



Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations

Tyler E. Miller^{1,2,3}, Caleb A. Lareau^{2,4,5,6}, Julia A. Verga^{2,3}, Erica A. K. DePasquale^{2,7}, Vincent Liu^{4,5}, Daniel Ssozi^{2,7}, Katalin Sandor⁴, Yajie Yin⁴, Leif S. Ludwig^{2,6,8}, Chadi A. El Farran^{2,3}, Duncan M. Morgan^{9,10}, Ansuman T. Satpathy⁴, Gabriel K. Griffin^{2,11}, Andrew A. Lane^{2,12,13}, J. Christopher Love^{2,9,10}, Bradley E. Bernstein^{2,3,13,14}, Vijay G. Sankaran^{2,6} and Peter van Galen^{2,7,13} ✉

The combination of single-cell transcriptomics with mitochondrial DNA variant detection can be used to establish lineage relationships in primary human cells, but current methods are not scalable to interrogate complex tissues. Here, we combine common 3' single-cell RNA-sequencing protocols with mitochondrial transcriptome enrichment to increase coverage by more than 50-fold, enabling high-confidence mutation detection. The method successfully identifies skewed immune-cell expansions in primary human clonal hematopoiesis.

Single-cell RNA sequencing (scRNA-seq) enables the unbiased assessment of cell states in health and disease^{1,2}. Combined acquisition of cell state and genetic information can provide additional insight, such as targeted enrichment of cancer driver mutations from single-cell transcriptomes^{3,4}. Separately, combining scRNA-seq with genetic cell barcodes (CBs) can reveal clonal relationships and the evolutionary dynamics of cells within organisms^{5,6}. However, this has largely been limited to experimental model systems that can be genetically manipulated to insert CBs. To infer clonal dynamics in primary human cells, recent methods have detected and utilized mitochondrial DNA (mtDNA) mutations as naturally occurring genetic CBs^{7–9}. The combination of scRNA-seq with mtDNA mutation detection can inform clonal relationships with high confidence but is currently restricted to expensive, low-throughput full-length transcript sequencing technologies like SmartSeq2 (refs. ^{7,10}). To enable the reconstruction of clonal relationships in complex human tissues, we developed a method that captures genetic variants from high-throughput scRNA-seq platforms, MAESTER (mitochondrial alteration enrichment from single-cell transcriptomes to establish relatedness) (Fig. 1a). MAESTER is compatible with the most common high-throughput scRNA-seq platforms, including 10x Genomics 3' protocols, Seq-Well S3 and Drop-seq (Supplementary Figs. 1–3) (refs. ^{11,12}). An intermediate

step in each of these platforms yields full-length cDNA transcripts, from which we enrich all 15 mitochondrial transcripts using pools of primers while maintaining cell-identifying barcodes (Fig. 1b and Supplementary Fig. 4). Standard next-generation sequencing with 250-bp reads is then used to obtain the sequence of the amplified mitochondrial transcripts (Fig. 1a). We developed a computational toolkit to call mtDNA variants from MAESTER data, the Mitochondrial Alteration Enrichment and Genome Analysis Toolkit (maegatk; Supplementary Fig. 5 and Methods). Building on previous tools that we developed⁸ for mtDNA variant detection from single-cell ATAC (assay for transposase-accessible chromatin) or SmartSeq2, maegatk specifically handles technical biases implicit in high-throughput transcriptomic libraries. maegatk uses unique molecular identifiers (UMIs) to collapse multiple sequencing reads of the same starting transcript, creating a consensus call for every nucleotide based on the most common call and base quality. This approach mitigates sequence errors introduced during PCR and sequencing and is essential to obtain high-confidence variant calls from high-throughput scRNA-seq protocols. We also incorporate indel calling and provide a resource to evaluate the potential functional impact of variants (Supplementary Fig. 6 and Supplementary Table 1). Alterations in mtDNA are then used to infer relatedness between cells.

To verify the recovery of variants in our approach, we sought to measure mtDNA and RNA from the same individual cells. To achieve this, we performed MAESTER on DOGMA-seq libraries, which enable the concomitant detection of accessible DNA (including mtDNA) and transcriptome-wide RNA via the 10x Genomics multiome kit¹³ (Fig. 1c). After identifying variants on mtDNA⁸, we examined the proportion of variants recovered by MAESTER. Our analyses revealed that MAESTER recovered 94.1% of variants at a single-cell heteroplasmy of >10% (Fig. 1c). These results confirm

¹Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. ²Broad Institute of MIT and Harvard, Cambridge, MA, USA. ³Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. ⁴Department of Pathology, Stanford University, Stanford, CA, USA. ⁵Department of Genetics, Stanford University, Stanford, CA, USA. ⁶Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ⁷Division of Hematology, Brigham and Women's Hospital, Department of Medicine, Harvard Medical School, Boston, MA, USA. ⁸Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. ⁹Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹⁰Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹¹Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA. ¹²Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. ¹³Ludwig Center at Harvard, Harvard Medical School, Boston, MA, USA. ¹⁴Departments of Cell Biology and Pathology, Harvard Medical School, Boston, MA, USA. ✉e-mail: pvangalen@bwh.harvard.edu

that MAESTER recovers true mtDNA variants from transcriptomic data.

We established the feasibility and efficiency of MAESTER in standard high-throughput scRNA-seq methods using human cell mixing experiments. Chronic myelogenous leukemia cells (K562) were mixed with brain tumor cells (BT142) and analyzed using Seq-Well S3 and 10x Genomics 3' v3 protocols. MAESTER dramatically increased the coverage of mitochondrial transcripts compared to scRNA-seq data alone (mean coverage per cell was 0.2- to 0.7-fold for RNA-seq and 52- to 217-fold for MAESTER; Fig. 1d and Supplementary Fig. 7a–d), enabling reliable mtDNA variant calling with high confidence across many of the transcripts. Using uniform manifold approximation and projection (UMAP) for dimensionality reduction, the two cell populations cluster based on mRNA expression data (Fig. 1e). MAESTER enabled the identification of six homoplasmic mtDNA variants that distinguished between cell types (Fig. 1e–g and Supplementary Fig. 7e,f). Combining data from all six informative variants cleanly separated cell types and demonstrated 100% concordance with mRNA clusters (Supplementary Fig. 7g,h). Of note, MAESTER identified the same six variants in the Seq-Well and 10x libraries (Supplementary Fig. 8a–d).

To benchmark MAESTER's ability to identify clonal structure at a more granular resolution, we performed a clonal expansion experiment. One hundred K562 cells were plated and allowed to expand for 14 days (doubling time ~24h), followed by scRNA-seq with MAESTER. We identified 21 informative mtDNA variants that revealed clonally related populations of K562 cells (Fig. 1h and Supplementary Fig. 8e), which were validated by orthogonal bulk ATAC sequencing (ATAC-seq) (Supplementary Fig. 8f). These data demonstrate the faithfulness of mtDNA variants enriched from mtRNA and the capacity of this method to resolve subpopulations within closely related cells.

We next applied MAESTER to derive clonal structure within primary human patient specimens. We first used a bone marrow aspirate from a patient with clonal hematopoiesis. The clonal hematopoiesis had evolved into blastic plasmacytoid dendritic cell neoplasm (BPDCN), as the patient had skin tumors at the time of collection. However, the concurrent bone marrow aspirate we utilized showed no tumor involvement (Methods). We performed 10x single-cell sequencing with MAESTER on this bone marrow aspirate and identified 9,346 high-quality cells, including all expected cell types, with an abundance of cytotoxic T lymphocytes (CTLs), likely due to hemodilution of the bone marrow sample with peripheral blood and possibly related to his evolving malignancy (Fig. 2a and Supplementary Fig. 9). We found that MAESTER coverage largely depends on the mtRNA content per cell (Supplementary Fig. 10) and tested different thresholds to select informative variants (Supplementary Fig. 11). We plotted the largest and most distinct 23 clones using 26 informative mtDNA variants (14.9% of cells were

assigned to these clones; Fig. 2b), indicating MAESTER can resolve clonal populations in primary human specimens.

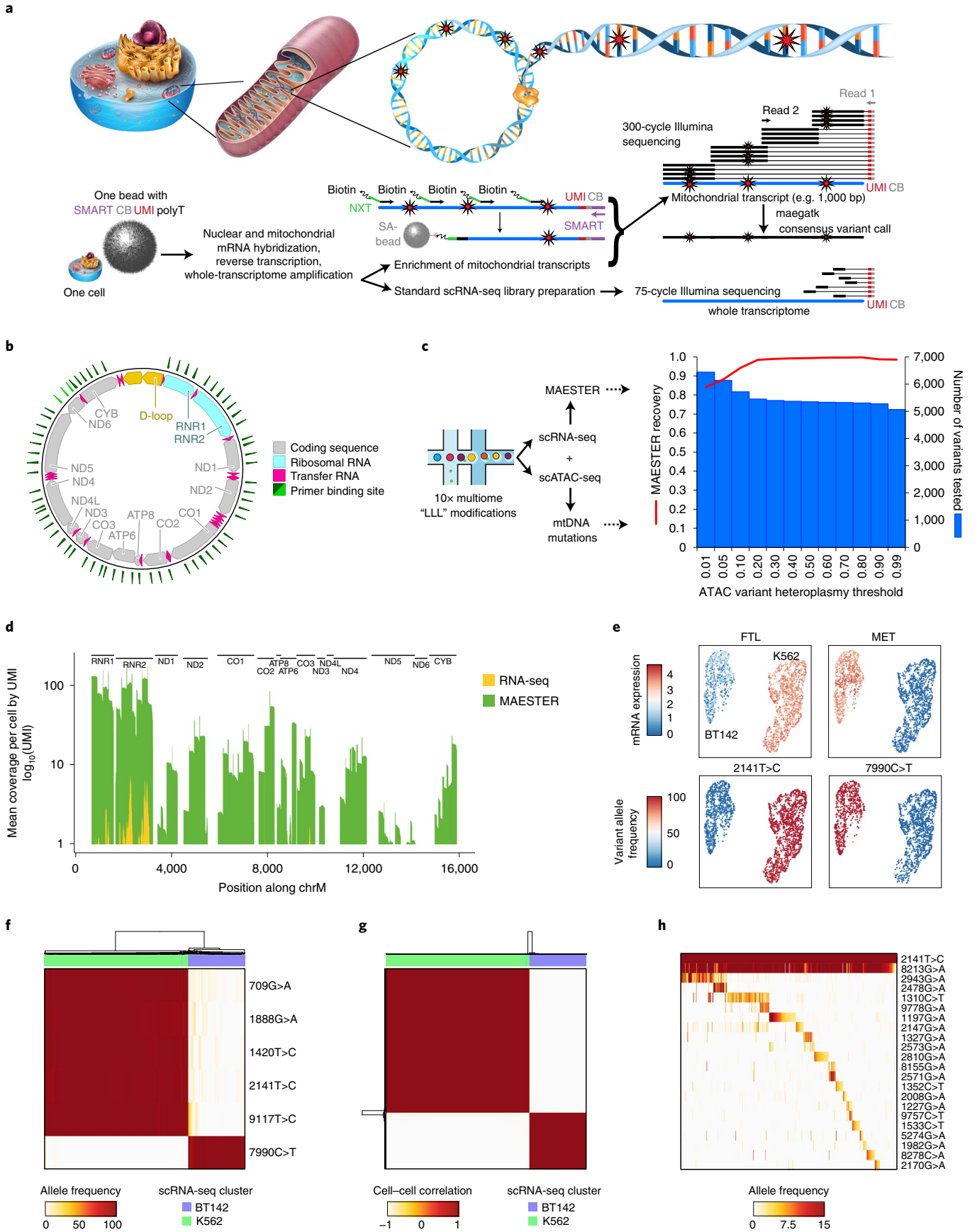
Many of the mtDNA clones clustered together in the RNA-based UMAP (Fig. 2c and Supplementary Fig. 12). Indeed, we found that many mtDNA clones were lineage biased, with 9 of 23 clones skewed toward CTLs and 2 clones with myeloid lineage bias (Fig. 2d and Supplementary Fig. 13).

The abundance of T cells in the sample provided an opportunity to validate the mtDNA clones with an orthogonal assessment of clonality using the T-cell antigen receptor (TCR) variable region. Building on a TCR enrichment method for Seq-Well¹⁴, we developed a protocol for TCR sequencing from 10x 3' scRNA-seq cDNA. We termed the protocol T-cell receptor enrichment to link clonotypes by sequencing (TREK-seq; Supplementary Fig. 14) and applied this to the bone marrow sample, adding an additional modality to the same single cells (Fig. 2e). *TRA* and *TRB* variable regions were detected in T cells, but not other lineages, and were highly concordant, confirming reliable TCR enrichment (Supplementary Fig. 15). When comparing *TRB* variable regions to mtDNA variants, we noted high overlap of the orthogonal clonal markers ($ARI=0.74$; Fig. 2f,g). The mtDNA clones that were skewed towards CTLs (e.g., 6205G>A–9164T>C), suggesting the mtDNA mutation occurred after TCR rearrangement, were largely restricted to a single T-cell state (Fig. 2h). In contrast, mtDNA clones with all hematopoietic cell types (e.g., 2593G>A), indicating the mtDNA mutation occurred within a multipotent hematopoietic stem cell, contributed to multiple T-cell states and clonotypes (Fig. 2h). Combining MAESTER, TCR sequencing and transcriptional states (Fig. 2i) can provide independent validation of clonal relationships and new opportunities to study T-cell biology.

We also identified two clones with myeloid lineage bias, identified by mtDNA alterations 2593G>A and 6243G>A (Supplementary Figs. 13 and 16a). Given the patient's clonal hematopoiesis, we sought to understand if these represented expanded clones. We utilized Genotyping of Transcriptomes (GoT)⁵ to identify the patient's known *ASXL1* and *TET2* loss-of-function mutations in single cells. We found that cells within the two myeloid-biased clones contained a high fraction of mutated transcripts (Supplementary Fig. 16b). Of cells in the 2593G>A clone, 44% and 40% had *TET2.S792X* and *TET2.Q1034X* mutations, respectively. This is consistent with biallelic *TET2* inactivation, a recurrent feature in myeloid malignancies¹⁵. No mutated transcripts were identified in cells from the other 21 clones we identified, providing evidence that the myeloid-biased clones identified by MAESTER represent cells derived from the patient's clonal hematopoiesis.

GoT only captured wild-type or mutant transcripts in 3.5% of all 9,346 cells for *TET2* and 0.4% for *ASXL1*. This relative lack of genotyping efficiency is related to gene expression, variant position and amplicon size and is similar to other protocols that genotype somatic

Fig. 1 | Targeted enrichment of mitochondrial transcripts enables discrimination between genetic clones. **a**, Schematic showing the procedures for lineage inference from single-cell transcriptomes using MAESTER. Following mRNA capture and whole-transcriptome amplification, part of the cDNA is used for standard scRNA-seq, and another part is used for PCR-based enrichment of mitochondrial transcripts. The 300-bp sequencing reads maximize mitochondrial genome coverage to call variants. **b**, Diagram depicting the circular mitochondrial genome with annotated features. The green triangles indicate where MAESTER primers bind. **c**, Barplot showing the number of mtDNA variants that were detected by ATAC-seq (blue bars) and their recovery by MAESTER (red line). DNA (ATAC) and RNA (MAESTER) were acquired from the same K562 cells using the 10x multiome workflow. **d**, Barplot showing coverage of the mitochondrial genome with and without amplification from Seq-Well libraries using MAESTER. Mean coverage of 2,482 K562 and BT142 cells is shown. Each transcript (UMI) was sequenced at least three times. **e**, UMAPs showing the detection of cell type-specific expression of *FTL* and *MET* genes from scRNA-seq (top) and homoplasmic mtDNA variants from MAESTER (bottom). **f**, Heatmap depicting separation of 1,523 K562 and BT142 cells (columns) based on six mtDNA variants (rows). Cell-type annotation from scRNA-seq is shown (top). **g**, Correlation matrix showing cell similarity based on the allele frequencies of six homoplasmic variants (rows and columns depict 1,523 cells). Unsupervised clustering identified two clusters that correlate with cell annotations from scRNA-seq. **h**, Heatmap showing variant allele frequency (VAF) of 21 mtDNA variants detected by MAESTER (rows) for 588 K562 cells (columns, 44.4% of all K562 cells) with informative variants. Homoplasmic K562 variant 2141T>C is shown for comparison. Heatmap is organized by clonal structure (Methods). For **e–h**, only cells with more than threefold coverage of the indicated variants are shown. SMART and NXT are specific primer binding sequences; SA, streptavidin.



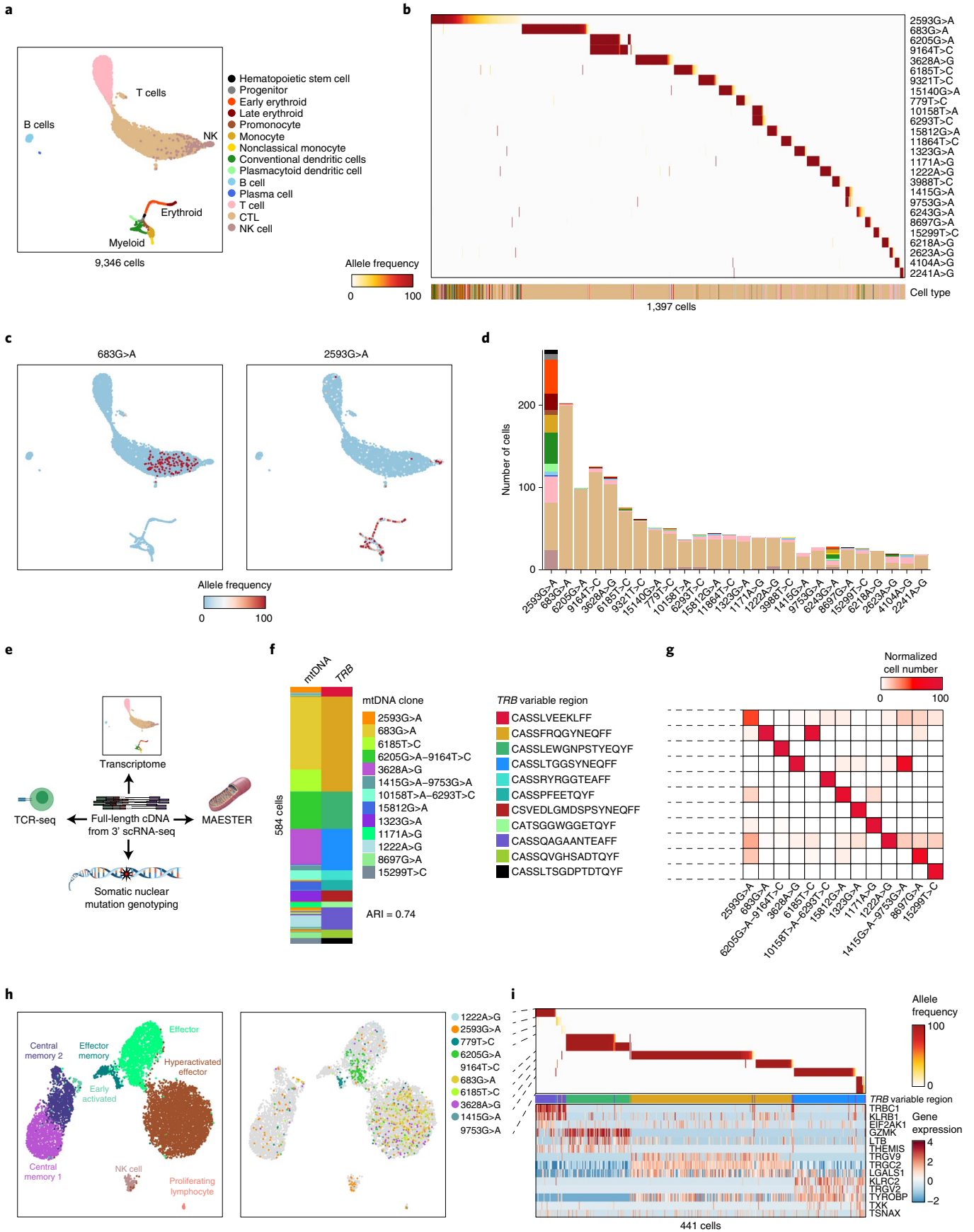


Fig. 2 | Genetic clones exhibit lineage bias in clonal hematopoiesis. **a**, UMAP of all 9,346 cells profiled by 10x scRNA-seq from a bone marrow aspirate from an individual with clonal hematopoiesis. **b**, Heatmap showing VAF of 26 informative mtDNA variants detected by MAESTER and maegatk (rows) for 1,397 cells (columns) with at least 1% VAF for one of the 26 variants. Heatmap is organized by clones and sorted by clone size. Cell type is listed on the bottom by color according to legend in **a**. **c**, UMAPs displaying VAF in each cell for mtDNA variants 683G>A (left) and 2593G>A (right). **d**, Stacked bar graph of the number of cells in each clone, with cell type denoted by color according to legend in **a**. **e**, Schematic depicts multimodal analysis we performed on the same single cells. **f**, Plot shows cells (rows) in which both mtDNA and *TRB* clonal markers were detected. Clones are indicated by colors and defined by mtDNA variants and the *TRB* variable region, respectively. **g**, Confusion matrix shows concordance between mtDNA clones and *TRB* clonotypes. **h**, UMAPs of 8,382 T cells in the clonal hematopoiesis sample, with state annotated by transcriptional signatures (left) and selected mtDNA clones (right). **i**, Heatmap of T cells (columns) in the selected mtDNA clones with mtDNA VAF (top), *TRB* sequence (middle) and state-defining transcript expression (bottom). ARI, adjusted Rand index; NK, natural killer.

mutations from high-throughput scRNA-seq libraries⁴. In contrast, MAESTER captured the mtDNA genotype at the 2593 position in 21,767 transcripts in 1,396 cells (99.9% of cells, an average of 16 transcripts per cell). Combining driver mutation with mtDNA variant detection facilitates phylogeny reconstruction (Supplementary Fig. 16c,d). This allowed us to explore cells marked by 2593G>A as clonally expanded cells with loss of *TET2*.

To interrogate this population further, we compared the myeloid differentiation trajectory of 2593G>A cells to other cells in the bone marrow using pseudotime analysis (Supplementary Fig. 16e) (ref. ¹⁶). We found the clonal population was skewed towards less mature cell types, consistent with hematopoietic stem cell expansion observed in *Tet2* knockout mice¹⁷. Although analysis of more cells and biological replicates is required to generalize these results, our data suggest the utility of MAESTER to identify and investigate premalignant cell expansions.

Finally, we applied MAESTER to a primary solid tumor tissue to demonstrate compatibility with complex tissues requiring cell dissociation. In tumor and peripheral blood samples from a patient with glioblastoma, we identified a mtDNA deletion in malignant cells that was absent in the blood, and we found region-specific clonal populations of malignant cells (Supplementary Fig. 17a–e). We also discovered tumor-associated myeloid cells that were derived from peripheral blood cells in the patient (Supplementary Fig. 17f).

In conclusion, MAESTER enables mtDNA variant detection in high-throughput 3'-biased scRNA-seq data, which was previously limited to ATAC-seq or full-length scRNA-seq. As with all methods that utilize mtDNA mutations to infer clonal relationships, there are limitations inherent to mitochondrial biology. It is currently not possible to track each cell division, as single-molecule mutations typically cannot be detected above background due to their low mtDNA heteroplasmy. For MAESTER, the VAF needs to reach >1% for confident detection. The mtDNA copy number per cell and rate of cell proliferation impact the time it takes to reach 1% VAF. Further improvements in mutation detection efficiency (mitochondrial or nuclear) will enable increasingly granular studies of clonal dynamics. Studies that require short-term, per-division tracking still require tunable and engineered lineage-tracing methods. In contrast, mtDNA variants are suitable to determine clonal relationships between subsets of cells that are more divergent, providing a tool to study in vivo cellular dynamics and human biology. In addition, due to the widespread use of 3'-biased scRNA-seq, the development of MAESTER makes mtDNA variant detection accessible to more research laboratories and a wider range of experimental contexts. The accompanying maegatk software uses UMIs to increase confidence in mtDNA variant calls, an advance over previous methods. MAESTER can be implemented on new or prior scRNA-seq datasets by using the amplified cDNA that is stored as a standard practice. The high-throughput nature of 3'-biased scRNA-seq and MAESTER enables the study of clonal relationships and evolutionary dynamics of cells within complex primary human tissues. The combination of MAESTER with other modalities such as TCR sequencing, nuclear mutation detection and RNA-seq creates synergies

that enable analyses and discoveries that are not possible with each method alone. By developing MAESTER, we democratize and expand the use of naturally occurring barcodes created by mtDNA alterations to enable discoveries in human biology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-022-01210-8>.

Received: 6 March 2021; Accepted: 6 January 2022;

Published online: 24 February 2022

References

- Giladi, A. & Amit, I. Single-cell genomics: a stepping stone for future immunology discoveries. *Cell* **172**, 14–21 (2018).
- Acosta, J., Ssozi, D. & van Galen, P. Single-cell RNA sequencing to disentangle the blood system. *Arterioscler. Thromb. Vasc. Biol.* **41**, 1012–1018 (2021).
- Nam, A. S. et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature* **571**, 355–360 (2019).
- van Galen, P. et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281 (2019).
- Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
- Liggett, L. A. & Sankaran, V. G. Unraveling hematopoiesis through the lens of genomics. *Cell* **182**, 1384–1400 (2020).
- Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339 (2019).
- Lareau, C. A. et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* **39**, 451–461 (2021).
- Xu, J. et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife* **8**, e45105 (2019).
- Velten, L. et al. Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. *Nat. Commun.* **12**, 1366 (2021).
- Hughes, T. K. et al. Second-strand synthesis-based massively parallel scRNA-seq reveals cellular states and molecular features of human inflammatory skin pathologies. *Immunity* **53**, 878–894 (2020).
- Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
- Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
- Tu, A. A. et al. TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures. *Nat. Immunol.* **20**, 1692–1699 (2019).
- Abdel-Wahab, O. et al. Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. *Blood* **114**, 144–147 (2009).
- Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
- Moran-Crusio, K. et al. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Methods

Cell lines and culturing. Human chronic myelogenous leukemia K562 cells (ATCC CCL-243) were cultured in RPMI 1640 Medium with GlutaMAX (Gibco, 61870127), supplemented with 10% fetal calf serum and penicillin–streptomycin. The BT142 gliomasphere line¹⁸ (ATCC, ACS-1018) was maintained in Neurobasal media supplemented with 20 ng ml⁻¹ recombinant EGF (R&D Systems), 20 ng ml⁻¹ FGF2 (R&D Systems), 1× B27 supplement (Invitrogen), 0.5× N2 supplement (Invitrogen), 3 mM L-glutamine and penicillin–streptomycin¹⁹; 25% conditioned media was carried over each passage. Cultures were confirmed to be mycoplasma-free, and their identity was verified by short tandem repeat analysis. K562 and BT142 cells from the same passage were used for Seq-Well scRNA-seq plus MAESTER, 10x 3' v3 scRNA-seq plus MAESTER and bulk ATAC-seq.

Primary human samples. The patients in this study consented to all study procedures under Dana-Farber Cancer Institute institutional review board-approved research protocols. The patient with clonal hematopoiesis had a history of leukopenia and thrombocytopenia. The bone marrow sample we analyzed was an aspiration in the context of evaluation for skin-only BPDCN. Histologic evaluation of the concurrent bone marrow core biopsy sample was normal and did not show involvement of malignant BPDCN cells. Targeted sequencing of the bone marrow aspirate identified alterations in *ASXL1* and *TET2* indicating clonal hematopoiesis (Supplementary Fig. 9B). Mononuclear cells were isolated from a bone marrow aspirate by density centrifugation and cryopreserved with 10% dimethylsulfoxide in liquid nitrogen. Cells were thawed using standard procedures; since viability (independently assessed by Trypan and propidium iodide staining) exceeded 90%, unsorted cells were used for scRNA-seq using the 10x 3' v3 protocol. A high proportion of CTLs were recovered, consistent with an expansion of large granular lymphocytes in the peripheral blood of this patient as demonstrated by routine clinical evaluation and confirmatory flow cytometry (Supplementary Fig. 9A). This specimen was likely hemodiluted with a contribution from the peripheral blood, as the bone marrow core biopsy sample did not contain this high fraction of T cells. The scRNA-seq data from this sample was also used in an independent study not involving the MAESTER technique and is under consideration elsewhere (patient 10, Griffin et al., manuscript under review).

scRNA-seq. For Seq-Well S³ experiments, cells were processed as described previously¹¹. A complete, updated protocol for Seq-Well S³ is hosted on the Shalek lab website (www.shaleklab.com). Briefly, an array with ~90,000 nanowells was first loaded with barcoded mRNA capture beads, and then 10,000–15,000 cells were added dropwise onto the surface of the array. After cells were allowed to settle into the wells, the array was sealed with a semipermeable polycarbonate membrane. Cells were lysed, and mRNA transcripts were hybridized to the bead contained within the same well at the polyT sequence of the barcoded oligonucleotides. The beads were then used to generate cDNA via reverse transcription. A second-strand synthesis step using a random octamer was performed to recover transcripts in which template switching during reverse transcription was not successful. Whole-transcriptome amplification (WTA) PCR was performed, and the product underwent a combination of tagmentation and PCR to generate dual-indexed sequencing libraries. Libraries were sequenced using a 75-cycle kit on the Illumina NextSeq 500 with custom read 1 (CRIP) and custom i5 primers (SW-Ci5P, Supplementary Table 2), 20 cycles for read 1 (CB + UMI), 56 cycles for read 2 (transcript sequence) and 2 × 8-bp library barcodes.

For 10x Genomics experiments, we used 3' Single Cell Gene Expression v3 reagents, following all manufacturer's recommendations. Briefly, 5,000 cells were loaded per well and captured in gel bead-in emulsions. Captured mRNAs were reverse transcribed into cDNAs and amplified to generate WTAs. Library construction involves fragmentation, adapter ligation and a sample index PCR. Libraries were sequenced using a NovaSeq SP 100-cycle kit with 28 cycles for read 1 (CB + UMI), 91 cycles for read 2 (transcript sequence) and an 8-bp library barcode.

For the cell line mixing experiments, we analyzed cells from the same passage using two Seq-Well S³ arrays and two 10x 3' v3 wells, yielding a similar number of cells and data quality. For the clonal hematopoiesis sample, we used four 10x 3' v3 wells.

Mitochondrial alteration enrichment for MAESTER from Seq-Well or

Drop-seq. Similar to a method we initially developed for the detection of somatic mutations⁵, the starting material for targeted amplification of mtDNA transcripts is the product of the Seq-Well WTA reaction (only a fraction of which is used for scRNA-seq). The general method consists of two PCR reactions with a streptavidin bead enrichment in between (Supplementary Figs. 1, 2). The first PCR reaction serves to add a biotin tag and Nextera adapter (NEXT) to mitochondrial transcripts while retaining the UMI and CB of the transcripts. The second PCR is used to append Illumina adapters (P5 and P7), dual index barcodes to identify the sample, and sequencing primer binding sites.

PCR1. We designed biotinylated primers to tile across the entire mitochondrial transcriptome. Twelve primer mixes were created using 2–11 of these primers

at a concentration of 1 μM each (10-fold relative to the final concentration). The SMART-AC primer, which is common to all PCR1 reactions, was included in each primer mix at 10 μM (Fig. 1b, Supplementary Fig. 4 and Supplementary Table 2).

As a template, WTA products from an individual sample were pooled and diluted to be used at 20 ng in a total volume of 10 μl per reaction. Next, 2.5 μl of the primer mix and 12.5 μl of KAPA HiFi Hotstart ReadyMix (Thermo Fisher Scientific, KK2602) were added to the template, and PCR was performed using the following conditions: initial denaturation at 95 °C for 3 min, followed by six cycles of 98 °C for 20 s, 65 °C for 15 s and 72 °C for 3 min, ending with a final extension at 72 °C for 5 min. There were 12 reactions in total for each sample, as each primer mix is used in a single reaction.

Following amplification, the PCR product is pooled and purified with 0.8× AMPure XP beads (Beckman Coulter, A63881). Pooling ratios of PCR1 products were empirically determined to obtain a more equal distribution of reads across the mitochondrial transcriptome (Supplementary Fig. 4d). Using Streptavidin-coupled Dynabeads, only biotinylated fragments containing the amplicons of interest were captured (following the manufacturer's instructions; Thermo Fisher Scientific, 60101). Dynabeads–DNA complex was eluted in 23 μl H₂O and used as a template for the second PCR.

PCR2. To add Illumina adapters (P7 and P5), index barcodes to identify the library (i7 and i5) and sequencing primer binding sites to the fragments, a second PCR was performed using 23 μl streptavidin-bound template, with 2 μl of a 5 μM primer mix (N70D_P7_BCXX and N70_P5_BCXX; Supplementary Table 2) and 25 μl plaque-forming units Ultra II HS 2× Master Mix (Thermo Fisher Scientific, Q32854). The parameters used for PCR2 were an initial denaturation at 95 °C for 2 min; six cycles of 95 °C for 20 s, 65 °C for 20 s and 72 °C for 2 min; and a final extension at 72 °C for 5 min. After the second PCR, the streptavidin beads were magnetized to collect the supernatant, from which DNA was purified with 0.7× AMPure XP beads. After elution in 22 μl TE buffer (10 mM Tris, 1 mM EDTA), the supernatant was transferred to a new tube and saved for sequencing.

The resulting libraries are similar to Seq-Well scRNA-seq libraries but with targeted integration of the sequencing primer binding site at the regions of interest. The libraries were generally 2–10 ng μl⁻¹, with sizes ranging from 250 to 1,000 bp. Libraries were sequenced on the Illumina NovaSeq SP 300 cycle kit with the forward strand workflow and the CRIP primer, using 20 cycles for read 1, 264 cycles for read 2 and 2 × 8-bp index barcodes.

Mitochondrial alteration enrichment for MAESTER from 10x Genomics.

Enrichment of mitochondrial transcripts from 10x Genomics 3' v3 cDNA was very similar to the protocol for Seq-Well or Drop-Seq described above. The main differences were the use of primer sequences specific to 10x and the omission of the biotin enrichment step (Supplementary Figs. 1 and 3).

PCR1. We designed primers to tile across the entire mitochondrial transcriptome. Twelve primer mixes were created using 2–11 of these primers at a concentration of 1 μM each (10-fold relative to the final concentration). A barcoded GoT-P5-i5-BCXX primer was included in each primer mix at 10 μM for sample indexing (Supplementary Fig. 4 and Supplementary Table 3).

As a template, cDNA products from an individual sample were pooled and diluted to be used at 20 ng in a total volume of 16 μl per reaction. Next, 4 μl of the primer mix and 20 μl KAPA HiFi Hotstart ReadyMix (Thermo Fisher Scientific, KK2602) were added to the template, and PCR was performed using the following conditions: initial denaturation at 95 °C for 3 min followed by six cycles of 98 °C for 20 s, 65 °C for 15 s and 72 °C for 3 min and ending with a final extension at 72 °C for 5 min. There were 12 reactions in total for each sample, as each primer mix is used in a single reaction.

Following amplification, the PCR product is pooled and purified with 1× AMPure XP beads to remove primers (Beckman Coulter, A63881). Pooling ratios of PCR1 products were empirically determined to obtain a more equal distribution of reads across the mitochondrial transcriptome (Supplementary Fig. 4d; all volumes multiplied by 1.6). After AMPure XP purification, the pooled PCR1 product was eluted in 20 μl H₂O.

PCR2. To add Illumina adapters (P7 and P5), index barcodes to identify the library (i7 and i5) and sequencing primer binding sites to the fragments, a second PCR was performed using 18 μl of the eluate, with 2 μl of a 5 μM primer mix (P5-generic and XV-P7-i7-BCXX; Supplementary Table 3) and 20 μl KAPA HiFi Hotstart ReadyMix (Thermo Fisher Scientific, KK2602). The parameters used for PCR2 were an initial denaturation at 95 °C for 3 min; six cycles of 98 °C for 20 s, 60 °C for 30 s and 72 °C for 3 min; and then a final extension at 72 °C for 5 min. After the second PCR, the DNA is purified with 0.8× AMPure XP beads. The DNA is eluted in 20 μl TE, and the supernatant is transferred to a new tube and saved for sequencing.

The resulting libraries are similar to 10x scRNA-seq libraries but with targeted integration at the regions of interest. The libraries were generally 2–100 ng μl⁻¹ with sizes ranging from 300 to 1,500 bp. Libraries were sequenced on the Illumina NovaSeq SP 300 cycle kit with 28 cycles for read 1, 256 cycles for read 2 and 2 × 8-p index barcodes. For the NovaSeq Forward Strand Workflow, no custom

sequencing primers are required. For the NovaSeq Reverse Complement Workflow, custom index primers should be used instead of the Illumina standards (10x-Ci7P and 10x-Ci5P; Supplementary Table 3).

10x Multiome sequencing. To assess the recall of MAESTER in identifying mitochondrial variants, we conducted an experiment to genotype both mtDNA and mtRNA in the same individual cells. We performed DOGMA-seq¹³ with LLL lysis on peripheral blood mononuclear cells from a consented healthy donor. As DOGMA-seq utilizes the 10x Genomics Multiome ATAC + Gene Expression kit to capture both DNA (via ATAC) and RNA from the same individual cells, our experimental framework provided the means to verify the detection of mutations on mtDNA via RNA. Though this lysis buffer yielded a low mtRNA copy number as previously described¹³, we amplified mitochondrial transcripts from the full-length cDNA using MAESTER. We sequenced the corresponding ATAC (containing mtDNA), gene expression and amplified mtRNA libraries separately.

Bulk ATAC-seq. K562 cells from the same passage used for the cell line mixing experiments were analyzed by bulk ATAC-seq for orthogonal validation of mtDNA variants. Cells were washed in PBS and pelleted by centrifugation, and ~12,000 cells were lysed and tagged in 1× TD buffer, 2.5 µl Tn5 (Illumina), 0.1% NP40 and 0.3× PBS in a 50-µl reaction volume as described previously²⁰. Samples were incubated at 37°C for 30 min at 300 rpm. Tagmented DNA was purified using the MinElute PCR kit (Qiagen). The complete eluate underwent PCR with initial extension and five cycles of preamplification using indexed primers and NEBNext High-Fidelity 2X PCR Master Mix (New England Biolabs). Then, the number of additional cycles was assessed by quantitative PCR using SYBR Green. Seven additional cycles were run. The final library was purified using a MinElute PCR kit (Qiagen). Libraries were sequenced on a NextSeq 500 instrument with paired-end 38-bp reads and dual library indices of 8 bp each.

TREK-seq. We adapted a previously described TCR sequencing protocol¹⁴, developed for Seq-Well, for use with 10x Genomics 3' v3 cDNA libraries (Supplementary Fig. 14). The modifications to the original protocol are as follows: in the TCR enrichment master mix, we added PartialRead1 and PartialTSO primers at a final concentration of 1.25 µM each (Supplementary Table 3). For amplification of TCR transcripts following enrichment, we used the same primers at a final concentration of 0.4 µM each. For the final PCR, to add the Illumina P5 and P7 sequences, we used UPS2-N70x and 10X_SI-PCR_P5 primers at a final concentration of 0.2 µM each. The libraries were sequenced using a 150-cycle kit on the Illumina MiSeq loaded at a final DNA concentration of 10 pM, aiming for a cluster density of roughly 450,000 per mm². A total of 28 cycles were used for read 1, which reads the CB and UMI; 150 cycles were used for index 1, which reads the TCR region. TCR-α and TCR-β-specific custom sequencing primers were used for index 1 at a final concentration of 2.5 µM (aTCR-Seq and bTCR-Seq, Supplementary Table 3).

scRNA-seq read processing. For Seq-Well, sequencing data were demultiplexed using bcl2fastq2. Read 1 yielded 20-bp reads (12-bp CB and 8-bp UMI), read 2 yielded 56-bp reads (transcript sequence) and i7 and i5 indices to identify the library were 8 bp each. Reads associated with CBs occurring less than 100 times were removed, and the list of remaining CBs was used to generate read 2 fastq files in which the library barcode, CB and UMI were appended to the read identifier. For 10x Genomics data, scRNA-seq data were processed using cellranger mkfastq to demultiplex into fastq files and cellranger count to quantify gene expression.

To generate the reference genome, we used hg38 sequences and annotations (version 99) from Ensembl with the addition of RNA18S and RNA28S annotations from the University of California, Santa Cruz. Annotations were filtered using cellranger mkgtf with recommended attributes as well as the gene biotypes gene_biotype: Mt_rRNA and gene_biotype: rRNA. The reference genome was then generated with cellranger mkref, which includes STAR indexing. To align Seq-Well scRNA-seq data to this reference, we used STAR with the options `--outSAMtype BAM SortedByCoordinate` and `--quantMode TranscriptomeSAM`. To align 10x Genomics scRNA-seq data to this reference, we used cellranger count, which implements STAR.

MAESTER read processing. MAESTER fastqs include read 1 encompassing the CB and UMI (20 bp for Seq-Well and 28 bp for 10x), read 2 covering mitochondrial transcript sequences (264 bp for Seq-Well and 256 bp for 10x) and 2 × 8 bp for dual-indexed library barcodes. We used Illumina bcl2fastq for demultiplexing with both indices. Reads associated with CBs occurring less than 100 times were removed, and the list of remaining CBs was used to generate read 2 fastq files in which the library barcode, CB and UMI were appended to the read identifier. We trimmed the first 24 bp from these fastqs using homerTools to avoid using primer binding sequences for variant calling. Next, we aligned the fastq files with STAR (`--outSAMtype BAM SortedByCoordinate`) to the same hg38 reference genome we used for scRNA-seq alignment above. More than 90% of MAESTER reads aligned to chrM. Supplementary Fig. 5 shows an overview of these procedures.

maegatk: mitochondrial genome variant calling. To facilitate the analysis of MAESTER data, we developed maegatk, a Python package, as an extension of

our previously described mgatk pipeline⁸. maegatk specifically handles technical biases implicit in high-throughput scRNA-seq to facilitate the identification of mtDNA variants. First, maegatk takes inputs of a single-cell BAM file following the 10x Genomics SAM tag conventions, a valid list of CBs and more than 20 customizable command-line arguments. Next, the software collapses duplicate reads based on UMI, start position and CB. Unlike most existing variant calling pipelines, including GATK and mgatk that select a representative read based on highest mean base quality score, maegatk identifies the most likely consensus nucleotide across sequencing read replicates via the CallMolecularConsensusReads (v1.1) tool from fgbio. Further, maegatk provides a `--min-reads` command-line argument that specifies the minimum number of sequence reads needed for a UMI to be considered for variant calling. This workflow minimizes artifacts due to PCR amplification and sequencing error compared to the standard Picard MarkDuplicates. By calling maegatk-indel, our software enables indel calling by implementing FreeBayes on a per-cell basis, which we validated using simulated mtRNA read data (Supplementary Fig. 6). After consensus read deduplication, per-cell, per-position nucleotide counts were enumerated and used for downstream analysis.

For use in our analysis with maegatk, we ensured that all BAM files contained the CB and UMI SAM tags according to 10x conventions (CB:Z and UB:Z). We selected reads aligning to chrM and merged BAM files from scRNA-seq and MAESTER (Aligned.sortedByCoord.out.bam from STAR or possorted_genome_bam.bam from cellranger). We generated a list of CBs by intersecting CBs with ≥100 alignments to chrM and high-quality CBs from scRNA-seq. maegatk was then executed with the options `--input merged.bam --mito-genome chrM.fa --barcodes CBs.txt --min-reads 3`. The last option specifies only UMIs with three reads are used to increase confidence in variant calls. Upon completion, maegatk saves mutation calls as a maegatk.rds file in the SummarizedExperiment format²¹ for convenient intersection with other modalities and downstream analysis in R.

Multiome analysis. Multiome libraries (ATAC and gene expression) were aligned using CellRanger-Arc to a modified hg38 reference genome with regions of mitochondrial genome homology hard-masked in the nuclear genome⁸. Mitochondrial DNA from the ATAC library was processed using the mgatk workflow⁸. In parallel, we applied MAESTER and maegatk to enrich mitochondrial variants from the mRNA library using an unmodified hg38 reference genome. Next, we identified a 'gold-standard' dataset of variants from the mgatk ATAC library and examined the concordance of variants that were covered at a minimum of five times in both DNA (ATAC) and RNA (MAESTER) within individual cells. For a heteroplasmy threshold value *X* (as shown on the *x* axis of Fig. 1c), we quantified the recovery of variants by MAESTER as the number of variants that exceeded *X* in the MAESTER library over the number of variants that exceeded *X* in the ATAC library.

scRNA-seq clustering and cell-type annotation. For Seq-Well and 10x scRNA-seq data alike, we filtered for cells with ≥2,000 UMIs, ≥1,000 genes, ≤20% alignment to rRNA genes and ≤20% alignments to genes on chrM. Genes from chrX and chrY were removed from the count matrix. Next, we used Seurat for standard scRNA-seq processing steps, including the functions `NormalizeData`, `FindVariableFeatures`, `ScaleData` and `RunPCA` (similar to https://satijalab.org/seurat/archive/v3.2/pbmc3k_tutorial.html)²². We implemented graph-based clustering with `FindNeighbors` with six PCA dimensions and `FindClusters` with a resolution of 0.05–0.1, as we only aimed to distinguish K562 and BT142 cells. We determined the top 10 cell type-specific genes by fold change using the `FindAllMarkers` function with `only.pos = TRUE`, `min.pct = 0.25` and `logfc.threshold = 0.25`.

For cell line mixing experiments, we used `decontX` from the R package `celda` to remove cells with high ambient RNA²³. We supplied the `decontX` function with the count matrix to calculate per-cell contamination scores. We also scored each cell for both cell type-specific gene signatures using the Seurat function `AddModuleScore`. Finally, we excluded cells that exceeded a contamination score of 0.05 and had a high module score for both cell type-specific signatures. For Seq-Well, this removed 218 of 2,525 cells (8.6%), with 1,387 K562 and 920 BT142 cells remaining. For 10x, this removed 112 of 2,778 cells (4.0%), with 1,310 K562 and 1,356 BT142 cells remaining.

For the clonal hematopoiesis sample, we used the cell-type annotations that were established using a random Forest classifier based on healthy donor populations (Griffin et al., manuscript under review). Further annotation of T and natural killer cell subsets was done by evaluating cluster-specific gene expression using the `FindAllMarkers` function with `logfc.threshold = 0.25`, `min.pct = 0.1`, `test.use = "roc"`, `return.thresh = 0.4`, `only.pos = TRUE` (cluster-identifying marker genes are on https://github.com/petervangalen/MAESTER-2021/tree/main/5_TCR-Seq). Filtering for high-quality cells, dimensionality reduction and removal of cells with ambient RNA were performed using similar procedures as we used for the cell line mixing experiments, except for `decontX`.

Identification of informative mtDNA variants. To establish clonal relationships between cells, we first select informative variants and then select cells that are positive for any of these informative variants (that is have a VAF of >1%). For

informative variant selection, we first calculated an allele frequency matrix of all possible variants (rows) and cells (columns) from the output of the maegatk software. The total number of possible variants is $3 \times 16,569 + 1 = 49,708$, because the mitochondrial genome (NC_012920) is 16,569 bp with three possible variants each, except base 3,107, which has four possible variants (A, C, T and G), because the reference is N. Next, we generated a table with features for every variant: the mean allele frequency, mean coverage, mean quality score and the VAF in percentiles of rank-sorted cells or the number of cells exceeding a chosen VAF. This allowed us to select informative variants by applying filters such as a mean coverage of >20 , mean quality of >30 , VAF of $<10\%$ in at least 10% of cells and a VAF of $>90\%$ in at least 10% of cells to distinguish cell lines. For the clonal hematopoiesis sample, we selected variants with a mean coverage of >5 per cell, mean quality of >30 , VAF of 0% in at least 90% of cells and a VAF of $>50\%$ in at least 10 cells (Supplementary Fig. 11). This allowed us to use 26 variants to identify the 23 largest clones, which together made up 14.9% of cells. Selected informative variants were highly enriched for transitions vs transversions, as expected ($>89\%$). Additional filters to remove artifacts were included for subclone identification (e.g., variants informing K562 subclones should be absent in BT142 cells). The classification of K562 and BT142 cells by mtDNA variants was determined by the sum of calls at all homoplasmic variants for either cell line (Supplementary Fig. 7g,h).

Having identified informative variants, we assessed their VAFs in single cells by UMAP visualization (Figs. 1e and 2c and Supplementary Figs. 8g and 12). We used a VAF threshold of 1% to consider a cell positive for an mtDNA variant. We combined highly correlated variants into groups or clones (e.g., 10158T>A and 6293T>C; Fig. 2b). To visualize clonal structures in VAF heatmaps, we sorted clones (rows) by their size, and we sorted cells (columns) within each clone from high to low VAF or clustered the cells by Pearson correlation (Figs. 1h and 2b,i and Supplementary Figs. 8e and 11c).

Bulk ATAC-seq analysis. For the analysis of ATAC-seq as an orthogonal validation of the presence of mtDNA variants in K562 cells, we aligned demultiplexed fastq files to hg38 using STAR. We subsetted the BAM file for alignments to the mitochondrial genome and used Picard-Tools MarkDuplicates to remove duplicates. Next, we ran BAM-readcount with the option `-w 5` to generate a table of read metrics for every position along the mitochondrial genome. From this table, we extracted informative K562 variants that were identified by MAESTER. We then compared the VAF determined from mRNA transcripts captured by MAESTER to the VAF determined from mtDNA fragments captured by bulk ATAC-seq, the latter being calculated as the number of reads supporting the variant allele over the total sequencing depth at that position.

TREK-seq analysis. Following sequencing of enriched TCR regions, the MiSeq run was demultiplexed using bcl2fastq v2.20.0.422 and a SampleSheet with 150xN as the index sequence. CDR3 sequences were aligned as outlined previously¹⁴. Briefly, Hamming errors in CBs were repaired using a whitelist of CBs generated from whole-transcriptome sequencing data and a single-base error tolerance. UMIs for each CB were then collapsed with a single-base error tolerance. The repaired reads were then aggregated by CB and UMI, and UMIs with fewer than 10 total reads were discarded. The remaining reads were mapped against TCRV and TCRJ IMGT (<http://imgt.org/>) reference sequences with IgBlast. CDR3 sequences were called by identifying the 104-cysteine and 118-phenylalanine according to IMGT references and translating the amino acid sequences in between. UMIs with a V-gene consensus frequency of less than 0.9 were discarded. Processed TCR sequences were paired with the single-cell transcriptome data by matching of CBs. We recovered *TRB* in 5,288 (63.1%) and *TRA* in 4,175 (49.8%) of the total 8,382 T cells. If multiple *TRA* or *TRB* sequences were detected for a single CB, then the corresponding sequence with the highest number of UMIs and raw reads was retained. Next, we stringently selected *TRB* calls present in at least 10 cells and subsequently selected *TRA* calls present in at least 10 cells. We visualized the overlap of cells with both *TRB* and *TRA* calls and calculated the ARI using the R package mclust (Supplementary Fig. 15). To test overlap with mtDNA variant calls, we selected *TRB* clones with an mtDNA variant in at least five cells and vice versa. These filtered *TRB* and mtDNA clonal determinations are visualized in Fig. 2f,g and intersected with cell-type classification and gene expression in Fig. 2h,i.

Genotyping of transcriptomes analysis. To analyze reads from the GoT analysis for the clonal hematopoiesis sample, we utilized IronThrone-GoT (<https://github.com/landau-lab/IronThrone-GoT>)³. We provided the IronThrone-GoT function with fastq files and a whitelist of CBs for 10x 3' v3 scRNA-seq. From the summTable, we selected high-confidence transcript calls by filtering for UMIs that were sequenced $\geq 3\times$ with $\geq 3\times$ more wild-type than mutant calls or vice versa. We then generated a table with CBs, the number of wild-type transcripts and the number of mutant transcripts per cell. These calls were intersected by CBs with cells of different subclones, as shown in Supplementary Fig. 16b-d.

Pseudotime analysis. We assigned pseudotime values to cells using the R slingshot library¹⁶. We selected relevant cell types for four differentiation trajectories from the clonal hematopoiesis sample and provided the slingshot function with the UMAP coordinates. The predicted trajectory (black curve) and assigned

pseudotime values (colors) are shown in Supplementary Fig. 16e, which is an enlargement of Fig. 2a. The same pseudotime values are used for the horizontal axis of adjacent density plots.

Functional annotation of mitochondrial variants. Mitochondrial genes encode factors involved in the respiratory chain. Variants that impact function are thus expected to alter energy metabolism. Because genes required for mitochondrial biogenesis and replication reside in the nuclear genome, mtDNA variants are less likely to impact the intrinsic replication rate of the mitochondrion in which the mutation occurs and hence the heteroplasmy level. However, beneficial/detrimental mutations could affect the fitness of the host cell, and the potential functional impact of variants that are used to establish lineage relationships should be assessed in this context. We generated a table of mitochondrial variants, their effects on the resultant proteins and potential functional consequences by merging the Ensembl Variant Effect Predictor (VEP) (<https://useast.ensembl.org/info/docs/tools/vep/index.html>) and MitoMap (<https://www.mitomap.org/MITOMAP>). First, a five-column table of variants was generated using the reference mitochondrial genome (https://www.genome.jp/dbget-bin/www_bget?-f+refseq+NC_012920), with columns as follows: chromosome number (MT), starting nucleotide, ending nucleotide, substitution (e.g., A/T) and strand (+), as described elsewhere (<https://useast.ensembl.org/info/website/upload/var.html>). This format is used as the default input for VEP. Second, VEP was run using the variant table as input via the web interface to predict the effect of variants on genes, proteins and regulatory regions. Additional configurations included 'Protein' and 'UniProt' (identifiers), 'Identify canonical transcripts' (transcript annotation) and 'Protein domains' (protein annotation). The same results can be acquired by running VEP on the command line with the following code:

```
./vep-af-appris-biotype-buffer_size 500-canonical-
check_existing-distance 0-domains-mane-polyphen
b-protein-pubmed-regulatory-sift b-species homo_
sapiens-symbol-transcript_version-tsl-uniprot-cache-
input_file [input_data]-output_file [output_file]
```

Third, the VEP output was further processed to retain only the following columns prior to integration with MitoMap data: Uploaded_variation, Location, Allele, Consequence, SYMBOL, Gene, Feature_type, Feature, BIOTYPE, Amino_acids, Codons, SIFT and PolyPhen. Finally, disease and polymorphism data were downloaded from the API resources in MitoMap in VCF format (<https://www.mitomap.org/foswiki/bin/view/MITOMAP/Resources>). These data were merged with the VEP data using the dplyr package in R.

Software used for analysis and figures. Geneious Prime version 2019.1.3 with Primer3 was used for primer design. Read processing was performed using command-line tools, including bcl2fastq v2.20.0, Samtools version 1.8, cellranger 3.1.0, homerTools 4.10, STAR version 2.6.0c and IronThrone-GoT version 1.0. Quality controls and downstream analyses were performed with R for Statistical Computing version 3.6.1 with RStudio version 1.2.5042. We used the tidyverse version 1.3.0 collection of packages, including ggplot2 version 3.3.2, Seurat version 3.2.2, SummarizedExperiment 1.24.0 for maegatk output, celda version 1.5.6 for deconvX and slingshot version 2.2.0 for trajectory analyses as described in the provided GitHub scripts. Mclust version 5.4.9 was used for ARI analysis. maegatk is a Python package, and therefore, Python was utilized as part of maegatk. FlowJo was used for FACS analysis. We also used GraphPad Prism, Microsoft Excel version 16.56 and Adobe Illustrator 2021 for additional statistical analyses and visualization.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed data have been deposited in the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182685>). Single-cell gene expression matrices, mtDNA variant calls and GoT results are available at <https://vangalenlab.bwh.harvard.edu/resources/maester-2021/>.

Code availability

maegatk is available at <https://github.com/caleblareau/maegatk>, and a table with functional annotation of all possible mtDNA variants is available at https://github.com/EDePasquale/Mitochondrial_variants. Computational analyses are described at <https://github.com/petervangalen/MAESTER-2021>.

References

- Luchman, H. A. et al. An in vivo patient-derived model of endogenous IDH1-mutant glioma. *Neuro. Oncol.* **14**, 184–191 (2012).
- Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
- 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**, 1213–1218 (2013).

21. Morgan, M., Obenchain, V., Hester, J. & Pagès, H. SummarizedExperiment: SummarizedExperiment container. R package version 1.24.0 <https://bioconductor.org/packages/release/bioc/html/SummarizedExperiment.html> (2019).
22. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
23. Yang, S. et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).

Acknowledgements

We thank patients for donating cells; C. Couturier and M. Villanueva from the A. Shalek lab for sequencing; A. Kreso, V. Hovestadt and A. A. Tu for helpful discussions; and P. Rogers for technical support. P.v.G., A.A.L. and B.E.B. are supported by the Ludwig Center at Harvard. P.v.G. and V.G.S. are supported by the Harvard Medical School Epigenetics & Gene Dynamics Initiative. P.v.G. is supported by the National Institutes of Health (NIH) R00 Award (CA218832), Gilead Sciences, the Bertarelli Rare Cancers Fund, and is an awardee of the Glenn Foundation for Medical Research and American Federation for Aging Research (AFAR) Grant for Junior Faculty. T.E.M. is supported by the American Brain Tumor Association Basic Research Fellowship in honor of Joel A. Gingras, Jr. T.E.M. and J.A.V. are supported by the UK Brain Tumour Charities Future Leaders Award (GN-000701). C.A.L. is supported by a Stanford Science Fellowship and Parker Scholar award. A.T.S. is supported by NIH grant U01CA260852, the Cancer Research Institute Technology Impact Award and a Pew-Stewart Scholars for Cancer Research Award. L.S.L. is supported by an Emmy Noether fellowship by the German Research Foundation (LU 2336/2-1). J.C.L. and D.M.M. were supported in part by the Koch Institute Support (core) NIH grant P30-CA14051 from the National Cancer Institute, as well as the Koch Institute–Dana-Farber/Harvard Cancer Center Bridge Project and the Food Allergy Science Initiative at the Broad Institute. V.G.S. is supported by the New York Stem Cell Foundation, a gift from the Lodish family to Boston

Children's Hospital and NIH grant R01 DK103794. V.G.S. is a New York Stem Cell Foundation–Robertson Investigator.

Author contributions

T.E.M., C.A.L., J.A.V., E.A.K.D., V.L., D.S., K.S., Y.Y., C.A.E.F., D.M.M., A.T.S. and P.v.G. conducted experiments and analyzed the data. T.E.M., C.A.L., L.S.L., G.K.G., A.A.L., J.C.L., B.E.B., V.G.S. and P.v.G. designed the study and interpreted the data. T.E.M. and P.v.G. wrote the manuscript. All authors edited the manuscript.

Competing interests

B.E.B. discloses financial interests in Fulcrum Therapeutics, HiFiBio, Arsenal Biosciences, Chroma Medicine and Cell Signaling Technologies. V.G.S. serves as an advisor to and/or has equity in Novartis, Forma, Cellarity, Ensoma and Branch Biosciences. T.E.M. discloses financial interests in Telomere Diagnostics and Reify Health. A.T.S. discloses financial interests in Immunai and Cartography Biosciences. J.C.L. has interests in Honeycomb Biotechnologies. J.C.L.'s interests are reviewed and managed under the Massachusetts Institute of Technology's policies for potential conflicts of interest. J.C.L. and the Massachusetts Institute of Technology have filed patents related to the single-cell sequencing methods used in this work. A patent application covering MAESTER has been filed by the Broad Institute of MIT and Harvard. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-022-01210-8>.

Correspondence and requests for materials should be addressed to Peter van Galen.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Variant calling and processing used maegatk version 0.1.1, and is available at <https://github.com/caleblareau/maegatk> and a table with functional annotation of all possible mtDNA variants is available at https://github.com/EDePasquale/Mitochondrial_variants. Computational analyses are described on <https://github.com/vangalenlab/MAESTER-2021>.

Other code or programs used to generate figures:

Geneious Prime version 2019.1.3 with Primer3 was used for primer design. Read processing was performed using command-line tools including bcl2fastq v2.20.0, Samtools version 1.8, cellranger 3.1.0, homerTools 4.10, STAR version 2.6.0c, and IronThrone-GoT version 1.0. Quality controls and downstream analyses were performed with R for Statistical Computing version 3.6.1 with RStudio version 1.2.5042. We used the tidyverse version 1.3.0 collection of packages including ggplot2 version 3.3.2, Seurat version 3.2.2, SummarizedExperiment 1.24.0 for maegatk output, celda version 1.5.6 for decontX, and slingshot version 2.2.0 for trajectory analyses as described in the provided Github scripts. Mclust version 5.4.9 was used for ARI analysis. Maegatk is a python package and therefore python was utilized as part of maegatk. FlowJo was used for FACS analysis. We also used GraphPad Prism, Microsoft Excel version 16.56, and Adobe Illustrator 2021 for additional statistical analyses and visualization.

Data analysis

The computational analysis is described in the appropriate methods sections.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability: Raw and processed data are deposited in the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182685>). Single-cell gene expression matrices, mtDNA variant calls and GoT results are available at <https://vangalenlab.bwh.harvard.edu/resources/maester-2021/>. Hg38 and annotations (v99) were utilized for mapping.

We generated a table of mitochondrial variants, their effects on the resultant proteins, and the potential functional consequences as described in by merging two databases: the Ensembl Variant Effect Predictor (VEP) (<https://useast.ensembl.org/info/docs/tools/vep/index.html>) and MitoMap (<https://www.mitomap.org/MITOMAP>) using R. The table and code are available at https://github.com/EDePasquale/Mitochondrial_variants.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical analysis were performed to pre-determine sample size. In this technology development paper, we utilized single clinical samples in order to demonstrate the ability of the technology to work on different types of tissues. The number of single-cells obtained were chosen based on prior experience with technology development, technical limitations, and budget considerations.
Data exclusions	Samples were not excluded. All filtering of single cells from the data are described in the figure legends and methods for each experiment.
Replication	Technology was replicated using two different single-cell methodologies on the same cell line mixing sample (10X Genomics and Seq-Well S3). We also performed the technology on two different clinical samples, once with 10X and once with SeqWell. All attempts were successful. Single-cell libraries were treated as independent measurements of the biological system used.
Randomization	Sample selection was not random, as experiments and data generated were used primarily to demonstrate feasibility and quality metrics. Single-cell libraries were inherently randomized during the cell collection and library construction process.
Blinding	Sample selection was not blinded, as all tissues were used with the same technology and for the same purpose - primarily to demonstrate feasibility and quality metrics. Single-cell libraries for cell-cell comparisons were inherently blinded during the cell collection and library construction phase.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Human chronic myelogenous leukemia K562 cells was obtained from ATCC (CCL-243). BT142 gliomasphere line was obtained from ATCC (ACS-1018).
Authentication	Cell line identify was verified by STR analysis.
Mycoplasma contamination	Cells were routinely tested for mycoplasma by PCR and were negative.
Commonly misidentified lines (See ICLAC register)	None.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Only two patient samples were used in this study. The clonal hematopoiesis sample was from a 73 year old male patient with a history of leukopenia, thrombocytopenia, and skin-only BPDCN. The glioblastoma sample was from a 61 year old female patient with glioblastoma. No Clinical Trial was conducted.
Recruitment	Only two patient samples were used in this study. Samples were collected for other studies and were utilized in this manuscript due to availability of the cDNA from single-cell sequencing libraries. No specific recruitment was conducted for this study.
Ethics oversight	The patients with clonal hematopoiesis and glioblastoma in this study consented to all study procedures under Dana-Farber Cancer Institute IRB-approved research protocols.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Mononuclear cells were isolated by density centrifugation. Cells were stained with antibodies and run for flow cytometry.
Instrument	BD LSR II Flow Cytometer
Software	FlowJo software was used
Cell population abundance	Shown in the manuscript in the single supplemental figure where flow cytometry was used (Supplemental Figure 9A). Flow cytometry plots of the same bone marrow aspirate used for 10x sequencing in Figure 2. Single lymphocytes were determined by forward and side scatter. CD2+, CD3+ cells represent T-cells (92.6%). CD2+, CD3- cells represent NK cells (2.6%). Of T-cells, 66% were CD8+.
Gating strategy	Shown in the manuscript in the single supplemental figure where flow cytometry was used (Supplemental Figure 9A). Single lymphocytes were determined by forward and side scatter. T-cells were gated on CD2 and CD3 expression, and then CD4 and CD8 expression.

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.