Supplementary Note

Single-cell RNA sequencing plus genotyping

Data generation: Cryopreserved bone marrow aspirate (BMA) was obtained from an individual who underwent routine arthroplasty for osteoarthritis and presented a *DNMT3A* mutation (*DNMT3A*-F755S) with 0.13 variant allele frequency as measured by targeted exon sequencing. Of note, this variant affected the catalytic domain (frequently impacted in clonal hematopoiesis[86]) with a damaging mutation as predicted by both Polyphen2[87] (maximal score of 1) and FATHMM[88] (probability of 0.96), and was previously reported in hematological malignancies[89,90] . The sample was collected under specimen acquisition and molecular profiling protocols approved by the Institutional Review Board of Memorial Sloan-Kettering Cancer Center were retrieved after a database search. The individual provided informed consent. Cryopreserved BMA were thawed and stained using standard procedures (10 min, 4°C) with the surface antibody CD34-PE-Vio770 (clone AC136, Miltenyi Biotec, Bergisch Gladbach, Germany) and DAPI (Sigma-Aldrich, San Luis, MI). Cells were then sorted for DAPI-negative, $CD34^+$ cells using BD Influx at the Weill Cornell Medicine flow cytometry core.

The standard 10x Genomics Chromium v.2 protocol was carried out according to manufacturer's recommendations (10x Genomics, Pleasanton, CA) until after emulsion breakage and recovery of first strand cDNA. During the amplification step, the 10x cDNA library underwent an extra cycle of PCR beyond the manufacturer's recommended number of cycles. After cleanup with SPRISelect, a small portion of the cDNA library, 3 μL (~10% of total) was aliquoted for targeted genotyping, and the remaining cDNA underwent the standard 10x protocol. The cDNA set aside for genotyping was amplified

for 3 to 4 additional cycles using KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA) and 10x primer mix to provide sufficient material for the enrichment step. After clean-up, locus-specific reverse primers and the generic forward SI-PCR were used to amplify the site of interest of the cDNA template (**Supplementary Table 6**) using 10 PCR cycles. The locus-specific reverse primers contain a partial Illumina read 2 handle, a stagger to increase the complexity of the library for optimal sequencing and a gene specific region to allow specific priming. The SI-PCR oligo (10x) anneals to the partial Illumina read 1 sequence at the 3' end of the molecule, preserving the cell barcode (CB) and UMI. After the initial amplification and SPRI purification to remove unincorporated primers, a second PCR was performed with a generic forward PCR primer (P5_generic) to retain the CB and UMI together with an RPI-x primer (Illumina, San Diego, CA) to complete the P7 end of the library and add a sample index. The targeted amplicon library was sequenced separately on MiSeq with v2 chemistry (Illumina, San Diego, CA). The cycle settings were as follows: 26 cycles for read 1, 98 cycles for read 2, and 8 cycles for sample index. Mutation calling was performed as previously described[47]. Briefly, to ensure correct priming, targeted amplicon reads (read 2) were screened for the presence of the primer sequence and the expected intervening sequence between the primer and the start of the mutation site. Subsequently, for reads that passed the priming step, the corresponding read 1 was screened for the presence of the 16 bp long cell barcode that matched the whitelist provided by 10x Genomics. For cell barcode reads that were 1-Hamming-distance away from the whitelisted cell barcode, the probability of that the observed barcode originated from the whitelisted cell barcode was calculated taking into account the base quality score at the differing base. The

whitelisted cell barcode with the highest probability was used to replace the observed cell barcode, only if the probability exceeded 0.99. For the duplicate reads with the same cell barcode and UMI, the genotype (WT vs. mutant) of the read with the highest base quality was assigned for the given UMI.

ATAC-Bseq protocol and data analysis

Data generation: Tn5 transposase (Lucigen, Middleton, WI) was loaded with modified adaptors as described in Qui Wang *et al*[91] (**Supplementary Table 6**). For performing ATAC-Bseq, 100,000 bone marrow LSK cells were purified by flow cytometry and then centrifuged for 5 minutes at 500 rcf at 4°C. Cells were washed in 1 ml of cold PBS and centrifuged at 500 rcf for 5 minutes at 4°C. Then, cells were then resuspended in lysis buffer (10mM Tris-HCl, 10mM $NaCl_2$, 3mM $MgCl_2$, 0.1% Igepal CA-630) and left on ice for 5 minutes. After incubation, cells were centrifuged at 500 rcf for 10 minutes and resuspended in tagmentation buffer (10mM TAPS-NaOH pH = 8.5, 10% di-methylformamide, 4mM $MgCl_2$). After adding Tn5 transposase loaded with methylated adaptors[91] (**Supplementary Table 6**), the tagmentation reaction was carried out for 30 minutes at 37°C in a shaker set at 300 rpm. Immediately after tagmentation, DNA was purified using (DNA clean and concentrator, cat# D4014, Zymo Research, Irvine, CA), and oligo displacement and gap repair were performed[91]. Next, DNA was subjected to bisulfite conversion (EZ methylation kit, cat #E5002, Zymo Research, Irvine, CA) and libraries were amplified using barcoded standard Illumina-compatible primers[91] with custom barcodes (**Supplementary Table 6**). Libraries were pair-end sequenced at 50 base pairs per read, spiking in 20% PhiX for increased library complexity.

ATAC-Bseq analysis pipeline: ATAC-Bseq data was aligned to mm10 genome and methylation calling was performed through Bismark[83] (version 0.14.4; parameters: --q --score-min L,0,-0.2 -p 4 --reorder --ignore-quals --no-mixed --no-discordant --maxins 500 Option '--directional' specified) running on bowtie2-2.2.8 aligner[84]. Accessibility peaks for each individual replicate were defined using MACS2[58] (v2.1.1.2); parameters: -f BAMPE --broad --broad-cutoff 0.1 -m 2 100). In order to retain high confidence calls, we took the intersection between the peaks called for each replicate within condition (e.g., WT, *Tet2* KO or *Dnmt3a* KO). The union of the peaks from each condition was used in downstream analysis. Peaks with $\log_{10}(\text{RPKM}) < 1$ were filtered out for replicate consistency. Differential accessibility analysis was performed using DEseq2[59] (version 1.18.1; parameters: minOverlap = 2, fragmentsize = 50, filter = 10, summits = TRUE, bCorPlot = TRUE) and differential methylation was performed using MethylKit[60] (version 1.4.1). After coverage normalization using the normalizeCoverage function with default parameters, we applied the calculateDiffMeth function with default parameters. *DNA binding motif analysis:* For motif analysis, position weight matrices (PWMs) were downloaded from the tf2dna[44] database. Motif sites in the genome were defined by using the scanMotifGenomeWide function from Homer[45] (v4.10.4). For position weight matrices obtained from tf2dna database, the initial Log odds detection threshold used to determine bound vs. unbound site was 5, and motif sites within the lower quartile were filtered out. For *de novo* motif finding, ATAC-Bseq peaks were annotated and defined as erythroid- or myelo-monocytic-associated peaks by intersecting the nearest gene with the list of differentially expressed genes (FDR<0.05) between WT clusters Mono-1 and Ery 1-3. Only accessible peaks within 10 kb of the closest annotated transcriptional start site

were kept for motif enrichment analysis. Peaks not associated with either erythroid or myelo-monocytic fates were used as background. Motif enrichment analysis was performed using the function findMotifsGenome from Homer (v4.10.4, parameters: -bg <custom> -len 12 -size 200 -noweight -mask). Of note, we further observed that CpG rich motifs at accessible enhancers were preferentially susceptible to DNAme changes, resulting in an increase, not only in the *number* of affected CpGs, but also in the *proportion* of affected CpGs (**Extended Data Figure 9f**). Such pattern is consistent with the known proclivity of Tet2 to bind to CpG rich genomic regions, through co-binding with the CXXC domain on IDAX gene[92]. Although bisulfite sequencing is unable to distinguish between 5mC and 5hmC, it has been shown that *Tet2* deletion results in reduction of 5hmC and accumulation of 5mC at enhancer sites[93]. Thus, our data may even underestimate the 5mC increase in CpG rich motifs in *Tet2* KO.

Exon Sequencing and Analysis. Targeted next generation sequencing of 585 genes (**Supplementary Table 1**) associated with malignancies was conducted by the Memorial Sloan Kettering integrated mutation profiling of actionable cancer targets MSK-IMPACT Heme platform[94]. Mutational calls were made by alignment to common unmatched normal samples. For mouse samples, mean target coverage was 600X. For the human clonal hematopoiesis sample mean target coverage was 824X, with 246X coverage for the DNMT3A-F755S mutation.

Single nucleus ATAC sequencing

Data generation and sequencing: Hematopoietic progenitors (Lin^-, $Sca1^+$, $c\text{-}Kit^+$) were sorted from WT (n = 2; 5,810 cells), *Tet2* KO (n = 2 mice; 7,214 cells) or *Dnmt3a* KO (n = 2 mice; 7,005 cells). Nuclei isolation was performed as suggested by the manufacturer (10x Genomics, Pleasanton, CA). Briefly, the cell suspension was centrifuged at 300 rcf for 5 minutes and the cell pellet was resuspended in 100 µl of lysis buffer (Tris-HCl pH 7.4, 10mM; NaCl 10mM; $MgCl_2$ 3mM; Tween-20 0.1%; Nonidet P40 substitute [Sigma #74385] 0.1%; Digitonin 0.01% and BSA 1%) and kept in ice for 5 minutes. Then, 1 ml of wash buffer (Tris-HCl pH 7.4, 10mM; NaCl 10mM; $MgCl_2$ 3mM; BSA 1% and Tween-20 0.1%) was added. The lysate was centrifuged for 5 min at 500 rcf, and the pellet was resuspended in Diluted Nuclei Buffer (10x Genomics, PN-2000153/2000207). Nuclei concentration was determined by hemocytometer, and processed as indicated by the manufacturer (10x Genomics, CG000209). Libraries were pair-end sequenced at a depth of 5,000 reads per nucleus. Nuclei barcodes with < 1,000 and > 60,000 fragments were filtered out to remove low-quality nuclei and potential doublets, and barcodes with higher than 20% of the fragments mapping to the mitochondrial genome were filtered out. We obtained a total of 20,029 nuclei (WT n = 2 mice, 5,810 nuclei; *Tet2* KO n = 2 mice, 7,214 nuclei and *Dnmt3a* KO n = 2 mice, 7,005 nuclei) for downstream analysis. Clustering was performed using cisTopic (v0.2.1), selecting a model containing 50 topics. We next generated a distance matrix by first calculating the z-score on the cell-topic distribution matrix followed by correlation distance matrix calculated as 1-Pearson correlation. We then generated communities by performing K-nearest neighbors (KNN; k = 400) on the distance matrix, followed by Louvain clustering using the iGraph (v1.2.4.1)

R package. For visualization, we performed UMAP dimensionality reduction on the cell-topic distribution matrix, retaining 50 topics. In order to annotate the defined clusters to known cell types, we utilized the ImmGen[61] bulk ATAC-seq profiles as reference, and included the megakaryocyte and erythroid profiles from ENCODE[95] (identifiers: ENCFF066SZX and ENCSR136XSY, respectively).

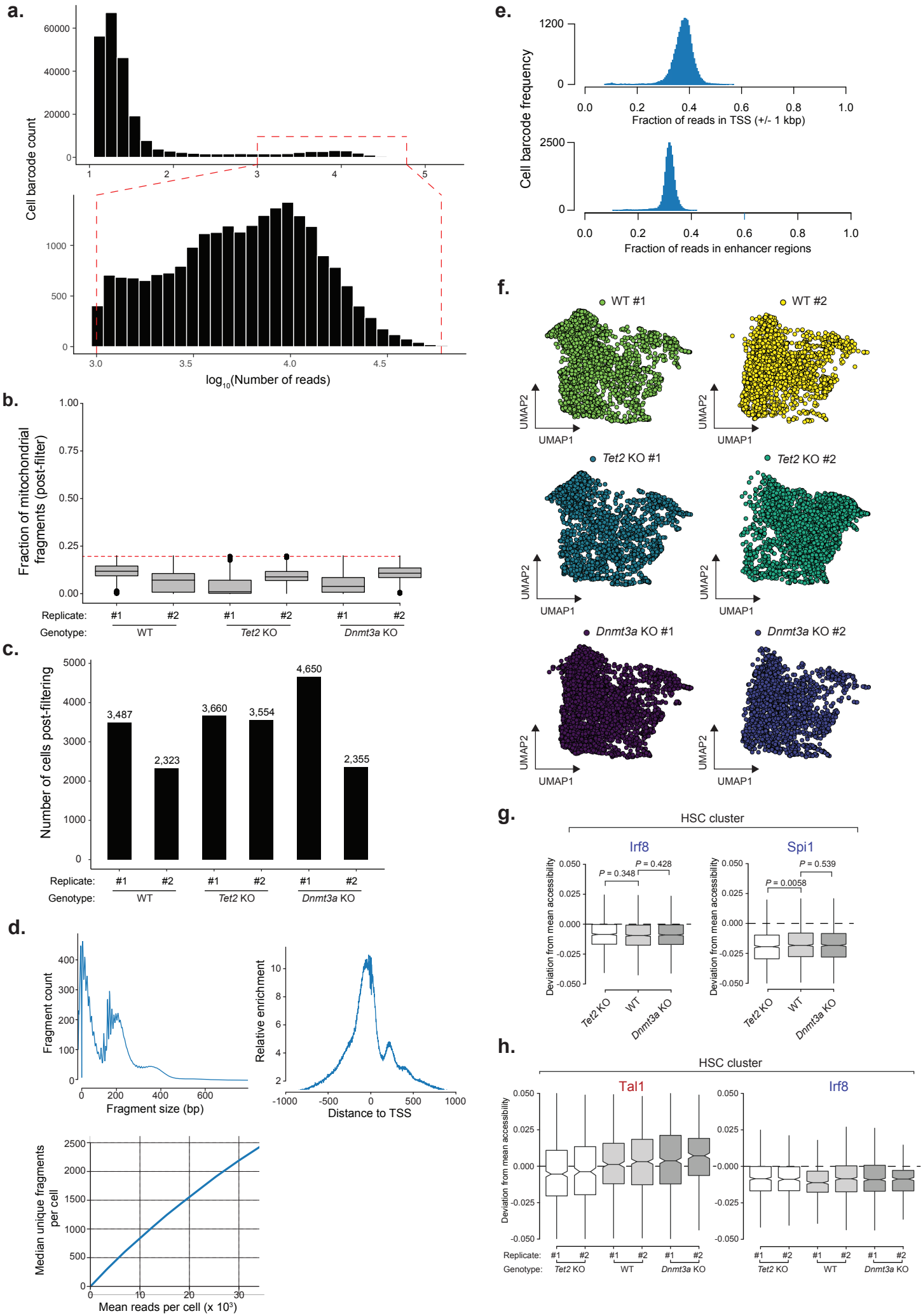
AUC Score Calculation: We scored each single nucleus as follows: we used the AUCell (v1.6.1) R package to calculate the enrichment of epigenetic signatures in the cells. We then multiplied the topic and cell assignments matrices to obtain the likelihood of each peak in each single cell. These distributions were used to determine the importance of a peak for a cell and to create the rankings used by AUCell. A GSEA-like recovery curve ranking-based approach is used to determine the significance of a signature in a cell. More information on the method is available in the cisTopic vignettes.

Pre-processing of the signature dataset: The regions in the ImmGen dataset were expanded 250 bp before and after the summit. Overlapping peaks were then filtered, keeping only the ones with the lowest p-values. Columns were merged to have broader cell types definition. Then, for each cell type, the regions were sorted by their scores and bed files were generated by taking only the top 60,000 regions. The same approach was used on the two ENCODE bed files. The resulting bed files were used as the signature for AUCell.

Motif accessibility and de novo motif enrichment: To identify the changes in transcription factor motif accessibility, we applied chromVar (v1.4.1) using the HOCOMOCO v11 mouse motif position weight matrix collection in Homer format with $P < 0.001$. In order to calculate the *de novo* motif enrichment in the HSCs cluster (n = 3,610 cells), we

generated pseudo-bulk profiles by genotype. Peak calling was then run on the *Dnmt3a* and *Tet2* KO profiles using MACS2 (v2.1.1.2) with parameters —broad --broad-cutoff 0.1 -m 2 100). Using the 137,963 peaks previously called by cellRanger-ATAC as background, we ran Homer (v4.10.4) for *de novo* motif enrichment (-size 200 -mask as parameters) on the bed files generated by MACS2 (v2.1.1.2). 27 de novo motifs were found for both *Dnmt3a* and *Tet2* KO profiles ($P < 0.05$). The 54 motifs were then ranked by their mean CpG frequency per base and split in quartiles.

Supplementary Figure 1



Supplementary Figure 1. ***Single nuclei ATAC sequencing (snATAC-seq) quality control.*** **a**) Distribution of reads per cell barcode across snATAC-seq samples. Cell barcodes containing between 1,000 and 60,000 total reads (dashed red line, n = 22,086 cells) were retained for downstream analysis. **b**) Fraction of reads mapping to the mitochondrial genome per sample (n = 6 biologically independent samples) after barcodes containing > 20% of reads mapping to the mitochondrial chromosome were removed. The red dashed line indicates the 20% cutoff threshold.**c**) Remaining number of cell barcodes per sample after filtering. **d**) Top left panel: Fragment length distribution obtained in the snATAC-seq experiment (representative example of n = 6 biologically independent samples). Top right panel: Representative example of the distribution of mapped reads +/- 1kbp from the transcriptional start sites (TSS). Bottom panel: Representative example of library saturation. The mean reads per cell relative to the median unique fragments per cell are shown. **e**) Distribution of cell barcodes according to the fraction of reads mapped to TSS (+/- 1kbp) or enhancer regions. **f**) Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction per sample for the snATAC-seq data for WT (n = 2 biologically independent animals), *Tet2* KO (n = 2 biologically independent animals) and *Dnmt3a* KO (n = 2 biologically independent animals), as obtained by cisTopic[68] (see **online methods**). **g**) Motif accessibility deviation scores comparison between *Tet2* KO (n = 1,173 cells), WT (n = 1,410 cells) and *Dnmt3a* KO (n = 1,305 cells) cells mapped to the HSC cluster (Wilcoxon rank sum test). **h**) Examples of motif accessibility deviation scores per replicate for *Tet2* KO (1,173 cells), WT (1,410 cells) and *Dnmt3a* KO (1,305 cells) cells mapped to the HSC cluster (two-sided Wilcoxon rank sum test).