
F
INPP4B gene activity
pDC/NKP Pseudotime
Mono/Neut Pseudotime
B Pseudotime
Clone
TAR
TARB1
TARB2
G
Within the LMPP-like cluster
chr2:156317347-156349348
Normalized ATAC signal Range: 0-0.21
TARB2
TAR
Peaks
NR4A2
H
NR4A2 gene activity
pDC/NKP Pseudotime
Mono/Neut Pseudotime
B Pseudotime
Clone
TAR
TARB1
TARB2
I
Within the LMPP-like cluster
Up in clone TARB1
Up in clone TAR
Differentially accessible
Not differentially accessible
−log10(FDR)
ERG
ETV6
MEF2D
FLI1
SMAD5
FOXP2
WT1
HOXA4
HOXD1
FOXO1
FOXO3
MEF2C
TCF7
NFE2
Motif accessibility Δz score
Clone TARB1 vs TAR
J
Within the LMPP-like cluster
Up in clone TARB2
Up in clone TAR
Differentially accessible
Not differentially accessible
IRF4
SMAD5
REL
NFKB1
NFKB2
RELA
HOXA4
HOXD8
HOXB3/5
NANOG
VENTX
RELB
HOXB8
HOXB7
HOXB2
BATF
MYCN
TCF7
BCL11A/B
ID1/2
MYCL1,MYC
CEBPA/D
FOS
FOSB
FOSL1
FOSL2
JUNB
JUN
JUND
Motif accessibility Δz score
Clone TARB2 vs TAR
K
Within all BCOR mutant leukemic cells
Up in clone TARB1
Up in clone TARB2
Differentially accessible
Not differentially accessible
SMAD5
MLLT10
BTG3
INPP4B
HOXA9
HOXA10
HOXA1
KIT
BMI1
NR4A2
HOXD11
CDK20
CDC20
Log2FC gene activity score
Clone TARB1 vs TARB2

Supplemental Figures and Legends.



Supplemental Figure 1. Related to Figures 2 and 3. Thresholds used for allelic drop-out (ADO) estimation and coverage around genotyped loci in K562

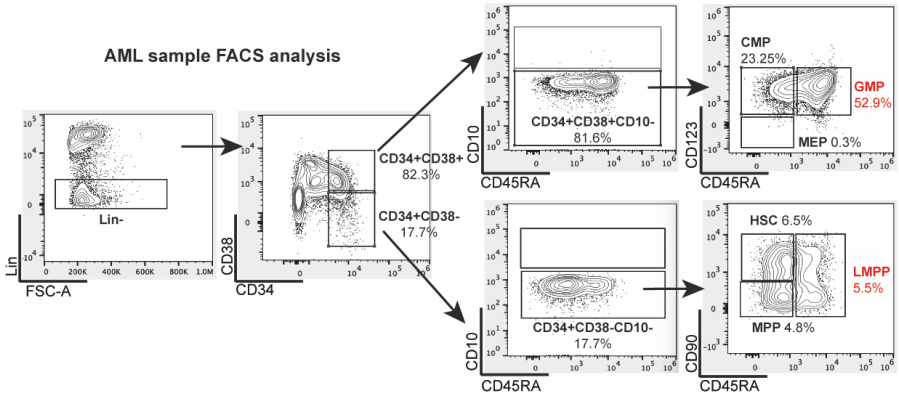
(A)-(J) scVAF (y-axis) versus the number of mutant or variant genotyping reads (x-axis) for nine heterozygous mutant or variant loci (A)-(I) and one homozygous mutant locus (J). Each dot represents a single cell. THP1 cells were used as homozygous WT control for each locus (cyan). Mutant cells used (red) are indicated above each graph. In each panel, the vertical black dotted line indicates the minimum number of mutant or variant reads required for a cell to be called mutant or variant (STAR Methods). The lower horizontal red dotted line indicates the minimum scVAF required for a cell to be called mutant or variant (STAR Methods). The upper horizontal red dotted line indicates the maximum scVAF for which a cell is considered heterozygous mutant or variant (STAR Methods). For heterozygous loci, dots outside of the azure rectangle are cells in which we estimate that ADO occurred.

(K) Percentage of K562 cells (x-axis) in which amplified loci were detected when either 4 (top), 6 (middle) or 8 (bottom) loci were amplified. The legend below shows the number of multiple amplicons detected. For each amplicon, at least 30 reads were required for the locus to be considered detected.

(L) Percentage of ATAC fragments (y-axis) as a function of fragment size (x-axis) showing data for K562 assayed by GTAC, scATAC alone, published K562 data obtained from a plate-based scATAC-seq approach (Chen plate-based) and from a commercial droplet-based approach (10X Chromium).

(M) UCSC genome browser tracks displaying aggregated reads (y-axis) across five genes containing heterozygous mutations/variants targeted in K562 cells (top to bottom: *NOTCH1*, *CBL*, *HOXA9*, *ASXL1*, *TET2*), comparing the 4 conditions (labelled on the left). Low signal in coding regions confirms that exonic variant profiling from scATAC data alone is unfeasible. Yellow rectangles indicate the position of the heterozygous mutations or variants analysed in genotyping.

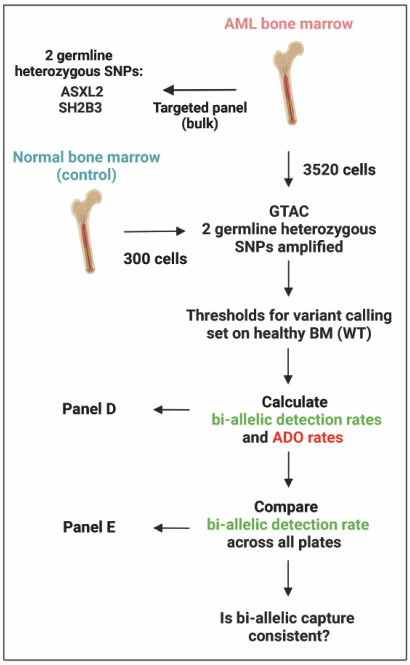
A



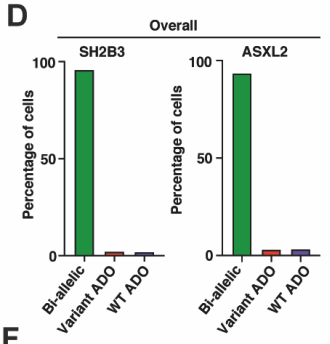
B

Population	AML Bone marrow	Normal Bone marrow*
Lin-CD34$^+$CD38$^-$	17.7%	16.1%
Lin-CD34$^+$CD38$^-$CD10$^-$	17.7%	15.8%
HSC	6.5%	11.4%
MPP	4.8%	3.7%
LMPP	5.5%	0.2%
Lin-CD34$^+$CD38$^+$	82.3%	83.9%
Lin-CD34$^+$CD38$^+$CD10$^-$	81.6%	78.4%
CMP	23.3%	34.7%
GMP	52.9%	17.8%
MEP	0.3%	14.6%

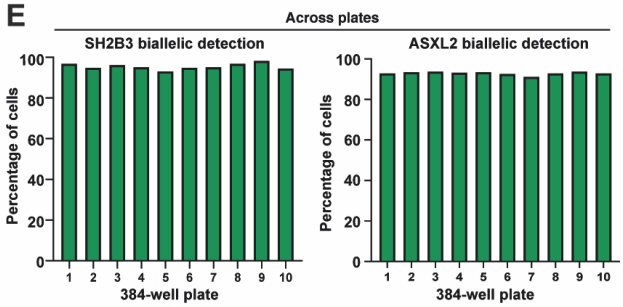
C



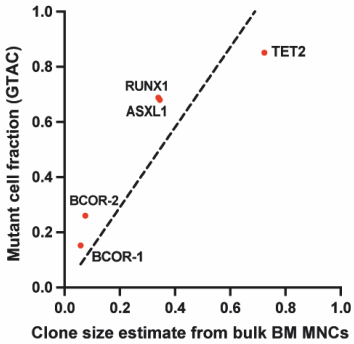
D



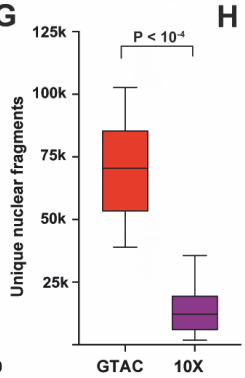
E



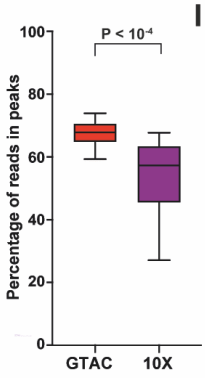
F



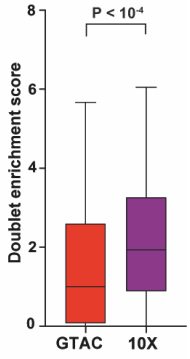
G



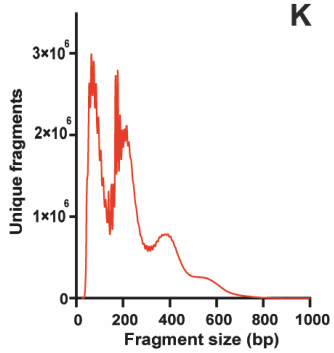
H



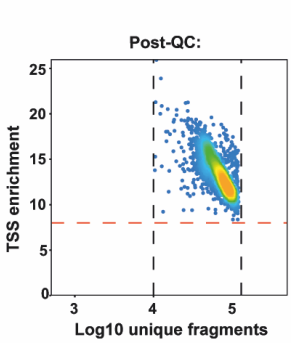
I



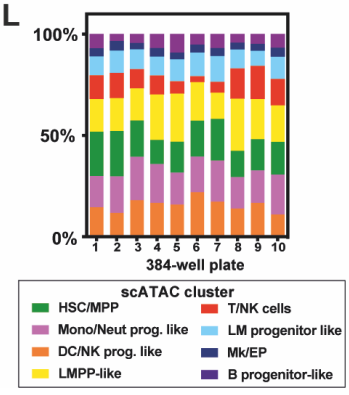
J

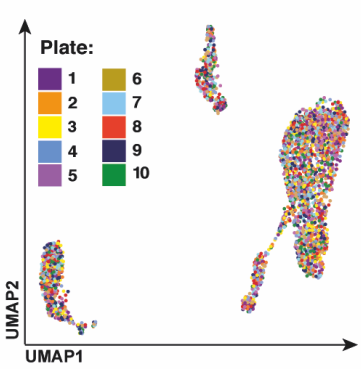


K



L





Supplemental Figure 2. Related to Figure 4. AML sample immunophenotype, ADO estimation and scATAC quality metrics.

(A) FACS analysis of the AML sample. Percentages of Lin-CD34$^+$ cells are indicated below populations. HSC: haematopoietic stem cell, MPP: multipotent progenitor, LMPP: lymphoid-primer multipotent progenitor, CMP: common myeloid progenitor, GMP: granulocyte-monocyte progenitor, MEP: megakaryocyte-erythroid progenitor.

(B) Table of % of immunophenotypic populations as a proportion of Lin$^-$CD34$^+$ in the AML sample (left) and normal bone marrow[S1], showing expansion of the GMP and LMPP immunophenotypic compartments in the AML sample.

(C) Overview of ADO estimation in the AML sample. Bone marrow cells were isolated from a newly diagnosed male AML patient. Next generation sequencing was performed on leukemic bone marrow mononuclear cells using a panel of genes recurrently mutated in myeloid malignancy. 2 heterozygous germline SNPs in *ASXL2* and *SH2B3* genes were identified. These two loci were co-amplified in the GTAC experiment together with the 5 somatic mutant loci. The same 2 loci were also amplified in 300 homozygous WT CD34+ human bone marrow cells used as control for setting ADO thresholds (STAR Methods).

(D) Percentage of cells (y-axis) in which both variant and WT alleles were detected (green) compared to cells where only the WT (red; i.e. there was variant ADO) or variant (purple; i.e. there was WT ADO) allele was detected, for the *SH2B3* and *ASXL2* loci. Percentages of cells in which there was bi-allelic detection, variant ADO, or WT ADO, respectively, are: for *SH2B3*: 95.8%, 2.2%, and 2.0%; for *ASXL2*: 93.5%, 3.1%, and 3.4%.

(E) Percentage of cells in which there was bi-allelic detection of *SH2B3* (left) and *ASXL2* (right) loci across ten 384-well plates.

(F) Fraction of mutant cells as estimated by GTAC (y-axis) versus the expected clone size inferred from bulk sequencing of bone marrow mononuclear cells (x-axis). The dotted line is the linear regression curve. Each locus is indicated by a red dot.

(G) Number of unique nuclear scATAC fragments obtained per cell, comparing data from the AML sample processed with GTAC (red, n=3285) with data from five mixed phenotype acute leukemia (MPAL) samples processed with 10X Chromium[S2] (purple, n=20059). Median values are: 70475 for GTAC; 12192 for 10X. Cells with >1000 and <120000 unique nuclear ATAC fragments were considered. *P* values obtained using two-tailed *t*-test.

(H) Percentage of ATAC reads in peaks, comparing data from the AML sample processed with GTAC (red, n=3285) with data from five mixed phenotype acute leukemia (MPAL) samples processed with 10X Chromium[S2] (purple, n=20059). Median values are: 67.83% for

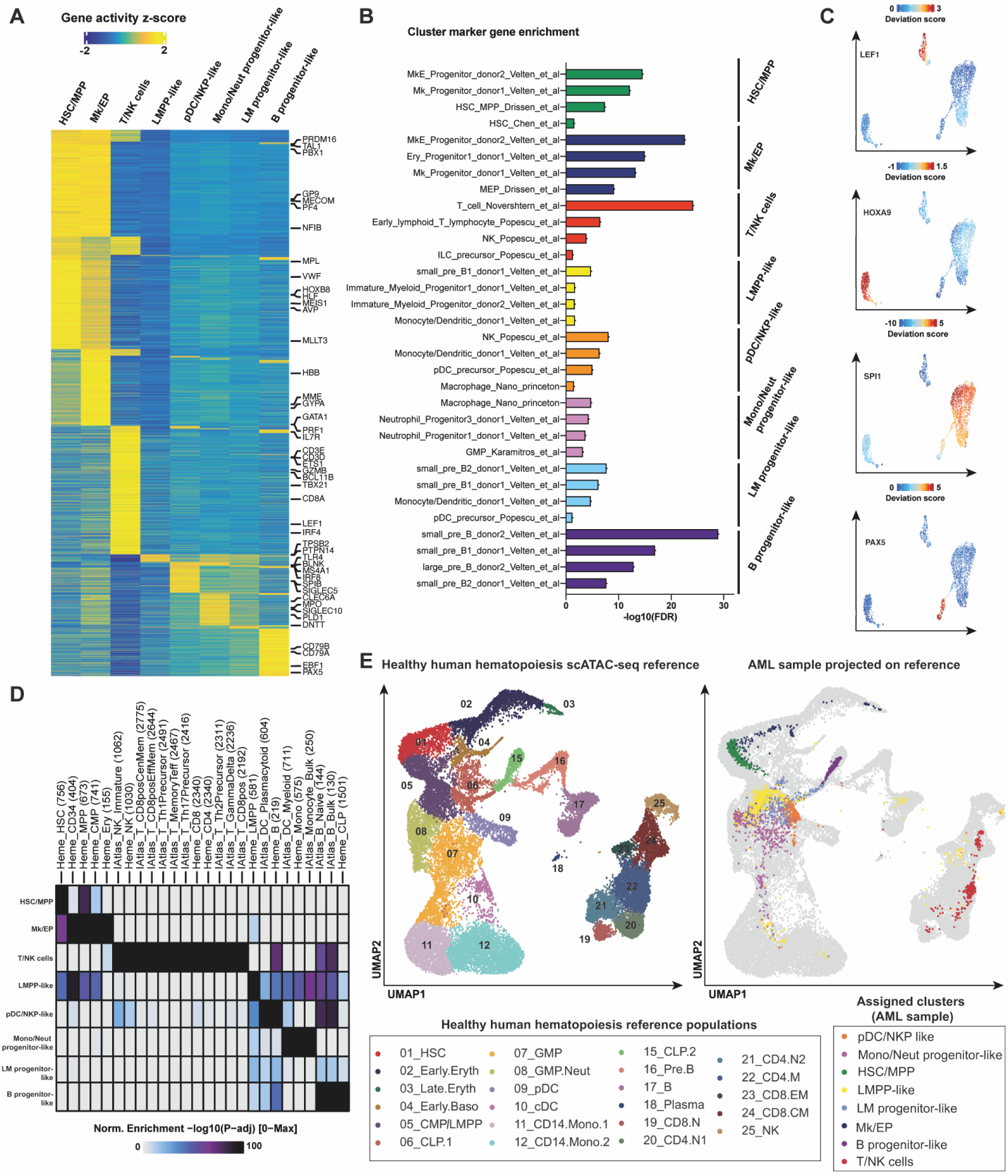
GTAC; 57.30% for 10X. Cells with >1000 and <120000 unique nuclear ATAC fragments were considered. *P* values obtained using two-tailed *t*-test.

(I) Doublet enrichment scores, comparing data from the AML sample processed with GTAC (red, n=3285) with data from five mixed phenotype acute leukemia (MPAL) samples processed with 10X Chromium[S2] (purple, n=20059). Median values are: 1.00 for GTAC; 1.93 for 10X. Cells with >1000 and <120000 unique nuclear ATAC fragments were considered. *P* values obtained using two-tailed *t*-test.

(J) Number of ATAC fragments (y-axis) as a function of fragment size (x-axis) aggregated across all cells from our AML sample.

(K) Scatterplot showing the $\log_{10}$ of the number of unique nuclear fragments (x-axis) versus the TSS enrichment score (y-axis) for each cell. Cells with TSS enrichment < 8 (red horizontal dotted line), less than 10000 (left vertical dotted line) and more than 120000 (right vertical dotted line) unique reads were excluded. Cells passing QC are displayed as dots. High cell density is indicated in orange.

(L) Top, bars indicating the contribution of each plate (x-axis) to each cluster (colour coded). Below, UMAP from Figure 4G in which each cell is colour-coded by plate.


A
Gene activity z-score
-2
2
HSC/MPP
Mk/EP
T/NK cells
LMPP-like
pDC/NKP-like
Mono/Neut progenitor-like
LM progenitor-like
B progenitor-like
PRDM16
TAL1
PBX1
GP9
MECOM
PF4
NFIB
MPL
VWF
HOXB8
HLF
MEIS1
AVP
MLLT3
HBB
MME
GYPA
GATA1
PRF1
IL7R
CD3E
CD3D
ETS1
GZMB
BCL11B
TBX21
CD8A
LEF1
IRF4
TPSB2
PTPN14
TLR4
BLNK
MS4A1
IRF8
SPIB
SIGLEC5
CLEC6A
MPO
SIGLEC10
PLD1
DNTT
CD79B
CD79A
EBF1
PAX5
B
Cluster marker gene enrichment
MkE_Progenitor_donor2_Velten_et_al
Mk_Progenitor_donor1_Velten_et_al
HSC_MPP_Drissen_et_al
HSC_Chen_et_al
MkE_Progenitor_donor2_Velten_et_al
Ery_Progenitor1_donor1_Velten_et_al
Mk_Progenitor_donor1_Velten_et_al
MEP_Drissen_et_al
T_cell_Novershtern_et_al
Early_lymphoid_T_lymphocyte_Popescu_et_al
NK_Popescu_et_al
ILC_precursor_Popescu_et_al
small_pre_B1_donor1_Velten_et_al
Immature_Myeloid_Progenitor1_donor1_Velten_et_al
Immature_Myeloid_Progenitor_donor2_Velten_et_al
Monocyte/Dendritic_donor1_Velten_et_al
NK_Popescu_et_al
Monocyte/Dendritic_donor1_Velten_et_al
pDC_precursor_Popescu_et_al
Macrophage_Nano_princeton
Macrophage_Nano_princeton
Neutrophil_Progenitor3_donor1_Velten_et_al
Neutrophil_Progenitor1_donor1_Velten_et_al
GMP_Karamitros_et_al
small_pre_B2_donor1_Velten_et_al
small_pre_B1_donor1_Velten_et_al
Monocyte/Dendritic_donor1_Velten_et_al
pDC_precursor_Popescu_et_al
small_pre_B_donor2_Velten_et_al
small_pre_B1_donor1_Velten_et_al
large_pre_B_donor2_Velten_et_al
small_pre_B2_donor1_Velten_et_al
0
10
20
30
-log10(FDR)
HSC/MPP
Mk/EP
T/NK cells
LMPP-like
pDC/NKP-like
Mono/Neut progenitor-like
LM progenitor-like
B progenitor-like
C
0
3
Deviation score
LEF1
-1
1.5
Deviation score
HOXA9
-10
5
Deviation score
SPI1
0
5
Deviation score
PAX5
D
Heme_HSC (756)
Heme_CD34 (404)
Heme_MPP (673)
Heme_CMP (741)
Heme_Ery (155)
IAtlas_NK_Immature (1062)
Heme_NK (1030)
IAtlas_T_CD8posCenMem (2775)
IAtlas_T_CD8posEffMem (2644)
IAtlas_T_Th1Precursor (2491)
IAtlas_T_MemoryTeff (2467)
IAtlas_T_Th17Precursor (2416)
Heme_CD8 (2340)
Heme_CD4 (2340)
IAtlas_T_Th2Precursor (2311)
IAtlas_T_GammaDelta (2236)
IAtlas_T_CD8pos (2192)
Heme_LMPP (581)
IAtlas_DC_Plasmacytoid (604)
Heme_B (219)
IAtlas_DC_Myeloid (711)
Heme_Mono (575)
IAtlas_Monocyte_Bulk (250)
IAtlas_B_Naive (144)
IAtlas_B_Bulk (130)
Heme_CLP (1501)
HSC/MPP
Mk/EP
T/NK cells
LMPP-like
pDC/NKP-like
Mono/Neut progenitor-like
LM progenitor-like
B progenitor-like
Norm. Enrichment −log10(P−adj) [0−Max]
0
100
E
Healthy human hematopoiesis scATAC-seq reference
AML sample projected on reference
UMAP2
UMAP1
Healthy human hematopoiesis reference populations
01_HSC
02_Early.Eryth
03_Late.Eryth
04_Early.Baso
05_CMP/LMPP
06_CLP.1
07_GMP
08_GMP.Neut
09_pDC
10_cDC
11_CD14.Mono.1
12_CD14.Mono.2
15_CLP.2
16_Pre.B
17_B
18_Plasma
19_CD8.N
20_CD4.N1
21_CD4.N2
22_CD4.M
23_CD8.EM
24_CD8.CM
25_NK
Assigned clusters (AML sample)
pDC/NKP like
Mono/Neut progenitor-like
HSC/MPP
LMPP-like
LM progenitor-like
Mk/EP
B progenitor-like
T/NK cells

Supplemental Figure 3. Related to Figure 4. Cluster annotation in GTAC data from the primary AML sample.

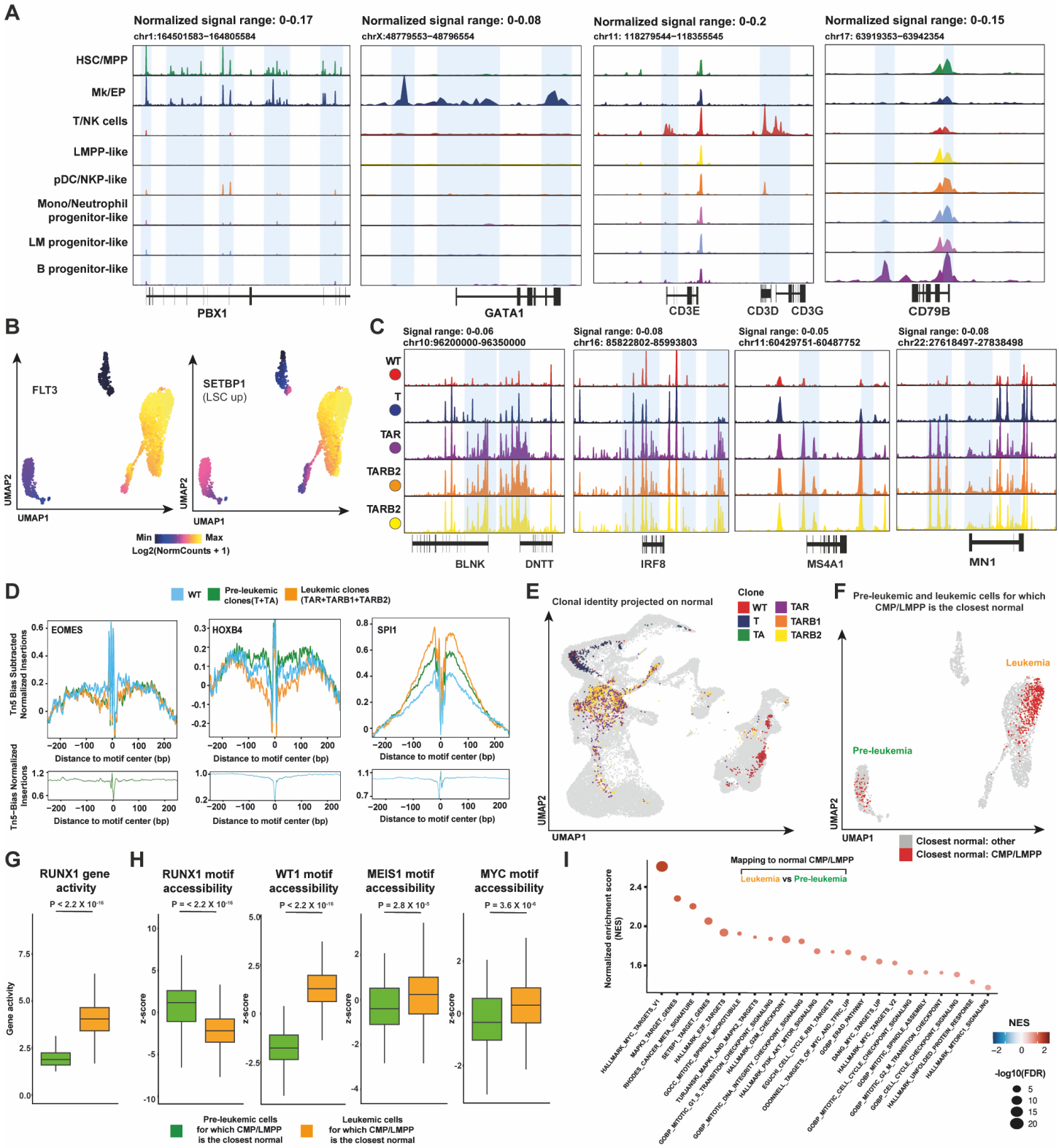
(A) Heatmap of gene activity scores for marker genes identified for each cluster (columns). Each row represents a gene. Marker genes with FDR < 0.05 and $\log_2$FC > 0.5 are shown. Several key marker genes are labelled on the right.

(B) Enrichment of haematopoietic gene signatures within marker genes identified for each cluster, based on gene activity scores. Here, marker genes were selected with FDR < 0.05 and $\log_2$FC > 0.25. Clusters are labelled on the right. Bars are colour-coded according to cluster, as in Figure 4G. Signatures were compiled using published scRNA-seq data[S1, S3-S7]. For each cluster, the top 4 enriched signatures are shown.

(C) Top to bottom: the UMAP from Figure 4G overlaid with per-cell deviation scores for motif accessibility for the LEF1, HOXA9, SPI1, and PAX5 motifs.

(D) Heatmap showing enrichment of curated peak sets from previously published bulk ATAC-seq data (provided by ArchR) within marker peaks identified for each cluster. Marker peaks were selected with FDR < 0.05 and $\log_2$FC > 0.25. The colour scale indicates the normalized peak set enrichment (-$\log_{10}$(adjusted *P* value)) on a scale of 0-100. The maximum -$\log_{10}$(adjusted *P* value) for each peak set is indicated by the numbers in brackets following the name of the enriched ATAC annotation.

(E) Left: UMAP based on previously published[S2] scATAC-seq data from healthy human hematopoiesis. Cluster labelling is indicated below the graphs. Numbers on the graph indicate the healthy reference bone marrow population. Clusters and labels were kept identical as in the original publication. Right: Projection of single cells from our AML dataset, coloured by assigned cluster, onto the healthy human hematopoiesis reference (grey).


A
Normalized signal range: 0-0.17
chr1:164501583–164805584
Normalized signal range: 0-0.08
chrX:48779553–48796554
Normalized signal range: 0-0.2
chr11: 118279544–118355545
Normalized signal range: 0-0.15
chr17: 63919353–63942354
HSC/MPP
Mk/EP
T/NK cells
LMPP-like
pDC/NKP-like
Mono/Neutrophil progenitor-like
LM progenitor-like
B progenitor-like
PBX1
GATA1
CD3E
CD3D
CD3G
CD79B
B
FLT3
SETBP1 (LSC up)
UMAP2
UMAP1
Min
Max
Log2(NormCounts + 1)
C
Signal range: 0-0.06
chr10:96200000-96350000
Signal range: 0-0.08
chr16: 85822802-85993803
Signal range: 0-0.05
chr11:60429751-60487752
Signal range: 0-0.08
chr22:27618497-27838498
WT
T
TAR
TARB2
TARB2
BLNK
DNTT
IRF8
MS4A1
MN1
D
WT
Pre-leukemic clones(T+TA)
Leukemic clones (TAR+TARB1+TARB2)
EOMES
HOXB4
SPI1
Tn5 Bias Subtracted Normalized Insertions
Tn5-Bias Normalized Insertions
Distance to motif center (bp)
E
Clonal identity projected on normal
Clone
WT
T
TA
TAR
TARB1
TARB2
UMAP2
UMAP1
F
Pre-leukemic and leukemic cells for which CMP/LMPP is the closest normal
Leukemia
Pre-leukemia
UMAP2
UMAP1
Closest normal: other
Closest normal: CMP/LMPP
G
RUNX1 gene activity
P < 2.2 X 10-16
Gene activity
H
RUNX1 motif accessibility
P = < 2.2 X 10-16
WT1 motif accessibility
P < 2.2 X 10-16
MEIS1 motif accessibility
P = 2.8 X 10-5
MYC motif accessibility
P = 3.6 X 10-6
z-score
Pre-leukemic cells for which CMP/LMPP is the closest normal
Leukemic cells for which CMP/LMPP is the closest normal
I
Mapping to normal CMP/LMPP
Leukemia vs Pre-leukemia
Normalized enrichment score (NES)
HALLMARK_MYC_TARGETS_V1
MAPK3_TARGET_GENES
RHODES_CANCER_META_SIGNATURE
SETBP1_TARGET_GENES
HALLMARK_E2F_TARGETS
GOCC_MITOTIC_SPINDLE_MICROTUBULE
TURUNSKI_MAPK1_AND_MAPK3_TARGETS
GOBP_MITOTIC_G1_S_TRANSITION_CHECKPOINT_SIGNALING
HALLMARK_G2M_CHECKPOINT
GOBP_MITOTIC_DNA_INTEGRITY_CHECKPOINT_SIGNALING
HALLMARK_PI3K_AKT_MTOR_SIGNALING
EGUCHI_CELL_CYCLE_RB1_TARGETS
ODONNELL_TARGETS_OF_MYC_AND_TFRC_UP
GOBP_ERAD_PATHWAY
DANG_MYC_TARGETS_UP
HALLMARK_MYC_TARGETS_V2
GOBP_MITOTIC_CELL_CYCLE_CHECKPOINT_SIGNALING
GOBP_MITOTIC_G2_M_TRANSITION_CHECKPOINT
GOBP_CELL_CYCLE_CHECKPOINT_SIGNALING
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_MTORC1_SIGNALING
NES
-log10(FDR)

Supplemental Figure 4. Related to Figure 4. A homozygous *RUNX1* mutation is associated with progression from preleukemia to leukemia and correlates with profound differences in chromatin states.

(A) Normalized scATAC sequencing read coverage (y-axis), comparing cluster pseudo-bulks (labelled on the left) around the *PBX1*, *GATA1, CD3E/D/G*, and *CD79B* loci. Signal range is displayed above each graph. Horizontal tracks represent cluster pseudo-bulks. Chromosome coordinates are indicated above. Gene annotations are indicated below.

(B) The UMAP from Figure 4G overlaid with gene activity scores for *FLT3* and *SETBP1*.

(C) Normalized scATAC sequencing read coverage (y-axis), comparing clone pseudo-bulks around the *BLNK/DNTT, IRF8*, *MS4A1* and *MN1* loci. Each horizontal track represents a different clone (labelled on the left). Rest of the panel as in (A).

(D) TF footprints for the EOMES, HOXB4 and SPI1 motifs comparing WT (azure, n=377 cells), preleukemic (green, n=595 cells) and leukemic clones (orange, n=2141 cells). Top, Tn5 bias-subtracted signal (y-axis) as a function of distance from the motif centre (x-axis). Bottom, Tn5 binding bias (y-axis) as a function of distance from the motif centre (x-axis), based on bias prediction computed by screening coverage of random hexamers within Tn5 insertion sites, using pseudo-bulk replicates. 1 indicates no bias. Colour legend is indicated above.

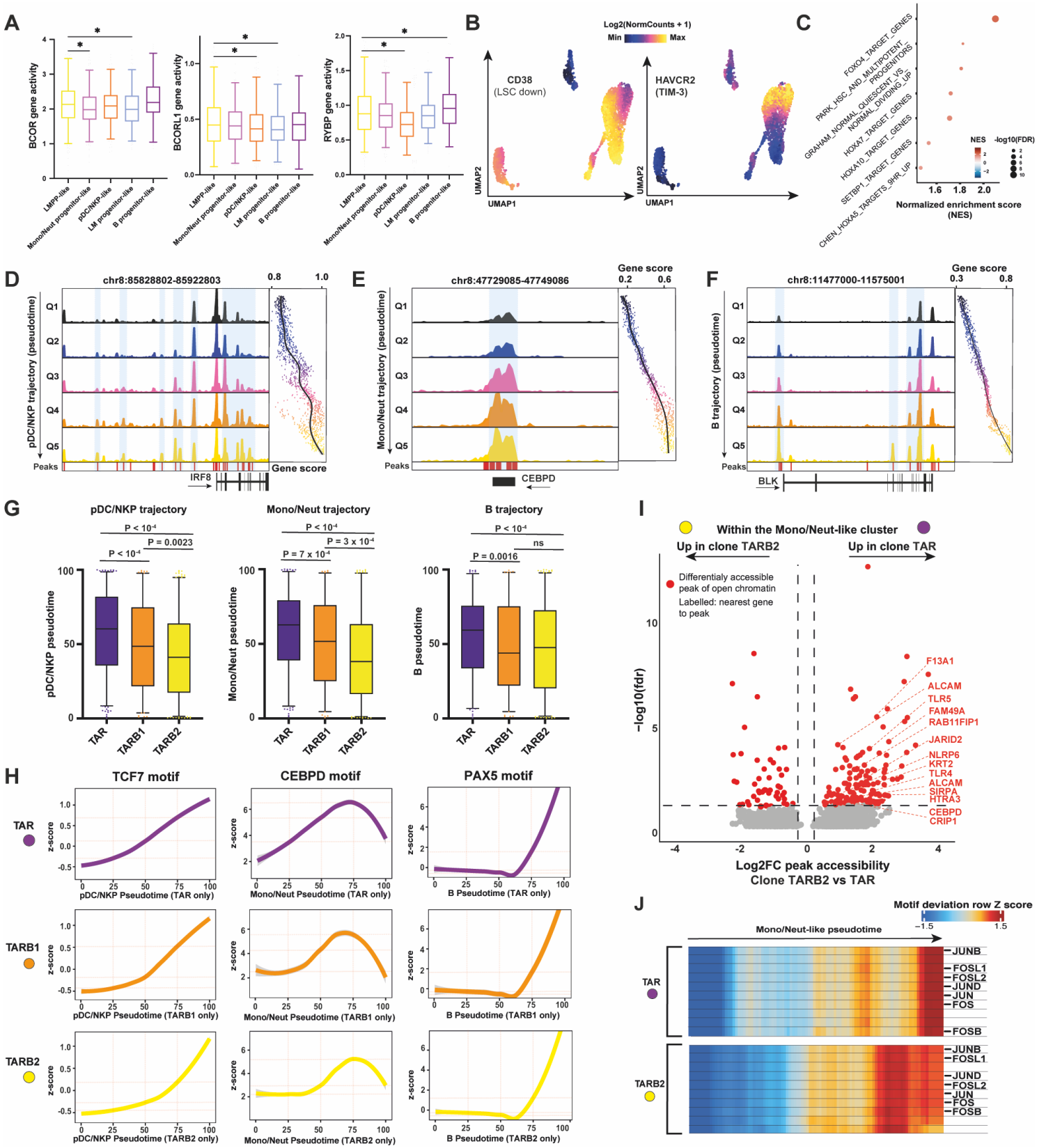
(E) Projection of single cells from our AML dataset, coloured by clone, onto the previously published[S2] scATAC-seq healthy human hematopoiesis reference (grey, same as in Figure S3E). Colour legend for clones is labelled in the top right corner (same as in Figure 4).

(F) UMAP from Figure 4G, highlighting cells that mapped onto normal CMP/LMPP in red (all other cells are in grey).

(G) Boxplots comparing *RUNX1* gene activity between preleukemic (green, n=97) and leukemic (orange, n=492) cells for which the projected closest normal population was CMP/LMPP. Wilcoxon rank sum test was used. The *P* value is reported above.

(H) Same as (G) but comparing RUNX1, WT1, MEIS1 and MYC binding motif accessibility between the two groups.

(I) Gene set enrichment analysis (GSEA) of genes with increased activity in leukemic cells compared to preleukemic cells within cells for which the closest normal population was CMP/LMPP. The y-axis plots the normalized enrichment score (NES). Dot size indicates $-\log_{10}$(FDR). Key pre-selected enriched pathways are indicated below (x-axis).


A
BCOR gene activity
BCORL1 gene activity
RYBP gene activity
LMPP-like
Mono/Neut progenitor-like
pDC/NKP-like
LM progenitor-like
B progenitor-like
B
Log2(NormCounts + 1)
Min
Max
CD38
(LSC down)
HAVCR2
(TIM-3)
UMAP1
UMAP2
C
FOXO4_TARGET_GENES
PARK_HSC_AND_MULTIPOTENT_PROGENITORS
GRAHAM_NORMAL_QUIESCENT_VS_NORMAL_DIVIDING_UP
HOXA7_TARGET_GENES
HOXA10_TARGET_GENES
SETBP1_TARGET_GENES
CHEN_HOXA5_TARGETS_9HR_UP
NES
-log10(FDR)
Normalized enrichment score (NES)
D
chr8:85828802-85922803
pDC/NKP trajectory (pseudotime)
Q1
Q2
Q3
Q4
Q5
Peaks
IRF8
Gene score
E
chr8:47729085-47749086
Mono/Neut trajectory (pseudotime)
CEBPD
F
chr8:11477000-11575001
B trajectory (pseudotime)
BLK
G
pDC/NKP trajectory
Mono/Neut trajectory
B trajectory
pDC/NKP pseudotime
Mono/Neut pseudotime
B pseudotime
TAR
TARB1
TARB2
H
TCF7 motif
CEBPD motif
PAX5 motif
z-score
pDC/NKP Pseudotime (TAR only)
Mono/Neut Pseudotime (TAR only)
B Pseudotime (TAR only)
pDC/NKP Pseudotime (TARB1 only)
Mono/Neut Pseudotime (TARB1 only)
B Pseudotime (TARB1 only)
pDC/NKP Pseudotime (TARB2 only)
Mono/Neut Pseudotime (TARB2 only)
B Pseudotime (TARB2 only)
I
Within the Mono/Neut-like cluster
Up in clone TARB2
Up in clone TAR
Differentialy accessible peak of open chromatin
Labelled: nearest gene to peak
−log10(fdr)
F13A1
ALCAM
TLR5
FAM49A
RAB11FIP1
JARID2
NLRP6
KRT2
TLR4
ALCAM
SIRPA
HTRA3
CEBPD
CRIP1
Log2FC peak accessibility
Clone TARB2 vs TAR
J
Motif deviation row Z score
Mono/Neut-like pseudotime
TAR
TARB2
JUNB
FOSL1
FOSL2
JUND
JUN
FOS
FOSB

Supplemental Figure 5. Related to Figure 5. GTAC identifies clone-specific chromatin dynamics throughout the leukemic hierarchy, highlighting a differentiation delay of *BCOR* mutant clones at an LSC-like state.

(A) Gene activity scores of genes encoding three PRC1.1 members (y-axis), namely *BCOR*, *BCORL1* and *RYBP*, across 5 leukemic clusters (labelled below). *P* values obtained using Kruskall-Wallis test with Dunnett's multiple comparisons. Significant differences are indicated with a star.

(B) The UMAP from Figure 4G overlaid with gene activity scores for *CD38* and *HAVCR2* (encoding the TIM-3 marker).

(C) Dot plot showing results of gene set enrichment analysis (GSEA), when LMPP-like cells were compared to all other leukemic cells. The x-axis plots the normalized enrichment score. Dot size indicates $-\log_{10}$(FDR). Pathways are indicated on the left (y-axis).

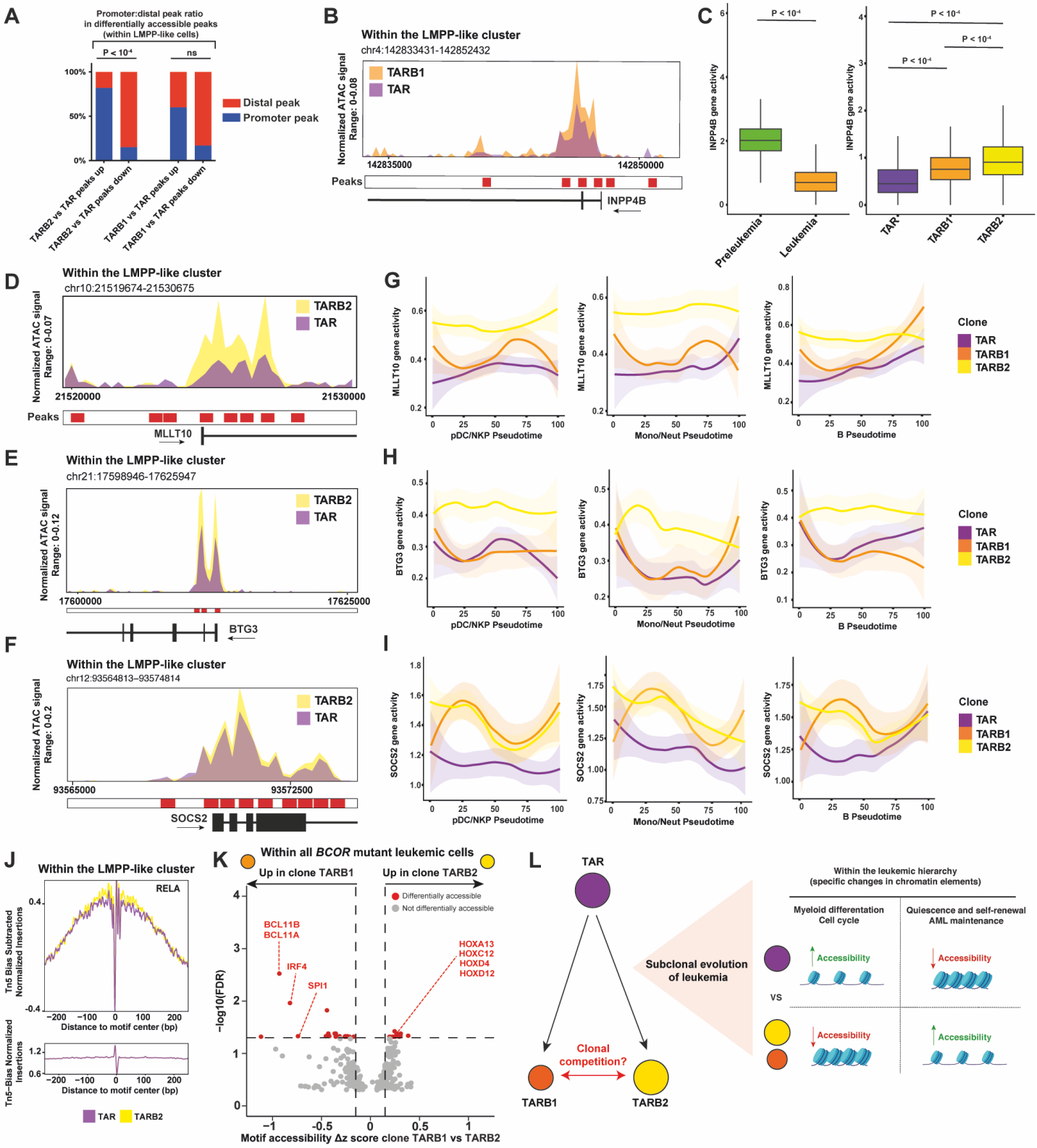
(D)-(F) Normalized scATAC sequencing read coverage around *IRF8* (D), *CEBPE* (E) and *BLK* (F) loci, across the pDC/NKP, Mono/Neut and B trajectory, respectively. Each trajectory was split into five quantiles (labelled Q1-Q5, colour-coded). Gene score as function of pseudotime is plotted on the right of each graph. Peaks and gene annotations are plotted below. Highlighted areas show variation across pseudotime.

(G) Boxplots depicting the per-cell pseudotime score (y-axis) across the three leukemic clones (colour-coded), for the three trajectories (left-to-right, indicated above each graph). *P* values obtained using Wilcoxon rank sum test.

(**H**) Same as in Figure 5E but showing TCF7, CEBPD and PAX5 motif accessibility.

(I) Volcano plot illustrating differentially accessible genomic peaks (FDR < 0.05 and $\log_2$FC > 0.25; labelled as red dots) between TARB2 (n=293 cells) and TAR (n=167 cells) clones within the Mono/Neut progenitor-like cluster. Vertical dotted lines indicate $\log_2$FC thresholds (-0.25 and 0.25), the horizontal dotted line indicates FDR=0.05. Labels indicate the closest gene to the peak. Labelled genes belong to monocytic and neutrophilic progenitor signatures.

(**J**) Heatmaps showing the variation of motif accessibility across clone-specific Mono/Neut pseudotime (x-axis), for a selected subset of most variable motifs across pseudotime (each row represents a motif). The upper heatmap relates to the TAR clone and the lower heatmap to the TARB2 clone. Motifs related to the AP-1 complex are labelled.


A
Promoter:distal peak ratio in differentially accessible peaks (within LMPP-like cells)
P < 10⁻⁴
ns
100%
50%
0%
Distal peak
Promoter peak
TARB2 vs TAR peaks up
TARB2 vs TAR peaks down
TARB1 vs TAR peaks up
TARB1 vs TAR peaks down
B
Within the LMPP-like cluster
chr4:142833431-142852432
Normalized ATAC signal Range: 0-0.08
TARB1
TAR
142835000
142850000
Peaks
INPP4B
C
INPP4B gene activity
P < 10⁻⁴
Preleukemia
Leukemia
TAR
TARB1
TARB2
D
Within the LMPP-like cluster
chr10:21519674-21530675
Normalized ATAC signal Range: 0-0.07
TARB2
TAR
21520000
21530000
Peaks
MLLT10
E
Within the LMPP-like cluster
chr21:17598946-17625947
Normalized ATAC signal Range: 0-0.12
TARB2
TAR
17600000
17625000
BTG3
F
Within the LMPP-like cluster
chr12:93564813–93574814
Normalized ATAC signal Range: 0-0.2
TARB2
TAR
93565000
93572500
SOCS2
G
MLLT10 gene activity
pDC/NKP Pseudotime
Mono/Neut Pseudotime
B Pseudotime
Clone
TAR
TARB1
TARB2
H
BTG3 gene activity
pDC/NKP Pseudotime
Mono/Neut Pseudotime
B Pseudotime
Clone
TAR
TARB1
TARB2
I
SOCS2 gene activity
pDC/NKP Pseudotime
Mono/Neut Pseudotime
B Pseudotime
Clone
TAR
TARB1
TARB2
J
Within the LMPP-like cluster
RELA
Tn5 Bias Subtracted Normalized Insertions
Distance to motif center (bp)
Tn5–Bias Normalized Insertions
Distance to motif center (bp)
TAR
TARB2
K
Within all BCOR mutant leukemic cells
Up in clone TARB1
Up in clone TARB2
Differentially accessible
Not differentially accessible
BCL11B
BCL11A
IRF4
SPI1
HOXA13
HOXC12
HOXD4
HOXD12
−log10(FDR)
Motif accessibility Δz score clone TARB1 vs TARB2
L
TAR
Clonal competition?
TARB1
TARB2
Subclonal evolution of leukemia
Within the leukemic hierarchy (specific changes in chromatin elements)
Myeloid differentation Cell cycle
Quiescence and self-renewal AML maintenance
Accessibility
VS

Supplemental Figure 6. GTAC reveals specific switches in chromatin accessibility within the LMPP-like differentiation stage upon subclonal acquisition of *BCOR* mutations.

(A) Ratio of differentially accessible peaks between *BCOR* mutant and *BCOR* WT LMPP-like cells which map to either distal elements (red) or promoters (blue). Two bars on the left represent the TARB2 versus TAR comparison. Two bars on the right represent the TARB1 versus TAR comparison. In each set of two bars, the left bar indicates peaks more accessible in *BCOR* mutant cells, and the right bar indicates peaks more accessible in *BCOR* WT cells. Fisher's exact test was performed to determine significance (ns: non significant).

(B) Same as Figure 6E but comparing the TAR and TARB1 clones.

(C) Left: Boxplots comparing *INPP4B* gene activity between genetic preleukemic and genetic leukemic cells. Wilcoxon rank sum test was used. The *P* value is reported above. Right: Boxplots comparing *INPP4B* gene activity between TAR, TARB1 and TARB2 leukemic clones. *P* values obtained using Kruskal-Wallis test with Dunn's test for multiple comparisons.

(**D**)-(**F)** Same as in Figure 5E but around the *MLLT10* (C)*, BTG3* (D), and *SOCS2* (E) gene loci, focusing on TSS-proximal regions.

(**G**)-(**I**) Same as in Figure 5F but relative to *MLLT10* (F), *BTG3* (G), and *SOCS2* (H) gene activity.

(J) TF footprint for the RELA motif comparing TARB2 (yellow) and TAR (purple) LMPP-like cells. Top, Tn5 bias-subtracted signal (y-axis) as a function of distance from the motif centre (x-axis). Bottom, Tn5 binding bias (y-axis) as a function of distance from the motif centre (x-axis), based on bias prediction computed by screening coverage of random hexamers within Tn5 insertion sites, using pseudo-bulk replicates. 1 indicates no bias.

(K) Volcano plot of differential TF motif accessibility between TARB1 (left) and TARB2 (right) clones across all leukemic clusters. Differentially accessible motifs (FDR < 0.05 and absolute $\Delta z > 0.20$) are labelled as red dots. Vertical dotted lines indicate Δz thresholds, the horizontal dotted line indicates the FDR threshold.

(L) Graphical depiction of the hypothesis regarding clonal dominance of *BCOR* mutant clones within the leukemic hierarchy, based on GTAC data from this patient.

Supplemental References

S1. Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J., et al. (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. Nat Immunol *19*, 85-97. https://doi.org/10.1038/s41590-017-0001-2.

S2. Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nature Biotechnology *37*, 1458-1465. https://doi.org/10.1038/s41587-019-0332-7.

S3. Popescu, D.-M., Botting, R.A., Stephenson, E., Green, K., Webb, S., Jardine, L., Calderbank, E.F., Polanski, K., Goh, I., Efremova, M., et al. (2019). Decoding human fetal liver haematopoiesis. Nature *574*, 365-371. https://doi.org/10.1038/s41586-019-1652-y.

S4. Chen, L., Kostadima, M., Martens, J.H.A., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033. https://doi.org/doi:10.1126/science.1251033.

S5. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nature Cell Biology *19*, 271-281. https://doi.org/10.1038/ncb3493.

S6. Drissen, R., Thongjuea, S., Theilgaard-Mönch, K., and Nerlov, C. (2019). Identification of two distinct pathways of human myelopoiesis. Science Immunology *4*, eaau7148. https://doi.org/doi:10.1126/sciimmunol.aau7148.

S7. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell *144*, 296-309. https://doi.org/10.1016/j.cell.2011.01.004.