# Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA

Jin Xu[1,2,3], Kevin Nuno[4,5], Ulrike M. Litzenburger[1,2,3], Yanyan Qi[1,2,3], M Ryan Corces[1,2,3], Ravindra Majeti[4,5] & Howard Y. Chang[1,2,3]

[1]Center for Personal Dynamic Regulomes, Stanford, CA 94305, USA.

[2]Dept. of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA.

[3]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA.

[4]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, California, USA.

[5]Division of Hematology, Department of Medicine, Stanford University School of Medicine, Stanford, California, USA.

Correspondence to: H.Y.C. (howchang@stanford.edu) and R.M (rmajeti@stanford.edu).

## Abstract

Simultaneous measurement of cell lineage and cell fates is a longstanding goal in biomedicine. Here we describe EMBLEM, a strategy to track cell lineage using endogenous mitochondrial DNA variants in ATAC-seq data. We show that somatic mutations in mitochondrial DNA can reconstruct cell lineage relationships at single cell resolution with high sensitivity and specificity. Using EMBLEM, we define the genetic and epigenomic clonal evolution of hematopoietic stem cells and their progenies in patients with acute myeloid leukemia. EMBLEM extends lineage tracing to any eukaryotic organism without genetic engineering.

## Introduction

Resolving lineage relationships between cells is necessary to understand the fundamental mechanisms underlying normal development and the progression of disease. In recent years, new methods have emerged to enable cell lineage tracking with increasing resolution, leading to substantial biological insights[1]. Specifically, genome editing of reporter constructs via CRISPR-Cas9 allowed synthetic reconstruction of cell

lineage relationships in model organisms, and has been coupled with transcriptome profiling to inform cell fates[2]. These prospective “mutate-and-record” methods provide powerful tools to resolve the developmental origin of cells in genetically engineered cells and organisms, but cannot be utilized in living humans, archival clinical samples, or any wild type organism[1]. Given these limitations, retrospective lineage tracing using endogenous genetic markers is an alternative solution. Recent advances in sequencing enable naturally occurring somatic mutations to be used as lineage markers, which usually required single-cell genome sequencing to capture the sparse genetic information[3,4]. Regions with high mutation rates, such as microsatellite repeats, retrotransposons, and copy-number variants, has been used to resolve the lineage relationship for normal or cancerous tissue samples[5,6]. These methods reduce the cost of whole genome sequencing, but still lack information on cell phenotypes.

Simultaneous measurement of the lineage relationship and cell fates is ultimately required to address many biomedical questions. Here we describe EMBLEM (Epigenome and Mitochondrial Barcode of Lineage from Endogenous Mutations), a strategy to track cell lineage using endogenous mitochondrial DNA variants in ATAC-seq data. The end result of EMBLEM is single-cell lineage information and rich global epigenomic profile from the same individual cells (Figure 1A and Figure 1-figure supplement 1).

We illustrate the utility of EMBLEM in human blood progenitor cells to clarify the process of pre-leukemic clonal evolution and the emerging biology of clonal hematopoiesis.

## Results

Assay of Transposase-Accessible Chromatin by sequencing (ATAC-seq) is a sensitive method used to study chromatin accessibility profiles in diverse cell types and organisms[7]. During DNA transposition and amplification in cells, mitochondrial DNA is also amplified at the same time (Figure 1A). Mitochondrial DNA (mtDNA) is a ~16kb circular genome with ~10-fold higher mutation rate compared to the nuclear genome. Hence, mtDNA incrementally accumulates unique, irreversible genetic mutations that are passed on to daughter cells even in healthy humans and may be used for lineage tracing[8,9]. The majority of somatic mtDNA mutations are noncoding and thought to be passenger[10]. Importantly, the number of mitochondria (and therefore mtDNA) range from

several hundreds to >10,000 per cell in different cell types, facilitating robust mtDNA analysis even from a single cell.

We first observed that ATAC-seq effectively enriches for mtDNA. While mtDNA is present in many kinds of DNA sequence libraries, it is substantially enriched in ATAC-seq libraries due to the fact that mtDNA is not chromatinized and is therefore highly accessible (Supplementary file 1). ATAC-seq enables a 17-fold or greater enrichment of mtDNA compared to exome sequencing or whole genome sequencing in GM12878 human B cells (Figure 1B), leading to an average ~18,000X coverage of mtDNA (Figure1--figure supplements 2A). With this coverage, we detected 27 mitochondrial variants from GM12878 cells (Figure 1C). 13 of these variants have a variant allele frequency (VAF) greater than 90%, which are known as homoplasmic variants (Figure 1--figure supplements 2B). We also detected 14 low frequency mitochondrial DNA variants, with VAFs ranging from 0.1% to 24% (Figure 1C and Figure 1-figure supplement 2C). Similar results for mtDNA enrichment were observed in human K562 cells (Figure 1-figure supplement 3, Supplementary file 1)

The VAF from bulk ATAC-seq data represents the average of the allele frequencies of the cell population. A 25% VAF may arise from 25% of cells in the population with a homoplasmic variant, or alternatively arise from 100% of cells all having a quarter of their mitochondria with the variant allele (Figure 1D). To distinguish between these two models, we analyzed single-cell ATAC-seq data from GM12878**.** For 4 mtDNA variants (VAF between 0.5%~24% at population level), we find that a mixture of both models is in action for different variants (Figure 1E). For instance, mtDNA mutation 3082 is widely spread among single cells, but at low frequency per cell. Because it is extremely unlikely (see METHOD) that the identical mutation arose independently in every single cell, cells sharing the same mitochondrial mutations are inferred to have descended from the same ancestral cell. These results suggest that even low frequency heteroplasmic mtDNA mutations can be exploited for lineage tracing.

To prove the principle that somatic mitochondrial mutations can track cells from the same ancestor and to quantify the power of lineage mapping, we next applied EMBLEM to primary blood cells from patients with acute myeloid leukemia (AML). Human AML is organized as a hierarchy: a hematopoietic stem cell first acquires an initiating mutation in one of a number of chromatin modifier genes, previously termed as "pre-leukemic" hematopoietic stem cell (pHSC)[11,13]. pHSCs are functionally normal and are not able to transplant AML, but upon accumulation of additional mutations, they give rise to leukemic stem cells (LSCs) that are able to self-renew and recapitulate AML disease upon transplantation[11,12]. Finally, LSCs give rise to the bulk leukemic blast cells in AML[12]. Targeted exome sequencing in these samples have identified somatic mutations in tumor suppressor genes and oncogenes that link the lineage relationship of pHSCs, LSCs and blasts, providing the ground truth for our analyses[13].

We applied EMBLEM to the ATAC-seq profiles of FACS-purified LSCs and leukemic blasts first (Supplementary file 1). Using high-confidence mtDNA mutations, detected both from bulk ATAC-seq and single cell ATAC-seq, we found the LSC and blast populations not only shared the same heteroplasmic variants, but also showed similar distribution and allele frequency at the cellular level (Figure 1-figure supplement 4). These results indicate the two populations are identical at the genetic level, but divergent at the epigenomic level, consistent with previous studies[14,15]. In patient SU353, we identified four diagnostic mtDNA mutations in the same cell **(**Figure 1F**)**, which indicates these four mitochondrial variants already co-existed in the ancestral cell (see METHOD). With the assumption that all these LSCs and blasts are clonal, we further quantified the detection rate of each mtDNA variant as a function of allele frequency and sequencing depth (Figure 1G). We found that when a single variant allele has a frequency greater than 20%, the detection rate can be up to 90% with >20X coverage (e.g. site 6776). In contrast, when the variant allele has a frequency lower than 1%, the detection rate drops to 20% when the coverage is below 100X (e.g. site 6705). While high drop-out rate is a common challenge for single-cell technologies[16], computational imputation of the missing information from single cell data can address this problem[17]. When multiple mtDNA variants are co-detected in multiple single cells, we can infer their origin and

linkage in the ancestral cell (see METHOD). Thus, cells containing any one of these variants will still inform their origin from the same lineage. With any combination of the four variants, 90% (sensitivity) of the cells can be unambiguously assigned to the correct lineage with just 20x mtDNA coverage (Figure 1H). Furthermore, two mtDNA mutations identified in other cells (e.g. pHSC specific site 2967,6268) were never detected (false positive=0) in LSCs and blasts (Figure 1-figure supplement 4), showing a high specificity of the method. Similar performance of single cell lineage tracing for another patient (SU070) are shown in Figure 1-figure supplement 5. These results demonstrate that somatic DNA mutations in the mitochondrial genome are a powerful endogenous marker to identify clonal cell populations.

To expand on these findings to additional different cell lineages, we applied EMBLEM to bulk ATAC-seq data from sorted blood cells from healthy human donors and patients with AML(Supplementary file 1)[14]. We identified heteroplasmic mtDNA mutations in multiple cell populations of primary blood cells from healthy donors and all AML patients (Figure 2-figure supplement 1A, Supplementary file 2). The heteroplasmic mtDNA mutations showed a similar mutant spectrum as observed by previous studies using cancer genomic data **(**Figure 2 -figure supplement 1B and C)[10].

Furthermore, EMBLEM, not only confirmed the previous lineage hierarchy of AML, but also extended the previous model of pHSC heterogeneity (Figure 2A**)**. In the AML cases with LSCs sequenced by ATAC-seq, the LSCs and their corresponding leukemic blasts have nearly identical heteroplasmic mtDNA mutations (Figure 2B-C and Figure 2-figure supplements 1D), suggesting a direct lineage relationship and short generation history between LSCs and blasts. We then examined whether any of the mtDNA variants present in LSCs can be seen in the pHSCs, where the first leukemia-associated protein-coding mutations have already occurred in functional normal hematopoietic stem cells[13,14]. We detected blast-associated mtDNA mutations in pHSCs in all 11 cases. Interestingly, we also detected additional heteroplasmic mtDNA mutations present specifically in pHSCs (Figure 2C). In the 11 cases we investigated, 7 cases have pHSC-unique heteroplasmic mtDNA mutations (Figure 2-figure supplement 1D and E), a previously

unrecognized level of pHSC heterogeneity. pHSCs are capable of long-term self-renewal and possess a clonal growth advantage, allowing them to clonally outcompete normal HSCs. Indeed, the clonal frequency of pHSCs is a poor prognostic factor for overall survival in AML[14]. Our discovery of pHSCs with distinct heteroplasmic mtDNA mutations suggests the existence of multiple distinct sub-clones of pHSCs in AML patients.

To validate the heterogeneity of pHSCs inferred from EMBLEM of bulk cell populations, we performed single-cell ATAC-seq of HSCs from AML patient SU353, which exhibited both a high burden of pre-leukemic somatic coding gene mutations and high frequency of pHSC-specific heteroplasmic mtDNA mutations[14]. We identified the heteroplasmic mtDNA variants from each single cell, which separated the HSCs into three lineages: Two clonal subpopulations termed “clone 1” (18 cells) and “clone 2” (104 cells), and a third population with no mtDNA variants despite sufficient mtDNA coverage (pHSC with WT mtDNA, 31 cells) (Figure 2D). Notably, clone 2 possessed pHSC-specific mtDNA mutations, while clone 1 possessed mtDNA mutations shared with LSCs, indicating clone1 is the lineage precursor of AML. These results confirm that multiple pHSC clones arise in AML patients, and one subclone eventually evolved to become the LSC (Figure 2-figure supplement 2A ).

Finally, we related the clonotype of pHSCs to their single-cell chromatin accessibility profiles. We interrogated the patterns of active DNA elements and enriched transcription factor motifs in sequential stages of AML development from the same patient, and contrasted with HSCs from normal donors using ChromVAR[15] (Figure 2E and Figure 2-figure supplement 2B). The chromatin accessibility profiles of pHSCs are more similar to HSCs than to LSCs or leukemic blasts. The greatest deviation between HSC and other cell types occurred at DNA binding motifs of the transcription factor Jun/Fos, a known key regulator of HSC biology[18] (Figure 2F). Furthermore, the three lineages of pHSCs revealed by mtDNA mutations also showed distinctive chromatin profiles (Figure 2G). Clone 1 pHSC, which gives rise to the LSC and AML leukemia, is already more similar to LSCs and blasts in its chromatin accessibility. In contrast, clone 2 that comprises the larger fraction of pHSCs exhibited variable chromatin profiles at the single-cell level that

spanned the range of normal HSCs, pHSC with WT mtDNA(WT cells), is also diverged from normal HSCs. Thus, both lineage tracing and single cell epigenomic states indicate clone 1 as the original stem cell of the AML in patient SU353. Supervised comparison of the chromatin accessibility profiles among these clonal sub-populations further identified distinct and significantly enriched transcription factor motifs (Figure 2H and Figure 2-figure supplement 2C-E). These results indicate the heterogeneity of HSCs from AML patients both on a genetic and epigenomic level.

## Discussion

We present a computational strategy to combine cell lineages tracing by endogenous mtDNA mutations and chromatin accessibility profiling in the same cell using single-cell ATAC-seq data. This approach is applicable to any eukaryote, does not require genetic engineering or genome editing, and is cost effective as the lineage information comes “for free” on top of epigenomic insights. The relative merits of mtDNA vs other genetic markers for lineage tracing are outlined in Supplementary file 3. An important advantage of EMBLEM is that we enable clonotype tracing in existing ATAC-seq data sets and hierarchical lineage construction from ATAC-seq that thousands of labs have already generated. All future ATAC-seq data acquired for other inquiries will also have the benefit of lineage information. EMBLEM may also be extended to other single cell technologies, in which mtDNA is sequenced. We show that EMBLEM is successful even with low frequency heteroplasmic mutations, detection of rare clones in a population, and authentic clinical samples. With advances in the throughput and depth of single-cell genomic technologies, we believe EMBLEM may be a powerful tool to bring insight for many biomedical questions, including development, regeneration, immunity, and cancer with integration of genotype and phenotype information from the same cell. During revision of this work, Ludwig et al. reported the feasibility of using mtDNA and single cell genomics for lineage tracing, which independently validates the potentially broad utility of this approach.[19]

Although powerful and broadly applicable, mtDNA lineage tracing also has its limitations. One limitation of this method is absence of mtDNA mutations in cells and

tissues of embryos and young animals, which precluded us from applying EMBLEM to published scATAC-seq data of early animal development. Moreover, the possibility of selective mitochondrial inheritance or intercellular mitochondria transfer may affect the accuracy of inferred lineages[20–23]. On the other hand, asymmetric transmission of mitochondria would not necessarily affect cellular lineage tracing, as long as the variant alleles are randomly segregated. Using scATAC-seq data from a mixing experiment with human and mouse cells[24], we found species-specific mtDNA always paired with species-specific nuclear genomic DNA (Figure 2-figure supplement 4 ). These results suggest that mitochondrial horizontal transfer is not a confounder of our study and does not universally occur between cells. The aforementioned two scenarios reflect the potential uncoupling of nuclear and mitochondrial genomes, which would be of interest to investigate by EMBLEM in combination with other gDNA tracing methods.

mtDNA lineage tracing produced new insights concerning the pHSC, the human hematopoietic stem cell that suffers the first oncogenic mutation in AML evolution. Our results add to the evidence that the pHSC population is heterogeneous, with evidence of multiple mtDNA clones. Unexpectedly, the pHSC lineage that gives rise to the subsequent acute myeloid leukemia is not the lineage with the best competitive potential among pHSCs, as the leukemogenic lineage is often in the minority. pHSC burden is a strong poor prognostic predictor of AML survival[14]. It is widely believed that the association between high pHSC burden and poor AML patient prognosis reflected the enhanced self-renewal and competitive ability of the mutant pHSC. Our analysis suggests that high pHSC burden may reflect the diversity of pHSCs or the underlying mutational processes. These alternative interpretations of the link between pHSC burden and poor clinical prognosis should be addressed in future studies.

## Material and Methods

**Public data accession.** Aligned bam files for GM12878 whole exome, low coverage whole genome, and PCR free whole genome sequence, were downloaded through phase 3 release of 1000 genomes (ftp://ftp.1000genomes.ebi.ac.uk)

The alignment files were accessed via the following ftp links:

- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878/exome_alignment/NA12878.mapped.ILLUMINA.bwa.CEU.exome.20121211.bam ./
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878/alignment/NA12878.mapped.ILLUMINA.bwa.CEU.low_coverage.20121211.bam ./
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878/high_coverage_alignment/NA12878.mapped.ILLUMINA.bwa.CEU.high_coverage_pcr_free.20130906.bam ./

ATAC-seq and single cell ATAC-seq data for GM12878 generated by Buenrostro et al. were downloaded through GEO with accession number GSE47753 and GSE65360, respectively[7,25]. Bulk ATAC-seq data from normal donors and AML patients generated by Corces et.al[14]., were downloaded through GEO with accession number GSE74912. Single cell ATAC-seq data for leukemia stem cell and leukemic blasts generated by the same study were downloaded through GEO with accession number GSE74310. Single cell ATAC-seq from normal HSC generated by Buenrostro et al., were downloaded through GEO with accession number GSE96772[26]. Supplementary file 1 summarized the detail information of all the datasets used in this study.

**Comparison of mitochondrial genome capture rate and coverage.** Sequencing reads from ATAC-seq were aligned to the reference genome by BWA alignment tool[27]. The same reference, GRCh37(used by 1000 genome) and human reference mtDNA sequence rCRS (revised Cambridge reference sequence), were used for ATAC-seq data processing. Samtools[28] was used for manipulating sequence reads and calculating sequence depth. For all the data sets, the aligned reads were further filtered with mapping quality (Q >30) and PCR redundancy was removed. The percentage of reads from mitochondrial genome compared to that of the nuclear genome were calculated after all

the clean-up steps. The mitochondrial genome coverage was calculated using bases with sufficient sequence quality score (q > 30). A strong depletion region around 3107 due to the sequencing error(3170N) in the reference genome was excluded in the coverage plot[10].

**Bulk ATAC-seq data process and mitochondrial DNA variants calling.** Most of the ATAC-seq pipelines remove mtDNA during their process. To rescue the genetic information from mtDNA, we modified our ATAC-seq pipeline and added SNP calling steps, which focuses on the mitochondrial genome. Briefly, adaptor sequences were trimmed from FASTQs using custom Python scripts. Paired-end reads were aligned to the reference genome using BWA. To improve the accuracy of heteroplasmic mutation calling, we followed the somatic mutation calling guidelines from GATK[29], with additional clean-up steps before variant calling. Reads mapped to mtDNA were extracted using Samtools[28] from the final bam files and variants were called using VarScan2[30] with "--min-var-freq 0.001" (Figure 1--figure supplements 1A). The heteroplasmic variants were further filtered through the following steps to exclude potential sequencing or mapping errors:

1. Thirteen frequent false-positive variants by misalignment due to extensive level of homopolymers in rCRS and due to sequencing error in the reference genome(reported in the previous study[10]), were also observed and removed in this study. The following sites were explicitly removed:

Misalignment due to ACCCCCCCTCCCCC (rCRS 302-315)

A302C, C309T, C311T, C312T, C313T, G316C

Misalignment due to GCACACACACACC (rCRS 513-525)

C514A, A515G, A523C, C524G

Misalignment due to 3107N in rCRS (ACNTT, rCRS 3105-3109)

C3106A, T3109C, C3110A

2. Strand imbalance is a potential feature of sequencing error with various causes. To remove the potential sequence error from Illumina NextSeq (with a known high error rate at A bases) and sequence error from DAN damage(G->T, C->A)[31], we required > 2 reads

detected from both the forward and reverse orientation, and strand is balanced (30%<forward/(forward + reverse)<70%).

3. Variant sites with VAF>0.9, but less than 1, were counted as homoplasmic variants. Although the germline polymorphic can be a back heteroplasmic mutations, the observation of these events is higher than expected, which implies the false positive calling due to mapping bias for non-reference allele and sequencing errors.

4. For bulk ATAC-seq data from AML patients, heteroplasmic mutations with variant allele frequency >1% were reported.

For all the AML cases(n=15) from Corces et.al[14], we selected the cases( n=12) with at least one confident heteroplasmic mtDNA mutation detected in any cell type for lineage relationship comparison. We found that in one patient (SU209), the number of heteroplasmic mutations (37) and their VAF are significantly higher than other patients. Most of these heteroplasmic mutations also overlapped with common variants present in the general human population (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chrMT.phase3_callmom-v0_4.20130502.genotypes.vcf.gz), which indicates potential sample contamination. Therefore, this case was excluded from lineage relationship comparison and 11 AML cases were finally shown in Figure 2--figure supplements 1.

**Single cell ATAC library resequencing.** To better evaluate the detection rate in single cell ATAC-seq data, we re-sequenced the previous libraries(LSCs and AML blasts from SU070 and SU373) from Corces et.al[14]. The re-sequenced data were uploaded to GEO and accession number is GSE122576.

**Human AML samples** Human AML samples were obtained from patients at the Stanford Medical Center with informed consent, according to institutional review board (IRB)-approved protocols (Stanford IRB, 18329 and 6453). Mononuclear cells from each sample were isolated by Ficoll separation, resuspended in 90% FBS + 10% DMSO, and cryopreserved in liquid nitrogen. All analyses conducted here on AML cells used freshly thawed cells.

**Cell Sorting.** Cell samples were first thawed and incubated at 37°C with 200 U/mL DNase in IMDM + 10% FBS. To enrich for CD34+ cells, magnetic bead separation was performed using MACS beads (Miltenyi Biotech) according to the manufacturer's protocol. For cell staining and sorting, the following antibody cocktail was used with the schema shown in Figure 2-- figure supplements 3

CD34-APC, clone 581, Biolegend, at 1:50 dilution.

CD38-PE-Cy7, clone HB7, Biolegend, 1:25 dilution.

CD19-PE-Cy5, clone H1B9, BD Biosciences, 1:50 dilution

CD20-PE-Cy5, clone 2H7, BD Biosciences, 1:50 dilution

CD3-APC-Cy7, clone SK7, BD Biosciences, 1:25 dilution

CD99-FITC, clone TU12, BD Biosciences, 1:20 dilution

TIM3-PE, clone 344823, R&D Systems, 1:20 dilution

CD45-KromeOrange, clone J.33, Beckman Coulter at 1:25 dilution

Samples were sorted using a Becton Dickinson FACS Aria II. pHSCs were re-suspended and kept in cold FACS buffer containing 1 ug/mL propidium iodide prior to and after sorting. Cells were then immediately prepared for single cell ATAC-seq.

**Single cell ATAC-seq from pHSC.** Cells were washed 2 times in C1 DNA Seq Cell Wash Buffer (Fluidigm). ~10K cells were then re-suspended in 6 mL of C1 DNA Seq Cell Wash Buffer, and were combined with 4 mL of C1 Cell Suspension Reagent, 7 mL of this cell mix was loaded onto the Fluidigm IFC. Cells at a concentration of 260-380 cells/µL were then assayed using scATAC-seq as previously described[25]. Briefly, single cells were captured using the C1 Single-Cell Auto Prep IFC microfluidic chips. Cells were permeabilized and accessible fragments were captured using 20 µL of Tn5 transposition mix (1.5x TD buffer, 1.5 µL transposease (Nextera DNA Sample Prep Kit, Illumina), 1x C1 Loading Reagent with low salt (Fluidigm), and 0.15% NP40) at 30 minutes at 37°C. In a 96-well plate, 7 µL of harvested libraries were amplified in 50 µL PCR for an additional 17 cycles (1.25 µM custom Nextera dual-index PCR primers in 1x NEBnext High-Fidelity PCR Master Mix using the following PCR conditions: 72°C for 5min; 98°C for 30 s;) using the following PCR conditions: 72°C for 5min; 98°C for 30 s; and thermocycling at 98°C

for 10 s, 72°C for 30 s, and 72°C for 1 min. The PCR products were pooled creating a final volume of ~4.8 mL. The pooled library was purified on a single MinElute PCR purification column (Qiagen). Libraries were quantified using qPCR prior to sequencing. The scATAC-seq libraries were sequenced by Illumina MiSeq. The sequence data was uploaded to GEO under the accession number GSE122577.

**Single cell ATAC-seq data processing and mitochondrial DNA variant calling.** Single cell ATAC-seq were processed similarly to the bulk ATAC-seq, taking each individual cell as one sample. Recalibration steps were not applied for single cell data, as the sequence depth is not sufficient to empirically adjust the quality scores. After cleaning the alignment, files from every single cell were merged and heteroplasmic variants were first called with the merged bam and filtered using the same criteria as bulk data. Heteroplasmic variants called from merged data or from bulk data were re-counted in each individual cell using Samtools with "-q 20 -Q 20". And the non-reference allele had to match the variants detected in merged or bulk data.

**Detection rate estimation.** In every single cell, if the variant allele detected in merged or bulk data were supported by any reads, it was considered positive; otherwise, it was counted as zero. A binary matrix was used to present the lineage relationship among single cells and plotted as a heat map. The intersections of the variants were quantified by the Upset R package[32]. The number of detected variants showed a correlation with sequencing depth and the number of cells with all variants (Figure 1--figure supplements 3 and 4) confirmed the variants already co-existed in the ancestral cell. Following this assumption, the detection rate can be measured as the proportion of cells with variants in the total number of cells. For each variant, cells were separated into different bins, increased by 10, according to the total sequencing depth at each variant. The detection rate for each variant site was then calculated in each bin. The combined detection rate was estimated by $1-(1-R_1)*(1-R_2)*(1-R_3)*(1-R_4)$, where $R_n$ is the detection rate for each variant.

**Lineage inference.** The probability of observing a mutation at a given site is $P_n=n*r$,

where *r* is the average mutation rate in the mitochondrial genome and *n* is the copy of mtDNAs in a single cell. *r* is estimated to be ~$10^{-7}$ per base[33], n is around 100~10000 per cell[34], so $P_n$ will be $10^{-5}$~$10^{-3}$. The probability of N cells sharing the same mtDNA mutations, but raising independently, will be $(P_n)^N$. Thus, when there are more than 3 cells in the population sharing a common mtDNA mutation, the probability of these independently occurring will be close to 0. Cells with common mtDNA mutations inherited the mutations from the same ancestral cell is more likely to explain the observation. Furthermore, when a set of mutations (more than 1) is detected in more than 1 cells, the null hypothesis (independently occurred) is rejected more confidently. The mutations within the ancestry cells can be inferred from the intersection of mutations. If a set of mutations are co-existed in the ancestral cell and the absence of mutations in the daughter cells are more likely caused by false detection in single cell libraries or genetic draft during cell replications. Then the observed cells with different intersections (e.g *V1+V2*) will be as expected by $P_{v1}*P_{v2}*N$, after normalized by sequencing depth. The exclusive of intersections from high-frequency mutations will infer the separation of mtDNA mutations and multiple cell lineage. The intersections of the variants were quantified by the Upset R package[32]. In the scATAC-seq from pHSCs from SU353, the intersection of variants showed most of the cells were separated by two sets of different variants (Figure 2D). But there are a few cells displaying a mixture of variants from the two sets. We suspected these may cause by the doublet of cells in the same well during single cell separation on C1 chip. We further separated the intersection map by the chip and observed the number of cells with mixture variants correlated to the concentration of cells loaded to C1 Chip. These cells were removed during subsequent analysis. Single cells with any variants in the two sets were kept and cells with more than 40X coverage on mtDNA, but no variants in the two sets were considered as wild-type HSCs. After all the filter steps, 153 cells had lineage information and were separated into three subgroups.

**Single cell ATAC-seq chromatin analysis.** ATAC sequences mapped to the nuclear genome were used for chromatin accessibility profiling. Bam files were merged for the same cell types and used as input files for chromVAR[15]. Peak files from Buenrostro et.al[26]

were used as open background regions to quantify the accessibility signal from every single cell. Cells with fewer than 200 unique reads or less than 25% of reads in peak regions were removed for chromatin analysis. chromVAR was applied to calculate TF motif-associated chromatin accessibility landscape changes and identify potential regulators of epigenomic variability. This approach quantifies accessibility variation across single-cells by aggregating accessible regions containing a specific TF motif, then compares the observed accessibility of all peaks containing a TF motif to a background set of peaks normalizing for known technical confounders. For determining differentially accessible motifs between different subpopulations, a Wilcoxon test was used to calculate the p values of the difference between the two groups.

**Code availability.** Custom analysis code can be downloaded from GitHub(https://github.com/ChangLab/ATAC_mito_sc)[35]

**Acknowledgements**

We thank C. Curtis, Ava Carter, Furqan Fazal, Kevin Parker and Chun-Kan Chen for insightful advice and assistance. Supported by US National Institutes of Health P50-HG007735 (to H.Y.C.), R01CA188055 (to R.M.), and R01HL142637 (to R.M.). R.M. is a Scholar of the Leukemia and Lymphoma Society. H.Y.C. is an Investigator of the Howard Hughes Medical Institute.

**Competing interests**

H.Y.C. is a co-founder of Accent Therapeutics and an advisor for 10x Genomics and Spring Discovery. Stanford University has filed a patent on ATAC-seq(US20160060691A1), on which H.Y.C. is named as an inventor.

**Figure Legends**

**Figure 1. EMBLEM reveals cell lineage from mtDNA mutations**

(A) EMBLEM workflow. Usings standard ATAC-seq data as input (left), an SNV calling step was added to enumerate all single nucleotide variants in mtDNA (middle). EMBLEM

identifies heteroplasmic mtDNA mutations in single cells, groups mutations into diagnostic sets, and infers cell lineage based on mtDNA variants, and overlays clonotype information on epigenomic profile of the same cells (right).

(B) ATAC-seq enriches for mtDNA reads compared to whole exome sequencing (WES), low coverage whole genome sequence (WGS_L), or PCR-free, high-coverage whole genome sequence (WGS_H).

(C) Bimodal distribution of variant allele frequency (VAF) of mtDNA mutations discovered using ATAC-seq. Yellow bar presents the homoplastic variants that can distinguish different individuals. Heteroplasmic variants can distinguish clonal cell populations within one indiviual.

(D) Two possible models for 25% mtDNA VAF in bulk: Homoplastic variants in a small proportion of cells (top) or heteroplasmic variants in nearly every single cell (bottom). Blue cells: cells with mutated mtDNA, blue dots: mtDNA with mutated allele.

(E) VAF of mtDNA mutations in single cell ATAC-seq data of human B cells. Each dot present the VAF (y-axis) in single cells, and rotated kernel density on each side present their distribution. The x-axis indicates the mutation site (the nucleotide position in mitochondrial genome).

(F) mtDNA mutations in human AML. Each row in the heap map is a single cell (LSC or AML blast); each column is a heteroplasmic mtDNA mutation. Blue color indicates the mtDNA variant is detected (>1 reads); white color indicates no mutation. The nucleotide position in mitochondrial genome for each mutation is indicated.

(G) Combined set of heteroplasmic mtDNA mutations improve cell lineage assignment in single cells. Cells were first separated into bins according to their mtDNA coverage (x-axis). The detection rate (y-axis) for each site (indicated by different color and shape) is calculated with the number of cells with that mutations divided by total number of cells in that bin. The detection rate of combining four sites (black line, METHOD) is substantially increased.

(H) Quantitation of mtDNA mutation detection rate as a function of sequencing depth and number of single cells. Cells were sorted in descending order by their sequencing depth and grouped into bins (10% of cells in each row). Distribution of sequencing depth is shown on the left panel. The black line and dark blue shade indicate mean ± standard

deviation, respectively. The light blue shade indicates remaining value of the bin. Cells with or without mtDNA variants are shown in blue and orange on the right panel, respectively.

**Figure 2. Clonal evolution of pre-leukemic HSCs inferred from joint lineage tracing and single cell chromatin accessibility.**

(A) Lineage hierarchy in acute myeloid leukemia based on EMBLEM and prior genetic information. mtDNA mutations reveals pHSC clonal heterogeneity. The clonal precursor of the leukemic stem cell is not the clone with most representation in the pHSC pool, but rather the clone with epigenomic bias towards the leukemic regulatory program, as depicted by related color schemes.

(B) EMBLEM deconvolutes AML clonal heterogeneity. Heteroplasmic mtDNA mutations in three cell populations from patients SU070 are shown. Mutations sites (in rows) in each FACS-sorted cell population (in columns) are shown, with size of each circle representing its VAF. Several mtDNA mutations (sites shown in purple) are detected in pHSCs and transmitted to LSCs and blasts, confirming those pHSC clones at the apex of leukemia lineage. LSCs accumulated additional mtDNA mutations (sites shown in green) and are transmitted to leukemic blasts in patient SU070. Allele frequency, sequencing depth and annotation of the variant allele are shown in Figure 2--figure supplements 1 and Supplementary file 2.

(C) Same plot as (B) shown for patient SU353. In addition to the shared mtDNA mutations in pHSCs, LSCs, and blasts (purple), two pHSCs-specific mtDNA mutations are also detected (yellow). Allele frequency, sequencing depth and annotation of the variant allele are shown in Figure 2--figure supplements 1 and Supplementary file 2.

(D) Heteroplasmic mutations in single pHSCs from one patient reveals clonal heterogeneity. Each column is a mtDNA nucleotide position; each row is one cell. Blue color indicates the presence of the mtDNA variant. Shown are cells with any mtDNA mutation detected, or cells with more than 40X coverage of the mitochondrial genome without any detected mutation(pHSC with WT mtDNA). The number of cells in each clonotype are indicated on the right.

(E) Landscape of single-cell chromatin accessibility of blood progenitor and leukemic cells in patient SU353. tSNE map using bias-corrected deviations from chromatin accessibility showing cluster of AML blasts, LSCs, pHSCs and normal HSC, colored by cell types.
(F) Chromatin accessibility of the FOS:JUN binding motif across the same single cells. tSNE map colored by deviation z-score for motif associated to FOS:JUN, the most variable TF motif.
(G) pHSC clones possess distinct epigenomic signatures. Clone 1 that gives rise to the AML has a chromatin accessibility profile that more resembles LSCs and leukemic blasts. "WT" pHSC refers to the pHSC with WT mtDNA. Clonotype information from EMBLEM is overlaid on the tSNE map defined by TF motif deviations, and colored by different lineal sub-populations defined by mtDNA mutations.
(H) Quantitation of distinct single-cell chromatin accessibility at FOS:JUN motifs among different pHSC clones defined by EMBLE. Clone 1 pHSCs tend to down regulate FOS:JUN accessibility, while clone 2 pHSC shows substantially greater cell-to-cell variability. pHSCs with no detectable mtDNA variants and normal HSCs are shown for comparison. TF deviation of single cells (black dots) is shown on the distribution box-plot. The statistical significant were indicated by "*" when $p<0.05$, "**" when $p<0.01$(Wilcoxon rank-sum test).

**Figure 1-figure supplement 1: EMBLEM workflow for SNP calling and lineage inference.**

(A) Workflow for mitochondrial DNA variant calling from ATAC-seq data. This workflow was applied to both bulk and single cell ATAC-seq. The steps indicated with dotted lines were not applied to single-cell data.
(B) Workflow for inferring lineage relationships from single-cell ATAC-seq data. BAM files from single cells were first merged and confident mtDNA variants were called. Mutated alleles from these variant sites were then counted for each single cell. The cell lineage was then inferred from mtDNA variants and analyzed alongside the chromatin profile for each cell.

**Figure 1-figure supplement 2: mtDNA coverage and variants from different sequencing libraries from GM12878 human B cells.**

(A) Mitochondrial genome coverage from each of four different sequencing libraries including WGS_H (high coverage PCR-free whole genome sequencing), WGS_L(low coverage whole genome sequencing), WES(whole exome sequencing), ATAC-seq. The Y axis shows coverage scaled in $\log_{10}$. 43M paired-end ATAC-seq reads(2x50bp) yielded the same coverage of mtDNA as 747M paired-end reads(2x250bp) from WGS-H data.

(B) Comparison of variants detected in sequencing data from four different library preparations. The number of variants detected in each library is shown on the bottom left. The intersection of different libraries (bottom-right) and the number of variants s are shown on the top. Homoplasmic variants are in yellow and heteroplasmic variants are in blue.

(C) Heteroplasmic mtDNA mutations detected by WGS_H( in blue) and ATAC-seq( in red). The X axis is the position of the mutation on mitochondrial genome and Y axis is the variant allele frequency in percentage.

**Figure 1-figure supplement 3: Heteroplasmic mtDNA mutation in K562 cells.**

(A) Percentage of mtDNA reads in ATAC-seq and whole genome sequence(WGS) libraries from human K562 cells. 4 millions mtDNA reads from 32 millions total mapped reads in ATAC, 7 millions mtDNA reads from 1775 millions total reads in WGS.

(B) The average coverage of the mitochondrial genome in ATAC and WGS form K562 cells.

(C) Number of heterplasmic mtDNA mutations detected in ATAC and WGS. The intersection size represents mutations detected by single or both methods.

(D) Variant allele frequency of mtDNA mutations and their correlation between ATAC and WGS. The red dots indicates the mutations detected by both ATAC and WGS, with the same criteria. The black dots indicates the mutations detected by ATAC or WGS only.

**Figure 1-figure supplement 4: Heteroplasmic variants in single cells from AML blasts and LSCs (SU353)**

(A) Heatmap showing variant mitochrondrial sites (columns) in each AML blast from patient SU353(rows). The color represents the number of reads supporting the variant

allele (log2(depth)). The first two sites are negative controls, which are detected in pHSCs only.

(B) Bar plot showing the number of cells in which we detect each mitochondrial variant. The last bar shows the number of cells with any one of the four variants detected.

(C) The top right shows the number of cells with each different combination of variants detected. The number of cells is shown on top of the bar. The combination of variants detected is annotated below the bar. The total number of cells with each variant site detected is shown to the left. The average coverage of the mitochondrial genome for each intersection group is shown below.

(D) VAF of mtDNA variants. The x-axis indicates the variant site notated by the nucleotide position in the mitochondrial genome. Each dot represents the VAF (y-axis) in single cells and the rotated kernel density on each side shows their distribution.

(E-H) Same as (A-D), for leukemia stem cells (LSCs) from patient SU353.

**Figure 1-figure supplement 5: Heteroplasmic variants in single cells from AML blasts and LSCs (SU070)**

(A-D) Same as Figure 1--figure supplement 3 A-D for AML blasts from SU070.

(E-H) Same as Figure 1--figure supplement 3 A-D for LSCs from SU070.

(I) Quantification of the detection rate for each heteroplasmic variant from mtDNA. Cells (both LSCs and AML blasts) were first separated into bins according to their coverage of mtDNA (x-axis). The detection rate (y-axis) for each site (notated by different color and shape) is calculated as the number of cells with that variant detected divided by the total number of cells in that bin.

(J) Quantitation of mtDNA mutation detection rate as a function of sequencing depth and the number of single cells. Cells were sorted in descending order by their sequencing depth and grouped into bins (10 cells in each row). Distribution of sequencing depth is shown on the left panel. Cells with or without mtDNA mutations are shown in blue or orange, respectively.

**Figure 2-figure supplement 1: Heteroplasmic mtDNA mutations detected in bulk ATAC-seq from AML patients.**

(A) The number of heteroplasmic variants detected using ATAC-seq data from normal primary blood cells and cancer cells from AML patients.
(B) The number of mtDNA variants identified from normal and cancer samples in different substitution classes are shown as a bar plot. Mutations from normal (gray) and cancer (yellow) samples are separated. The C>T and T>C signature in cancer mtDNA has been observed in previous studies and it's equivalent to the one that has been operating during the evolution of human germline mtDNAs.
(C) Annotation of mtDNA mutations and the proportion of mutations in coding and non-coding regions. Coding mutations are divided into synonymous, nonsynonymous, and gain of stop codon. Heteroplasmic mutations detected from cancer samples show a similar distribution as those from normal samples, with a slightly higher proportion falling within coding regions.
(D) Heteroplasmic mutations in three cell stages for each AML patient. Variant allele (in rows) in each cell population (in columns) are shown with a circle, with size indicating their variant allele frequency. Sequencing depth of the variant allele is indicated by the color of the circle (in $\log_2$ scale). pHSC specific mutation sites are in red. Allele frequency and annotation of mtDNA mutations were shown in Supplementary file 2
(E) Heteroplasmic mutations in two cell stages for each AML patient. Variant allele (in rows) in each cell population (in columns) are shown with a circle, with size indicating their variant allele frequency. Sequencing depth of the variant allele is indicated by the color of the circle (in $\log_2$ scale). pHSC specific mutation sites are in red. Allele frequency and annotation of mtDNA mutations were shown in Supplementary file 2

**Figure 2-figure supplement 2: Single cell chromatin accessibility**

(A) Phylogenetic relationship of cells from SU353 was inferred using the Neighbor-Joining method. The phylogenetic tree is drawn to scale, with branch lengths in the units of the number of base difference per site. The clade in purple matched to clone 1 and the clade in yellow matched to clone2 in Figure 2D. The cell type and mtDNA variants in each single cell are shown on the right. 229 single cells with at least one of the six heteroplasmic mtDNA mutations were included.

(B) Heat map showing clusters of pHSCs from SU353 and normal HSCs from a healthy donor, based on the z-score of TF deviation. The Z-scored deviation is shown for individual cells (columns) for each TF (rows). Clone information is shown on the top of the heat map. Top 50 most variable motifs were used in this heat map.
(C) Volcano plot showing the difference in chromatin accessibility for transcription factor binding motifs between Clone 1 and Clone 2. The x-axis shows the mean difference of bias-corrected deviations and the y-axis shows the p-value (in $\log_{10}$ scale). The most significant differential motifs are annotated with TF names.
(D)Same as in (C) for Clone 1 vs. WT cells.
(E)Same as in (C) for Clone 2 vs. WT cells. No significantly differential motifs were detected.

**Figure 2--figure supplements 3: Sorting Scheme for pHSCs.**
Scheme of FACS sorting of the pHSC population from AML patient SU353. Initial sort (top panel) and post-sort purity (bottom panel) are shown.

**Figure 2-figure supplement 4: Investigation of horizontal mitochondrial transfer using mixing experiment from mouse and human cells.**
(A) Scatter plot shows the number of unique reads mapped to human and moue nuclear genome(gDNA). Red circle indicates cell doublet. Sequence reads from each single cell were mapped to human and mouse combined reference genome. Unique mapped reads on gDNA and mtDNA were counted respectively.
(B) Scatter plot shows the number of unique reads mapped to human and moue mitochondrial genome(mtDNA). Red circle indicates cell doublet.
(C) Species-specific score for gDNA and mtDNA. The species-specific score was calculated with (Chuman/(Chuman+Cmouse)-0.5. “-0.5” or “0.5” indicate 100% alignment to mouse or human reference. The positive correlation between gDNA and mtDNA indicates the species-specific mtDNA always paired with species-specific gDNA.

**Supplementary file 1: Information of datasets utilized in this study.**

**Supplementary file 2: Heteroplasmic mtDNA mutations detected in each AML patient.** Allele frequence, seqeunce coverage and annotation information of the variants are provided.

**Supplementary file 3: Relative merits of mtDNA vs. other genetic markers for lineage tracing**

## References


1. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: Genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18,** 230–244 (2017).
2. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. *Nat. Biotechnol.* **36,** 469–473 (2018).
3. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561,** 473–478 (2018).
4. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512,** 155–160 (2014).
5. Evrony, G. D. *et al.* Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. *Neuron* **85,** 49–60 (2015).
6. Biezuner, T. *et al.* A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res.* **26,** 1588–1599 (2016).
7. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–8 (2013).
8. Morris, J. *et al.* Pervasive within-Mitochondrion Single-Nucleotide Variant Heteroplasmy as Revealed by Single-Mitochondrion Sequencing. *Cell Rep.* **21,** 2706–2713 (2017).
9. Fellous, T. G. *et al.* A methodological approach to tracing cell lineage in human epithelial tissues. *Stem Cells* **27,** 1410–1420 (2009).
10. Ju, Y. S. eo. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3,** 1–28 (2014).
11. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4,** 149ra118 (2012).
12. Thomas, D. & Majeti, R. Review Series Biology and relevance of human acute myeloid leukemia stem cells. (2017). doi:10.1182/blood
13. Corces-Zimmerman, M. R., Hong, W.-J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl. Acad. Sci.* **111,** 2548–2553 (2014).
14. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48,** 1193–1203

(2016).
15. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14,** 975–978 (2017).
16. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).
17. Zhang, L. & Zhang, S. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **PP,** 1 (2018).
18. Santaguida, M. *et al.* JunB Protects against Myeloid Malignancies by Limiting Hematopoietic Stem Cell Proliferation and Differentiation without Affecting Self-Renewal. *Cancer Cell* **15,** 341–352 (2009).
19. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* 1–15 (2019). doi:10.1016/j.cell.2019.01.022
20. Mishra, P. & Chan, D. C. Mitochondrial dynamics and inheritance during cell division, development and disease. *Nat. Rev. Mol. Cell Biol.* **15,** 634–646 (2014).
21. Moschoi, R. *et al.* Protective mitochondrial transfer from bone marrow stromal cells to acute myeloid leukemic cells during chemotherapy. *Blood* **128,** 253–265 (2016).
22. Marlein, C. R. *et al.* NADPH oxidase-2 derived superoxide drives mitochondrial transfer from bone marrow stromal cells to leukemic blasts. *Blood* **130,** 1649–1660 (2017).
23. Hayakawa, K. *et al.* Transfer of mitochondria from astrocytes to neurons after stroke. *Nature* **535,** 551–555 (2016).
24. Satpathy, A. T. *et al.* Transcript-indexed ATAC-seq for precision immune profiling. *Nat. Med.* **24,** 1 (2018).
25. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).
26. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173,** 1535–1548.e16 (2018).
27. Li, H., Li, H., Durbin, R. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
30. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).
31. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (80-. ).* **355,** 752–756 (2017).
32. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties.

doi:10.1093/bioinformatics/btx364
33. Coller, H. A. *et al.* High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.* **28,** 147–150 (2001).
34. Miller, F. J., Rosenfeldt, F. L., Zhang, C., Linnane, A. W. & Nagley, P. Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Res.* **31,** e61 (2003).

35. Jin Xu. 2019 ATAC_mito_sc. Github https://github.com/ChangLab/ATAC_mito_sc.

1c1946a

Figure 1

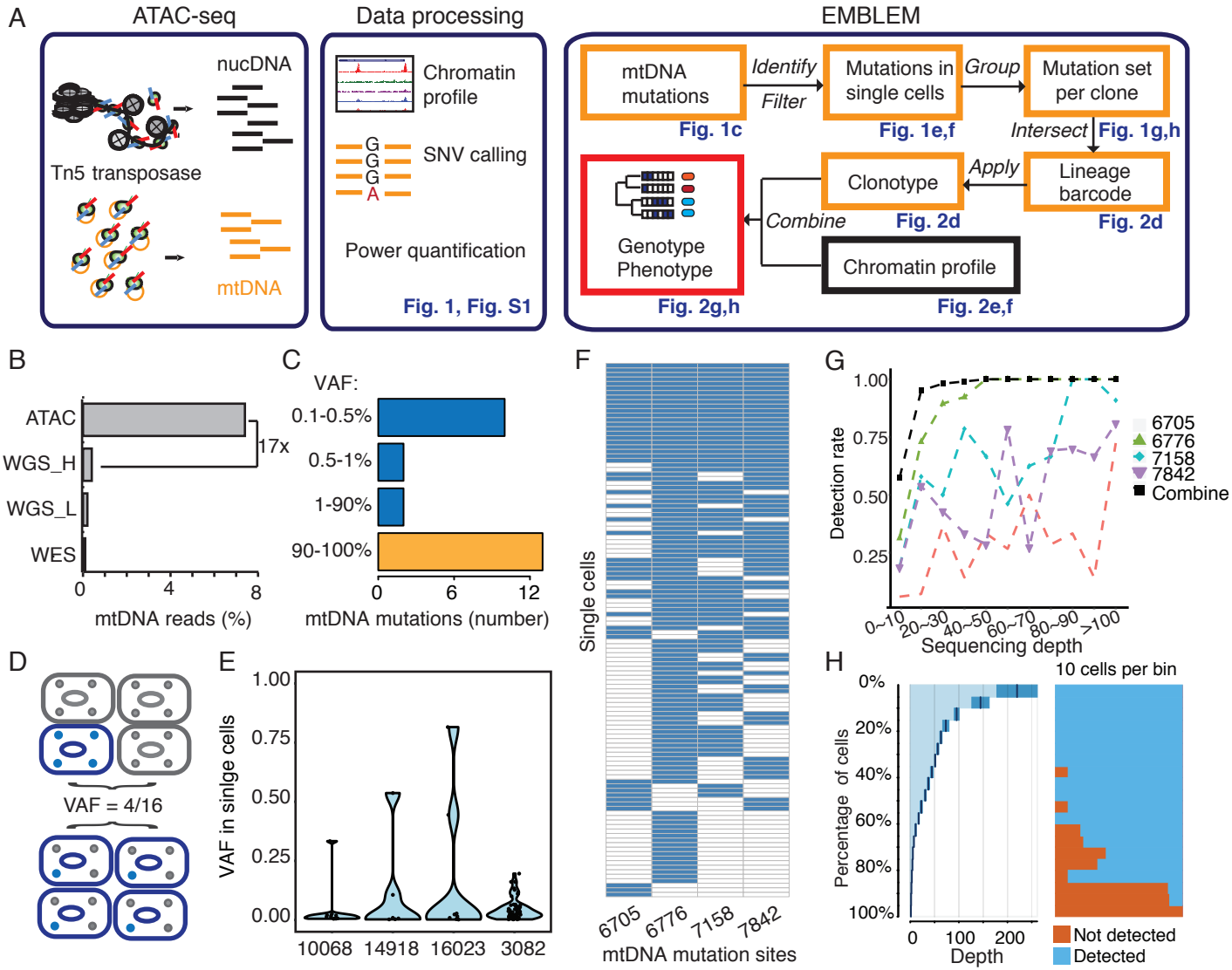

A
ATAC-seq
nucDNA
Tn5 transposase
mtDNA
Data processing
Chromatin profile
SNV calling
Power quantification
Fig. 1, Fig. S1
EMBLEM
mtDNA mutations
Fig. 1c
Identify
Filter
Mutations in single cells
Fig. 1e,f
Group
Mutation set per clone
Intersect
Fig. 1g,h
Apply
Lineage barcode
Fig. 2d
Clonotype
Fig. 2d
Combine
Genotype Phenotype
Fig. 2g,h
Chromatin profile
Fig. 2e,f
B
ATAC
WGS_H
WGS_L
WES
17x
0
4
8
mtDNA reads (%)
C
VAF:
0.1-0.5%
0.5-1%
1-90%
90-100%
0
6
12
mtDNA mutations (number)
D
VAF = 4/16
E
VAF in sinlge cells
1.00
0.75
0.50
0.25
0.00
10068
14918
16023
3082
F
Single cells
6705
6776
7158
7842
mtDNA mutation sites
G
Detection rate
1.00
0.75
0.50
0.25
0~10
20~30
40~50
60~70
80~90
>100
Sequencing depth
6705
6776
7158
7842
Combine
H
10 cells per bin
Percentage of cells
0%
20%
40%
60%
80%
100%
0
100
200
Depth
Not detected
Detected

Figure 2

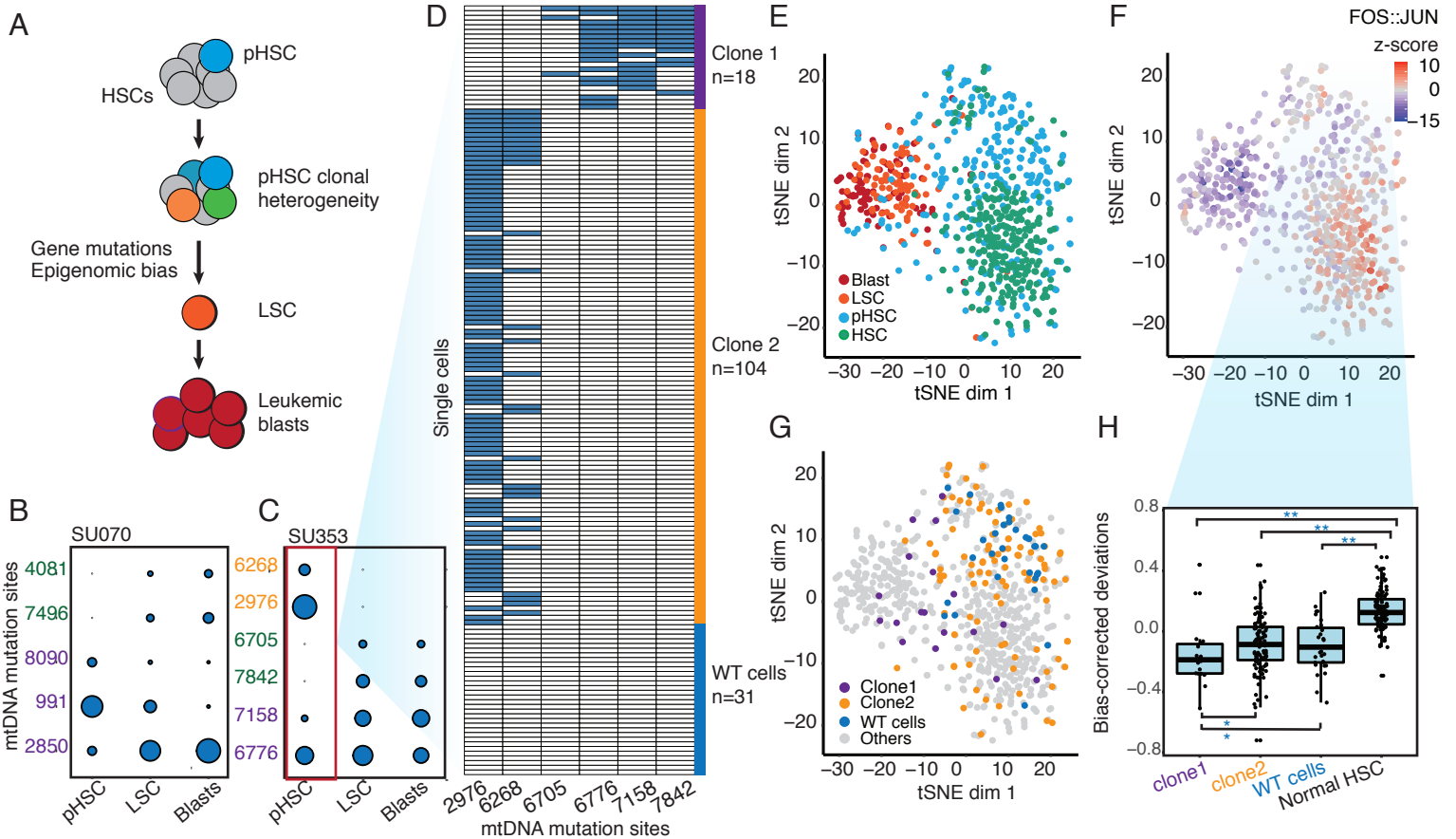

A
HSCs
pHSC
pHSC clonal heterogeneity
Gene mutations
Epigenomic bias
LSC
Leukemic blasts
B
SU070
mtDNA mutation sites
4081
7496
8090
991
2850
pHSC
LSC
Blasts
C
SU353
6268
2976
6705
7842
7158
6776
pHSC
LSC
Blasts
D
Single cells
Clone 1
n=18
Clone 2
n=104
WT cells
n=31
2976
6268
6705
6776
7158
7842
mtDNA mutation sites
E
tSNE dim 2
tSNE dim 1
Blast
LSC
pHSC
HSC
F
FOS::JUN
z-score
10
0
−15
tSNE dim 2
tSNE dim 1
G
tSNE dim 2
tSNE dim 1
Clone1
Clone2
WT cells
Others
H
Bias-corrected deviations
clone1
clone2
WT cells
Normal HSC

Figure 1-S1

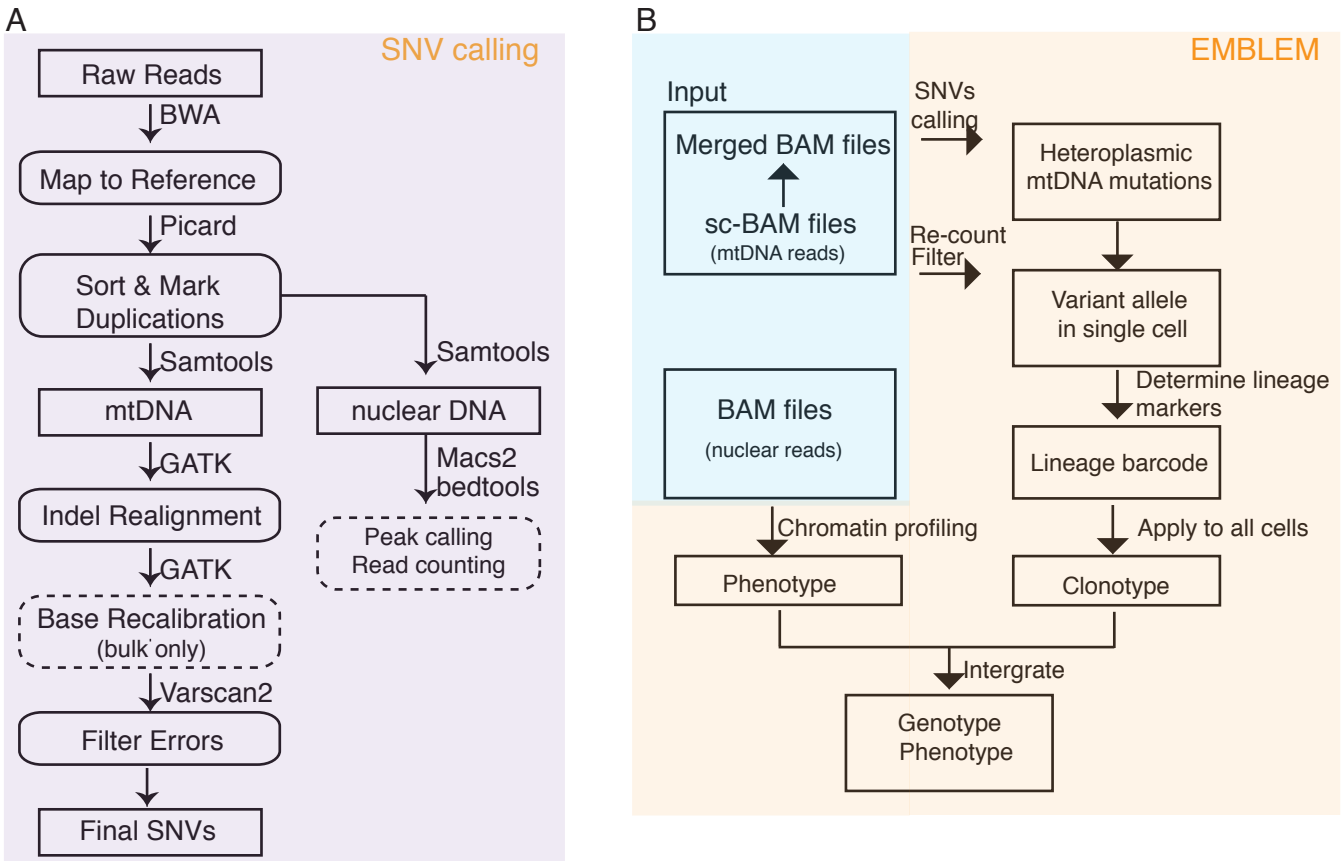

A
SNV calling
Raw Reads
BWA
Map to Reference
Picard
Sort & Mark Duplications
Samtools
mtDNA
GATK
Indel Realignment
GATK
Base Recalibration (bulk only)
Varscan2
Filter Errors
Final SNVs
Samtools
nuclear DNA
Macs2 bedtools
Peak calling Read counting
B
EMBLEM
Input
Merged BAM files
sc-BAM files (mtDNA reads)
BAM files (nuclear reads)
SNVs calling
Re-count Filter
Heteroplasmic mtDNA mutations
Variant allele in single cell
Determine lineage markers
Lineage barcode
Chromatin profiling
Phenotype
Apply to all cells
Clonotype
Intergrate
Genotype Phenotype

Figure 1-S2

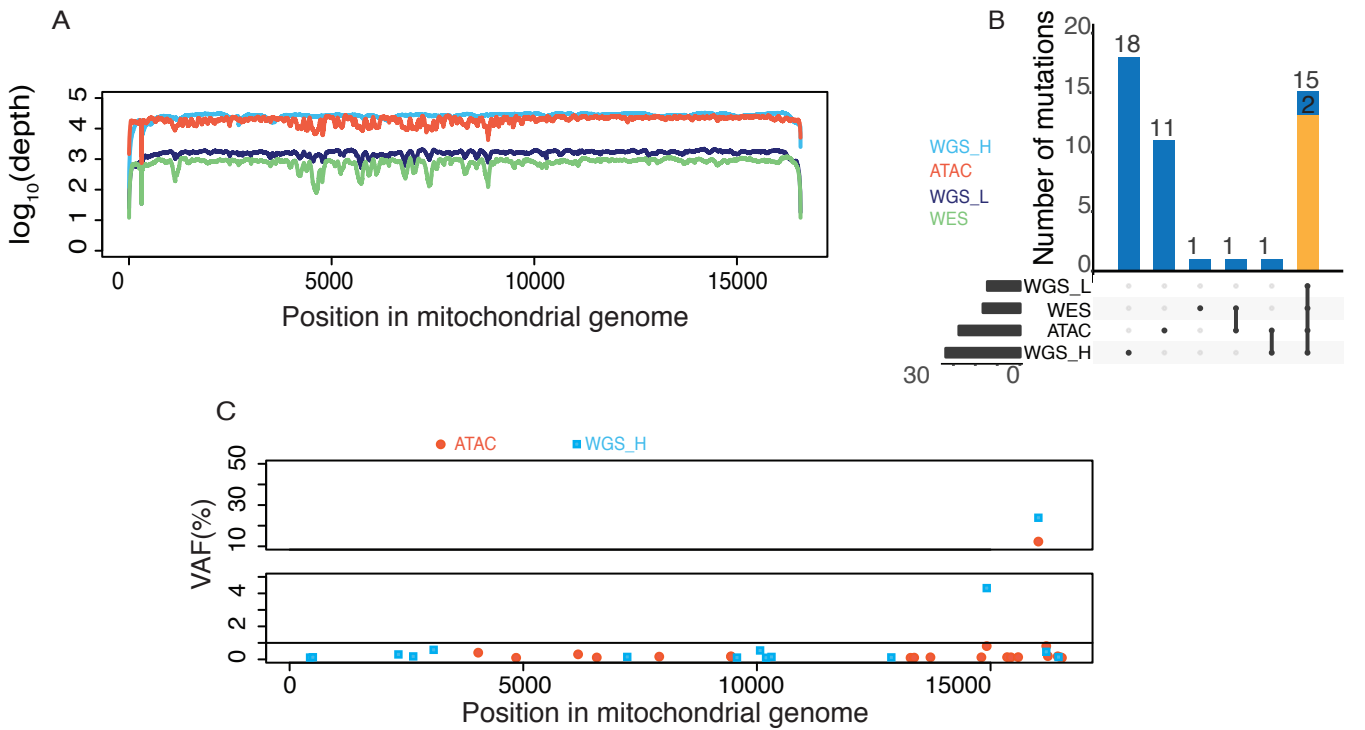

A
log10(depth)
0 1 2 3 4 5
0
5000
10000
15000
Position in mitochondrial genome
B
Number of mutations
20
15
10
5
0
18
11
1
1
1
15
2
WGS_H
ATAC
WGS_L
WES
WGS_L
WES
ATAC
WGS_H
30
0
C
ATAC
WGS_H
VAF(%)
50
30
10
4
2
0
0
5000
10000
15000
Position in mitochondrial genome

Figure 1-S3

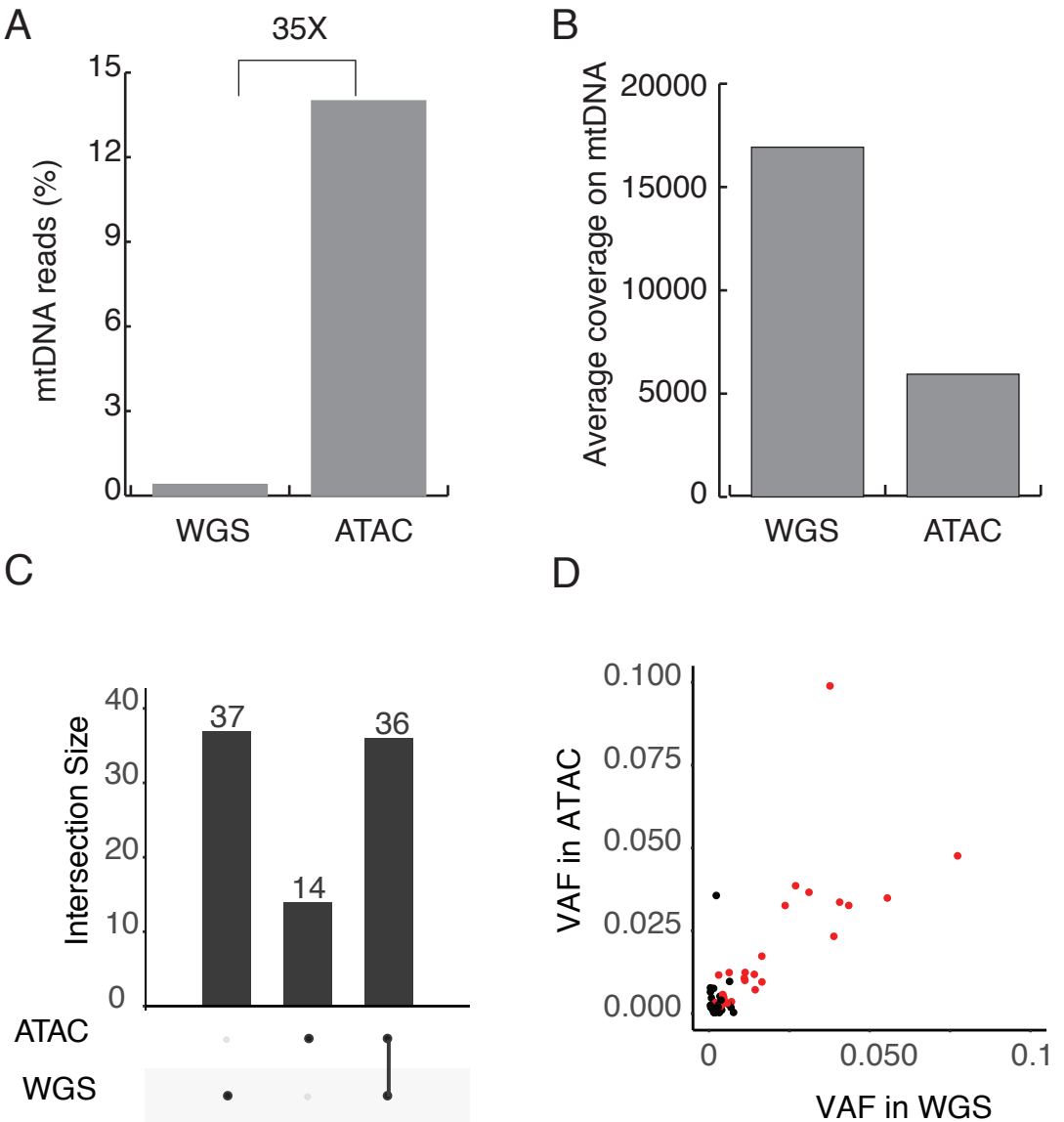

A
35X
mtDNA reads (%)
15
12
9
6
3
0
WGS
ATAC
B
Average coverage on mtDNA
20000
15000
10000
5000
0
WGS
ATAC
C
Intersection Size
40
30
20
10
0
37
14
36
ATAC
WGS
D
VAF in ATAC
0.100
0.075
0.050
0.025
0.000
0
0.050
0.1
VAF in WGS

Figure1-S4

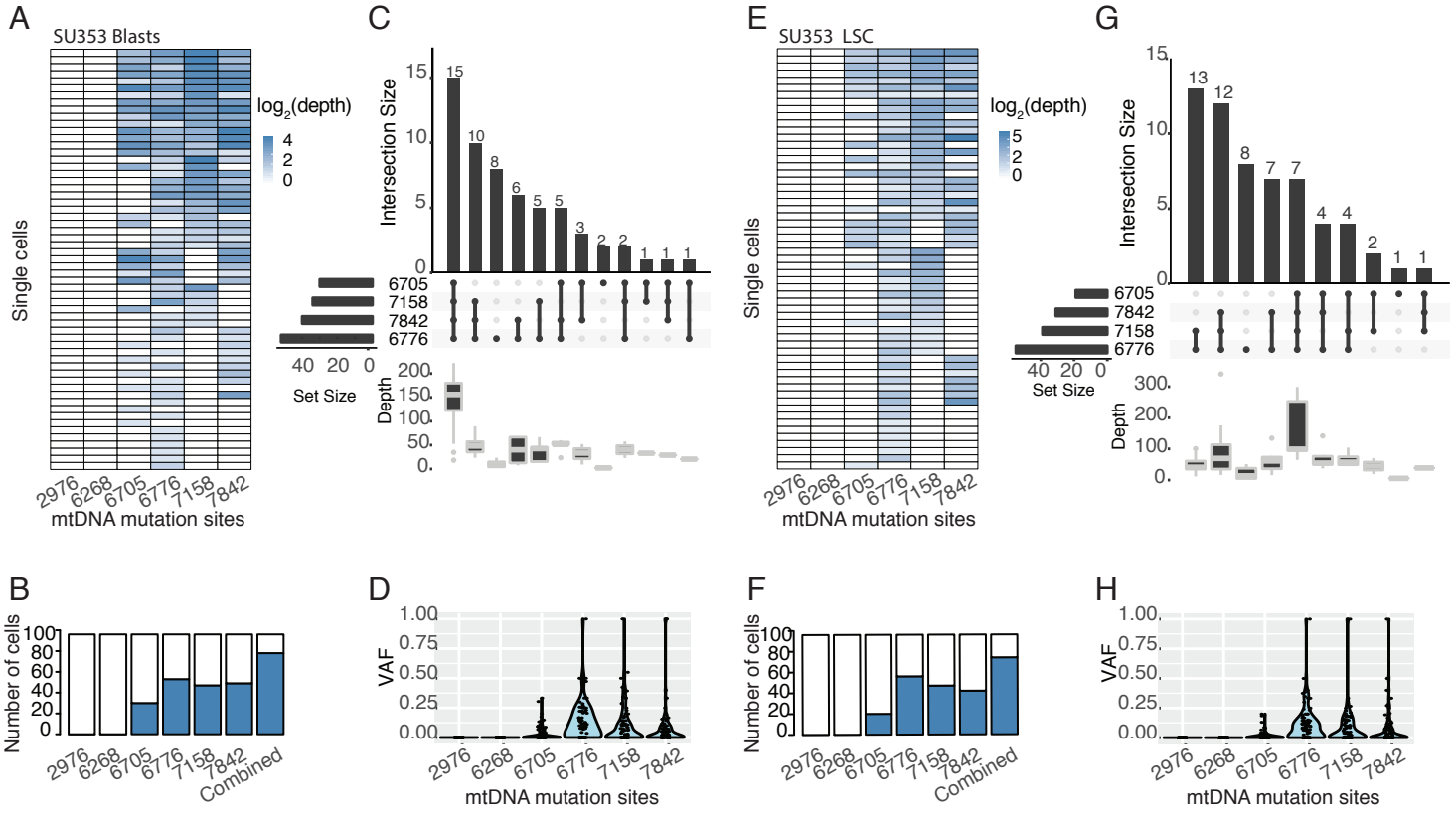

A
SU353 Blasts
Single cells
log2(depth)
4
2
0
2976
6268
6705
6776
7158
7842
mtDNA mutation sites
C
Intersection Size
15
10
8
6
5
5
3
2
2
1
1
1
6705
7158
7842
6776
40 20 0
Set Size
Depth
200
150
100
50
0
E
SU353 LSC
Single cells
log2(depth)
5
2
0
mtDNA mutation sites
G
Intersection Size
15
13
12
8
7
7
4
4
2
1
1
6705
7842
7158
6776
Set Size
Depth
300
200
100
0
B
Number of cells
100
80
60
40
20
0
Combined
D
VAF
1.00
0.75
0.50
0.25
0.00
mtDNA mutation sites
F
Number of cells
H
VAF
mtDNA mutation sites

Figure1-S5

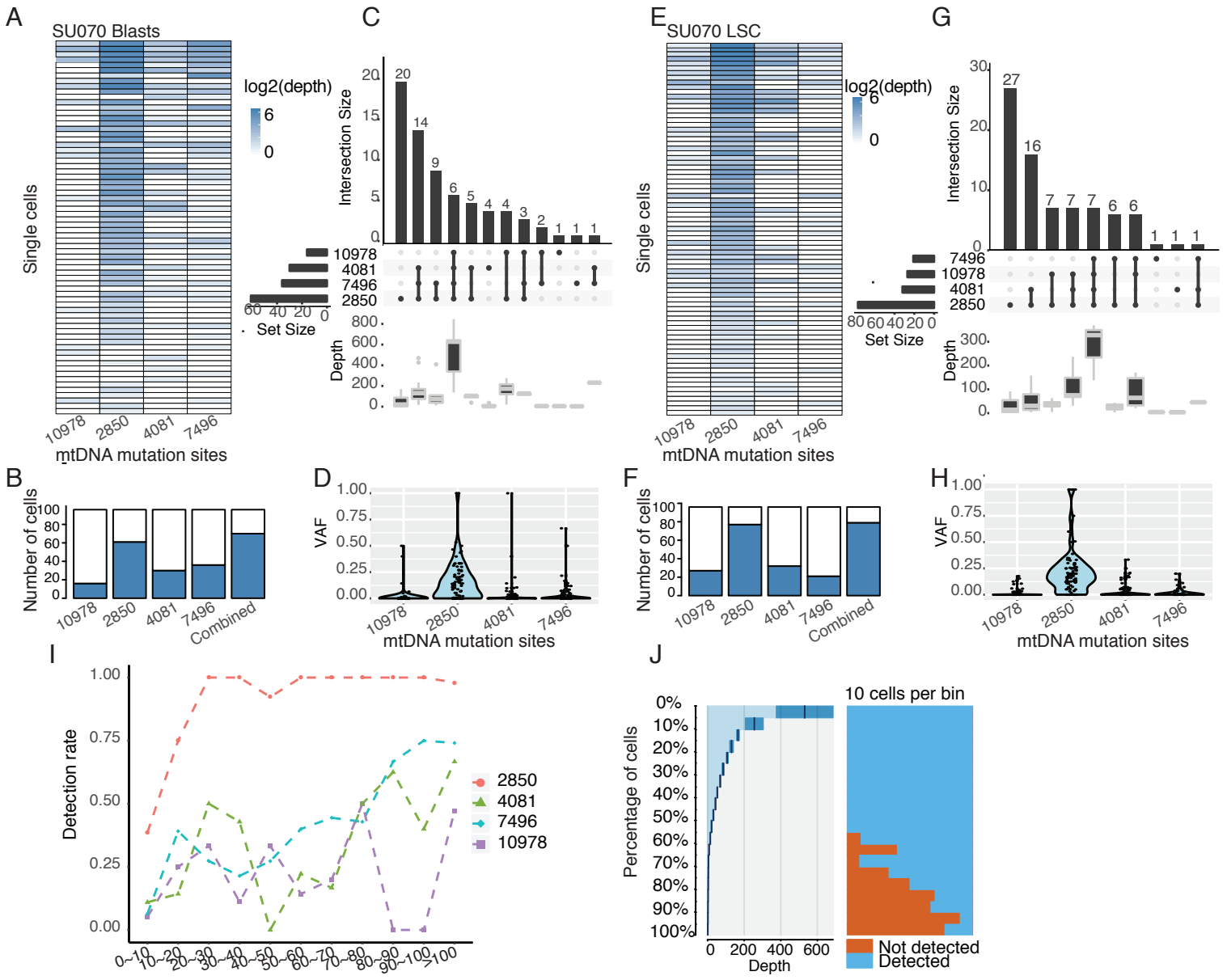

A
SU070 Blasts
log2(depth)
6
0
Single cells
10978
2850
4081
7496
mtDNA mutation sites
C
Intersection Size
20
14
9
6
5
4
4
3
2
1
1
1
10978
4081
7496
2850
60 40 20 0
Set Size
Depth
800
600
400
200
0
E
SU070 LSC
log2(depth)
6
0
Single cells
10978
2850
4081
7496
mtDNA mutation sites
G
Intersection Size
30
20
10
0
27
16
7
7
7
6
6
1
1
1
7496
10978
4081
2850
80 60 40 20 0
Set Size
Depth
300
200
100
0
B
Number of cells
100
80
60
40
20
0
10978
2850
4081
7496
Combined
D
VAF
1.00
0.75
0.50
0.25
0.00
10978
2850
4081
7496
mtDNA mutation sites
F
Number of cells
100
80
60
40
20
0
10978
2850
4081
7496
Combined
H
VAF
1.00
0.75
0.50
0.25
0.00
10978
2850
4081
7496
mtDNA mutation sites
I
Detection rate
1.00
0.75
0.50
0.25
0.00
0~10
10~20
20~30
30~40
40~50
50~60
60~70
70~80
80~90
90~100
>100
2850
4081
7496
10978
J
Percentage of cells
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0
200
400
600
Depth
10 cells per bin
Not detected
Detected

Figure2-S1

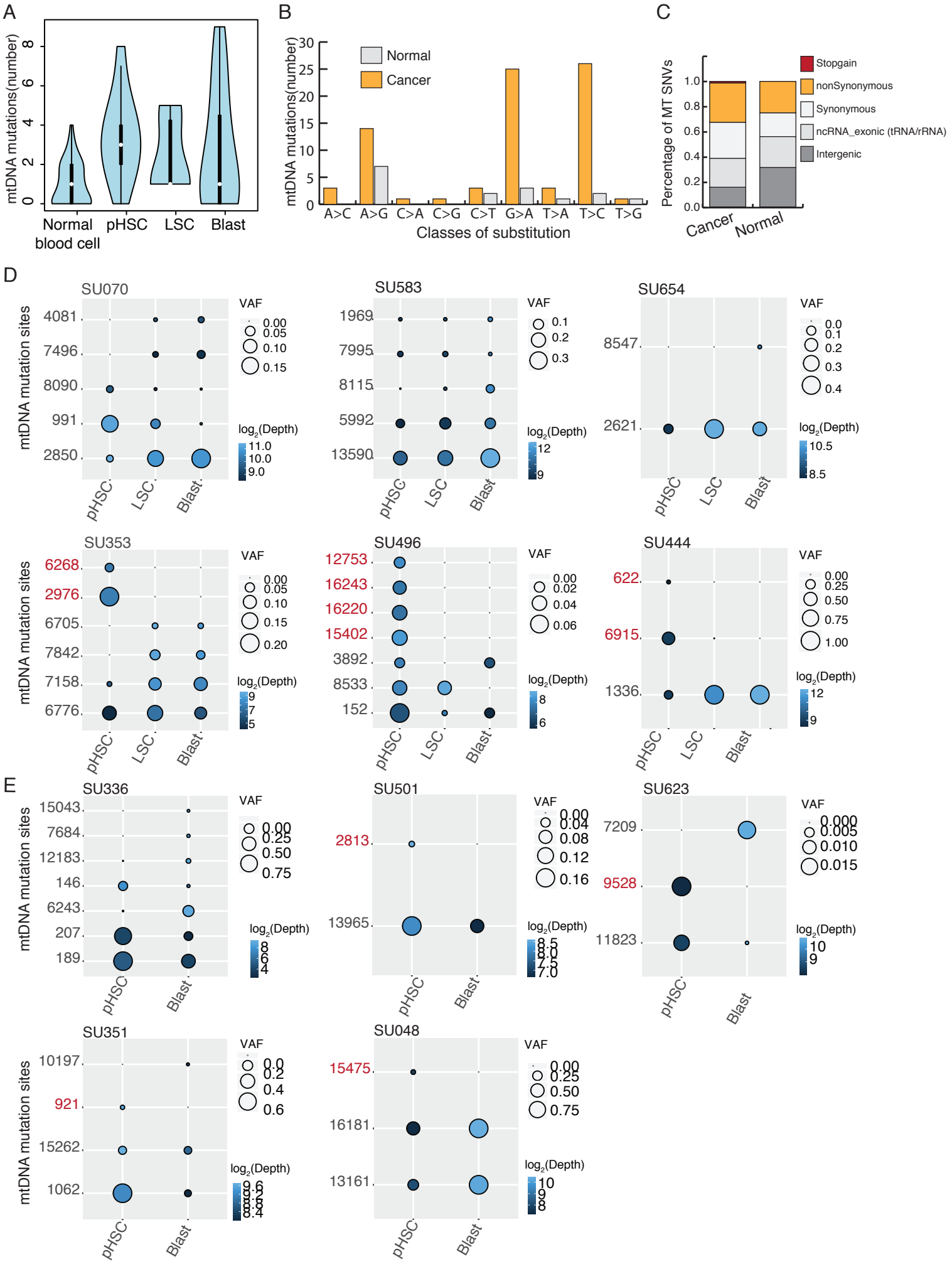
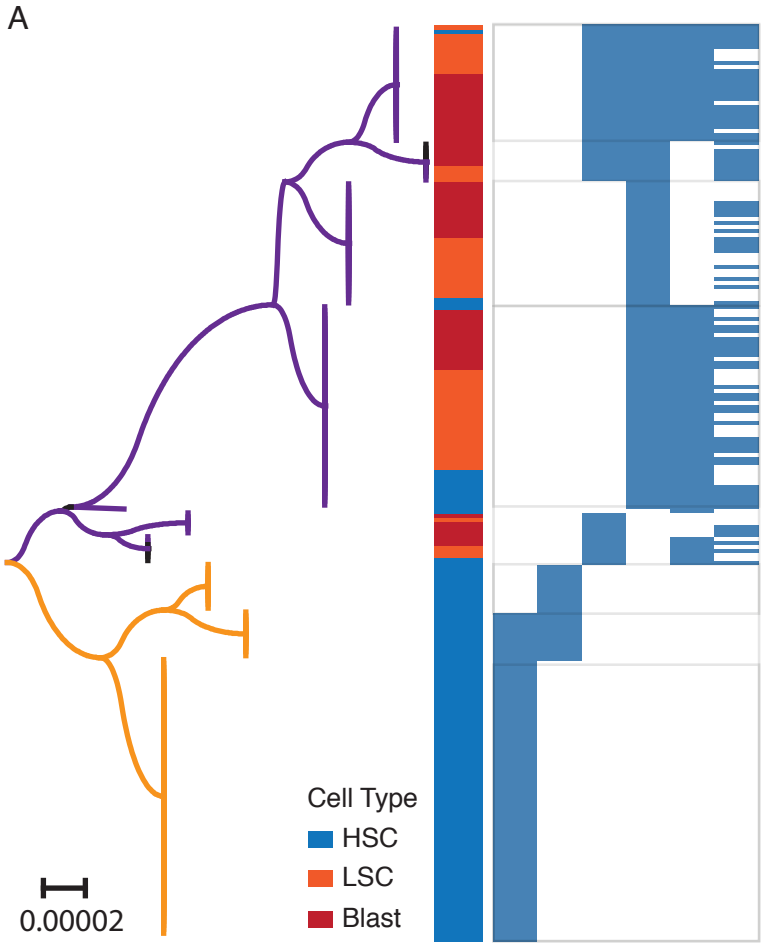
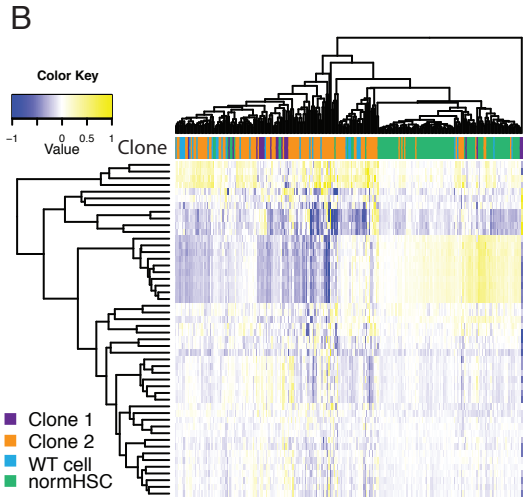
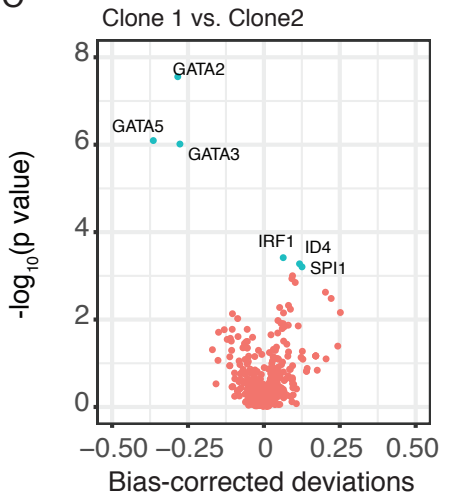
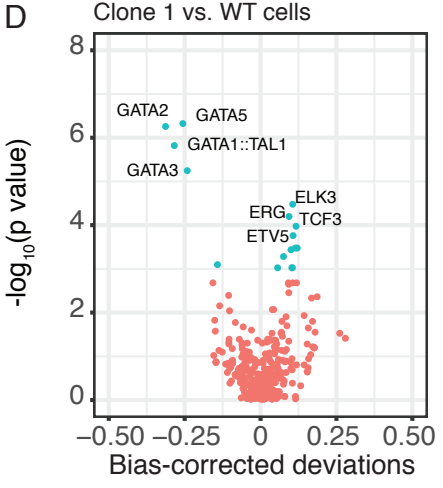

A
mtDNA mutations(number)
Normal blood cell
pHSC
LSC
Blast
B
mtDNA mutations(number)
Normal
Cancer
A>C A>G C>A C>G C>T G>A T>A T>C T>G
Classes of substitution
C
Percentage of MT SNVs
Stopgain
nonSynonymous
Synonymous
ncRNA_exonic (tRNA/rRNA)
Intergenic
Cancer
Normal
D
SU070
SU583
SU654
SU353
SU496
SU444
mtDNA mutation sites
VAF
log2(Depth)
pHSC
LSC
Blast
E
SU336
SU501
SU623
SU351
SU048

Figure2- S2


A
0.00002
Cell Type
HSC
LSC
Blast


B
Color Key
Value
Clone
Clone 1
Clone 2
WT cell
normHSC


C
Clone 1 vs. Clone2
GATA2
GATA5
GATA3
IRF1
ID4
SPI1
-log10(p value)
Bias-corrected deviations


D
Clone 1 vs. WT cells
GATA2
GATA5
GATA1::TAL1
GATA3
ELK3
ERG
TCF3
ETV5
-log10(p value)
Bias-corrected deviations

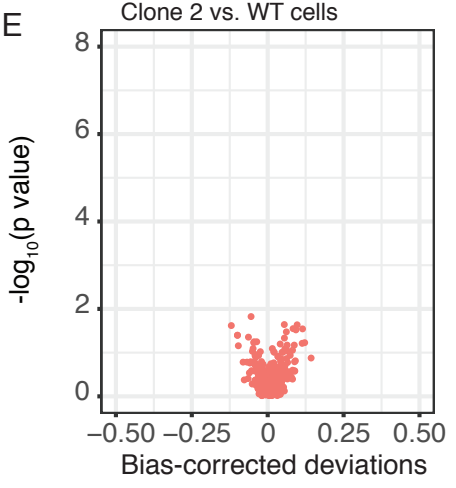
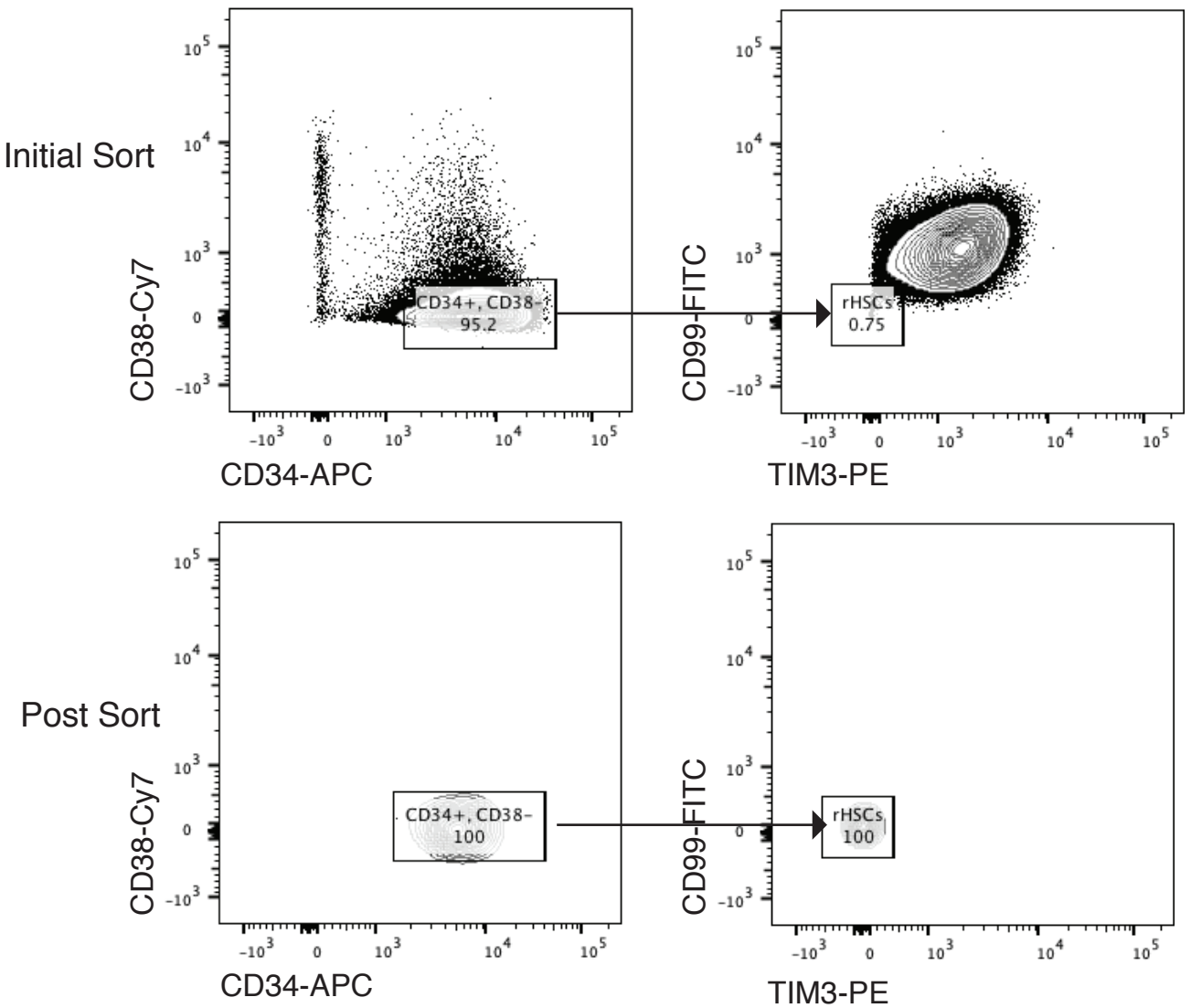
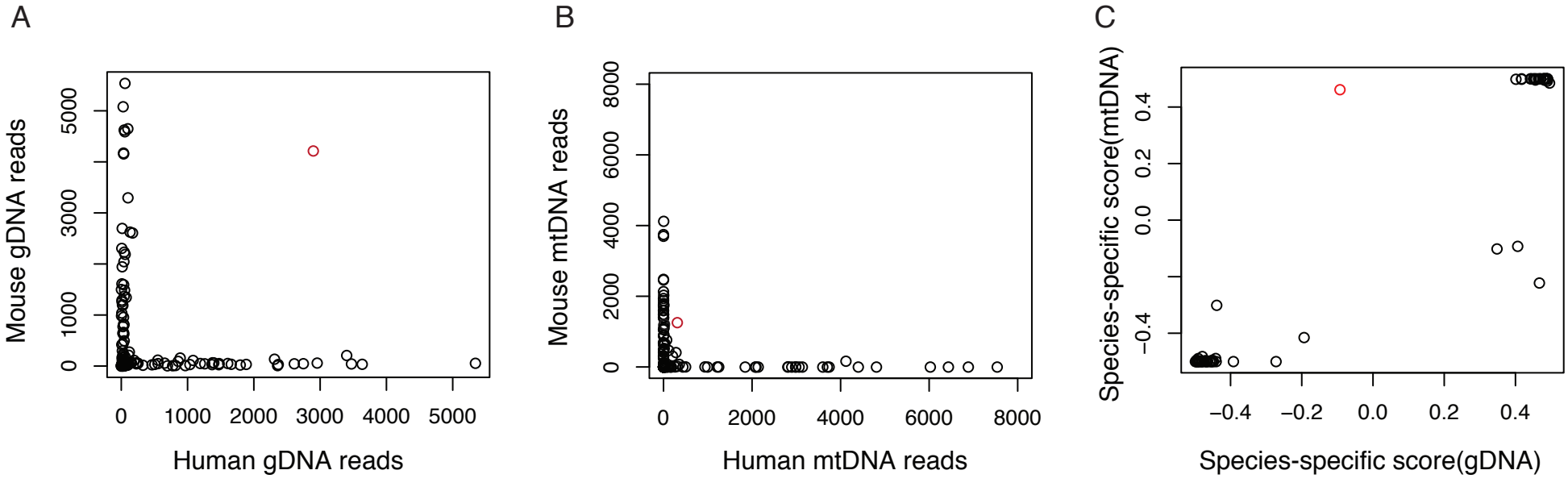

E
Clone 2 vs. WT cells
-log10(p value)
Bias-corrected deviations

Figure2- S3


Initial Sort
CD38-Cy7
CD34-APC
CD34+, CD38-
95.2
CD99-FITC
TIM3-PE
rHSCs
0.75
Post Sort
CD38-Cy7
CD34-APC
CD34+, CD38-
100
CD99-FITC
TIM3-PE
rHSCs
100


A
Mouse gDNA reads
0
1000
3000
5000
0
1000
2000
3000
4000
5000
Human gDNA reads
B
Mouse mtDNA reads
0
2000
4000
6000
8000
0
2000
4000
6000
8000
Human mtDNA reads
C
Species-specific score(mtDNA)
−0.4
0.0
0.2
0.4
−0.4
−0.2
0.0
0.2
0.4
Species-specific score(gDNA)