Scalable Approaches for Inferring Chromatin States and Lineages of Human Cells

Scalable approaches for inferring chromatin states and lineages of human cells

A DISSERTATION PRESENTED

BY

Caleb Andrew Lareau

TO

The Division of Medical Sciences

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

IN THE SUBJECT OF

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

April 2020

Dissertation advisor: Professor Vijay G. Sankaran Caleb Andrew Lareau

Scalable approaches for inferring chromatin states and lineages of human cells

Abstract

The human hematopoietic system is a paradigm for stem cell biology wherein a heterogeneous tissue (blood) is established and maintained by a small pool of stem and progenitor cells. Herein, this dissertation represents a collection of new approaches, both computational and technical, to chart cell fate transitions and clonal properties of the hematopoietic system. I present specific innovations that enable the massive-scale inference of chromatin accessibility in single cells as well as their clonal relatedness within humans. Importantly, these concepts, technologies, and innovations are broadly applicable to understanding human tissue biology in other systems.

Chapter 1 introduces the concept of charting lineal relationships between cells (*i.e.* lineage tracing) in human tissue by utilizing somatic mitochondrial DNA (mtDNA) mutations as clonal markers via single-cell genomics technologies. Further, I show that this concept enables scalable lineage tracing at a greater throughput (~1,000x) than other approaches for human cells. In Chapter 2, I demonstrate that somatic mtDNA mutations can be propagated longitudinally *in vivo* over ~3 years and in lineage-restricted progenitors. Together, these chapters provide the theoretical basis for scalable lineage tracing of hematopoietic cells.

Next, Chapter 3 introduces a droplet microfluidics platform that enables profiling accessible chromatin in hundreds of thousands of single cells. I show how this approach can be utilized to dissect multi-lineage non-coding regulatory logic of hematopoietic tissue in response to stimuli. In Chapter 4, I identify and correct a previously uncharacterized artifact termed 'barcode multiplets' in single-cell data. Importantly, I show that, if uncorrected, barcode multiplets artificially inflate clonality estimates. These two chapters provide a technical basis for accurate, large-scale profiling and clonal estimation of human cells.

Chapter 5 synthesizes these advances (mtDNA-based lineage tracing and droplet-based single-cell genomics) into one assay, termed mtscATAC-seq. Importantly, this multimodal approach provides a technical basis to simultaneously infer both contemporary cell state (via accessible chromatin) and cell fate (via somatic mutation lineage tracing), altogether enabling the dissection of complex tissues and stem cell hierarchies *in vivo*.

Taken together, this body of work summarizes several key advances that uniquely enable the study of developmental and regenerative processes in native human tissue.

Contents

Listing of figures

Acknowledgments

Beyond what is written in this document, I have learned a tremendous amount in the past five years, both as a scientist and as a human. For this, I am indebted to hundreds of my colleagues, mentors, friends, and family for their roles in this effort. Thank you for the love and support; I must acknowledge a few of you by name.

First, to my fiancé, Isabel: I love you, and every day, I am thankful to have you in my life. To my parents, Chad and Julie, my sister Maci, Aunt Lorie, Pa, and Grandma Melba, and many others: thank you for the unconditional support and encouragement; I most certainly would not be here without you.

To my advisor, Vijay, thank you for being my biggest proponent from the first day that I joined your group until the last. I will always be grateful for all that you have done to enable me. To Martin, thank you for developing the skills, attentiveness, and understanding that I have today. To Jason, I am ever-grateful for the synergy of your mentorship and our friendship. To Brad, thank you for pushing me to new heights after every interaction.

To my friends and colleagues, I am humbled to have worked with each and every one of you. To Leif, the magnitude of your intellectual contribution to my training is evident in these pages, but the totality of your friendship means so much more. To Christoph, Erik, Jacob, and Jeffrey– thank you; because of you, this arduous process became perhaps the most fun years of my life. To Fabiana and Sai, your kindness and eagerness made collab-

oration truly enjoyable; I hope to always work with friends like you. To Bob, Cameron, Caitlin, Colby, Hannah, Harry, Josh, Kathy, Kelvin, Lindsay, Mollie, Sheila, Tai, and Zack, old friends and new, this work was embolden by your support– thank you.

Finally, these chapters would not be possible without the extensive collaborations that are uniquely facilitated by completing a doctorate at Harvard. A full list of individuals who contributed to each chapter, as they appear in peer-reviewed manuscripts, is shown below:

Chapter 1: Leif S. Ludwig*, Caleb A. Lareau*, Jacob C. Ulirsch*, Elena Christian, Christoph Muus, Lauren H. Li, Karin Pelka, Will Ge, Yaara Oren, Alison Brack, Travis Law, Christopher Rodman, Jonathan Chen, Genevieve Boland, Nir Hacohen, Orit Rozenblatt-Rosen, Martin J. Aryee, Jason D. Buenrostro, Aviv Regev, Vijay G. Sankaran

Chapter 2: Caleb A. Lareau*, Leif S. Ludwig*, Vijay G. Sankaran

Chapter 3: Caleb A. Lareau*, Fabiana M. Duarte*, Jennifer Chew*, Vinay Kartha, Zach D. Burkett, Andrew S. Kohlway, Dmitry Pokholok, Martin J. Aryee, Frank J. Steemers, Ronald Lebofsky, Jason D. Buenrostro

Chapter 4: Caleb A. Lareau, Sai Ma, Fabiana M. Duarte, Jason D. Buenrostro

Chapter 5: Caleb A. Lareau*, Leif S. Ludwig*, Christoph Muus, Satyen H. Gohil, Tongtong Zhao, Zachary Chiang, Jeffrey M. Verboon, Wendy Luo, Elena Christian, Danial Rosebrock, Gad Getz, Fei Chen, Jason D. Buenrostro, Catherine J. Wu, Martin J. Aryee, Aviv Regev, Vijay G. Sankaran

* = co-first authors

You don't have to see the whole staircase, just take the first step.

Martin Luther King, Jr.

0

Introduction

The genesis and maintenance of multicellular organisms requires the differentiation of cells through a hierarchy of fate-decisions, leading to a functional end-state. In 1957, Conrad Waddington posited a model for this developmental process by conceptualizing cellular differentiation as a ball rolling down a three-dimensional surface with many bifurcations (Waddington, 1957). Since Waddington's schematic, thousands of studies have helped define the key molecular effectors underlying theoretical bifurcations, including lineage-specific transcription factors. An extensively characterized paradigm of this differentiation hierarchy is the human hematopoietic system, wherein a relatively small number (~$10^4 - 10^6$) of hematopoietic stem and progenitor cells (HSPCs) are responsible for constituting trillions (~10^{14}) of cells with vastly different functions in the human body (Lee-Six et al., 2018; Orkin & Zon, 2008). Despite being the best characterized model of multi-lineage cellular differentiation in adult humans, fundamental questions underlying the hematopoietic differentiation remain unanswered. These include: 1) How many HSPCs actively constitute blood production at a given time? 2) To what extent do HSPCs contribute equally or unequally across all hematopoietic lineages (so-called "lineage-bias")? 3) How do clonal dynamics become altered in diseased or pre-diseased state? 4) What are the distinct molecular programs that emerge during differentiation from multipotentency toward a terminal state?

In order to study these fundamental questions underlying human hematopoiesis, hematopoietic stem cells, and other multipotent cell types in complex tissues in the human body, I purpose that two pieces of information about a cell are essential to determine. First, a characterization of a cell's state defines where a cell exists along some differentiation trajectory. Though canonically estimated by the presence or absence of cell surface antigens, recent approaches using single-cell transcriptomics (Weinreb et al., 2020) and epigenomics (Buenrostro et al., 2018) similarly provide a high-dimensional readout characterizing the molecular phenotype of the cell. Second, an account of a cell's lineage defines the relationships between a cell and its progenitor ancestors as well as other related cells. Molecular techniques termed ”lineage tracing” enable the mapping of cell-cell relatedness but have been difficult to ascertain in human cells without genetic manipulation. Together, a combined inference of cell state and cell lineage will aid in our understanding of the properties underlying cellular differentiation and clonal composition of human tissues, including in the hematopoietic system.

0.1 Ascertaining single-cell states

As cell fate decisions occur at a single-cell level, methods and technologies that provide this resolution are essential to study differentiating systems, such as the human hematopoietic system. Fundamentally, these approaches rely on the variable transcription of genes, which is an essential feature underlying the establishment and maintenance of cellular identity (Shema et al., 2019). Here, genetic transcription is regulated by a diverse set of transcription factors and non-coding genomic elements that enable transcriptional machinery to actively transcribe a locus. Importantly, complementary advances in inferring the activity of these non-coding elements (via chromatin accessibility) and active genes (via mRNA sequencing) have provided molecular frameworks to infer cell states in a massively-parallel fashion. Many such technologies exist for obtaining these measurements in single-cells, all relying on the physical separation biological nucleic acids and subsequent barcoding with technical oligonucleotides (Klein & Macosko, 2017). When these technologies are paired with sophisticated computational tools, comprehensive state maps of cellular hierarchies across a range of complex tissues and organisms have been attainable (Weinreb et al., 2020). As single-cell epigenomic assays provide an opportunity to define regulators, and thus putative mechanisms, of gene regulatory logic underlying cell fate transitions (Shema et al., 2019), continued innovation of techniques that profile the non-coding genome will be a focus of this work.

0.2 Lineage tracing approaches

A variety of lineage tracing techniques have been established that broadly fall into two categories. The first class of lineage tracing techniques tag individual cells (mostly in model organisms) with heritable genetic markers, including fluorescent reporter genes, high diversity DNA barcode libraries, or CRISPR-based genetic scars (Woodworth et al., 2017). As these engineering techniques are generally not applicable to humans, a second class of lineage tracing studies utilized in humans have relied on the detection of naturally occurring somatic mutations, including single nucleotide variants (SNVs), copy number variants (CNVs), and microsatellites (Woodworth et al., 2017; Lodato et al., 2015). This second class of techniques utilizes the principle that acquired somatic mutations are propagated to daughter cells but are absent in distantly related cells, providing a naturally-occurring means for lineage reconstruction. To detect these mutations, these approaches often rely on techniques such as single-cell whole-genome sequence (scWGS) or WGS derived from single-cell colonies to survey cells. In either case, these strategies remain expensive and provide only modest throughput. Furthermore, no technology readily pairs scWGS with a concomitant measure of cell state at any reasonable throughput, making the dissection of complex human tissues challenging with either approach. Thus, further work is needed to develop new approaches that enable paired state and lineage inference in single-cells via a scalable technology.

0.3 Somatic mtDNA mutations for lineage tracing

A core hypothesis underlying advances in this dissertation is that mitochondrial DNA (mtDNA) sequence variation could provide an innate and natural barcode from which to infer clonal relationship in human cells. The mtDNA genome is 16.6 kb, which is small enough for cost-effective sequencing, but still provides a substantial target for somatic genetic diversity. Futhermore, mitochondrial genomes have high copy number (100s-1,000s), and mutations in mtDNA often reach high levels of heteroplasmy (defined as the proportion of mitochondrial genomes containing a specific mutation) due to a variety of factors (Stewart & Chinnery, 2015; Wallace & Chalkia, 2013). Notably, the utility of mtDNA mutations for clone tracking has already been indirectly demonstrated in various tissues using various cell staining procedures, primarily for active/inactive mtDNA-associated proteins (Taylor et al., 2003; Teixeira et al., 2013). Thus, the inference of somatic mutations in mtDNA may provide an efficient target for clonal lineage tracing in native human cells.

Importantly, for application in the hematopoietic system, a critical feature of this approach is the fact that mtDNA replicates independently of cell cycle (Mishra & Chan, 2014). Thus, even quiescent long-term hematopoietic stem cells will continue to accumulate somatic mtDNA mutations even without dividing. Utilizing this fact, I hypothesize that many (if not a majority) of HSCs will accumulate sufficient independent mtDNA mutations to be individually barcoded. As such, progeny cells from each HSC can be identified by measuring the mtDNA variation in single cells. Finally, as mtDNA and mtRNA are already detected by many single-cell genomics assays, I hypothesize that a scalable platform that can simultaneously infer cell state and cell lineage (via somatic mtDNA mutations) is achievable through innovations in single-cell genomics techniques.

Every cell in your body has the same DNA. Except it doesn't.

Carl Zimmer

1

Lineage tracing in humans enabled by mitochondrial mutations & single-cell genomics

Abstract

Lineage tracing provides key insights into the fate of individual cells in complex organisms. While effective genetic labeling approaches are available in model systems, in intact humans most approaches require detection of nuclear somatic mutations, which have high error rates, limited scale, and do not capture cell state information. Here, we show that somatic mutations in mitochondrial DNA (mtDNA) can be tracked by single cell RNA or ATAC sequencing. We leverage somatic mtDNA mutations as natural genetic barcodes and demonstrate their utility as highly accurate clonal markers to infer cellular relationships. We track native human cells both *in vitro* and *in vivo*, and relate clonal dynamics to gene expression and chromatin accessibility. Our approach should allow clonal tracking at a 1,000-fold greater scale than with nuclear genome sequencing, with simultaneous information on cell state, opening the way to chart cellular dynamics in human health and disease.

1.1 Introduction

Recent innovations in single cell genomics have enabled insights into the heterogeneity of human cell populations and have redefined concepts about lineage commitment and development (Giladi & Amit, 2018). While all cells in the human body are derived from the zygote, we lack a detailed map integrating cell division (lineage) and differentiation (fate). As a result, we have a limited understanding of how cellular dynamics play a role in physiologic and pathologic conditions for any given tissue.

Two classes of methods have been developed to study cellular relationships and clonal dynamics in complex tissues of vertebrates. In model organisms, most approaches to date rely on an engineered genetic label to tag individual cells with heritable marks (Woodworth et al., 2017; Kester & van Oudenaarden, 2018), such as fluorescent reporter genes, high diversity DNA barcode libraries, mobile transposable elements, Cre-mediated recombination, or CRISPR-based genetic scars (McKenna et al., 2016; Pei et al., 2017; Sun et al., 2014; Yu et al., 2016). Recent studies have combined several of these tracing methods with single cell RNA-seq (scRNA-seq) to interrogate both lineage relationships and cell states (Raj et al., 2018; Spanjaard et al., 2018).

However, the genetic manipulations required for such approaches cannot be applied in intact humans (Biasco et al., 2016). Limited lineage tracing studies in humans have relied on the detection of naturally occurring somatic mutations, including single nucleotide variants (SNVs), copy number variants (CNVs), and variation in short tandem repeat sequences (microsatellites or STRs), which are stably propagated to daughter cells, but are absent in distantly related cells (Ju et al., 2017). Detection of nuclear somatic mutations by whole genome sequencing in individual cells remains costly, is difficult to apply at scale, and has substantial error rates (Tao et al., 2017; Lodato et al., 2015). Moreover, most methods have not been combined with approaches that provide information about cell type and state based on gene expression or epigenomic profiles. As a result, we have had a limited ability to study cellular dynamics in humans in health and disease.

We hypothesized that mitochondrial DNA (mtDNA) sequence variation could provide an innate and natural barcode from which to infer clonal relationships. This sequence variation has several promising attributes for its utility in clonal and lineage tracing. The 16.6 kb long genome provides a substantial target for genetic diversity, but is sufficiently small for cost-effective sequencing. Although there is some variation in the measurements, mtDNA mutation rates are estimated to be 10- to 100-fold higher than for nuclear DNA (Stewart & Chinnery, 2015). Mitochondrial genomes have high copy number (100s-1,000s), and mutations in mtDNA often reach high levels of heteroplasmy (proportion of mitochondrial genomes containing a specific mutation) due to a combination of vegetative segregation, random genetic drift, and relaxed replication (Figure 1.1.A) (Stewart & Chinnery, 2015; Wallace & Chalkia, 2013). Indeed, the utility of mtDNA mutations for clone tracking has already been indirectly demonstrated in various tissues (Taylor et al., 2003; Teixeira et al., 2013).

Critically, mtDNA sequences and genetic variation are detected by existing methods, including the single cell assay for transposase accessible chromatin-sequencing (scATAC-seq) and single cell RNA-seq (scRNA-seq). While sequencing reads mapping to the mitochondrial genome are often treated as an experimental nuisance, we reasoned that they can open an opportunity to trace cellular hierarchies at scale. To demonstrate the utility of mtDNA variation for clonal tracing, we must show that heteroplasmic mtDNA mutations (1) can be reliably detected in single cells; (2) are propagated in daughter cells; (3) can be used to accurately determine clonal relationships; (4) can be combined with cell state measurements to learn meaningful biology; and (5) can be applied to study human samples.

Here, we investigate these properties, provide evidence that scRNA- and scATAC-seq provide reliable measurements of mtDNA genetic variation, and demonstrate how these mutations can be used as endogenous genetic barcodes to retrospectively infer cellular relationships in clonal mixtures of native hematopoietic cells, T lymphocytes, leukemia, and solid tumors.

1.2 Results

1.2.1 mtDNA genotyping with ATAC-seq allows accurate clone tracking and association with chromatin state

To test if mtDNA genotypes can correctly identify clonal relationships we performed a proof-of-principle experiment, where we derived and propagated sub-clones of the hematopoietic TF1 cell line (Figure 1.1.B). We generated a "ground truth" experimental lineage tree of 65 individual sub-clonal populations over 8 generations (generation time ~3 weeks between two consecutive bottlenecks) (Figure 1.1.C). For each generation, we isolated single cells from the parental colony and expanded each clone to derive sub-clones in an iterative process. The original population and each expanded sub-clone were profiled by ATAC-seq, which captures the full mitochondrial genome as an unwanted by-product (Figure A.1.A). On average, the 16.6 kb mitochondrial genome was covered at 3,380-fold per million mapped reads. We determined high-confidence heteroplasmic mitochondrial genotypes with a computational variant-calling pipeline that utilizes individual per-base, per-allele base quality (BQ) scores and verified that our calls were reproducible across sequencing runs (Figure A.1.B,C; Appendix A).

The large range of detected mutations included clone- and sub-clone-specific mutations that were propagated over generations (Figure 1.1.D and Figure A.1.D). Most mutations were C>T transitions, consistent with previous reports (Ju et al., 2017). Although some somatic mutations were shared among multiple first-generation clones and their progeny (*e.g.*, Figure 1.1.D 8003 C>T), nearly all progeny of an individual clone shared muta-

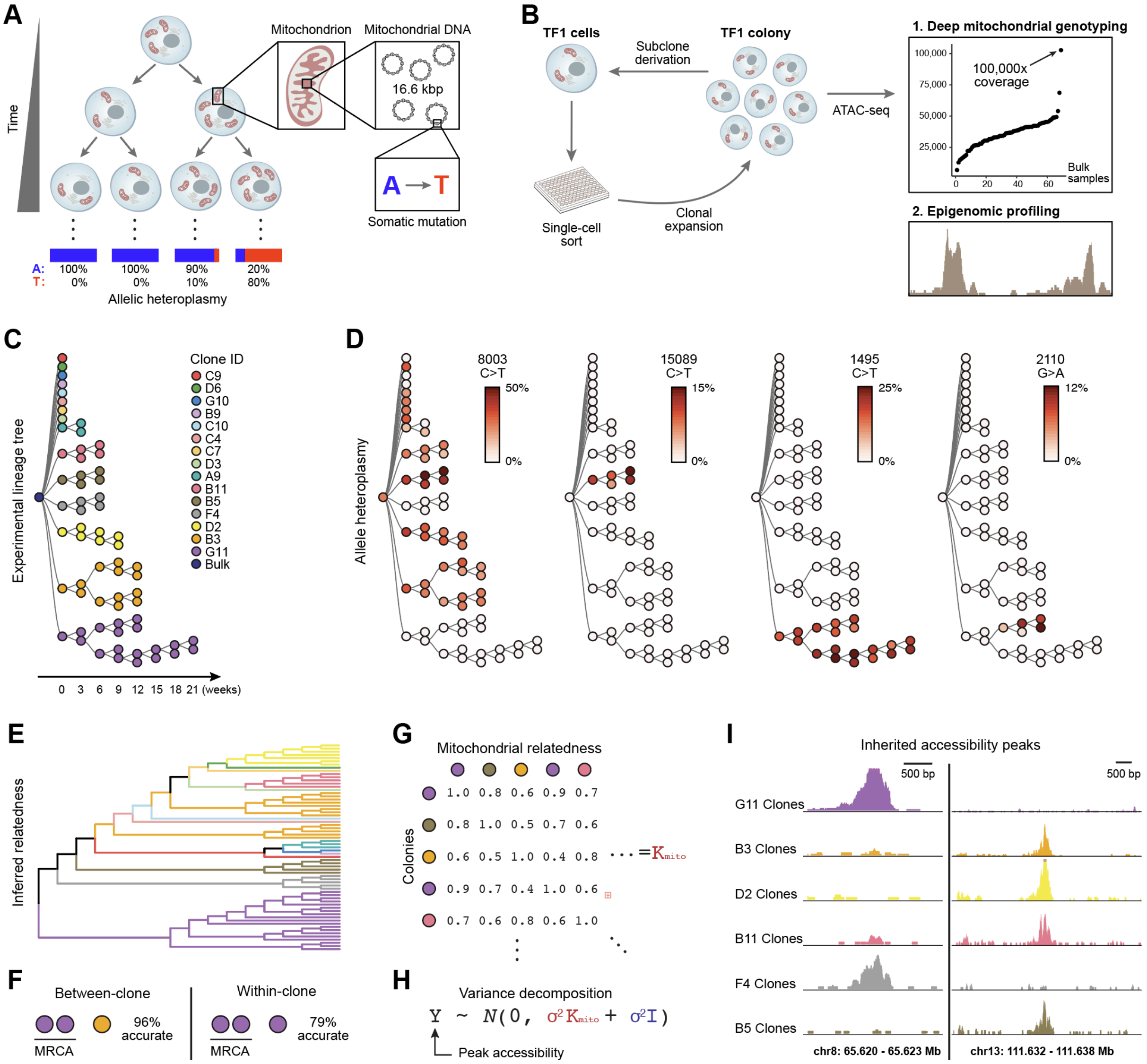

Figure 1.1: Mitochondrial mutations are stably propagated in human cells *in vitro*. **(A)** Dynamics of mtDNA heteroplasmy in single cells. Each cell has multiple mitochondria, which in turn contain many copies of mtDNA that may acquire somatic mutations over time. **(B)** Proof-of-principle design. Each TF1 cell clone and sub-clone is assayed with ATAC-seq. **(C)** Supervised (true) experimental TF1 lineage tree. Colors indicate each primary clone from initial split. **(D)** Allelic heteroplasmy of four selected variants reveals stable propagation and clone-specificity. Color bar: allelic heteroplasmy (%). **(E)** Unsupervised hierarchical clustering of TF1 clones. Color: primary clones as in **(C)**. **(F)** Between-clone and within-clone accuracy of identifying the most-recent common ancestor (MRCA) per trio of clones based on mtDNA mutational profile. **(G)** Schematic of mitochondrial relatedness matrix K_{mito} where each pair of clones is scored based on mitochondrial genotype similarity. **(H)** Random effects model for variance decomposition of epigenomic peaks. **(I)** Two examples of peaks inherited in clonal lineages. Peaks represent the sum of open chromatin for the clones with the most samples.

tions that were unique and stably propagated over the course of the experiment (*e.g.*, Figure 1.1.D 15089 C>T, 1495 C>T, Figure A.1.D). Furthermore, we detected new somatic mutations that arose within sub-clones and were stably propagated (Figure 1.1.D, 2110 G>A; Figure A.1.D).

We used these high-confidence mtDNA mutations to reconstruct clonal relations with high accuracy (Figure 1.1.E,F). Ordinal hierarchical clustering on individual samples grouped nearly all (sub-)clones belonging to a single clonal family correctly (Figures 1.1.C,E). Specifically, we accurately identified the most recent common ancestor (MRCA) at 96% between first-generation clones and 79% within sub-clones derived from first-generation clones (Figure 1.1.F and Figure A.1.E; Appendix A). Moreover, we correctly inferred clonal contributions to heterogeneous bulk populations comprised of three clones at various concentrations (Figure A.1.F; Appendix A).

We next paired mitochondrial genotypes with chromatin state information for each clone and identified differences in chromatin state that follow inferred clonal relationships. We approximated the pairwise clone-clone mitochondrial relatedness (Figure 1.1.G; Appendix A) and performed a random effects variance decomposition of each chromatin accessibility peak in our TF1 clones (Figure 1.1.H), asking how "heritable" a chromatin feature is in a population. Of 91,607 peaks tested, 8,570 peaks were highly heritable (> 90% variance explained; Figure 1.1.I and Figure A.1.G). Overall, this demonstrates the utility of ATAC-seq for mtDNA genotyping to enable accurate clone tracing, while simultaneously providing information on cell state.

1.2.2 Successful detection of mtDNA heteroplasmy using single cell genomics

Because the mitochondrial genome is almost completely transcribed (Figure 1.2.E), we hypothesized that heteroplasmic mitochondrial mutations might be detected by scRNA-seq. Across six scRNA-seq protocols (Ziegenhain et al., 2017), full length scRNA-seq methods showed more extensive coverage of the mtDNA genome than 3' end directed scRNA-seq (Figure 1.2.A and A.2.A,B). Importantly, there was a high concordance between heteroplasmic allele frequency estimates from scRNA-seq and whole genome sequencing from the same cell (Han et al. (2018a); Figure 1.2.B). However, several highly heteroplasmic mutations were specific to mtRNA (Figure 1.2.B): some likely reflect RNA-editing, including one that has been previously validated (2619 A>G)

(Bar-Yaacov et al., 2013), but many others are observed at low frequencies (<20%) and reflect either RNA transcription errors or technical errors in scRNA-seq (Venteicher et al., 2017).

We systematically compared our ability to detect clones from mtDNA mutations at various levels of heteroplasmy in three TF1 cell clones (Figure 1.2.C: clones C9, D6, G10) using bulk and scATAC-seq, bulk and scRNA-seq (SMART-seq2), and a newly developed single cell mtDNA sequencing protocol based on rolling circle amplification (scMito-seq; Figure 1.2.C and A.2.C, Appendix A). We observed high concordance in the frequencies of RNA and DNA-derived mitochondrial genotypes across all methods (in addition to RNA-specific mutations, as described above; Figure 1.2.D and A.2.E). As expected, scATAC- and scMito-seq had more uniform and deeper coverage of the mitochondrial genome than SMART-seq2 (Figure 1.2.E and A.2.D). Data from every method allowed us to detect the previously identified unique clonal allele for 95.4% (210/220) of cells and to accurately infer clonal relationships by hierarchical clustering (Figure 1.2.F and A.2.F).

1.2.3 Mitochondrial mutation clones match those from lentiviral barcoding

To compare mitochondrial mutations to an exogenous gold standard of clone detection, we used a lentiviral barcoding approach. We infected TF1 cells with a modified Perturb-seq lentiviral construct (Dixit et al., 2016) expressing a mNeonGreen gene carrying a 30 bp random nucleotide sequence in its untranslated region (Figure A.3.A). We sorted 25 mNeonGreen+ cells and expanded them, followed by bulk ATAC-seq and scRNA-seq of 158 quality-controlled cells (Figure 1.3.A). Notably, there was no correlation between the number or types of barcodes discovered and mitochondrial coverage (Figure A.3.B). The 158 cells included 15 informative barcodes that mapped cells to one of 11 non-overlapping groups (Figure A.3.A). To filter any artefactual mitochondrial mutations from scRNA-seq (Figures 1.2.B, 1.3.C and A.2.E, A.6.F,G), we restricted our analysis to the 20 variants that were present in the bulk ATAC-seq at allele frequencies > 0.5% and which had high per-allele base quality scores in bulk and in the sum of single cells (Figures 1.3.B,C, Appendix A).

Hierarchical clustering by these 20 mitochondrial mutations correctly inferred clonal structure in single cells in a comparable manner with gold-standard exogenous barcodes (Figure 1.3.D). Of note, specific mutations were

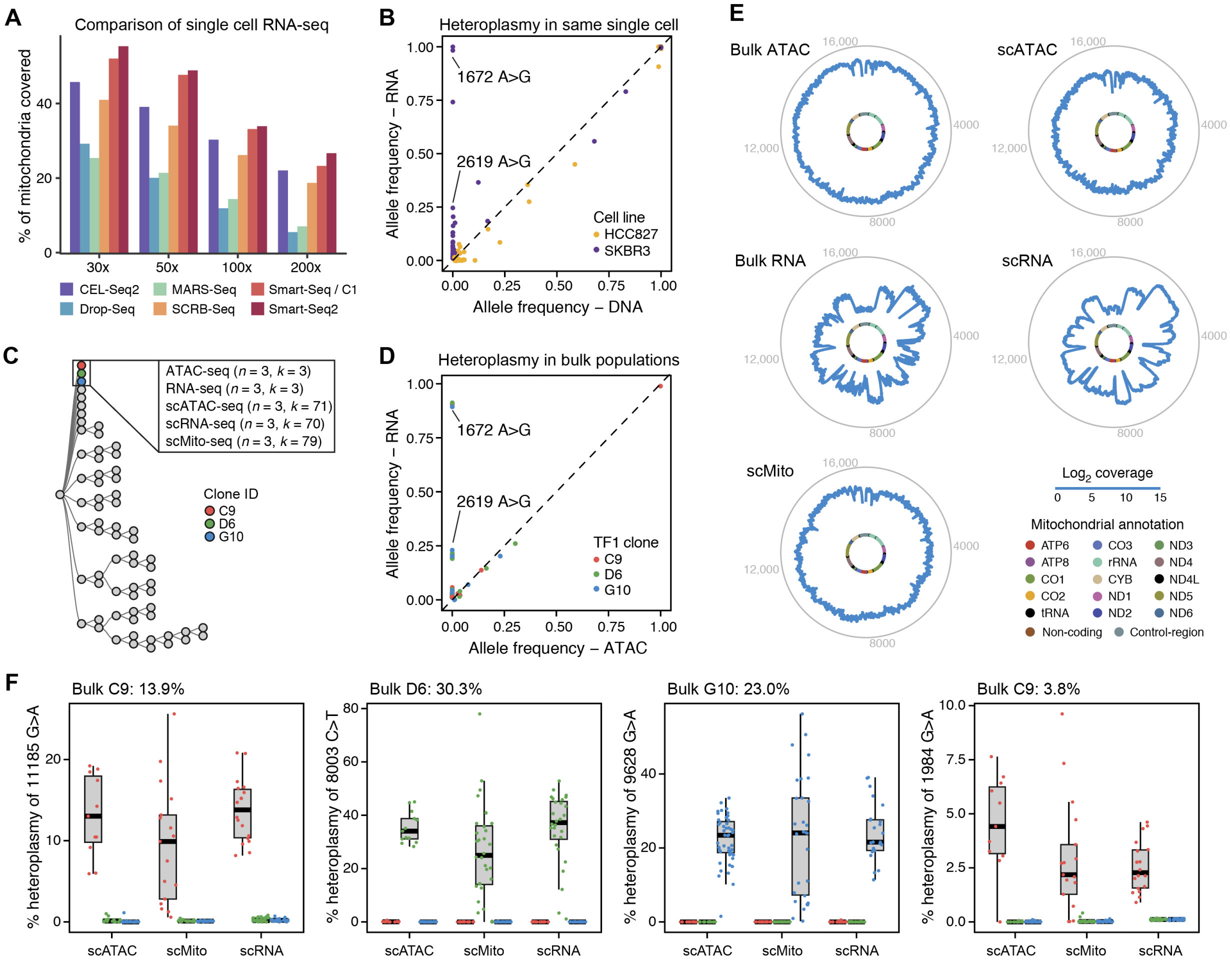

Figure 1.2: Mitochondrial mutations are detected using single cell genomics (A) Coverage of mouse mitochondrial genome by six scRNA-seq methods. Shown is the fraction (%) of the mitochondrial genome (y axis) covered by reads from each of six methods (color code), at different levels of coverage (x axis). **(B)** Agreement in allelic heteroplasmy estimates from single cell whole genome sequencing (WGS) and scRNA-seq from the same single cells. Shown is the allele frequency for scRNA- (y axis) and scWGS-seq (x axis) based estimates for two cell lines (HCC827: orange; SKBR3: purple). Two examples of RNA-specific changes are highlighted. **(C-F)** Identification of mitochondrial mutations by scRNA-, scATAC- and scMito-seq in three TF1 clones. **(C)** Bulk and single cell data collected for three TF1 clones (boxed). Each clone ($n = 3$) was processed with variable numbers of single-cell libraries (k). **(D)** Agreement in allelic heteroplasmy estimates from bulk ATAC- (x axis) and bulk RNA-seq (y axis) from three indicated TF1 clones (as in **(C)**). Two examples of RNA-specific changes are highlighted. **(E)** Coverage of the mitochondrial genome of the TF clone G10 by each indicated assay. Inner circle: mitochondrial genome; middle blue outline: coverage; outer grey circle: genome coordinates. For single cell assays, coverage is the sum of single cells. **(F)** Four clone-specific mutations that are reliably detected by various single-cell assays with heteroplasmies as low as 3.8%. Each boxplot shows the % heteroplasmy (y axis) of one mutation across scATAC-, scMito- and scRNA-seq in the three TF1 clones (color code as in **(C)**). Dots: individual cells.

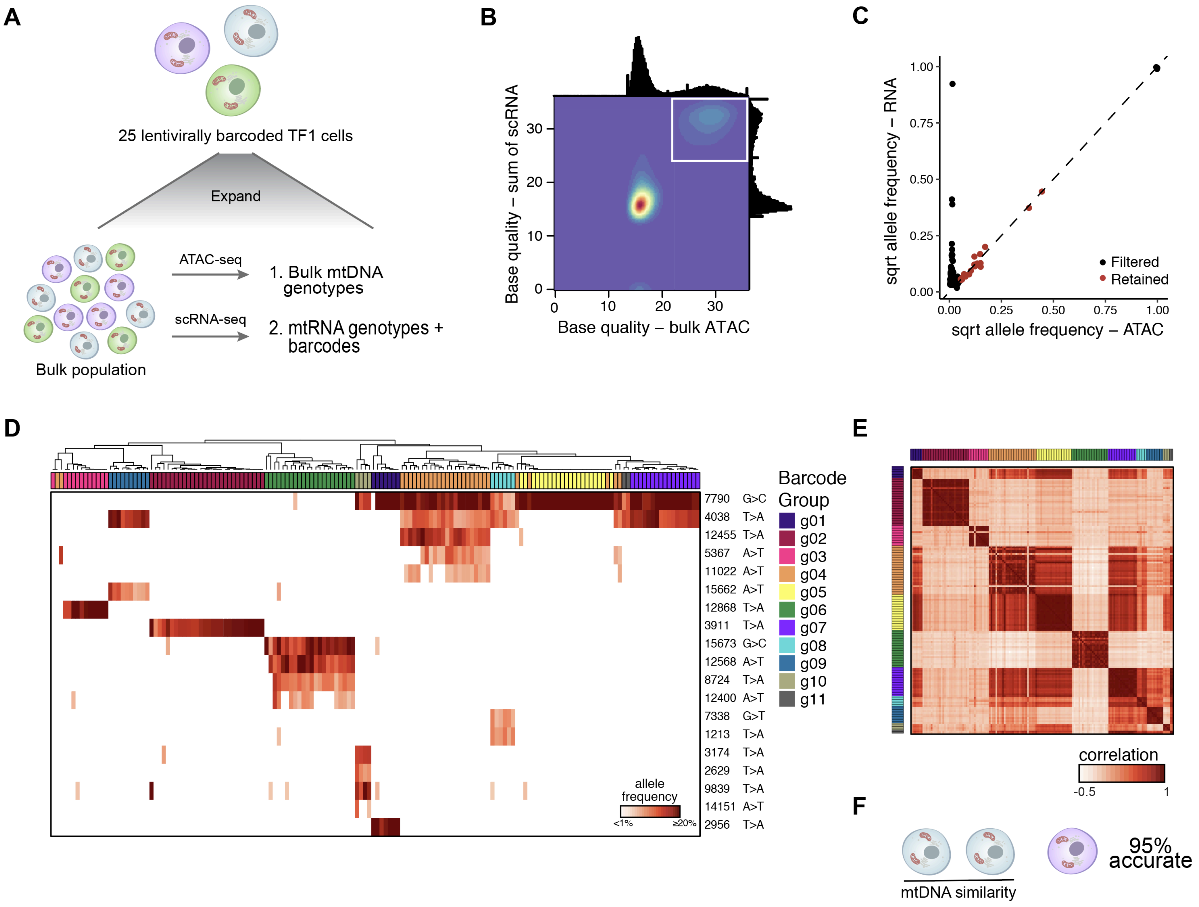

Figure 1.3: Validation of mitochondrial mutations as clonal markers in single cells using lentiviral barcoding. (A) Experimental overview. TF1 cells were infected with a lentiviral vector expressing the mNeonGreen gene and a 30bp random barcode in the untranslated region (Figure 1.3.A). 25 cells were sorted and expanded, followed by bulk ATAC-seq and scRNA-seq. **(B)** Filtering of high confidence mutations. Base quality (BQ) scores from scRNA- (y axis) and from bulk ATAC-seq (x axis). White box: high-confidence variants detected by both technologies (BQ >20) (Appendix A). **(C)** Allele frequencies determined by the sum of single cells from scRNA-seq (y axis) and bulk ATAC-seq (x axis). Black – filtered; red – retained. **(D-F)** mtDNA inferred clones agree with barcode-based clones. **(D)** Hierarchical clustering of TF1 mitochondrial genotyping profiles (rows) for cells assigned to annotated barcode groups (columns) (from Figure 1.3.A). Color bar: Heteroplasmy (% allele frequency). **(E)** Cell-cell similarity from mitochondrial mutations called in **(C)**. Column and rows are annotated by barcode group. **(F)** Between-group accuracy of identifying the most-similar pair per trio of clones based on mtDNA mutational profile using detected barcodes as a true positive.

shared among a number of barcode groups (7790 G>C and 4038 T>A), suggesting these may reflect common sub-clonal structure in the original population. A cell-cell similarity matrix using a Pearson correlation distance metric of the 20 mutations (Figure 1.3.E) effectively classified pairs of cells within the same barcode group (area under receiver operating characteristic curve (AUROC): 0.96; area under the precision recall curve (AUPRC): 0.84; Figures S3C,D). Cells that were most similar based upon mitochondrial genotypes correctly predicted shared barcode pairs in a trio analysis with 95% accuracy (Figure 1.3.F). In this context, mitochondrial mutations provided a significantly more accurate measure of shared clonality than alterations in copy number variants (CNVs) inferred from scRNA-seq (FiguresA.3.E,F).

1.2.4 mtDNA mutation diversity across human tissues

To assess the broader applicability of mitochondrial genotyping, we examined mtDNA mutations across diverse human tissues, similar to previous studies that have shown widespread inter- and intra-individual diversity of heteroplasmic mtDNA mutations (Ye et al., 2014). We analyzed mitochondrial genotypes from bulk RNA-seq of 8,820 individual samples in the GTEx project, spanning 49 tissues with at least 25 donors, as well as 462 donors with at least 10 tissues (Consortium et al. (2017);Figure 1.4.A; Appendix A). There was significant variation in the proportion of mitochondrial reads mapping to the mitochondrial transcriptome across tissues, consistent with known differences in the absolute numbers of mitochondria and levels of mitochondrial gene expression in each tissue (Figures 1.4.B,C and A.4.A). After stringent filtering to remove artifacts related to RNA-seq (Appendix A), we identified 2,762 mutations that were tissue-specific within an individual donor at a minimum of 3% heteroplasmy (Figures 1.4.D-G and A.4.B), revealing a diverse spectrum of mutations. The majority of mutations were C>T (G>A) or T>C (A>G) transitions (Figure 1.4.E), consistent with previous reports (Ju et al., 2014).

Each of the 49 tissues examined had at least one tissue specific mutation across all donors, only 28 non-polymorphic mutations were shared between any two tissues from any one donor (minimum heteroplasmy of 5%), and no non-polymorphic mutations were shared between three such tissues, indicating that these mutations

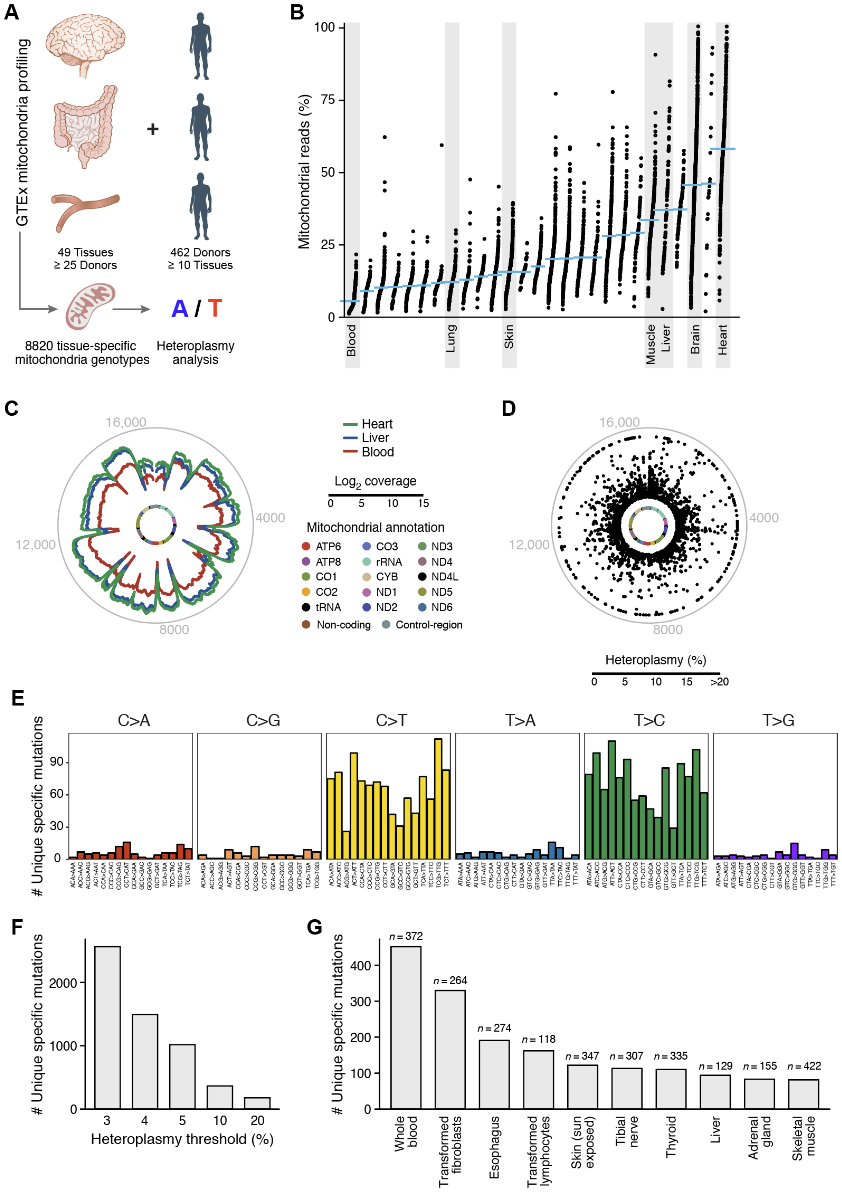

Figure 1.4: Tissue-specific mitochondrial heteroplasmic mutations. **(A)** Analysis overview. **(B)** Proportion of aligned reads that map to the mitochondrial genome for each tissue. **(C)** Mitochondrial genome coverage for different tissues. Inner circle: mitochondrial genome; middle circular tracks: mean coverage for heart (green), liver (blue), and blood (red); outer grey circle: genome coordinates. **(D-G)** Tissue-specific heteroplasmic mutations (> 3% heteroplasmy) in GTEx RNA-seq data. **(D)** Distribution along the mitochondrial genome. Inner circle: mitochondrial genome. Dots: % heteroplasmy of each tissue specific mutation; outer grey circle: genome coordinates. **(E)** Number of observed tissue-specific heteroplasmic mutations (y axis) in each class of mononucleotide and trinucleotide change. **(F)** Number of tissue-specific heteroplasmic mutations (y axis) at different allele frequency thresholds (x axis). **(G)** Number of tissue-specific heteroplasmic mutations (y axis) across the 10 tissues (x axis) with the largest number of tissue specific mutations in GTEx.

arose somatically and in a tissue-specific manner. However, this is likely an underestimate of the true extent of heteroplasmy at the level of individual cells, due to measurement of bulk populations (Kang et al., 2016). Most of the predicted deleterious mutations (Appendix A) did not show an appreciable difference in median heteroplasmy compared to the benign ones (Figures A.4.C-D), although high heteroplasmic (>20%) mutations were present at 3.6- to 4.4-fold fewer than expected (Figures A.4.E-F). Of note, these levels are substantially below the estimated biochemical threshold of 60-90% heteroplasmy, where deleterious mtDNA mutations are generally thought to have an effect (Stewart & Chinnery, 2015). Thus, even predicted damaging mutations appear to be tolerated at heteroplasmy levels suitable for lineage tracing, although high-throughput functional studies of mtDNA mutation and large population genetic studies are needed to refine these definitions. Overall, this diversity of mitochondrial mutations within individual humans indicates that these can be leveraged to probe questions related to cellular relationships across a range of healthy tissues and cell types.

1.2.5 Stable propagation of heteroplasmic mtDNA mutations in primary hematopoietic cells

We next tested if mtDNA mutations are clonally propagated in primary human cells. We plated CD34+ hematopoietic stem and progenitor cells (HSPCs) from two independent donors in semi-solid media, derived 65 erythroid and myeloid colonies, and profiled 8-16 cells per colony by scRNA-seq for a total of 935 cells that passed quality metrics (Figure 1.5.A). Cells composing any individual colony are derived from a single, distinct hematopoietic progenitor cell. As expected, based on expression profiles, the cells partitioned into two major clusters, corresponding to erythroid and myeloid cells, consistent with colony morphology and irrespective of donor (Figures 1.5.A-D and A.5.A,B). Conversely, the mtDNA mutation profile separates single cells according to their donor of origin, as well as their single cell-derived colony of origin based on highly heteroplasmic mutations (Figures 1.5.E-G and A.5.A,B).

Supervised analysis shows that colony-specific mutations within each donor are faithfully propagated (Mann-Whitney U Test p-value $< 10^{-10}$), a significant subset of which distinguishes most cells in each colony from all

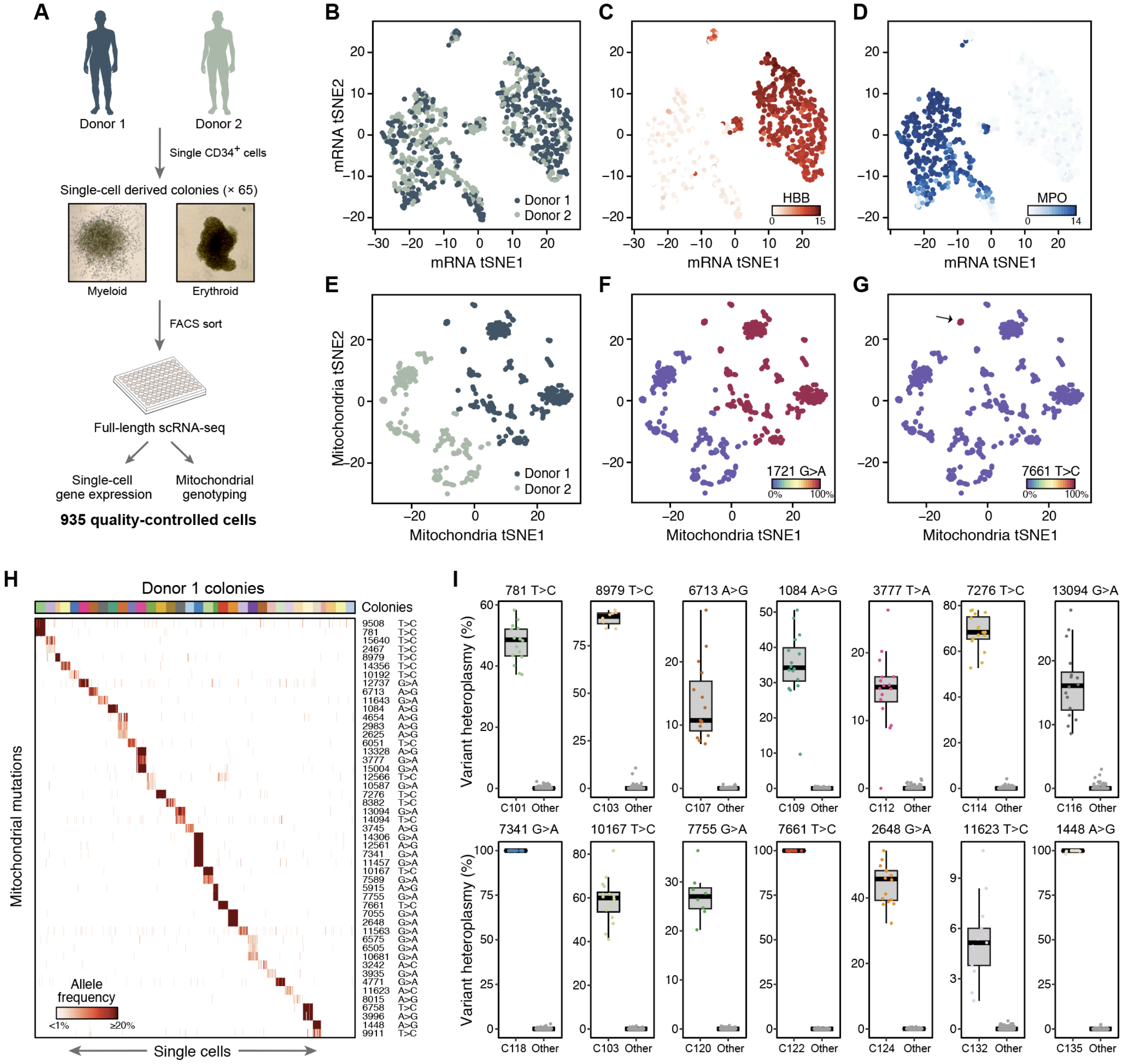

Figure 1.5: Mitochondrial mutations are stably propagated in primary hematopoietic cells. (A) Overview of experiment. Hematopoietic colonies are derived from single primary CD34+ HSPCs in semi-solid media, which were then picked and sorted before performing scRNA-seq. **(B-D)** Expression profiles separate cells by types and not by donor. t-Stochastic Neighborhood Embedding (tSNE) plots of cells' expression profiles, labeled by donor **(B)** or by expression of HBB (**C**, marking erythroid cells) or MPO (**D,** marking myeloid cells). **(E-G)** Mitochondrial mutation profiles separate cells by donor. tSNE plots of mitochondrial mutation profiles, with cells labeled by donor **(E)**, a polymorphic mutation unique to donor 1 **(F)**, or a heteroplasmic mutation present only in a specific colony **(G)**. **(H)** Colony-specific mutations for Donor 1. Shown are the allele frequencies and base pair change of mutations (rows) that are found by supervised analysis as specific to the cells (columns) in each colony (sorted by colony membership; colored bar on top), color bar: allelic heteroplasmy (%). **(I)** 14 selected colony-specific mutations in Donor 1 colonies. Box plots show the distribution of heteroplasmy (%, y axis) in cells of a specific colony for the indicated mutation, and in the cells in all other colonies. Dots: individual cells.

other cells from the same donor (Figures 1.5.H and A.5.C). Specifically, we identified unique clonal mutations in 71% of colonies for donor 1 and 47% for donor 2, each detected at similar frequencies in at least 80% of cells of the same colony (Figure 1.5.F; Appendix A), although certain experimental challenges, such as mixing between adjacent colonies (Figures A.5.D,E) likely result in an underestimate. The extent of heteroplasmy varied considerably, including multiple mutations that nearly achieved homoplasmy (Figure A.5.I). We observed similar mutational diversity with bulk ATAC-seq of colonies similarly derived from two other donors (Figure A.5.H), and in 268 sorted phenotypic hematopoietic stem cells (HSC) from three additional donors from a published scATAC-seq study (Figure A.5.I; Buenrostro et al. (2018)). Importantly, the colony-specific mitochondrial mutations do not overlap between donors in the scRNA-seq analysis (Figure A.5.G) and show very limited overlap between donors in the scATAC-seq analysis (Figure A.5.J). Thus, adult human HSPCs show a large spectrum of mtDNA mutational diversity and these mutations are stably propagated in daughter cells at a level that allows for lineage or clonal tracing studies of *in vivo* human hematopoiesis.

1.2.6 mtDNA mutations from scRNA-seq and scATAC-seq allow inference of clonal structure in primary human cells

To assess our ability to accurately infer clonal structures in complex primary human cell populations, we obtained 30 primary CD34+ HSPCs from donor 2, expanded them into a single large population over 10 days, and processed cells by bulk ATAC-seq and either scATAC- or scRNA-seq (Figure 1.6.A and A.6.A). We used probabilistic k-medoids clustering of these mtDNA mutation profiles to cluster individual cells (Appendix A). Our clustering assigned cells with high-confidence to 10 clusters consisting of 3-36 cells per cluster, with cells in each cluster sharing one or two heteroplasmic mutations at comparable frequencies (Figures 1.6.B and A.6.E), consistent with expectations under a simulated setting (Figure 1.6.C, Appendix A). Notably, when all RNA-based mtDNA mutations (including the artefactual variants) were included, we could not readily discern clusters (Figures A.6.F,G). Applying this approach to cells with mtDNA mutations called from scATAC-seq, we were similarly able to assign 95 of 148 cells (64%) to 9 different clusters (Figures 1.6.C and A.6.B,D,H) and identify clone-specific regions of open chromatin (Figure 1.6.D, Appendix A).

1.2.7 Somatic mtDNA mutations are consistent with and further refine human T lymphocyte clones defined by TCR rearrangements

As a test of the ability of mtDNA mutations to correctly resolve human cell clones *in vivo*, we turned to T lymphocytes, where T cell receptor (TCR) rearrangements are frequently used as natural markers of clonality. We applied our method to tumor-infiltrating T lymphocytes from human lung and liver cancers (Zheng et al., 2017a; Guo et al., 2018). Supervised analysis of T lymphocytes sharing a unique TCR sequence revealed shared specific mtDNA mutations that were absent from other T lymphocytes (Figure 1.6.E). In some instances, mtDNA mutations in T lymphocytes with the same TCR rearrangement further classified cells into subpopulations (Figure 1.6.F). These mutations may have arisen after TCR rearrangement as subpopulations underwent stimulation and proliferation, or the TCR may have developed independently from clonally distinct T lymphocyte progenitor cells. Moreover, some mtDNA mutations were shared across T lymphocytes with unique TCR sequences, suggesting they shared a common ancestor prior to V(D)J recombination (Figure 1.6.G). These findings further demonstrate that mtDNA mutations are reliable clonal markers *in vivo*.

1.2.8 Somatic mtDNA mutations reveal subclones in primary human colorectal cancer

To test our approach in solid tissues and tumors, we analyzed EPCAM+ cells from a colorectal adenocarcinoma primary tumor resection by bulk ATAC-seq and scRNA-seq (Figure 1.7.A). To derive the non-cancer mtDNA genotype, we processed EPCAM+ cells from two adjacent, presumed healthy sites by bulk ATAC-seq. We identified 11 mtDNA mutations specific to the tumor and absent in adjacent healthy tissue (Figure 1.7.B). Across 238 cells from the tumor sample, we were able to partition 107 cells (45%) into 12 distinct clusters by mtDNA mutations (Figures 1.7.B,C and A.7.A), suggesting the presence of clonal heterogeneity. We annotated the clusters by

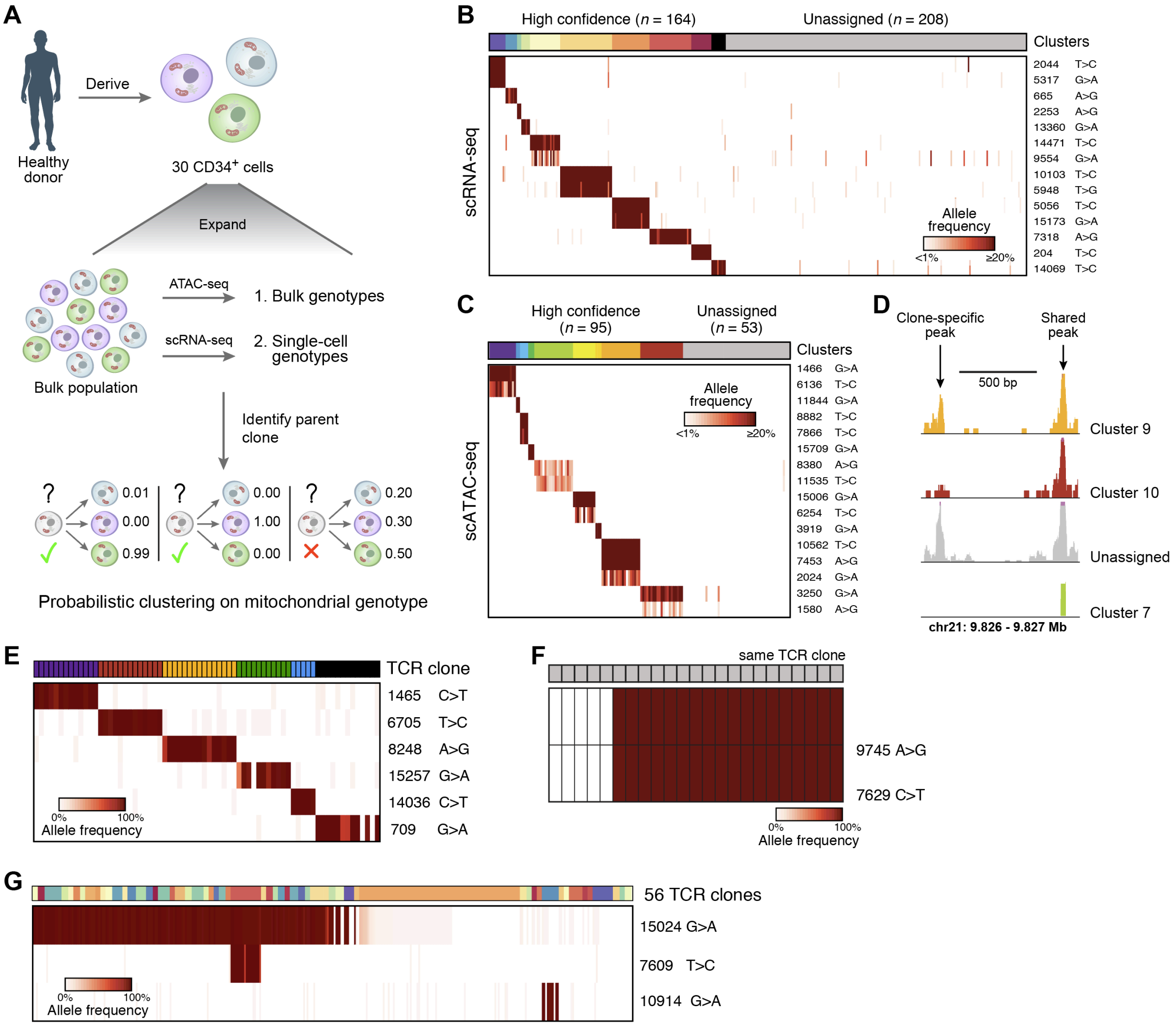

Figure 1.6: Mitochondrial mutations identify clonal contributions in polyclonal mixtures of human cells. (A-D) Determination of clones in primary hematopoietic cells. **(A)** Overview of experiment. CD34+ HSPCs are expanded, genotyped in bulk and single cells, and clonal origin is inferred. **(B, C)** Identification of confident cell subsets based on retained heteroplasmic mutations by unsupervised clustering of scRNA- or scATAC-seq using probabilistic k-medoids. Cells (columns) are sorted by unsupervised clustering on the variants (rows). Clusters: colored bar on top; grey: unassigned cells; color bar: allelic heteroplasmy (%). **(D)** Example locus with one clone-specific (left) and one shared (right) open chromatin peak recovered by mitochondrial clustering. **(E-G)** Relationship between mitochondrial mutations and TCR clones in human T lymphocytes. Each panel shows data from independent patients. **(E)** Shown are the allele frequencies of heteroplasmic mutations (rows) that are concordant with individual TCR clones (columns, color code). **(F)** Sub-clonal relations within a single TCR clone. Heteroplasmic mutations (rows) that differ between cells within a single TCR clone (columns). **(G)** Heteroplasmic mutations (rows) shared among a variety of TCR clones (columns, color code). Color bar: allelic heteroplasmy (%).

known markers of colonic epithelial cells (Figures 1.7.D,E and A.7.A-C, Dalerba et al. (2011)). Of note, 28/30 (93%) of the tumor cells expressing the stem cell marker LGR5 shared the 9000 T>C mutation (Figures 1.7.D,F). Expression of the proliferation marker MKI67 was particularly high in these cells, potentially explaining the large contribution of this population to the tumor tissue (Figures 1.7.E,F).

1.2.9 Somatic mtDNA mutations as stable clonal markers in CML in humans *in vivo*

To further validate the utility of our approach *in vivo*, we focused on chronic myelogenous leukemia (CML). Using our mitochondrial genotyping pipeline we analyzed scRNA-seq data from 2,145 cells profiled across 49 samples from 31 CML patients, collected at the time of diagnosis, when CML clones predominate, and at 3 and 6 months of therapy, when malignant clones are expected to decrease in frequency relative to benign HSPCs (Giustacchini et al., 2017). Since neither bulk ATAC-seq, nor DNA-seq were available, we applied particularly conservative quality thresholds (Appendix A).

The mitochondrial genotypes robustly separated donors by unsupervised analysis (Figures 1.7.G and A.7.D,J), consistent with our observations of mtDNA variation across humans (Figure 1.4), and, in some patients, further partitioned cells in a manner consistent with disease stage (Figures 1.7.H,I and A.7.E,F). In one striking example, three heteroplasmic mtDNA mutations were nearly exclusive to BCR-ABL positive cells, but absent in non-leukemic cells from the same donor (Figure 1.7.J). Importantly, integration of these mtDNA mutations appears to improve stratification of malignant cells vs. benign cells compared to the BCR-ABL genotyping assay alone, resulting in 100% concordance with transcriptional signatures (Figure 1.7.K, boxed cells). Interestingly, although the frequency of BCR-ABL positive cells decreased with treatment (compare cells in cluster 1 and 2 to cells in cluster 3), one mitochondrial mutation (6506 T>C) present in the majority of BCR-ABL positive cells at diagnosis continued to mark BCR-ABL positive cells post-treatment, thereby validating the stable propagation of mtDNA mutations over extended periods of time*in vivo* (Figure 1.7.K). On the other hand, BCR-ABL positive cells with the 4824 T>C mutation (that also harbor the 6506 T>C mutation) were depleted, implying that this sub-clone was likely more susceptible to therapy.

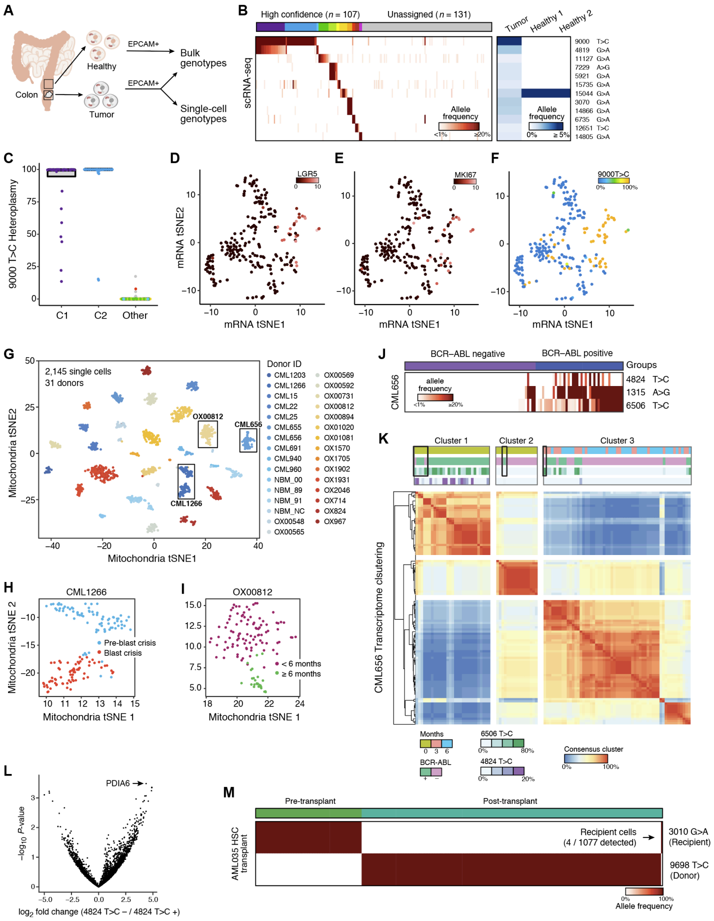

Figure 1.7: Application of mitochondrial mutation tracking in human cancer *in vivo*. (A-F) Identification of clones in human colorectal cancer. **(A)** Cells from tumor and adjacent normal tissue are sorted based on EPCAM+ surface marker expression and genotyped using bulk ATAC-seq and scRNA-seq. **(B)** Identification of clonal subsets based on heteroplasmic mutations (rows) across cells (columns), sorted by unsupervised clustering (clusters: colored bar on top; grey: unassigned cells). Right: allele frequencies in the bulk healthy and tumor populations. **(C)** Heteroplasmy levels per single-cell. Colors and clusters are from panel **B**. **(D-F)** Clone of predominantly LGR5+ cells. tSNE of scRNA-seq profiles from the tumor, colored by expression for **(D)** LGR5 **(E)** MKI67 (Color bar: log2 counts per million) and **(F)** heteroplasmy of the 9000 T>C allele (color bar: % allelic heteroplasmy). **(G)** Near-perfect separation of donors based on mitochondrial genotypes. tSNE of mitochondrial mutation profiles of 2,145 single cells from 31 donors with CML, colored by donor ID. Boxes: Donors analyzed for sub-clones in **(H-L)**. **(H,I)** Identification of putative sub-clonal structure within donors. tSNE of mitochondrial mutation profiles of cells from donor CML1266 **(H)**, sampled at pre- (blue) and during (red) blast crisis, and for donor OX00812 **(I)**, sampled at diagnosis and <6 months of treatment (magenta) or >6 months treatment (green). **(J)** Shown are the allele frequencies of three highly heteroplasmic mutations (rows) across BCR-ABL positive vs. negative cells (columns). Color bar: allelic heteroplasmy (%). **(K)** Consensus clustering of CML656 transcripts suggests variable annotation in BCR-ABL positive cells at diagnosis. Heatmap showing proportion of times (red/blue) that two cells (columns, rows) belong to the same cluster. Color bars denote from top to bottom: time of collection, BCR-ABL status, and allele frequencies (6506T>C, 4824T>C). Boxes indicate cells where mitochondrial mutations suggest that the BCR-ABL status was incorrectly determined by the BCR-ABL genotyping assay alone. **(L)** Differentially expressed genes (x-axis) between cells in Cluster 1 comparing cells with and without the 4824 T>C mutation. P-value (y-axis) is from an empirical Bayes moderated t-test. **(M)** mtDNA mutations distinguish recipient- and donor-specific cells after HSCT in AML.

Unsupervised clustering by expression profiles partitioned this patient's cells into three clusters. Clusters 1 and 2 were comprised of cells from the initial sample at diagnosis, but separated by BCR-ABL status as well as by mitochondrial genotype. Cluster 3 was comprised of cells obtained 3 and 6 months after the start of treatment (Figure 1.7.K). Differential expression analysis of Cluster 1 cells stratified by the 4824 T>C mutation status (Figure 1.7.L) identified the induction of PDIA6, a gene implicated in cancer cell proliferation (Gao et al., 2016), in cells lacking the mutation, suggesting that it may be associated with the observed variation in sub-clone frequencies. Thus, mitochondrial genetic analysis can improve stratification of malignant cells and enhance understanding of clonal evolution and therapy resistance.

1.2.10 In vivo chimerism inferred from mtDNA mutations

Mitochondrial genotyping has the potential to allow efficient tracking of donor and recipient chimerism during HSC transplantation (HSCT). We analyzed scRNA-seq profiles of peripheral blood mononuclear cells (PBMCs) from an AML patient before and after HSCT, which were profiled with 3' directed massively parallel scRNA-seq (Zheng et al., 2017b). Although such approaches have substantially reduced coverage of mtDNA (Figures A.2.A and A.7.G,H), we reasoned that a small number of homoplasmic mutations should be detectable. Indeed, our analysis revealed two homoplasmic mitochondrial alleles distinguishing the donor and recipient cells (Figure 1.7.M), and inferred that 99.6% of cells sampled post-transplant were donor-derived, but four recipient cells were still present. These results demonstrate the potential of using mitochondrial mutations to measure the dynamics of donor chimerism in transplantation settings. Such approaches may demonstrate even greater sensitivity in conjunction with currently employed approaches (Zheng et al., 2017b).

1.3 Discussion

Here, we describe an approach for high-throughput and unsupervised tracing of cellular clones and their states at single cell resolution in native human cells by mtDNA mutation detection. This approach is likely to be broadly useful and immediately applicable, since mtDNA mutations can be readily detected by commonly employed single cell genomic methods, including scRNA-seq and scATAC-seq, which concomitantly provide readouts of cell state. We show that somatic mtDNA mutations with levels as low as 5% heteroplasmy can be stably propagated and serve as clonal markers in primary human cells. We additionally provide an improved mutation detection framework, where mutations are first identified based on a DNA-based bulk sample (lower threshold 0.5%), and then called in scRNA-seq data, allowing for accurate mutation detection in RNA-based measurements. Overall, in our validation experiments, mitochondrial genotypes correctly inferred clonal lineage with ~95% accuracy (Figures 1.1-1.3), achieving similar accuracy as widely applied genetic labeling methods.

Our approach has three key advantages: (1) it is highly scalable; (2) it is directly applicable to human tissues; and (3) it is combined with assays to profile a cell's state at the chromatin or transcriptome level. Conversely, single cell whole genome sequencing can be applied in human tissues, but is neither scalable nor combined with a functional state profile, whereas exogenous genetic barcoding cannot be applied to native human samples. For example, 18,000 individual cells' mitochondrial genomes can be sequenced at 100-fold coverage for the sequencing cost of a single nuclear genome at 10-fold coverage, a depth not sufficient for confident mutation calling (Lodato et al., 2015).

Our approach can be further enhanced in several ways. First, additional assays devised to focus on directly measuring mitochondrial genomes can reduce cost and increase coverage (Figure 1.2.D). For example, we developed a scMito-seq protocol (Figure 1.2.C), potentially providing a higher fidelity of mitochondrial mutation detection based on rolling-circle amplification (Ni et al., 2015) that could be also used in combination with scRNA-seq (Macaulay et al., 2015). Currently, massively parallel scRNA-seq data from droplet based approaches have limited coverage of the mitochondrial genome (Figures A.2.A and A.7.G,H), restricting their immediate utility and application, though a combined enrichment and capture of mitochondrial transcripts could improve this approach (Zemmour et al., 2018; Dixit et al., 2016). Finally, mtDNA sequencing could be combined with nuclear DNA sequencing strategies to detect SNVs, CNVs, and microsatellites to further increase the fidelity and reach of current single cell clonal tracing applications.

One potential limitation with using mtDNA mutations for clone detection or lineage inference may arise from the horizontal transfer of mitochondria between cells, which has been described in specific contexts, but the extent and physiologic relevance of such a process remains unclear. The transfer of organelles appears to be primarily triggered by various stress responses, is restricted to specific cell types, and can be a feature of malignant cells, but the extent of organelle transfer appears to be limited (Caicedo et al., 2015; Griessinger et al., 2017; Marlein et al., 2017; Moschoi et al., 2016; Torralba et al., 2016). Moreover, such transfer would have to be extensive to significantly confound the analysis (Figure 1.7.I) and we have been unable to detect evidence of such transfer in our data (Figures 1.3, 1.6, 1.7.J,K and 1.7.M). Another limitation is that we are currently unable to account for phenotypic effects of the mtDNA mutations used for clonal tracing. Although most mutations likely have at most small effects at the heteroplasmy levels investigated here, accurate maps of allele heteroplasmy and cellular function will be an important area for further investigation.

Overall, we show that measuring somatic mitochondrial mutations provides a powerful and scalable approach to assess cellular dynamics of native human cells. Mitochondrial mutations readouts are readily compatible with single cell measurements of cell state to provide a potent means to relate stem and progenitor cells with their differentiated progeny that should facilitate probing the molecular circuits that underlie cell fate decisions in health and disease. Clonal tracking using mitochondrial mutations opens up a novel avenue to infer critically needed relationships in large-scale efforts, such as the Human Cell Atlas or in tumor cell atlases, to better understand the mechanics of homeostasis and development across a reference map of human tissues (Regev et al., 2018).

1.4 Acknowledgements

We thank members of the Regev and Sankaran laboratories for valuable comments and the Broad Institute Flow Cytometry Facility for assistance with cell sorting. We are grateful to Leslie Gaffney for assistance with illustrations. Funding: CAL is supported by F31 CA232670 from the National Institutes of Health (NIH). JCU is supported by 5T32 GM007226-43 from the NIH. JDB received support from the Broad Institute Fellows program and the Allen Institute Distinguished Investigator Award. JHC is supported by 5T32CA207021-03 from the NIH. KP is supported by a research fellowship of the German Research Foundation (DFG). This research was supported by a Broad Institute SPARC grant (AR and VGS), the New York Stem Cell Foundation (NYSCF, VGS), the Howard Hughes Medical Institute (AR), the Klarman Cell Observatory (AR), a gift from Arthur, Sandra, and Sarah Irving (NH), and NIH grants R33 CA202820 (AR), R01 DK103794 (VGS), R33 HL120791 (VGS), and R01 CA208756 (NH). VGS is a NYSCF-Robertson Investigator.

1.5 Author contributions

Conceptualization, LSL with input from CAL, JCU, AR, and VGS; Methodology, LSL, CAL, JCU, KP, AR, and VGS; Formal Analysis, CAL with input from JCU, LHL, LSL, MJA, JDB, AR, and VGS; Investigation, LSL, CAL, EC, CM, KP, WG, AB, TL, and CR; Resources, AR, VGS, MA, JDB, JHC, GMB, NH, and OR; Writing – Original Draft, LSL, CAL, JCU, AR, and VGS with input from all authors; Writing – Review and Editing, LSL, CAL, JCU, AR, and VGS with input from all authors; Visualization, CAL, JCU, LSL, and LHL; Supervision, AR and VGS; Project Administration, OR, AR, and VGS; Funding Acquisition, AR and VGS.

1.6 Code and data availability

All sequencing data generated in this work is available on the gene expression omnibus (GEO) accession GSE115218, along with tables that contain variant calls and heteroplasmy estimates for all primary data generated in this study. Custom code used for producing mtDNA genotypes from single-cell data are available here at https://github.com/sankaranlab/mito-genotyping.

There is nothing permanent except change.

Heraclitus

2

Longitudinal assessment of clonal mosaicism in human hematopoiesis via mtDNA tracing

Abstract

Our ability to track cellular dynamics in humans over time*in vivo* has traditionally been limited. Here, we demonstrate how somatic mutations in mitochondrial DNA (mtDNA) can be used to longitudinally track the dynamic output of hematopoietic stem and progenitor cells in humans. Over the course of three years of blood sampling in a single individual, our analyses reveal somatic mtDNA sequence variation and evolution reminiscent of models of hematopoiesis established by genetic labeling approaches. Furthermore, we observe fluctuations in mutation heteroplasmy, coinciding with specific clinical events, such as infections and further identify lineage-specific somatic mtDNA mutations in longitudinally sampled circulating blood cell subsets in individuals with leukemia. Collectively, these observations indicate the significant potential of how tracking of somatic mtDNA sequence variation presents a broadly applicable approach to systematically assess hematopoietic clonal dynamics in human health and disease.

2.1 Introduction

Recent studies have described the application of lineage tracing in model organisms (Rodriguez-Fraticelli et al., 2018; Pei et al., 2017) and genetically modified cells in humans undergoing gene therapy (Scala et al., 2018; Biasco et al., 2016). These studies have provided insights into clonal dynamics in complex tissues. In the hematopoietic system, such inferences have provided previously unappreciated knowledge on the contributions of hematopoietic stem and progenitor cells (HSPCs) to blood cell production (Jacobsen & Nerlov, 2019). However, as most methods rely on the introduction of exogenous genetic labels (*e.g.* lentiviral- and transposon-based barcoding or Cre-loxP based recombination), these techniques are not readily amenable to broadly study physiologic and pathologic processes in humans. Assessing the dynamics of and outputs from HSPCs in an unperturbed setting in humans represents a methodological challenge, leaving open questions about their frequency, functionality, and longevity (Scala & Aiuti, 2019). This raises the important question of how we can effectively and longitudinally study clonal dynamics in humans.

While somatic mutations in the nuclear genome have been leveraged to perform clonal lineage tracing in humans, these approaches are expensive and often error-prone in single-cells, limiting broader or routine applications (Lee-Six et al., 2018; Lodato et al., 2015). Recently, we and others have demonstrated the utility of somatic mitochondrial DNA (mtDNA) mutations as natural genetic barcodes that may be stably propagated across cell divisions (see Chapter 1). Importantly, common genomic techniques, including the assay for transposase-accessible chromatin sequencing (ATAC-seq) and RNA sequencing (RNA-seq), provide the means to concomitantly assess cell type and state with mtDNA genotypes. As our previous work demonstrated substantial somatic mtDNA mutational diversity within HSPCs, we reasoned that tracking these mutations would enable assessment of clonal contributions to blood production. Specifically, as progenitor cell-specific mutations would be propagated to differentiated circulating blood cells, we hypothesized that fluctuations of mtDNA mutations should be reflective of the clonal output of progenitor cells over time. However, the utility of this approach to evaluate longitudinal clonal dynamics remains unexplored.

2.2 Results

We reasoned that assessment of somatic mtDNA mutations in data from recent studies that have longitudinally profiled human peripheral blood using genomic approaches could enable clonal inferences in circulating blood and immune cells (Figure 2.1.A; B.1.A; Rendeiro et al. (2020); Chen et al. (2018a)). As virtually the entirety of human mtDNA is transcribed, we reasoned that we could examine patterns of somatic mutation dynamics from bulk RNA-seq data. To these ends, we processed 57 RNA-seq datasets that had been sampled over the course of 161 weeks serially from a single individual (Chen et al., 2018a). Using our genotyping pipeline from Chapter 1, we were able to identify numerous high-confidence mtDNA mutations and illustrate their dynamics over nearly 3 years of peripheral blood sampling (Figure 2.1.B). These mutations were selected as they did not show evidence of RNA-editing or other known biases. For example, while the 10000A>G allele was gradually lost over the course of three years, the 295C>T allele increased in heteroplasmy during this time. In contrast, the 13636T>C allele appeared to be stably propagated over the full three years *in vivo*. Other mutations such as 829A>G and 10278A>C became more prominent in discrete windows spanning several months. Collectively, these observations support distinct models of hematopoiesis, including those involving clonal succession (progressive recruitment of distinct clones, marked by specific mtDNA mutations) and others involving stability of specific clones over periods of time (Yu et al., 2016; Scala et al., 2018). Considering all available alternative allele frequencies, we observed a decay in the Spearman correlation of mutation frequencies comparing baseline to subsequent time points (Figures 2.1.C and B.1.B), further reflecting the dynamic evolution of mitochondrial mutations in the sampled circulating blood and immune cells.

As we previously observed highly-heteroplasmic mtDNA mutations in clonal lymphocytes (defined by T-cell receptor rearrangements; see Chapter 1), we hypothesized that a subset of mutations may reflect clonal expansion of lymphocytes in response to foreign pathogens (Figure B.1.C). Indeed, we observed a rare mutation (2394T>A) emerge specifically when the donor was exposed to human respiratory syncytial virus (RSV; Figure 2.1.D), noting the heteroplasmy was 0.0% at the previous timepoint (34 days prior). We confirmed the occurrence of this specific mutation in matched whole-genome bisulfite sequencing (WGBS) data comparing time points where viral infections were detected (Figure 2.1.E). These results paired with our previous observations suggest that a subset of lymphocytes carrying the 2394T>A allele clonally expanded upon RSV infection and persisted at detectable frequencies in peripheral blood for at least 10 days. Furthermore, we note the recurrence of two mutations (1575A>G and 14250G>C) at times of clinically documented infection with adenovirus (ADV) and human rhinovirus (HRV; Figure 2.1.F), respectively, suggesting virus-specific proliferation of distinct clonal lymphocyte populations in response to these infections. Together, our association of heteroplasmic variation with these clinical infections indicate that mtDNA heteroplasmy can enable the assessment of clonal dynamics and would be of particular value in settings where other clonal markers (such as lymphocyte receptor sequences) are unavailable.

As HSPCs can give rise to multiple lineages, an extension of our results from bulk peripheral blood measurements would be to examine the relative contributions of HSPCs to specific blood cell lineages, marked by the

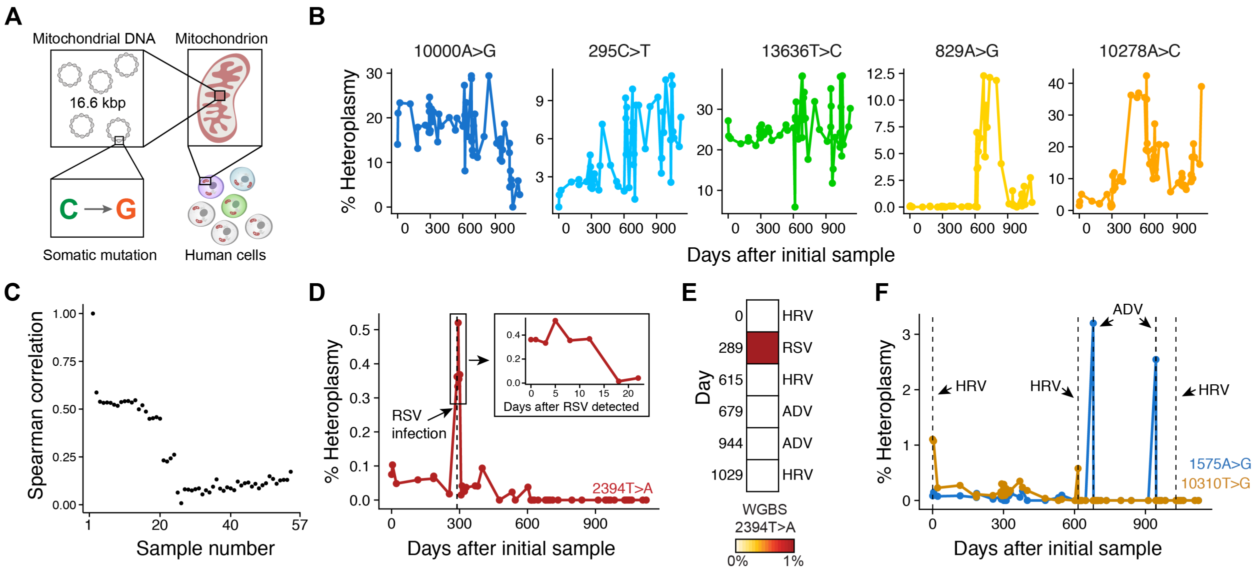

Figure 2.1: Evidence of clonal mosaicism from mtDNA mutations over 3 years *in vivo*. (A) Schematic of somatic mtDNA mutations in human cells. Each cell contains multiple mitochondria, which in turn contain multiple copies of mitochondrial DNA (mtDNA). **(B)** Examples of variable mutations *in vivo* across three years of observation that reflect clonal mosaicism in one donor. **(C)** Spearman correlation of 57 time points (ordered by relative time of sampling) across time points sampled. Correlation value is measured with the baseline sample. **(D)** Heteroplasmy of 2394T>A allele, which is associated with human respiratory syncytial virus (RSV) detection in this donor; inset shows heteroplasmy levels for 2394T>A for ~23 days after the initial detection of RSV. **(E)** Corroboration of 2394T>A allele in time of RSV infection using WGBS data at the six time points at specific days (left) that correspond to infection (right). **(F)** Heteroplasmic mutations specific the day of detection for adenovirus (ADV, 1575A>G) and human rhinovirus infections (HRV, 10310T>G).

presence of distinct somatic mtDNA mutations that are absent in other lineages (Figure 2.1.A). To explore this concept, we reanalyzed 188 ATAC-seq profiles from surface phenotype-sorted circulating blood cell populations from a cohort of eight patients with CLL that were collected up to 40 weeks following the initiation of ibrutinib treatment (Rendeiro et al., 2020). Importantly, as mtDNA is nucleosome-free and therefore is highly susceptible to transposon insertion, ATAC-seq provides a facile approach for capturing somatic mutations in mtDNA. Strikingly, we observed many instances of recurrently detected lineage-specific mutations across the sampled time points, suggesting the presence of these somatic mtDNA mutations in a lineage-biased progenitor, including 1496T>C in CD4+ T-lymphocytes (Donor CLL7), 10685G>A in CD8+ T lymphocytes (Donor CLL5), and 822G>A in NK cells (Donor CLL1; Figure 2.2.B). Alternatively, some of these may represent mtDNA mutations in clonally expanded and long-lived T lymphocytes. The persistence of these three mutations over the course of sampling is distinguished from 6453T>C, a CD19+CD5- B-lymphocyte-specific mutation, that declined over >20 weeks of sampling (Figure 2.2.C). Furthermore, we identified mutations that were shared among multiple lineages, indicating that these mtDNA mutations may exist in multipotent progenitor populations (2.2.D). The incidence of these mutations in both CD19+CD5+ leukemic cells and CD19+CD5- B-lymphocytes further support the notion that mtDNA mutations could be informative to trace sub-clonal structure in response to targeted therapies, such as ibrutinib9. Indeed, we observed instances of mutations (2885T>C and 7496T>C) decreasing in frequency with treatment, suggesting that particular subclones carrying these alleles are sensitive to the administered therapy (Figure 2.2.E).

To further verify the utility of our approach in potentially tracking clonal evolution in response to treatment, we processed an additional 81 bulk ATAC-seq samples from patients with cutaneous T cell lymphoma (CTCL) treated with histone deacetylase (HDAC) inhibitors (Qu et al., 2017). Reanalysis of these longitudinally collected samples confirmed the detection of mtDNA sequence variation, further highlighting the utility of these mutations to track clonal dynamics in response to therapies, including putative treatment sensitive and resistant clones (Figure B.2). While our analyses largely elucidated specific examples of heteroplasmic mutations and their dynamics, the bulk nature and relative sparsity of mtDNA transcriptome/ genome coverage of these data (RNA-

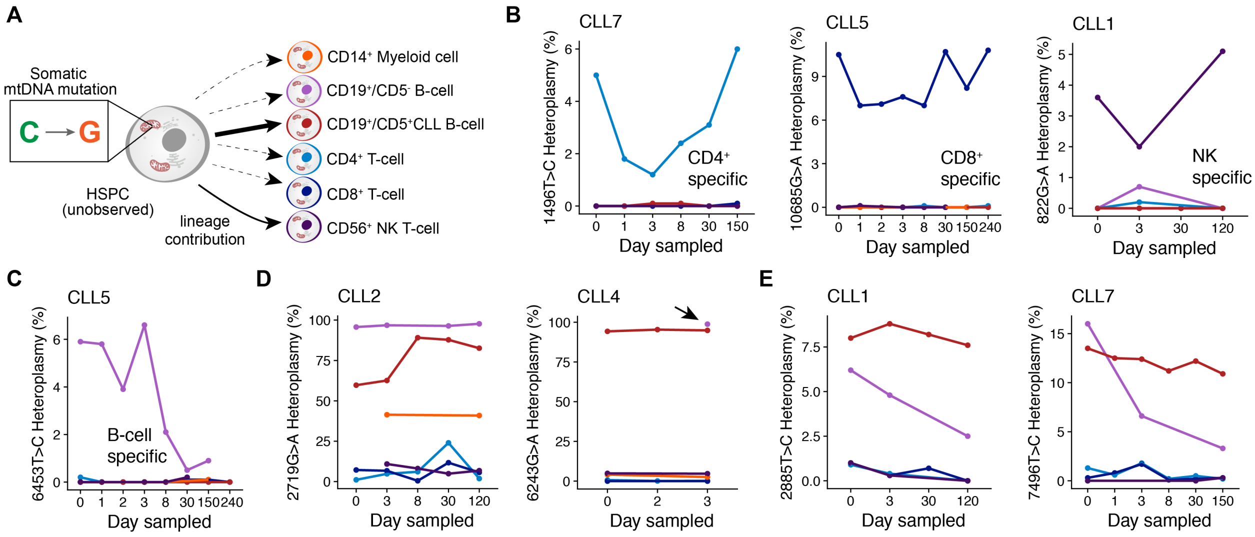

Figure 2.2: Inference of putative multi-lineage contributions from HSPCs. (A) Schematic of hematopoietic stem and progenitor cell (HSPC; unobserved) and six FACS-sorted populations. **(B)** Examples of cell type-specific mutations *in vivo* across up to 240 days of evaluation. Donor is indicated at the top of panel (*e.g.* CLL7). **(C)** Heteroplasmy of 6453T>C allele in donor CLL5, a CD19+CD5- B-cell-specific mutation that decreases in frequency over 150 days of observation. **(D)** Mutations in two donors that are present in both CD19+CD5+ (CLL) and CD19+CD5- B-cells. Arrow highlights one observed CD19+CD5- sample. **(E)** Examples of shared CLL and B-cell mutations that decay at different rates for two donors.

seq for Figure 1; single-end sequencing ATAC-seq for Figure 2) limit confident detection of low frequency variants/ clones that would enable more comprehensive analyses. We suggest that future work utilize complementary bulk and single-cell genotyping assays optimized for mtDNA sequence capture, as we have previously shown that this can increase the resolution of inferences for clonal HSPC population dynamics (see Chapter 1).

2.3 Discussion

Overall, our results illustrate the potential to leverage somatic mtDNA mutations to longitudinally study clonal dynamics and somatic mosaicism in human hematopoiesis *in vivo*, and we hope this further stimulates the design of such prospective studies in this poorly charted area of biomedical research. For example, such studies could enable assessments of cellular dynamics and responses to stressors, such as infections or acute blood loss, or complement existing strategies to track subclonal evolution in leukemia via bulk and single-cell analyses. While these results reflect a multitude of scenarios where bulk heteroplasmy changes could reflect clonal mosaicism, we note that mtDNA heteroplasmy has been described to drift over time. However, our previous work has shown that mtDNA mutations, depending on heteroplasmy, may be stably propagated to daughter cells over many cellular generations9. In this respect, we emphasize the need for systematic longitudinal studies with single-cell technologies and computational tools to comprehensively model and reliably infer clonal dynamics for future analyses. Taken together, our analyses illustrate a broadly applicable strategy to facilitate our understanding of clonal dynamics in human health and disease.

2.4 Acknowledgements

We are grateful to members of the Sankaran laboratory for valuable discussions. CAL is supported by F31 CA232670 from the National Institutes of Health (NIH). This work was supported by NIH grants R01 DK103794 and R33 HL120791, as well as the New York Stem Cell Foundation (NYSCF; to VGS). VGS is a NYSCF-Robertson Investigator.

2.5 Author Contributions

CAL, LSL, and VGS conceived and designed the study. CAL performed analyses. CAL, LSL, and VGS wrote the manuscript. VGS the work.

2.6 Code and data availability

No new sequencing data was generated as part of this work. Raw sequencing reads from previous works were were downloaded from Gene Expression Omnibus (GEO) accessions GSE33029, GSE85853, GSE111015, and GSE111405. Code for generating mtDNA heteroplasmy is available here: https://github.com/sankaranlab/mito-genotyping.

Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.

Sydney Brenner

3

Droplet-based combinatorial indexing for massive scale single-cell chromatin accessibility

Abstract

Although recent technical advancements have facilitated the mapping of epigenomes at single-cell resolution, the throughput and quality of these methods have limited their widespread adoption. Here, we describe a high-quality (10^5 nuclear fragments per cell) droplet microfluidics–based method for single-cell profiling of chromatin accessibility. We use this approach, named 'droplet single-cell assay for transposase accessible chromatin' (dscATAC-seq), to assay 46,653 cells for the unbiased discovery of cell types and regulatory elements in the adult mouse brain. We further introduce combinatorial indexing to this droplet platform (dsciATAC-seq), resulting in lower coverage and higher throughput measurements. We demonstrate the utility of this approach by measuring chromatin accessibility across 136,463 resting and stimulated human bone marrow derived cells to reveal changes in the cis- and trans-regulatory landscape across cell types and upon stimulation conditions at single-cell resolution. Altogether, in this work we describe a total of 510,123 single-cell profiles, demonstrating the scalability and flexibility of this droplet-based platform.

3.1 Introduction

Although the primary sequence of the eukaryotic genome is largely invariant across cells in an organism, the quantitative expression of genes is tightly regulated to define the functional identity of cells. Eukaryotic cells use diverse mechanisms to regulate gene expression, including an immense repertoire ($> 10^6$) of DNA regulatory elements (Roadmap Epigenomics Consortium et al., 2015; Thurman et al., 2012). These DNA regulatory elements are established and maintained by the combinatorial binding of transcription factors (TFs) and chromatin remodelers, which together function to recruit transcriptional machinery and drive cell-type specific gene expression (Spitz & Furlong, 2012; Calo & Wysocka, 2013). DNA regulatory elements, characterized by their functional roles (promoter, enhancer, insulator, etc.), are marked by a diverse array of histone and DNA modifications (Calo & Wysocka, 2013). Both classical observations (Weintraub & Groudine, 1976) and recent genome-wide efforts (Thurman et al., 2012) have shown that active regulatory elements are canonically nucleosome free and accessible to transcriptional machinery. Thus, methods that measure chromatin accessibility based on sensitivity to enzymatic digestion followed by sequencing (Boyle et al., 2008; Hesselberth et al., 2009; Buenrostro et al., 2013) provide an integrated map of chromatin states which encompass a diverse repertoire of functional regulatory elements.

Methods to assay chromatin accessibility genome-wide have been used for a variety of applications including the discovery of i) cell type-specific cis-regulatory elements, ii) master TFs that shape the regulatory landscape, or iii) mechanisms for disease-relevant non-coding genetic variation (Gerstein et al., 2012; Maurano et al., 2012; Thurman et al., 2012). However, these "epigenomic" approaches are generally applied to bulk samples, limiting their resolution into the regulatory diversity underlying heterogeneous cell populations. In parallel, methods to measure the transcriptomes of single-cells have been used to discover new cell types (Plasschaert et al., 2018) and new functional cell states (Tirosh et al., 2016; Wagner et al., 2016), and provide additional motivation for the development of tools to measure chromatin regulation at single-cell resolution (Kelsey et al., 2017).

Technological innovations have enabled the development of single-cell epigenomic methods (Kelsey et al., 2017; Jin et al., 2015; Rotem et al., 2015); however, these approaches remain relatively low-throughput and high-cost. Assay for Transposase Accessible Chromatin (ATAC-seq; Buenrostro et al. (2013)) is particularly promising for single-cell studies due to the relative simplicity of the experimental protocol, and widespread use. Previous efforts have adapted ATAC-seq to profile chromatin accessibility in single-cells, either by individually isolating cells (Buenrostro et al., 2015) or by combinatorial addition of DNA barcodes (Cusanovich et al., 2018), to enable *de novo* deconvolution of cell types and the discovery of cell-type specific regulatory factors. However, these current methods for single-cell ATAC-seq (scATAC-seq) remain either relatively low-throughput (100s to 1,000s of cells/experiment) or provide low-complexity data (1,000s of fragments per cell). Therefore, new methods for sensitive, scalable, and high-throughput profiling are needed to measure the full repertoire of regulatory diversity across normal and diseased tissues.

To meet the challenges of assaying chromatin states in the breadth and depth of complex cell populations within tissues, we report the development of a droplet-based scATAC-seq assay. In brief, our approach utilizes a droplet microfluidic device to individually isolate and barcode transposed single-cells. We demonstrate that this approach results in substantially higher data quality than existing methods, and describe an approach to improve cell throughput and cell capture efficiency by super-loading barcoded beads into droplets. Further, we extend this droplet barcoding approach by combining this method with barcoded transposition (Cusanovich et al., 2018) followed by super-loading cells into droplets, to develop droplet-based single-cell combinatorial indexing for ATAC-seq (dsciATAC-seq), providing chromatin accessibility profiles at significantly improved throughput. We apply these approaches to generate accessibility profiles of 510,123 cells, which includes i) a reference map of chromatin accessibility in the mouse brain (46,653 cells), and ii) an unbiased map of human hematopoietic states in the bone marrow (60,495 cells), isolated cell populations from bone marrow and blood (52,873 cells), and bone marrow cells in response to stimulation (75,958 cells). These unbiased chromatin accessibility profiles provide new insights into the regulators defining cells within these tissues. Further, we find that pooled stimulus of human bone marrow derived cells uncovers mechanistic insights driving genetic variants leading to human disease. Overall, this new approach for high-throughput single-cell epigenomics charts a clear course towards obtaining an epigenomic atlas across normal tissues, and provides new opportunities for single-cell epigenomic profiling at a massive scale.

3.2 Results

3.2.1 DscATAC-seq implemented on a droplet microfluidic device

In this work we describe a method for single-cell chromatin accessibility profiling using droplet microfluidics and ATAC-seq. Consistent with previously described methods for bulk ATAC-seq, nuclei are first transposed using Tn5 transposase to integrate sequencing adapters into regions of open chromatin (Buenrostro et al., 2013). Importantly, previous studies have described that transposed nuclei and DNA remain intact following transposition (Buenrostro et al., 2013; Amini et al., 2014). We therefore leverage this finding and use intact transposed nuclei as input material into a droplet microfluidics device, which co-encapsulates transposed chromatin with PCR reagents and barcoded beads into a single droplet (Figure 3.1.a). Each bead contains clonal copies of oligonucleotides that encode a common PCR primer sequence and a bead-specific DNA barcode. Following co-encapsulation, we perform droplet PCR to add cell-identifying DNA barcodes to transposed chromatin, and the resulting pool of PCR products are then collected and prepared for sequencing. We refer to this droplet-based scATAC-seq platform as dscATAC-seq.

To develop a robust and high-sensitivity platform, we optimized the concentration of Tn5 transposase (Figure 3.1.b; C.1.a-c). We found that increasing the total abundance and concentration of Tn5, notably the same enzyme contained within a widely available commercial product (Appendix C), significantly improves the total number of nuclear fragments, including improvements to the fraction of reads at transcription start sites (TSS) and distal elements (Figure 3.1.b; C.1.a-c). Furthermore, we also adapted previously described transposition methods to reduce the proportion of mitochondrial reads (Corces et al. (2016, 2017); Appendix C). Altogether, these optimizations combined with droplet encapsulation and PCR, provide a platform for high-yield and high-efficiency single-cell epigenomic profiling.

To optimize cell capture and throughput, we developed a joint experimental and computational strategy to super-load beads into droplets. Our computational strategy, which we call the bead-based scATAC processing (bap), determines bead barcodes with a high overlap of Tn5 insertion positions along the genome to identify and merge barcodes within a common droplet (Figure C.1.d, Figure C.2.a-b). This analytical approach enables loading beads at higher density, to increase the number of droplets with one or more beads, by identifying single-cells barcoded with more than one bead barcode (Figure C.2.c-d; see Appendix C). To validate our approach we included a library of random oligonucleotides to a dscATAC-seq experiment, enabling us to define true-positive bead pairs based on the overlap of these exogenous sequences (Figure C.2.b,e-j). Using these orthogonal readouts, the unique Tn5 insertions across single-cells and the random oligonucleotides introduced in this experiment, we computed precision-recall and receiver operating curves to verify the accuracy and precision of the bap approach (mean area under the receiver-operating curve (AUROC)=1.000 and mean area under the precision recall curve (AUPRC)=0.997) (Figure C.2.k, Appendix C). We also found consistent experimental results across a range of

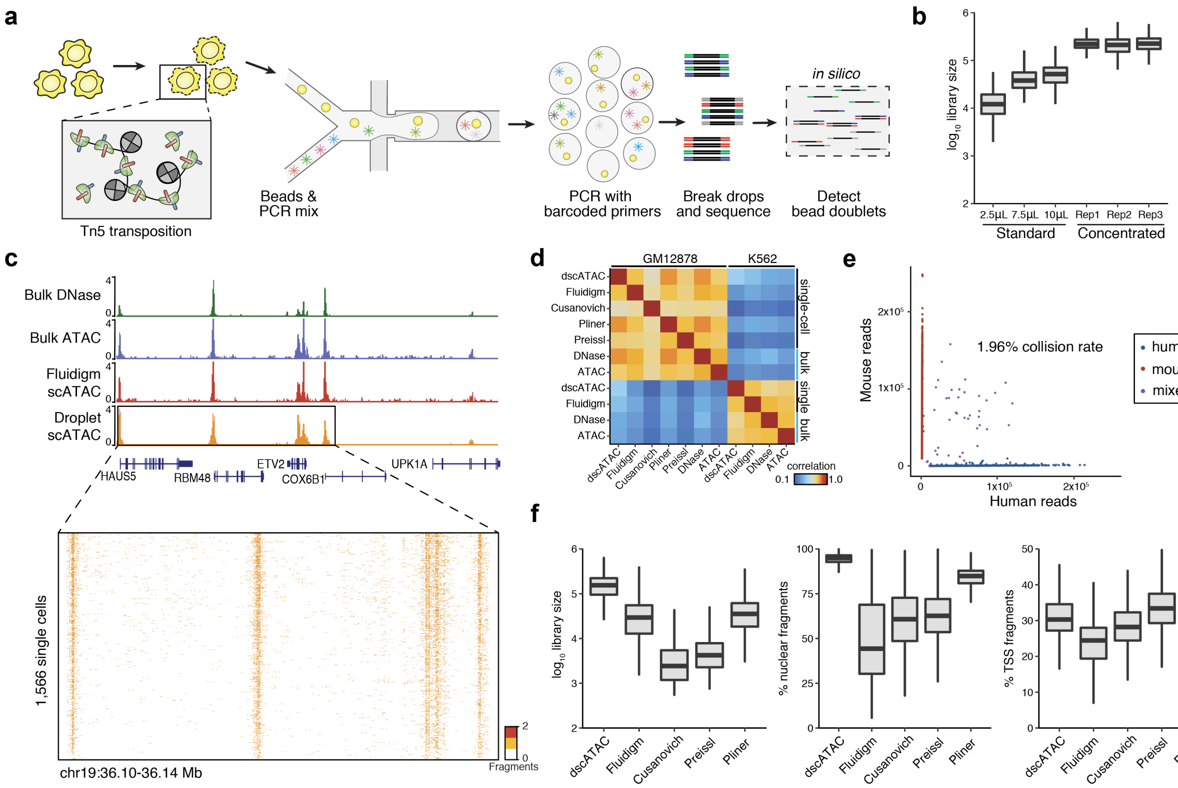

Figure 3.1: dscATAC-seq enables high-resolution characterization of open chromatin regions in single cells. **(a)** Schematic of technology. Cells are transposed with Tn5 transposase, transposed chromatin is then barcoded and amplified in a microfluidic device. **(b)** Comparison of per-cell library sizes using different Tn5 conditions using K562 cells. Three replicates (Rep1, Rep2, Rep3) are noted for the concentrated enzyme mixture (n=500 cells for each). **(c)** Comparison of the aggregate chromatin accessibility profiles from GM12878 cells using different technologies, and visualization of single-cell chromatin accessibility profiles from dscATAC-seq. Aggregate chromatin accessibility profile from dscATAC-seq is representative of ≥10 replicates. **(d)** Spearman correlation of reads in chromatin accessibility peaks across bulk and single-cell technologies for GM12878 and K562 cells (n=1 for each). **(e)** The number of unique fragments aligning to human or mouse genomes using human (GM12878) and mouse (3T3) cells at 800 beads/µL. **(f)** Quality metrics of scATAC-seq methods for GM12878 cells. Median library size for dscATAC-seq was 165,204 reads (left panel, all reads reported are passing quality filters), compared to profiles generated from the Fluidigm C118 (50,443 reads) and a recently optimized sciATAC-seq method (Pliner et al., 2018). Median fraction of mapped nuclear fragments for dscATAC-seq is 95% (middle panel). Boxplots: center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. Sample size for each method is shown in Figure C.3.c.

bead concentrations without loss of data quality (Figure C.2.l-o). To compare the efficacy of our approach, we uniformly processed cell line data (GM12878 and K562) generated using dscATAC-seq and from four other recently published approaches (Buenrostro et al., 2018; Preissl et al., 2018; Pliner et al., 2018). We found that bulk ATAC-seq (Buenrostro et al., 2013), DNase-seq (Thurman et al., 2012) and the aggregate chromatin accessibility across the different single-cell technologies (Buenrostro et al., 2018; Preissl et al., 2018; Pliner et al., 2018) were highly correlated (Figure 3.1.c,d). We also observed <2% collision rate (defined by >10% alternate species) with 800 beads/µL and 5,000 beads/µL (Figure 3.1.e and Figure C.2.n-o). Notably, this estimated collision rate (<2%) is considerably lower than other previously described high-throughput sciATAC-seq methods (>5%; Cusanovich et al. (2018)) (Figure C.3.a). Our dscATAC-seq method achieved improved library complexity per cell and number of cells per experiment without compromising the proportion of reads mapping to the nuclear genome (Figure 3.1.f and Figure C.3.b,c) – common quality metrics for scATAC-seq experiments. Notably, dscATAC-seq recapitulates known variation in TF motif activity across single GM12878 cells as previously reported (Figure C.3.d). Taken together, our new methodology provides an approach for high-resolution profiling of chromatin accessibility across thousands of single cells.

3.2.2 Epigenomic diversity of the adult mouse brain

We sought to determine whether our approach could be applied to large-scale efforts to identify cell types within complex tissues *de novo*. Thus, we applied the dscATAC-seq platform to whole brain tissues derived from two mice using our super-loaded bead concentration (5,000 beads/µL). Over 12 experimental libraries, we observed a median cell capture of 5,324/5,600 (95%), consistent with our theoretical expectation (Figure C.2.f). Cells passing additional stringent quality filters had a median of 34,046 unique nuclear reads, 58.8% reads in peaks and an average of 2.5 bead barcodes per cell for 46,653 total cells (Figure C.4.a).

To characterize differences in chromatin accessibility across cell types, we first reduced dimensionality of our mouse brain profiles by computing k-mer deviation scores (7-mers) using the chromVAR algorithm (Schep et al., 2017). Cell clusters are identified using the Louvain modularity method built from a cell nearest-neighbor graph

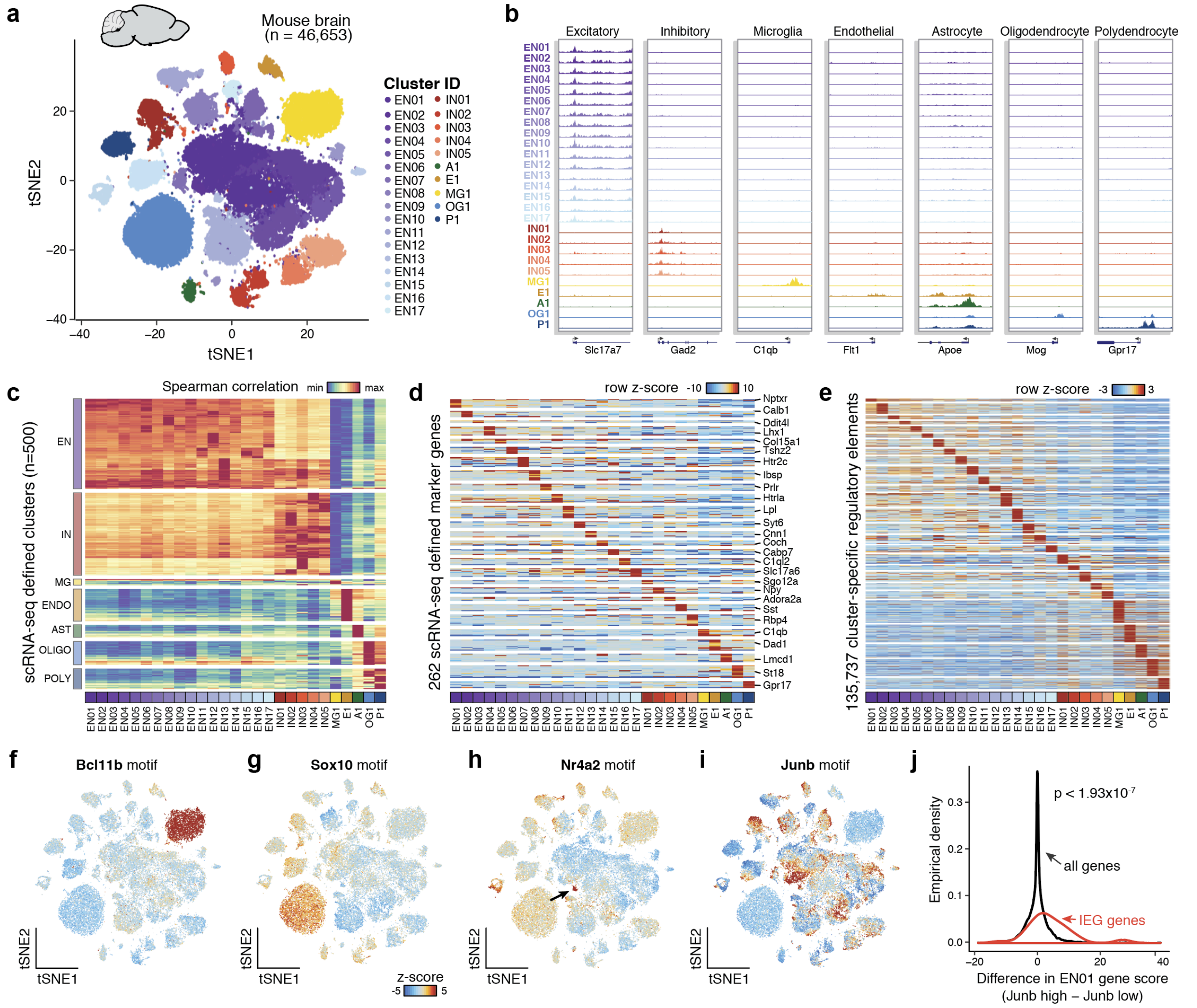

Figure 3.2: *De novo* classification of cell types in the mouse brain. **(a)** A t-SNE visualization of cells (n=46,653) derived from two mouse whole brains across 12 experimental batches. Cells are colored by their identity across 27 clusters. **(b)** Aggregate chromatin accessibility profiles per cluster surrounding the promoter region of known marker genes. Aggregate profiles were combined over 12 experimental libraries. **(c)** Correlation matrix of mouse brain clusters defined by scRNA-seq (Zeisel et al., 2018) with dscATAC-seq clusters. Margin labels indicate cell class. Each row is min/max normalized from the Spearman correlation of scRNA-seq derived marker genes. **(d)** Promoter region chromatin accessibility scores for previously defined mouse brain marker genes. Per-cluster marker genes are denoted. **(e)** Chromatin accessibility signal across 135,737 cell type-specific peaks within clusters defined in the mouse brain. **(f-h)** Cluster-specific activity of known TF regulators in the mouse brain, each panel depicts the chromVAR deviation score for each transcription factor motifs, including **(f)** Bcl11b (microglia), **(g)** Sox10 (oligodendrocytes), and **(h)** dopaminergic neurons. **(i)** Within-cluster variation shown by the Junb motif. **(j)** Comparison of promoter region chromatin accessibility scores between Junb motif high cells and low cells in EN01. Empirical densities of 47 annotated IEG compared to all annotated genes is shown. Statistical test: two-tailed, two-sample Kolmogorov-Smirnov ($n = 24,360$ genes).

using the 7-mer scores, which uncovered 27 cell clusters. We then use these 7-mer features to subsequently map each cell to a two-dimensional representation using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Figure 3.2.a). Importantly, these clusters are largely uncorrelated with known technical confounders (Figure C.4.b-d), and we observe a largely consistent pattern when compared to dimensionality reduction and clustering using latent semantic index (LSI) of our dscATAC-seq data as has been previously performed (Cusanovich et al. (2018); Figure C.4.e). For comparison with previous techniques, we also analyzed published sciATAC-seq data from two mouse brains, where we identified 13 clusters using this computational approach (Figure C.4.f), which we attribute to the fewer cells assayed (5,744 cells), lower library complexities (median 14,681), and a smaller fraction of reads in peaks per cell (median 30.0%) (Figure C.4.g,h).

To annotate these clusters, we calculated per-cluster promoter region accessibility scores (weighted-sum of chromatin accessibility ±100 kb around the TSS) (Figure C.5.a). Of note, dimension reduction using promoter region chromatin accessibility scores for all genes resulted in reduced resolution of neuronal subclusters (Figure C.5.b). We therefore used previously annotated mouse brain marker genes to correlate our promoter region accessibility scores to a recently described single-cell transcriptomic atlas of cell types across 9 regions of the adult mouse brain (Zeisel et al., 2018). We then used the highest correlation to the scRNA-seq clusters to partition these dscATAC-seq clusters into the major mouse brain cell types. These clusters include microglia (MG1), oligodendrocytes (OG1), oligodendrocyte progenitor cells (OPC; P1), astrocytes (A1), endothelial cells (E1), inhibitory neurons (IN01-IN5) and excitatory neurons (EN01-EN17) (Figure 3.2.a). Pooled ATAC-seq signal (Figure 3.2.b) and promoter region accessibility scores (Figure C.5.c-d) at known cell type specific gene markers further validated the cluster assignments. Interestingly, we also observed consistently higher library complexity and a higher ratio of distal to promoter reads per cell for annotated neurons compared to other cell types, suggesting that neurons may have overall increased chromatin accessibility at distal regulatory elements (Figure C.5.e-g).

To refine cluster annotations, we employed an optimal matching algorithm to link our promoter accessibility scores to two published scRNA-seq datasets (Saunders et al. (2018); Zeisel et al. (2018); Figure 3.2.c). Here, we identified multiple scRNA-seq clusters to be highly correlated with each of our scATAC-seq clusters, likely reflecting the nature of the annotations which classify cell types both by expression signatures and from regions of the brain. To define most likely pairs, we employed the Gale-Shipley algorithm to maximize the global correlation (Spearman) of our cluster assignments to scRNA-seq clusters (Appendix C). Differentially enriched genes in each scATAC-seq cluster provided further insights into the putative cell identities (Figure 3.2.d). For instance, chromatin accessibility enrichment of Sst in IN04 suggests that this cluster corresponds to Sst+ (Somatostatin-expressing) neurons, a defined subset of GABAergic inhibitory neurons with high levels of spontaneous activity (Urban-Ciecko & Barth, 2016). Further, Syt6, a marker of layer 6 pyramidal neurons (Ullrich & Südhof, 1995), is enriched in EN12. Htr1a and Htr2c, which encode serotonin receptors and are known markers of serotonin neurons (Meneses, 2015), are enriched in EN10 and EN07, respectively. Lhx1, a TF enriched in the suprachiasmatic nucleus that maintains synchrony among circadian oscillator neurons (Hatori et al., 2014) is enriched in EN04. In addition to the inference of cell types, our approach also enabled the unbiased identification of 135,737 cell type-specific chromatin regulatory elements (Figure 3.2.e), which further validates the unique identity of each cell cluster and provides a general resource for defining regulatory elements to drive cell type-specific reporters in effort to better understand the mouse brain (Visel et al., 2013).

To further utilize the underlying chromatin data of our resource, we sought to further examine cell type-specific TF regulators within each cluster using transcription factor motif deviations. We observed strong enrichments of the Bcl11b (Figure 3.2.f) and Sox10 (Figure 3.2.g) motifs in the microglia and oligodendrocyte clusters, respectively. These known master regulators further validate their cluster assignment using our approach. Next, we identified a highly-specific activity of the Nr4a2 motif (Figure 3.2.h), suggesting that the EN13 cluster is comprised of dopaminergic neurons, given the critical role of this TF in the development and maintenance of the dopaminergic system (Kadkhodaei et al., 2009). In addition to observing cluster-specific transcription factors, we found considerable within-cluster variability of the Junb TF motif specific to neuron clusters (Figure 3.2.i). We hypothesized that this variability may reflect neural activity driving immediate early gene (IEG) expression (Yap & Greenberg, 2018). Indeed, this hypothesis was supported by a statistically-significant enrichment of 47 previously annotated IEG genes between the Junb-high (z-score >0) and Junb-low (z-score <0) cells within the

EN01 cluster (two-sample Kolmogorov-Smirnov test; p=1.93x10-7; Figure 3.2.j). Altogether, we observe that the dscATAC-seq platform provides a powerful means for defining and annotating cell types and states while further identifying cell type-specific chromatin features.

3.2.3 Droplet-based sciATAC-seq for massive-scale profiling

Although the dscATAC-seq approach can be scaled to generate data for large cell numbers by simply performing the experiment across many replicates, as shown above (Figure 3.2), we reasoned that we could further increase cell throughput by surpassing Poisson loading of cells in droplets (one cell per droplet). We therefore sought to combine this approach with combinatorial indexing (Cusanovich et al., 2015) to improve throughput and enable multiplexing of multiple samples in a given experiment. To achieve this, we developed a method for droplet-based sciATAC-seq (dsciATAC-seq), wherein we load Tn5 transposase with barcoded DNA adapters to add well-specific DNA barcodes to open chromatin. Following barcoded transposition, transposed cells are pooled and loaded at high density to co-encapsulate multiple Tn5 barcoded cells with multiple beads in each droplet (Figure 3.3.a,b). Here, each individual cell may be identified by both the per droplet bead barcode and the well-specific Tn5 barcode, enabling an increase in cell throughput proportional to the initial number of Tn5 barcodes used in the experiment. Thus, using our droplet-based platform with barcoded Tn5 reactions increases the number of theoretical barcode combinations to enable a greater cell or sample throughput, if cells originate from different samples.

First implementing this technology with 24 transposase barcodes, we generated high-quality chromatin accessibility profiles for up to 50,000 cells in a single well of the device representing one experimental sample. Species mixing analysis (using Tn5 barcode-aware parsing; Appendix C) confirmed that we could increase cell throughput approximately 10-fold while maintaining a collision rate lower than 5% using 24 transposase barcodes (Figure 3.3.c-e and Figure C.6.a) and confirmed a further reduction in the overall detected collision rates at large cell inputs with 48 barcodes (Figure C.6.b). Altogether, in this cell titration experiment, we generated 274,144 single-cell profiles demonstrating the massive scalability of this approach. Notably, to perform this experiment, we sepa-

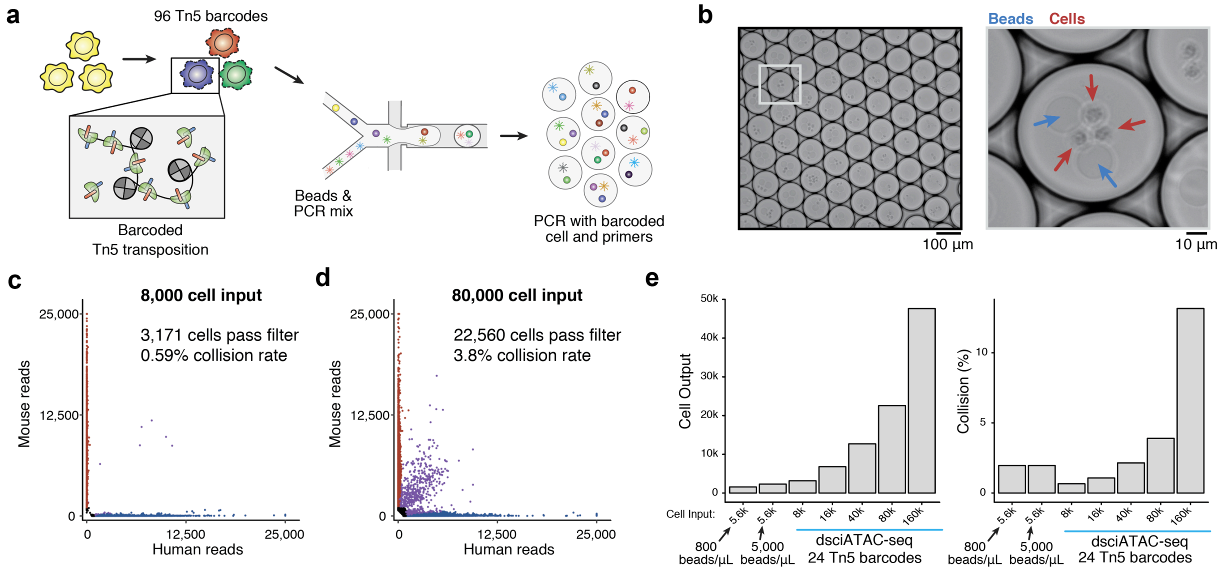

Figure 3.3: dsciATAC-seq enables massive scale single-cell experiments. **(a)** Schematic of dsciATAC-seq. Cells are transposed with barcoded Tn5, pooled, and then further processed through the droplet PCR microfluidic device. **(b)** Representative image of droplets containing multiple beads and cells. Blue arrows indicate beads and red arrows indicate transposed nuclei. **(c,d)** The number of unique reads aligning to the mouse or human genome from dsciATAC-seq profiles of human (K562) and mouse (3T3) cells with **(c)** 8,000 and **(d)** 80,000 cell input. **(e)** Summary of species mixing and cell yield results at variable cell inputs.

rately purified and *in vitro* assembled barcoded Tn5 (Picelli et al., 2014). As this proof-of-concept experiment did not utilize the optimized Tn5 concentration described in Figure 3.1.b, we observed fewer reads per cell but maintained a high fraction of reads in peaks (72.2%). Together, these experiments demonstrate that barcoded Tn5 can enable super-loading of cells into droplets to achieve a significantly greater throughput for generating epigenomic profiles from 10^4 to 10^5 single-cells per experiment.

3.2.4 Chromatin accessibility profiling of human bone marrow

Barcoded Tn5 transposition enables a significantly increased cell throughput and the opportunity to multiplex scATAC-seq to multiple conditions or samples. Notably, tissue-scale perturbations (Bendall et al., 2011) have been used to uncover diverse cell response dynamics (Bodenmiller et al., 2012). We therefore reasoned that pooled stimulation across heterogenous cell types within bone marrow mononuclear cells (BMMCs) would provide unique avenues to understand the functional roles of epigenomic diversity within the human bone marrow. To achieve this, we used dsciATAC-seq using 96 transposase barcodes to profile BMMCs from two human donors before (untreated controls) and after stimulation, producing chromatin accessibility profiles for a total of 136,463 cells passing quality filters (Figure 3.4.a and Figure C.7.a-f).

The reference map of 60,495 resting cells (untreated controls) revealed the major lineages of hematopoietic differentiation *de novo*. To analyze these reference datasets, we projected the untreated BMMCs onto hematopoietic development trajectories using a reference-guided approach, whereby single-cells are scored by principal components trained on bulk sorted hematopoietic ATAC-seq profiles (Corces et al. (2016); Figure 3.4.a; Appendix C). With this approach, we are able to visualize and predict cell labels given the bulk reference map of epigenomic states (Figure 3.4.b). Further, using the Louvain modularity method, we identified 15 distinct clusters from the 60,495 resting cells, which recapitulate the major constitutive cell types in the human hematopoietic system (Figure 3.4.c). These *de novo* derived single cell clusters reflect changes in chromatin accessibility mediated by key lineage-specific transcription factor motifs, including those associated with the step-wise progression of B-cell development from hematopoietic stem cells (HSCs) to mature B-cells (Figure 3.4.c,d). While our embedding and

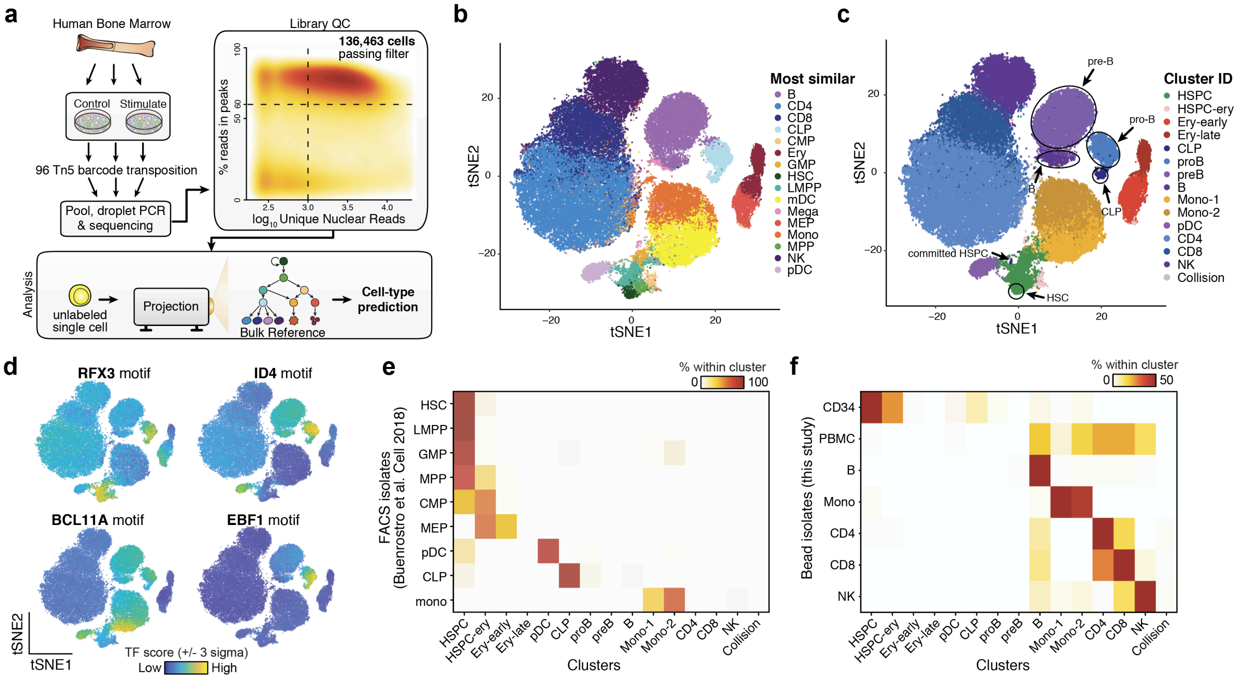

Figure 3.4: Profiling human bone marrow cells with dsciATAC-seq reveals the major lineages of hematopoietic differentiation. (a) Schematic of experimental and computational workflow used to assess bone marrow mononuclear cells dsciATAC-seq data. 96 Tn5 transposase barcodes are used to define different donors and stimulation conditions. Library QC box displays the summary of data passing quality filters across all assayed cells. 136,463 cells were identified passing filters of 60% reads in peaks and 1,000 unique nuclear reads. **(b,c)** Two dimensional t-SNE embedding of single bone marrow mononuclear cells without stimulation (n=60,495 cells). Cells are colored by **(b)** the most correlated cell type from a bulk ATAC-seq reference or **(c)** 15 *de novo* defined cluster assignments covering known hematopoietic cell types. Cell types covering the B-cell differentiation trajectory are highlighted. **(d)** Single-cells are colored by TF motif accessibility scores, computed using chromVAR, for the motifs RFX3, ID4, BCL11A and EBF1. **(e,f)** Confusion matrix showing the percent overlap between **(e)** published scATAC-seq data and **(f)** isolated subsets collected in this study.

cell clustering were defined from bulk projections in keeping with our previous work, we note a considerable concordance of these data using our *de novo* k-mer strategy (Figure C.7.g-i). Furthermore, we observed unexpected epigenomic heterogeneity across transcription factor motifs, including CEBPD and BCL11A, within monocyte clusters (Mono-1 and Mono-2), which likely reflects the heterogenous developmental transitions from myeloid progenitors to mature monocytes, myeloid dendritic cells (mDCs) and granulocytes (Figure C.8.a).

To validate the clusters and cell type annotations from our approach, we assigned previously profiled FACS-sorted single-cell ATAC-seq profiles from progenitors in human bone marrow and peripheral blood monocytes (2,034 cells; see Buenrostro et al. (2018)) to clusters defined here. We classified these published single-cell data to clusters based on the minimum Euclidean distance of a single-cell profile to a cluster medoid. We observed significant overlap between each isolated subset and its corresponding cluster in the dsciATAC-seq data, validating the progenitor cell type annotations (Figure 3.4.e). Furthermore, we performed dscATAC-seq on CD34+ bone marrow progenitor cells, peripheral blood mononuclear cells (PBMCs), and from bead-enriched subpopulations from PBMCs to derive a total of 52,873 cells, which validated our cluster label assignments for mature cell types (Figure 3.4.f and Figure C.8.b). We also used an orthogonal approach to visually validate these findings by dimensionality reduction using the uniform manifold projection (UMAP) algorithm (Becht et al., 2018), which allows for data to be projected onto the dsciATAC-seq base dimensionality (Figure C.8.c-f). Collectively, we have used this approach to define a reference epigenomic atlas of cell states within hematopoietic cells in the human bone marrow, highlighting the applicability of our combinatorial approach to generate accurate large-scale epigenomic maps to define cell types within primary human tissues.

3.2.5 Regulatory consequences of multi-lineage stimulation

Our multiplexed, droplet-based sciATAC-seq method further provides a unique opportunity to decipher regulatory consequences of perturbation without concerns for batch-effects confounding experimental results. To characterize the response of each immune cell cluster to stimulation conditions, we explored the differences between our untreated control cells and *ex vivo* cultured and LPS-stimulated BMMCs (Figure 3.4.a). To determine trans-acting regulators altered in response to these perturbations, we developed an analytical strategy wherein we compute differential TF scores by i) defining a K-nearest neighbor map connecting stimulus to control conditions, and ii) computing differential TF scores by calculating the difference in TF scores between each cell and the average of 20-nearest stimulus cells (Figure 3.5.a).

Interestingly, we found significant and highly correlated epigenomic responses to both *ex vivo* culture and LPS stimulation (Figure C.9.a-e), suggesting that the effects of *ex vivo* culture dominates those induced by LPS. For clarity we simply refer to these conditions as "stimulation" for downstream analysis. With this stimulation data representing the full spectrum of bone marrow hematopoietic cell states, we found cell type-specific induction of a diverse repertoire of TF motifs (Figure 3.5.b-d and Figure C.9.f-j). This list of differential TFs included induction of the Jun and NFkB motifs, largely localized to human hematopoietic stem and progenitor cells (HSPCs) (Figure 3.5.b,c), depletion of the SPIB motif in myeloid cell types (Figure 3.5.d) and relatively weak induction of MAFF (myeloid) and IRF8 motifs (MEP and CLP to pre-B) (Figure C.9.i,j). Interestingly, Jun and NFkB were largely correlated in HSPCs, with the exception of CLP and early erythroid differentiation, wherein cells appeared to respond exclusively by NFkB motif induction (Figure 3.5.b).

Next, we examined the cis-regulatory consequences of stimulus across our multi-lineage defined cell states. To compute differential chromatin accessibility peaks within each cluster, we devised a permutation test per peak, permuting control and perturbation cell labels, which allowed us to improve the robustness of our statistical methods by considering each cell as an independent observation (Figure C.9.k-l; Appendix C). This analysis revealed a total of 9,638 distinct stimulus-responsive chromatin accessibility peaks (FDR 1%). Interestingly, we broadly observed a gain in the total number of chromatin accessibility peaks, represented by the Mono-1 cluster with 2,114 peaks gained compared to 1,264 peaks lost (binomial p-value < 2.2e-16) (Figure 3.5.e). A global gain in chromatin accessibility upon stimulation was also corroborated by an approximate 20% gain in the average library complexity per-cell. The most prominent cell types that responded to the stimulation treatment included the two monocyte and CD4 T-cell clusters. Unexpectedly, we also observed 501 chromatin accessibility peaks gained in the HSPC cluster, and approximately 34% of these HSPC gained peaks were unique to HSPCs (Figure

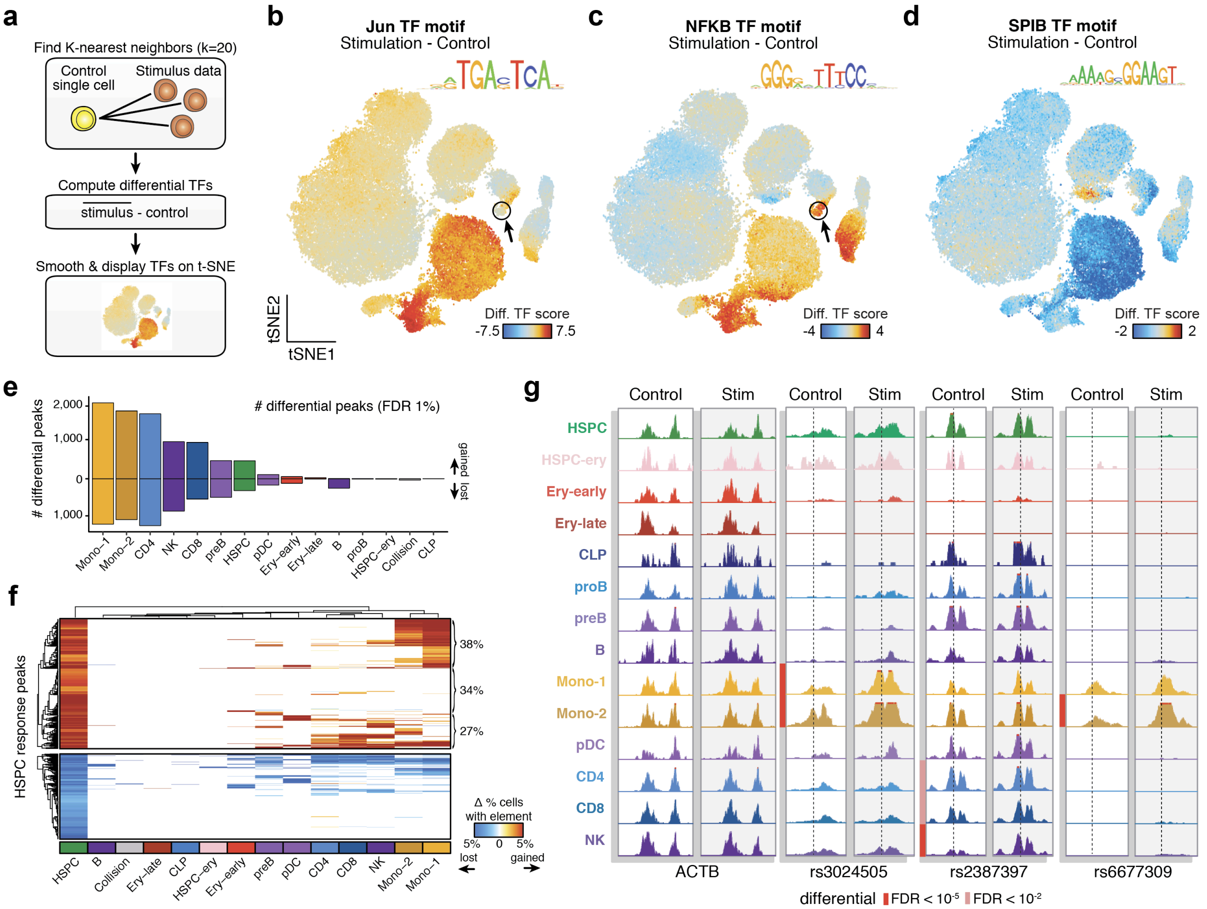

Figure 3.5: Identification of stimulus-response regulators in human bone marrow. **(a)** Schematic depicting the computational workflow for comparing stimulus versus control single-cell data. **(b-d)** Differential TF deviation scores for **(b)** Jun, **(c)** NF-κB and **(d)** SPIB motifs in response to stimulation for n=60,495 resting cells. **(e)** Summary of the number of differential chromatin accessibility peaks across each cluster at a false discovery rate (FDR) of 1% after a two-sided permutation test. Bars above the zero line represent gained chromatin accessibility peaks, bars below the zero line represent lost chromatin accessibility peaks. **(f)** Hierarchical clustering of peaks (top) gained or (bottom) lost across clusters, restricted to the differential peaks identified in HSPCs. **(g)** Locus specific views of the ACTB promoter and three fine-mapped variants identified through genome-wide association studies. The dotted line represents the location of the SNP in each window.

3.5.f), thus uncovering an HSPC-specific stimulus response signature. Altogether, considering the TF motif and peak-specific analyses, we find that HSPCs respond to stimulus using the NFkB and Jun TF motifs to drive an HSPC-specific stimulus response. This finding supports reports suggesting that HSPCs are responsive to interferon-mediated immune signaling (Essers et al., 2009; Espín-Palazón et al., 2014), and may be used to further characterize the regulatory basis of interferon signaling in HSPCs to nominate chemical inhibitors to facilitate *ex vivo* expansion and gene editing of HSCs for hematopoietic stem cell transplantation (HSCT).

We further hypothesized that this approach to uncover cell type-specific stimulation changes could elucidate mechanisms of relevant cell types and regulatory regions for variants implicated in genome-wide association studies (Farh et al., 2015). Towards this effort, we observed stimulation response chromatin accessibility peaks near the IL10 locus in monocytes overlapping the pleiotropic rs302405 variant locus associated with Type 1 Diabetes (posterior probability (PP) = 0.38), Crohn's Disease (PP=0.40), and Ulcerative Colitis (PP=0.41), as well as chromatin accessibility gains at the variant rs2387397 associated with Celiac Disease (PP=0.32) within the natural killer (NK) and T-cell clusters (Figure 3.5.g). Additionally, we observed a Mono-2 stimulation-specific peak overlapping rs6677309, a fine-mapped variant associated with multiple sclerosis (PP=0.49), near the CD58 locus (Figure 3.5.g). Interestingly, CD58 presentation by activated monocytes has been shown to expand CD56+ NK cells (Lopez et al., 2001), which may promote an autoimmune response in multiple sclerosis (Laroni et al., 2016). Overall, this single experiment comprising 60,495 resting and 75,968 stimulated cells enabled the unbiased discovery of regulatory changes across stages of hematopoietic differentiation, and the unbiased identification of the regulatory consequences of *ex vivo* perturbation across multiple lineages, providing new opportunities to better define cell types within complex tissues and their relationship to stem cell therapy and autoimmune disease.

3.3 Discussion

In the genomics era of cell atlases, a major goal of single-cell methods is to provide an unbiased classification of cell types and the epigenomic, transcriptomic and proteomic features that define them (Regev et al., 2018). We find that scATAC-seq maps can provide information rich measurements of cells (10^5 fragments per cell), which enables the identification of cell types and their underlying regulatory elements. Further, previous work has suggested that regulatory element activity may be a more accurate reflection of cell potential and perhaps more cell type-specific than gene expression measurements (Corces et al., 2016). The scATAC-seq approach described here produces single-cell profiles at higher-throughput, improved yield and higher sequencing efficiency than previous scATAC-seq methods, providing a robust platform for identifying new cell types within heterogeneous tissues. We expect that the combination of this scATAC-seq approach with scRNA-seq profiling will provide a more accurate definition of cell types and further integration of these data (Stuart et al., 2019) will enable opportunities to define mechanistic gene regulatory models to understand their function.

We present a series of technological innovations leading to a high-throughput epigenomic profiling approach that enables super-loading loading of cells and beads into microfluidic droplets. To achieve this, we have developed a computational approach to identify droplets with multiple barcoded beads and paired this approach with combinatorial indexing by barcoded transposition to add multiple cells per droplet. Combining these approaches dramatically improves cell throughput to approximately 25,000 cells per well (100,000 cells per droplet device), which we expect may be further improved with optimizations of the approach and additional Tn5 barcodes. More generally, we expect this conceptual framework of combinatorial indexing coupled with a microfluidics device may be compatible with other methods for high-throughput PCR and other single-cell genomics assays leveraging combinatorial indexing for cell barcoding (Cao et al., 2018; Mulqueen et al., 2018).

This approach allows for multiplexing of many samples in a single experiment. In this work, we multiplex control and perturbation conditions across an entire tissue, enabling us to define shared and cell type-specific regulatory changes induced upon stimulation across diverse cell types. These advances for multiplexing experiments along with advances in high-throughput sequencing, opens new opportunities to define not only cell type-specific chromatin accessibility, but also changes across diverse genetic and environmental conditions. As such we expect this approach to be used to profile epigenomic variation across healthy individuals or from cohorts of diseased patients to determine the functional roles of both regulatory elements and cell types underlying common traits or in disease (Rakyan et al., 2011). Altogether, these advances enable a new era of single-cell epigenomic studies at a massive scale, providing a powerful new tool to connect the vast repertoire of DNA regulatory elements to function.

3.4 Acknowledgements

We thank members of the Buenrostro lab for useful discussions and critical assessment of this work. We thank D. Norton at Bio-Rad, the SureCell ATAC-Seq Library Prep Kit product manager, for enabling the collaboration with the Broad Institute and Harvard University. Other Bio-Rad colleagues that we would also like to acknowledge for establishing and providing droplet related consumables are D. Do, B. Zhang, P. Pattamatta, S. Cater, L. Frenz, D. Greiner, and J. Agresti. We also recognize L. Christiansen, A. Yunghans, and L. Watson from the Illumina Assay Development team, in addition to F. Zhang and F. Schlesinger at Illumina for their bioinformatics contributions. We are grateful to the Zhang lab (Broad Institute) for providing the Tn5 for combinatorial experiments. JDB, CAL, FMD, and VKK acknowledge support by the Allen Distinguished Investigator Program through The Paul G. Allen Frontiers Group. This work was further supported by the Chan Zuckerberg Initiative (CZI). CAL is supported by an NIH F31 grant (F31CA232670).

3.5 Author Contributions

CAL, FMD, and JGC conceived and designed the study with supervision from RL and JDB. FMD, JGC, and ASK performed experiments and generated the data. CAL developed the bead merging computational approach. CAL lead analyses of data with VKK and ZDB. FJS proposed the droplet scATAC-seq approach and oversaw the proof-of-concept studies performed by DP. MJA assisted in the development of computational resources. CAL, FMD, and JDB wrote the manuscript with input from all authors. RL and JDB jointly supervised this work.

3.6 Code and data availability

Raw sequencing files and processed files for all data generated in this study were deposited at Gene Expression Omnibus (GEO) under accession number GSE12358. Complete code and documentation for the software suite developed in this study (bap - bead-based ATAC-seq processing tool) is available on GitHub under the following weblink: https://github.com/caleblareau/bap. Scripts corresponding to the analyses contained in this paper are further provided at: https://github.com/buenrostrolab/dscATAC_analysis_code.

Reality is merely an illusion, albeit a very persistent one.

Albert Einstein

4

Inference & effects of barcode multiplets in droplet-based single-cell assays

Abstract

A widespread assumption for single-cell analyses specifies that one cell's nucleic acids are predominantly captured by one oligonucleotide barcode. Here, we show that ~13 − 21% of cell barcodes from the 10x Chromium scATAC-seq assay may have been derived from a droplet with more than one oligonucleotide sequence, which we call "barcode multiplets". We demonstrate that barcode multiplets can be derived from at least two different sources. First, we confirm that approximately 4% of droplets from the 10x platform may contain multiple beads. Additionally, we find that approximately 5% of beads may contain detectable levels of multiple oligonucleotide barcodes. We show that this artifact can confound single-cell analyses, including the interpretation of clonal diversity and proliferation of intra-tumor lymphocytes. Overall, our work provides a conceptual and computational framework to identify and assess the impacts of barcode multiplets in single-cell data.

4.1 Introduction

Droplet-based partitioning systems have become an essential tool for single-cell genomics research. In contrast to plate-based single-cell assays, droplet-based methods, including scRNA-seq (Klein & Macosko, 2017; Zheng et al., 2017b) and scATAC-seq (see Chapter 3 and Satpathy et al. (2019)) enable profiling of thousands of cells in a single experiment. The marked increase in throughput is achieved by parallel barcoding of cellular nucleic acids with beads containing high-diversity DNA barcodes. Critically, downstream computational analyses assume that one barcode sequence equates to one cell.

In this work, we provide multiple lines of evidence that indicate that cells often associate with multiple barcodes by (i) multiple beads occurring within the same droplet or (ii) heterogeneity of oligonucleotide sequences within a single bead (Figure 4.1a). Here, we refer to these instances whereby multiple DNA barcodes occur within the same droplet as "barcode multiplets". We find that barcode multiplets can considerably impact single-cell analyses and demonstrate that rare cell events (*e.g.*, the analysis of cell clones) can be particularly affected by this artifact. Further, we provide a computational solution to identify these barcode multiplets in existing single-cell datasets, particularly from the scATAC-seq platform. Finally, we provide recommendations to mitigate these biases in existing assays.

4.2 Results

4.2.1 Bead multiplets quantified through imaging

While cell doublet rates are routinely quantified by species-mixing analyses, analogous multiplet rates for bead loading are scarcely discussed. Importantly, commonly used droplet-based assays (*e.g.* the 10x Chromium platform) leverage a close-packing ordering of beads (Abate et al., 2009) to load predominantly one bead per droplet and achieve "sub-Poisson" loading. First, we sought to test this assumption and empirically quantify bead loading within droplets. To achieve this, we loaded hydrogel training beads into droplets following recommended guidelines and imaged the resulting solution. Beads were readily visible and quantifiable per droplet (Figure 4.1.b; Figure D.1.a-d), enabling empirical estimates of the number of beads per droplet. A total of 3,865 droplets spanning 30 total fields of view (FOV) over three experimental replicates were quantified (Appendix D). Importantly, while the training beads do not differ from those used in the regular protocol, the training buffer is required to visualize beads after loading.

On average, we found that 16.1% of droplets contained no beads, 80.0% contained exactly one bead, and 3.9% had two or more beads (Figure 4.1.c). These results were consistent with the previously reported results of this platform (Zheng et al., 2017b) and confirm the sub-Poisson loading of beads into droplets (compare to Figure D.1.e for optimal Poisson loading). While the mean of the bead loading was consistent with previous reports, we note considerable run-to-run variability from our imaging replicates, ranging from 0.8% to 8.4% (Figure D.1.f). Furthermore, we noted occurrences of large droplets with multiple beads (Figure D.1.g) that likely originated from the errant merging of several individual droplets, yielding another source of potential barcode multiplets. While our imaging results indicate that the occurrence of bead multiplets likely varies between machines and individual runs, we note that the training kits are only a proxy for the reagents used in producing single-cell data,

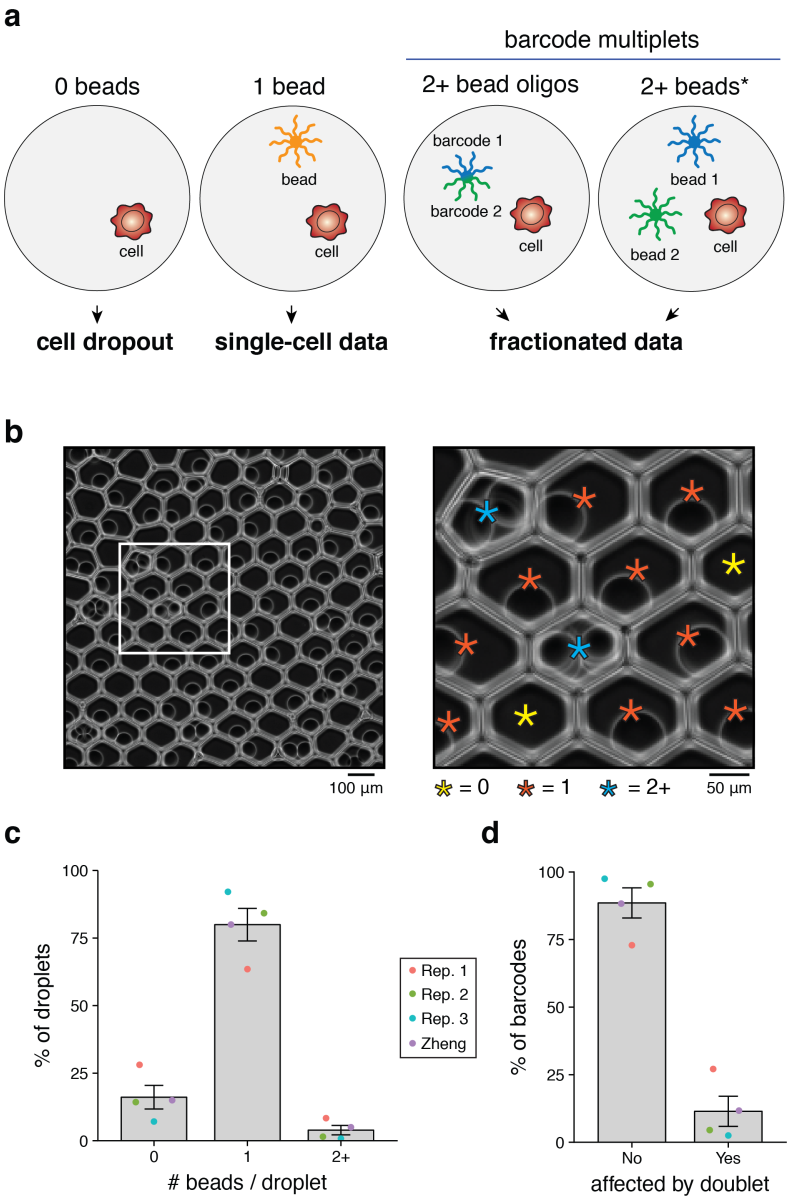

Figure 4.1: Quantification of barcode multiplets from multiple beads in 10x **Chromium platform. (a)** Schematic of bead loading variation and phenotypic consequences. Droplets with 0 beads fail to profile nucleic acid from the loaded cell ("dropout") whereas barcode multiplets fractionate the single-cell data. Barcode multiplets can be generated by either heterogeneous barcodes on an individual bead or two or more beads loaded into the same droplet. The * indicates the bead multiplet that can be quantified via imaging. **(b)** Representative example of beads loaded into droplets from the 10x Chromium platform. The white box is magnified 3x for the panel on the right, revealing multiple beads loaded into droplets. Stars indicate beads (except 0) and are colored by the number of beads contained in the droplet. The image is representative of a total of 30 fields of view taken from 3 independent experiments. **(c)** Empirical quantification of number of bead barcodes based on image analysis over 3 replicates with previously published data (Zheng et al., 2017b). **(d)** Percent of barcodes associated with multiplets under the distribution observed in **(c)**. Error bars represent standard error of mean over the experimental replicates.

and may reflect a higher rate of bead doublets. Though imperfect, our results suggest that multiple beads may co-occur in droplets and motivates additional computational analysis to determine potential barcode multiplets

While our estimate of the occurrence of multiple beads in droplets confirms previous reports2, we emphasize that this problem is exacerbated when considering potential barcodes in single-cell data. On average, we estimate that 11.4% of barcodes would represent barcode multiplets, reflecting droplets with heterogeneous oligonucleotide sequences (Figure 4.1.d; Appendix D). Moreover, we note that imaging provides a lower-bound estimate for the true occurrence of barcode multiplets for two reasons. First, droplets with four or more beads were assigned a count of four since the exact number of beads could not be reliably determined in these instances (*e.g.* Figure D.1.d). Second, imaging cannot evaluate the possibility of heterogeneous beads, a second class of artifact that leads to barcode multiplets (Figure 4.1.a). Despite the alarmingly high prevalence of barcode multiplets, the effect of this confounding phenomenon has not been systematically considered in single-cell analyses. Intuitively, these observed barcode multiplets fractionate data from the cell to multiple barcodes, resulting in a reduction of data per cell and the substantial overestimation of the total number of cells sequenced by artificial synthesis of barcodes reflecting the same single cell. With this artifact could be confirmed by imaging, we sought to further understand its properties and effects in single-cell data.

4.2.2 Identifying barcode multiplets in 10x scATAC-seq data with bap

Recently, we developed a computational framework called bead-based ATAC processing (bap), which identifies instances of barcode multiplets in droplet single-cell ATAC-seq (dscATAC-seq; see Chapter 3). Critically, we discriminate between multiple true cells and barcode multiplets by considering the Tn5 insertion sites, noting that barcode multiplets would amplify the same exact fragments (Figure 4.2.a; Figure D.2). Thus, our computational approach leverages the molecular diversity of Tn5 insertion sites across the genome to identify pairs of barcodes that share more insertion sites than expected and merge these corresponding barcode pairs (Figure 4.2.a). Previously, we utilized bap to facilitate super-loading beads into droplets to achieve a ~95% cell capture rate with a mean 2.5 beads/droplet (see Chapter 3). Here, we reasoned that bap may identify barcode multiplets in 10x

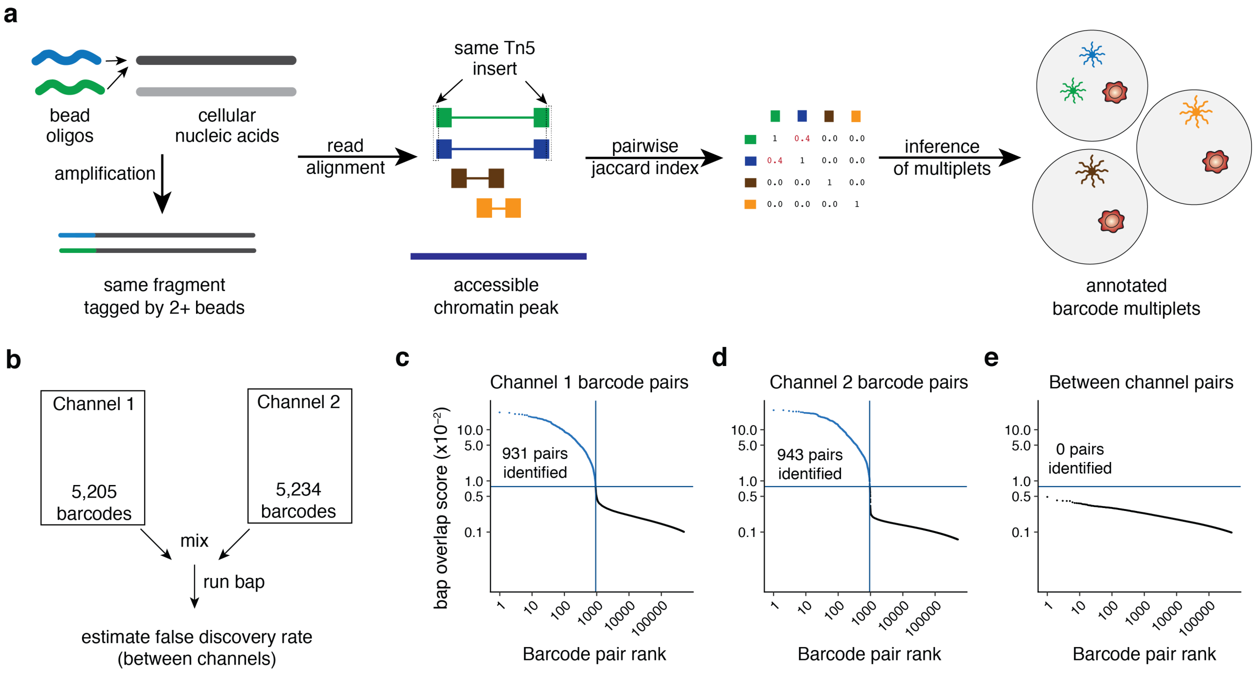

Figure 4.2: Verification of bap to identify barcode multiplets using 10x **scATAC-seq data. (a)** Schematics of methodology to detect barcode multiplets whereby cellular nucleic acids are tagged by two different oligonucleotide sequences and later inferred from sequencing a scATAC-seq library from the same Tn5 insertions per fragment. **(b)** Schematic of mixing experiment. Two channels were combined and the resulting merged files were analyzed with bap. **(c-e)** Knee plots comparing the top 500,000 barcode pairs from **(c)** only channel 1, **(d)** only channel 2, and **(e)** between channels. The number of pairs calls is indicated by the number of points above the blue horizontal line (Appendix D).

scATAC-seq data.

After updating bap to facilitate processing of the 10x scATAC data (Figure D.2; Appendix D), we conducted an initial *in silico* experiment in order to verify the applicability of our approach to 10x scATAC-seq data. Here, we combined two channels from a similar biological source (~5,000 cells of peripheral blood mononuclear cells; PBMCs) and executed bap on the resulting combination (Figure 4.2.b; Appendix D). As any barcode pairs merged between channels would be false positives, our approach facilitated an estimation of the false positive rate of our approach in 10x data. After executing bap with the default parameters, 1,874 barcode pairs were identified as sharing an unusual number of shared transposition events. Specifically, 931 pairs from channel 1 (Figure 4.2.c) and 943 pairs from channel 2 (Figure 4.2.d) were identified. However, zero pairs were identified between channels (Figure 4.2.e), indiciating a very low false positive rate for bap when applied to this assay. Moreover, the shape of the ranked-ordered barcode pair curves for the channels separately were distinct from the between-channel curve (Figure 4.2.c-e). Overall, these results support the utility of bap in inferring barcode multiplets from the 10x platform.

After establishing the applicability of bap for 10x scATAC-seq data, we sought to better understand the properties of barcode multiplets determined by bap, focusing on two datasets ("This Study" and "Public"; Appendix D) of ~5,000 human PBMCs (Figure 4.3.a). Overall, we estimated the percentage of barcodes in multiplets were 13.2% (This Study; Figure D.3.a) and 17.6% (Public; Figure 4.3.b). These cell barcodes were identified from the high-quality, error-corrected barcode sequences from CellRanger with abundant reads in peaks. Additionally, since individual barcodes in the space of all possible barcodes are separated by a minimum Hamming distance of three in the 10x platform, the high prevalence of barcode multiplets is unlikely to be caused by sequencing errors. Importantly, these implicated barcodes are normally considered in downstream analyses, including cell clustering and clonotype abundance estimates. Furthermore, we suggest that additional multiplets are present in the library but likely did not pass thresholds for reads detected due to the fractionation of data associated with these barcodes (Figure D.3.b; Appendix D).

Surprisingly, from these experiments, we observed instances in both datasets where barcode multiplets contained at least 7 distinct barcodes. In particular, we observed two instances of multiplets containing 9 unique barcodes in the Public dataset. Here, each implicated barcode contained a restricted longest common subsequence (rLCS) of 9 (Figure 4.3.c; Appendix D). As such, we suggest that these barcode multiplets likely reflect error during barcode synthesis resulting in a single bead with multiple barcodes, resulting in a "complex bead" (Figure 4.1.a). Visualization of these barcode multiplets from dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE) confirmed these barcodes reflect markedly similar chromatin accessibility profiles (Figure 4.3.d; D.3.c). Overall, barcode multiplets generally co-localized with barcode singlets and do not dramatically alter the interpretation of cell types in an embedding (Figure 4.3.e). However, we find that certain regions of the t-SNE embedding contained a disproportionate concentration of barcode multiplets, which may lead to

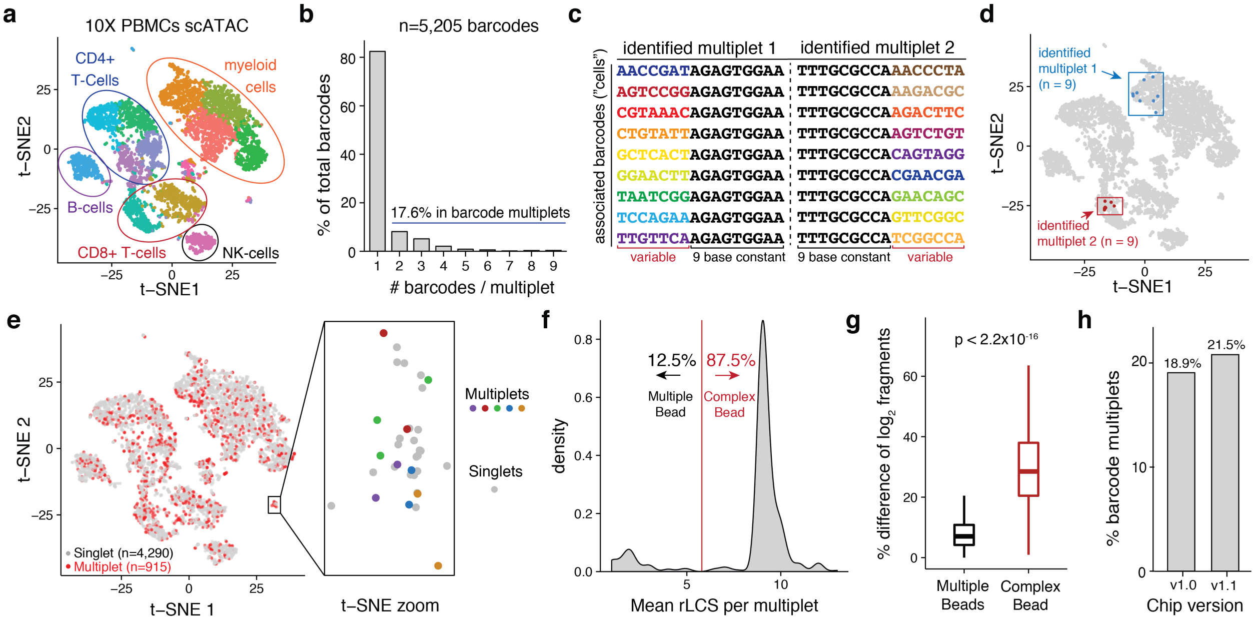

Figure 4.3: Inference and effect of barcode multiplets in single-cell ATAC-seq data. **(a)** Default t-SNE depiction of public scATAC-seq PBMC 5k dataset. Colors represent cluster annotations from the automated CellRanger output. **(b)** Quantification of barcodes affected by barcode multiplets for the same dataset (identified by bap). **(c)** Depiction of two multiplets each composed of 9 oligonucleotide barcodes. Barcodes in each multiplet share a long common subsequence, denoted in black. **(d)** Visualization of two barcode multiplets from **(c)** in t-SNE coordinates. **(e)** Visualization of all implicated barcode multiplets from this dataset. The zoomed panel shows a small group of cells affected by five multiplets, indicated by color. **(f)** Empirical distribution of the mean restricted longest common subsequence (rLCS) per multiplet. A cutoff of 6 was used to determine either of the two classes of barcode multiplets. **(g)** Percent difference of the mean log2 fragments between pairs of barcodes within a multiplet. The reported p-value is from a two-sided Kolmogorov–Smirnov test. Boxplots: center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. **(h)** Overall rates of barcode multiplets from additional scATAC-seq data comparing v1.0 and v1.1 (NextGEM) chip designs.

errant identification of presumed rare cell types (*e.g.* 5 unique multiplets shown in Figure 4.3.e).

To further elucidate these identified barcode multiplets, we annotated these barcodes with graph-based Louvain clusters (produced using the default CellRanger execution). As expected, we observed a significant enrichment of barcode multiplet pairs occurring in the same cluster (91.1% for This Study; 74.1% for Public) compared to a permuted background (11.6% and 8.6% respectively; Figure D.3.d; Appendix D). We note that barcode multiplets not within the same cluster largely reflect barcodes split between multiple clusters of the same cell type (*e.g.* myeloid cells; see Multiplet 5 in Figure D.3.c). Additionally, we observed a statistically-significant association between the Louvain cluster assignment and inferred barcode multiplet status for both This Study (p=0.0065) and Public datasets (p=2.46e-05; chi-squared test; Appendix D). These results indicate that the barcode multiplets can occur in clusters unevenly, potentially confounding inferences regarding cell-type abundance. Additionally, through iteratively downsampling and re-executing bap, we confirmed the stability of our metric with sequencing depths as low as a median 10,000 fragments detected per barcode (Figure D.3.e; Appendix D), confirming the broad utility of this approach. Overall, as these barcode multiplets represent quasi-independent observations of the accessible chromatin landscape of the same single cell, we suggest that these identified barcode multiplets may be utilized in a variety of different useful applications. Examples include determining sequencing saturation, inferring sequencing biases, and benchmarking bioinformatic clustering approaches. Furthermore, these barcode multiplets can be merged to improve data quality (see Chapter 3).

4.2.3 Contributions of types of barcode multiplets

Having verified the overall detection of the effects of barcode multiplets in these datasets, we sought to determine the relative contributions of each source of barcode multiplets to the overall abundance (Figure 4.1.a). To achieve this, we established a null distribution by computing the rLCS for random pairs of barcodes from the 10x whitelist (Appendix D). Over 1,000,000 sampled pairs, we determined that pairs with an rLCS ≥6 were extremely uncommon assuming an independent co-occurrence (<0.5% probability of co-occurring; Figure D.3.f). Thus, for inferred multiplets with a mean rLCS ≥6, we interpret these to be most likely caused by heterogeneous barcodes within a single bead. After computing the mean rLCS between pairs of barcodes per multiplet, we determined that 87.5% of multiplets were likely caused by these complex or heterogeneous beads in the Public dataset (Figure 4.3.f). Using this classification, we could further estimate the prevalence of these complex beads to be 6.41% in this dataset (Appendix D). Parallel analyses for This Study dataset yielded similar results (83.5% of barcode multiplets were due to complex beads; 4.95% of beads were heterogenous beads). Interestingly, the percent difference between the log2 number of valid fragments for these two classes of multiplets showed greater variability in the number of fragments per barcode for the complex beads than for barcode multiplets presumably caused by two beads (Figure 4.3.g; Appendix D). This result supports the idea that there may be a predominant individual barcode sequence on these complex beads though there is detectable heterogeneity. Finally, as 10x recently released their v1.1 "NextGem" design, we processed two additional datasets that were run with the two different chip designs in parallel. Our results confirm that the abundance of barcode multiplets persists across both of these two different chip designs (Figure 4.3.f) as well as the rates of complex beads and multiple beads underlying the multiplets (Figure D.3.g).

4.2.4 External corroboration of barcode multiplets

In response to a pre-print version of this chapter, 10x Genomics released a letter a software solution to identify multiplets from the output of the CellRanger-ATAC pipeline. In principle, their approach similarly utilizes the molecular diversity of Tn5 cut sites to identify putative barcode multiplets. After obtaining this script, we evaluated our two well-characterized PBMC datasets and determined that the rates of barcode multiplets were extremely similar as >98% of barcodes were concordantly classified as belonging to a barcode multiplet or not (Figure D.3.h; Appendix D). As a solution to the barcode multiplet artifact, the 10x method discards the lower abundance barcodes per multiplet. While further analysis is required to determine the optimal strategy for handling barcode multiplets, these results corroborate our estimates inferred and reported from bap.

4.2.5 Confounding of clonal lymphocytes estimation from barcode multiplets

We suggest that many applications of the 10x Chromium platform are unlikely to be impacted by bead multiplets. However, droplet single-cell approaches are now employed for purposes requiring increasingly precise quantitation, such as highly multiplexed perturbations (Dixit et al., 2016), clonal lymphocyte analyses (Simone et al., 2018), or diagnostics (Haque et al., 2017). Thus, for analyses of rare events, such as those routinely quantified in CRISPR perturbations or in clonal analyses of cells, the surprisingly high prevalence of barcode multiplets may become particularly problematic. As one example, we hypothesized that barcode multiplets may significantly alter quantitation of cell clones distinguished by unique B-cell receptor (BCR) and T-cell receptor (TCR) sequences in a tumor microenvironment (Figure 4.4.a). Though there is no current approach to define bead multiplets in scRNA-seq data, we reasoned that certain abundant BCR and TCR clonotypes may be explained by complex beads representing one true cell (similar to Figure 4.3.c). To test this, we reanalyzed a publicly available dataset generated using the 10x V(D)J platform that analyzed lymphocytes from a non-small-cell lung carcinoma (NSCLC) tumor (Figure 4.4.a). Indeed, we observed two instances of a BCR clone with four or more cells that could be more parsimoniously interpreted as barcode multiplets derived from a single B-cell (Figure 4.4.b). In particular, all presumed cells from these clones shared an rLCS of ≥9, an extremely unlikely event assuming true clonal cells would be randomly assigned barcode sequences (Figure D.3.f; D.1.a). Indeed, the distribution of the rLCS across all BCR clonotypes indicated a detectable bias indicative of barcode multiplets (Figure D.1.a; Appendix D). Furthermore, we identified additional clones that were depicted with a more complex heterogeneous structure that still broadly reflected bead synthesis errors (Figure 4.4.c).

Having established the clear possibility of barcode multiplets occurring in these data, we sought to determine how the interpretation of the overall clonality would be changed when accounting for the barcode multiplets. Using conservative estimates of barcode multiplets from the scATAC-seq analyses, we conducted a series of simulations (Appendix D). Overall, the percentage of cells associated with a clonotype comprised of at least two cells decreases considerably for both BCR (24.5% to 18.6%; Figure 4.4.d) and TCR (23.6% to 17.9%; Figure 4.4.e)

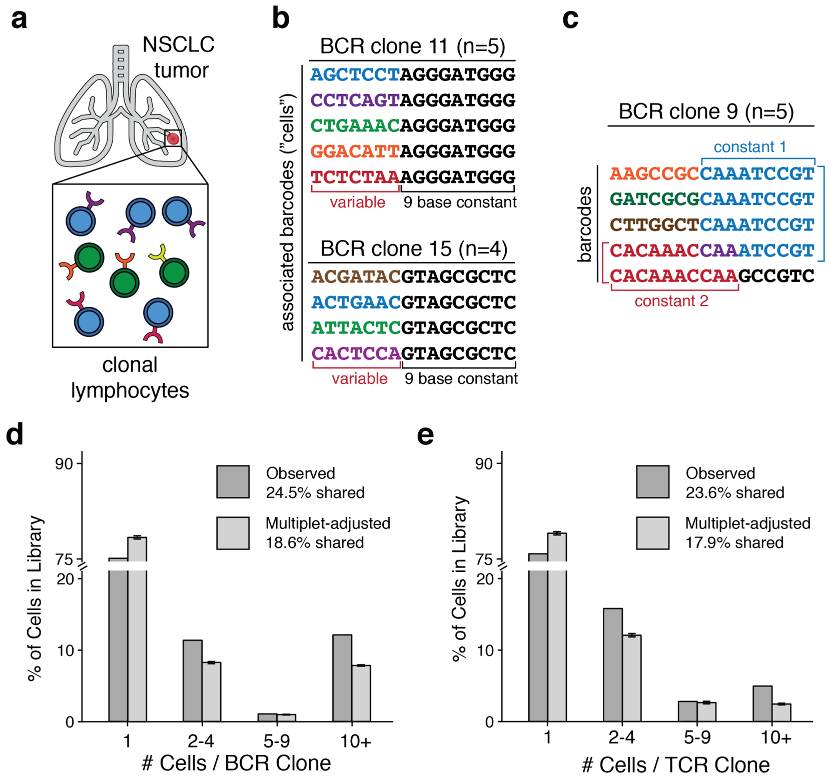

Figure 4.4: Confounding of intratumor clonal lymphocytes inference from barcode multiplets. **(a)** Schematic of intra-tumor lymphocytes identified from single-cell V(D)J sequencing on the 10x platform. **(b)** Identification of two presumed clonotypes composed of 5 and 4 barcodes. These clonotypes are likely to have been derived from one cell observed multiple times via barcode multiplets. **(c)** Example of a presumed clone composed of 5 barcodes with multiple constant sequences. **(d,e)** Overall summary of prevalence of **(d)** B-cell and **(e)** T-cell clone size before and after adjusting for observed rates of barcode multiplets in single-cell data. Error bars represent standard errors of the mean across 100 permutations.

clonotypes. Further analyses indicated a clone false discovery rate as high as 23.5% (BCR) and 22.5% (TCR) in these data (Appendix D), painting a much more conservative picture of clonality within NSCLC tumors. The results from these simulations indicate that bead multiplets may significantly confound clonal analysis and that this quantifiable discrepancy may falsely lead to conclusions of clonal expansion of lymphocytes in primary tumors.

4.3 Discussion

Overall, our work provides a new perspective to consider barcode multiplets in single-cell data. Though the exact chemistry of the training beads and reaction is different than what is typically employed in the 10x single-cell reactions, our imaging results confirm detectable bead multiplets as previously reported (Zheng et al., 2017b). Additionally, we show that bap, a computational algorithm designed to infer barcode multiplets, can be applied to sequenced scATAC-seq data from the 10x platform and confidently identify barcode multiplets. As the rates inferred from imaging and from bap are derived from distinct sources (*i.e.* bead/droplet counting versus sequencing), discretion is required when comparing between the detection modalities. Further analyses of multiplets identified by bap indicate that putative heterogeneity of beads in the 10x reaction is the predominant driver of the surprisingly high rates of multiplets in these datasets. Our analyses of clonal cells marked by BCRs and TCRs further suggest that bead sequence heterogeneity may be an artifact present across multiple sources of 10x single-cell data.

Conceptually, the presence of heterogeneity in beads is unlikely to be caused by an on/off process and instead likely exists as a spectrum across all beads used in these assays. As the estimated number of complex beads relies on sufficient amplification and detection of lower-frequency barcodes inside of droplets, the proportion of barcodes affected by this artifact becomes a function of the read depth (Figure D.3.e) and the barcode threshold (Figure D.3.b), which are in turn functions of the underlying chemistry of the assays. While our estimation of the clone false discovery rate assumed comparable rates for barcode multiplets for scATAC-seq and scRNA-seq methods, technical differences across these assays could also result variable barcode multiplet abundances. As such, our work motivates further investigation into the relationship between barcode multiplets and clonal diversity across various technical platforms.

As single-cell approaches move toward the precise quantification of rare cell types, trajectories, perturbations, and clones, an understanding of potential artifacts is essential as their confounding effects may become exacerbated in large datasets. Additionally, as these measurements move toward clinical applications9, particularly in tumors where TCR repertoire may serve as a prognostic biomarker (Cui et al., 2018), barcode multiplets may significantly confound interpretation. In some analyses (with <15% clones), we anticipate that many identified clonal cells may arise from bead multiplets. While our existing computational approach (bap) can facilitate the identification of barcode multiplets in scATAC-seq data, further experimental and computational tools are needed to more broadly identify these effects in RNA or genome sequencing droplet-based assays. We envision a combination of dense exogenous barcodes via cell hashing (Stoeckius et al., 2018) and evolved by CRISPR-Cas9 (Raj et al., 2018) or intrinsic features such as clonal mutations, rearrangements, or highly correlated abundances with barcode sequence similarity metrics could be leveraged to better infer barcode multiplets. Such approaches would complement existing tools that robustly identify cell doublets (Wolock et al., 2019; McGinnis et al., 2019) and empty droplets (Lun et al., 2019) from droplet-based scRNA-seq and further mitigate hidden confounders in single-cell data. Until then, we suggest that inferences regarding rare cell events should be corroborated across multiple channels or technologies to validate interpretation.

Taken together, our estimation and identification of barcode multiplets has a wide range of potential applications and confounding effects that influence widely-used droplet-based single-cell assays.

4.4 Acknowledgements

We thank J. Ulirsch and members of the Buenrostro lab for insightful comments. We are grateful to A. Labade and L. Ludwig for technical assistance. We thank Z. Burkett and R. Lebowsky of Bio-Rad for helpful conversations. We acknowledge a useful blog post from L. Pachter discussing sub-Poisson bead loading. J.D.B., C.A.L., S.M., and F.M.D. acknowledge support by the Allen Distinguished Investigator Program through the Paul G. Allen Frontiers Group. This work was further supported by the Chan Zuckerberg Initiative. C.A.L. is supported by F31 CA232670 from the NIH.

4.5 Author contributions

C.A.L. and J.D.B. conceived and designed the study. C.A.L. implemented the software and performed analyses. S.M. and F.M.D. performed experiments and aided analyses. J.D.B. supervised the work. C.A.L. wrote the manuscript with input from the authors.

4.6 Code and data availability

Software associated with the barcode multiplet identification and merging algorithm is available at https://github.com/caleblareau/bap. Code and data to reproduce the main findings of this study are available at https://github.com/caleblareau/barcode-multiplets. The public 10x scATAC-seq datasets are available for download at https://support.10xgenomics.com/single-cell-atac/datasets and the public NSCLC clontypes at https://www.10xgenomics.com/solutions/vdj/. Sequencing data generated as part of this work is available at the Gene Expression Omnibus under accession GSE143197.

By the end of several generations, all the descendants of the tribe, male or female, might track their mitochondrial ancestry...

Siddhartha Mukherjee, *The Gene*

5

Massively parallel single-cell mtDNA genotyping & chromatin profiling in human cells

Abstract

Natural mitochondrial DNA (mtDNA) sequence variation enables the inference of clonal relationships among human cells, and, in the case of pathogenic mutations, can contribute to human diseases. Unlike other genotyping approaches, mtDNA can be profiled along with measures of cell state, but has not yet been combined with the massively parallel approaches needed to tackle the complexity of human tissue. Here, we introduce a high-throughput, droplet-based mitochondrial single-cell Assay for Transposase Accessible Chromatin with sequencing (mtscATAC-seq) protocol and computational framework that facilitate high-confidence mtDNA mutation calling in thousands of single cells with their concomitant high-quality accessible chromatin profile. This enables the paired inference of individual cell mtDNA heteroplasmy, clonal relationships, cell state, and accessible chromatin variation. Our multi-omic analyses reveal single-cell variation in heteroplasmy of a pathologic mtDNA variant (m.8344A>G), which we associate with intra-individual chromatin variability and clonal evolution. Moreover, using somatic mtDNA mutations, we clonally trace thousands of differentiating hematopoietic cells *in vitro* and in patients with chronic lymphocytic leukemia, linking epigenomic variability to subclonal evolution *in vivo*. Taken together, our approach enables the study of cellular population dynamics and clonal properties of human cells *in vivo* in health and disease.

5.1 Introduction

Mitochondria play a central role in cellular metabolism and are unique organelles, carrying their own genome - often in high copy number - encoding a subset of proteins, tRNAs, and rRNAs essential to their function. Mutations in the mitochondrial genome are associated with a multitude of clinical phenotypes that are estimated to affect ~1 in 4, 300 individuals, making them among the most common inherited metabolic disorders (Stewart & Chinnery, 2015). Critically, the fraction of mitochondrial genomes carrying a specific variant - heteroplasmy - may dictate the degree of severity in an organ system in affected patients (Stewart & Chinnery, 2015; Shoffner & Wallace, 1992; Elliott et al., 2008). Furthermore, the high mutation rate (~2-10x that of nuclear DNA), leads to accumulation of somatic mtDNA mutations with time that may contribute to aging phenotypes (Stewart & Chinnery, 2015). While genomic approaches are emerging to quantify the level of heteroplasmy, the large majority of sequencing assessments have been based on bulk cell populations, limiting detection of somatic mutations in individual cells (Morris et al., 2017; Kang et al., 2016).

Recently, we (see Chapter 1) and others (Xu et al., 2019) have shown that commonly used single-cell profiling approaches can detect heteroplasmic or homoplasmic mutations, which we further leveraged as natural genetic markers in clone and lineage tracing of human cells along with their cell state. Due to the relatively small size of the mitochondrial genome (16.6 kb) and its higher copy number per cell, retrospective inference of cellular relationships by somatic mtDNA mutations is significantly more cost-effective and robust compared to mutation detection in the nuclear genome by single cell whole-genome sequencing. Moreover, single-cell RNA- and ATAC-seq (scRNA/ATAC-seq) allow concomitant mtDNA mutation detection along with the transcriptional or accessible chromatin cell state. While this presents a powerful system for larger scale clonal / lineage tracing in humans *in vivo*, only modest-throughput single-cell genomic assays had sufficient coverage of mitochondrial sequences for reliable mutation detection, whereas the massively parallel methods needed to draw meaningful conclusions on many biological systems had insufficient mitochondrial coverage. Therefore, additional innovations are required to increase the scale and scope of joint single-cell mtDNA genotyping in conjunction with cell state measurements.

As recently reported droplet-based scATAC-seq techniques enable the profiling of accessible chromatin in thousands of cells per experiment (see Chapter 3), we hypothesized that with appropriate modification, they may facilitate the enrichment of non-chromatinized (and thus readily transposase-accessible) mtDNA. However, these droplet-based protocols rely on processing of nuclei, thereby depleting mitochondria and resulting in only ~1% of reads mapping to mtDNA, compared to 20-50% in the original ATAC-seq protocol (Buenrostro et al., 2013); a level that is inadequate for single-cell mutation calling and clonal inferences.

Here, we establish mtscATAC-seq, a massively parallel protocol for high and uniform single-cell mitochondrial genome coverage that retains high-quality chromatin accessibility data concomitantly, and combine it with robust computational methods to identify rare, clonal mtDNA mutations in healthy and diseased cells. We demonstrate the wide applicability of mtscATAC-seq to quantify single-cell mitochondrial genotypes in the context of mitochondrial disease and clonally trace thousands of native human cells *in vitro* and *in vivo*. Given the multi-omic nature, we envision the broad utility and applicability of mtscATAC-seq to enhance our understanding of mtDNA genotype-phenotype correlations and reconstruct clonal dynamics across diverse areas of human health and disease.

5.2 Results

5.2.1 Development and validation of mtscATAC-seq

To develop mtscATAC-seq, we modified the droplet-based scATAC-seq workflow of the widely used 10x Genomics Chromium controller to improve mtDNA yield and genome coverage. As most scATAC-seq protocols use nuclei, depleting cytoplasmic mitochondria, we turned to processing whole cells to retain mtDNA. We further reasoned that mild lysis or permeabilization of cells would be required for the Tn5 enzyme to integrate adapters into accessible nuclear chromatin and mtDNA. Moreover, as cells contain multiple mitochondria, which may be more readily released upon cell lysis or permeabilization, we reasoned that fixation should minimize leakage of mtDNA between cells. Finally, we aimed to identify conditions retaining high-quality chromatin accessibility data.

We systematically tested for conditions that satisfy all of these desired features in a mixture of two human hematopoietic cell lines (GM11906 and TF1; Figure 5.1.a) by evaluating mtDNA abundance, specificity (*i.e.*, mtDNA fragments associated with its corresponding nuclear genome), and fragment complexity (for mtDNA and chromatin). Because each cell line harbored private homoplasmic mutations, we could sensitively detect mtDNA abundance, cell doublets, and possible mtDNA crosstalk due to cell lysis or permeabilization and tagmentation that occurs in a pool prior to droplet-mediated separation of cells. Omitting digitonin and tween-20 in the lysis and wash buffers (“Condition A”) yielded substantially more mtDNA fragments per single-cell (median 21.5%) than the recommended lysis protocol (1.9%; Figure 5.1.b; Appendix E), consistent with earlier observations (Corces et al., 2016). These modified conditions retain high-quality chromatin accessibility data: while per-cell complexity of nuclear fragments slightly decreased (Figure E.1.a), other metrics associated with scATAC-seq data quality improved, such as the fraction of reads in annotated DNase hypersensitivity peaks (from 74.1% to 79.7%; Figure 5.1.c) and fraction overlapping transcription start sites (TSS) (from 27.9% to 33.2% Figure E.1.b). BioAnalyzer traces confirmed an increased ratio of nucleosome free to mononucleosome fragments, consistent with the increased recovery of mtDNA (Figure E.1.c). Based on 43 high-confidence homoplasmic mtDNA variants private to each cell line (Appendix E), ~8.7% of barcodes carried otherwise cell type-specific homoplasmic variants at intermediate (60%-90%) heteroplasmy, indicating contamination of mtDNA fragments across cells (Figure 5.1.d; Figure E.1.d). Because this contamination may occur due to the release of mitochondria during processing, we added a fixation step with formaldehyde (FA), consistent with other scATAC-seq workflows (Chen et al., 2018b). Indeed, fixation with 0.1 or 1% FA led to a ~3x reduction in mtDNA fragment cross-contamination (Figure 5.1.e,f; Figure E.1.d), a 55% increase in mitochondria fragment complexity (Figure E.1.e), and restoration of chromatin library complexity (Figure E.1.f). After removing cell doublets (Appendix E), the empiric rate of contamination was 0.16% (Figure 5.1.f), which is consistent with the order of magnitude for short-read sequencing error (Ross et al., 2013). Importantly, FA treatment did not introduce additional mtDNA mutations as shown by comparison of variant allele frequencies of unfixed and fixed aggregated cell data (Figure E.1.f).

Furthermore, we observed regions of lower coverage across the mitochondrial genome per single-cell, which we determined were due to high homology (and thus low mappability) to nuclear mitochondrial DNA segments (NUMT). We reasoned that due to the high mtDNA copy number and the high Tn5 accessibility of mtDNA, ambiguous fragments could be confidently assigned to the mitochondrial genome with a low false positive rate. Indeed, we estimated that only ~1 accessible fragment from NUMTs would be detected per cell (by analyzing a compendium of DNase hypersensitivity data (Roadmap Epigenomics Consortium et al., 2015; ENCODE Project Consortium, 2012) and additional public scATAC-seq data; Appendix E), such that these are unlikely to

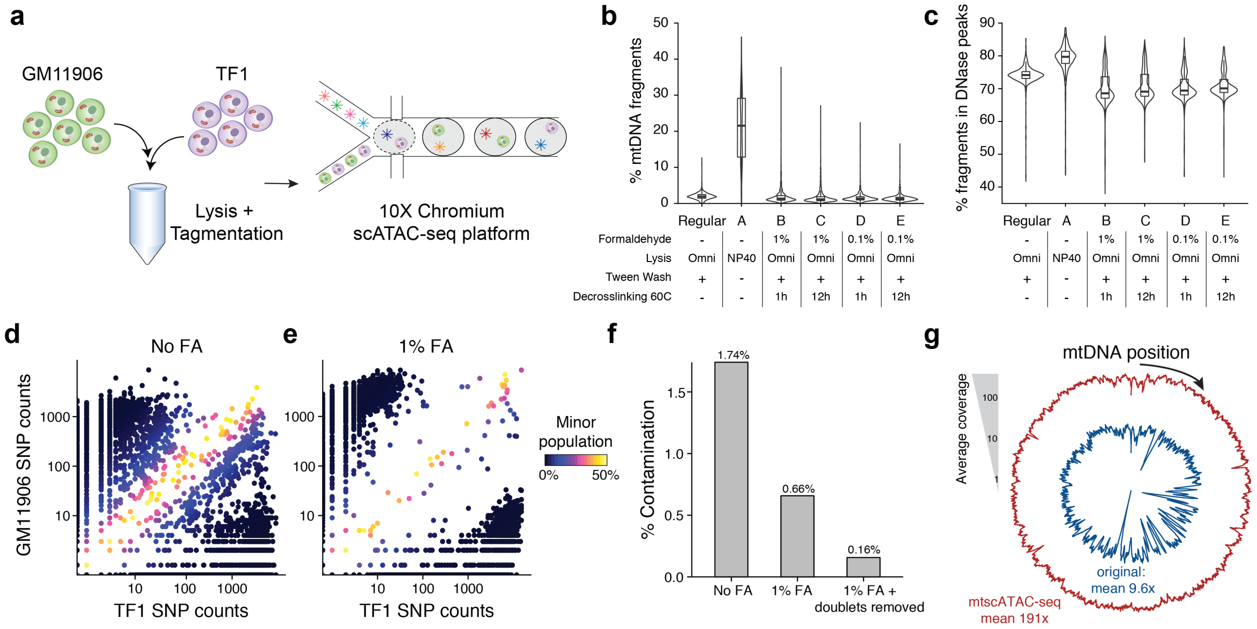

Figure 5.1: Optimization of a high-throughput single-cell mitochondrial DNA genotyping platform with concomitant accessible chromatin measurements. **(a)** Schematic of cell line mixing experiment between indicated two human hematopoietic cell lines. **(b)** Distribution of percentage of mtDNA reads per single cell for screened conditions. **(c)** Distribution of percentage of reads mapping to annotated DNase hypersensitivity peaks (nuclear reads only) per single cell. **(d)** Mitochondrial SNP mixing depiction of variants for the TF1 or GM11906 cell line for "Condition A" as in **(b)**. Both axes are log transformed. **(e)** Same as **(d)** but for "Condition A" with 1% FA treatment. **(f)** Summary of contamination (percent of reads from minor cell population) for FA treated and untreated comparison. **(g)** Depiction of overall mitochondrial genome coverage improvements from three biotechnical and computational optimizations (mtscATAC-seq) compared to the original protocol. Boxplots: center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range.

be a confounding element in heteroplasmy estimation. We therefore developed a computational approach that effectively assigns reads that map to both the mitochondrial and nuclear genome strictly to mtDNA, facilitating near-uniform coverage without altering chromatin complexity (Figure 5.1.g; Figure E.1.g-i).

Overall, mtscATAC-seq combines our modified lysis, cell fixation, and computational analysis of multi-mapping reads, leading to a ~20-fold increase in mean mtDNA coverage per cell (from 9.6x to 191.0x; Figure 5.1.g) and in fraction of mtDNA reads (median per cell from 1.9% to 36.8%; Figure E.1.h) with only modest reduction in chromatin complexity (median per cell from 87,569 to 73,864; Figure E.1.e) and in reads mapping to pre-annotated DNase hypersensitivity peaks (from 74.1% to 72.3%), retaining cell type-specific accessible chromatin peaks (93.8% of 777,704 peaks; Figure E.1.j; Appendix E).

5.2.2 Single-cell features of pathogenic mtDNA mutations

We used mtscATAC-seq to identify pathogenic mtDNA mutations in individual cells, and gain insights into their impact. The GM11906 lymphoblastoid cell line used in the mixing experiment (Figure 5.1) was derived from a patient diagnosed with myclonic epilepsy with red ragged fibers (MERFF), a mitochondrial disorder that in 80-90% of cases is caused by a 8344A>G mutation that alters tRNA function2 (Figure 5.2.a). Bulk ATAC-seq analyses of these cells estimated a population heteroplasmy of 44% for the 8344A>G allele, consistent with previous reports (Dames et al., 2013). We retained 818 high-quality data GM11906 cells with at least 50x single-cell mtDNA coverage and 40% reads in ATAC peaks (Figure 5.2.b). Interestingly, we observed a broad range of heteroplasmy values (0% to 100%) for the 8344A>G allele, with a median of 38%, consistent with the heteroplasmy estimation from bulk ATAC-seq (Figure 5.2.c), and from previous family studies of this mutation (Wallace & Chalkia, 2013). We independently replicated the distribution of heteroplasmy levels with 70 high-quality cells from the Fluidigm C1 scATAC-seq platform (Buenrostro et al., 2015) and *in situ* hybridization (Lee et al. (2015); Figure 5.2.c-e, Figure E.2.a). Thus, our mtscATAC-seq approach enables reliable single-cell genotyping of mitochondrial variants, including those causing disease.

Analysis of matched chromatin profiles highlighted specific loci and TF activities that are associated with dif-

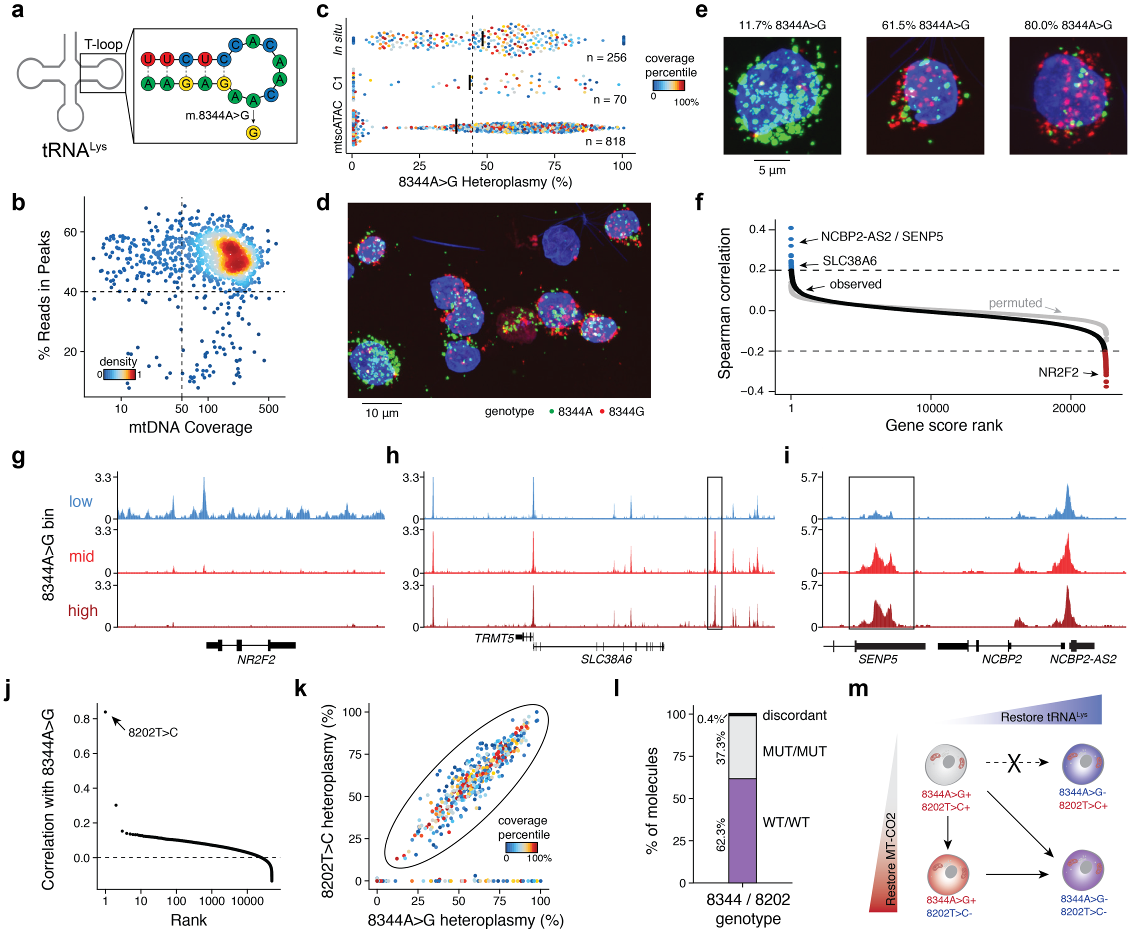

Figure 5.2: Pathogenic mtDNA variability and clonal evolution in cells derived from a patient with MERRF. (a) Schematic of the mitochondrial lysine tRNA secondary structure with sequence and the pathogenic single nucleotide variant (8344A>G). **(b)** Quality control filtering for GM11906 single cells based on mean mtDNA genome coverage and percentage of nuclear reads in chromatin accessibility peaks. **(c)** Quantification of 8344A>G heteroplasmy variability in single GM11906 cells across three technologies. Numbers (n) of cells plotted are shown. Color represents the within-assay coverage percentile. Black bars indicate the median heteroplasmy per technology; the dotted line presents the mean heteroplasmy as determined for bulk ATAC-seq. **(d)** Field of view for in situ genotyped GM11906 cells, highlighting **(e)** single cells with low, medium, and high heteroplasmy as indicated for the pathogenic allele. **(f)** Per-gene score Spearman correlations with the 8344A>G allele heteroplasmy. The grey dots show values for a permutation. Pseudo bulk accessibility track plots are shown for the **(g)** NR2F2, **(h)** TRMT5, and **(i)** SENP5/ NCBP2-AS2 loci. Pseudo-bulk groups were binned based on 0-10% (low), 10-60% (mid), and 60-100% (high) 8344A>G heteroplasmy. **(j)** Per-mutation heteroplasmy correlation with 8344A>G allele. The 8202T>C mutation is highlighted as the most correlated mutation. **(k)** Single-cell heteroplasmy for two indicated mutations. The circled population represents a double-positive population for both mutations. **(l)** Abundances of each variant on single sequencing reads in the double positive population. **(m)** Schematic of the co-evolution of two subclonal populations marked by indicated mutations detected based on single-cell genotyping data. Putative cell transitions are indicated with solid arrows that may be a result of selective pressure of the pathogenic variant and/ or genetic drift.

ferent levels of the 8344A>G allele. First, promoter accessibility scores (see Chapter 3) of 32 and 94 genes were confidently positively or negatively correlated, respectively, with single-cell 8344A>G heteroplasmy, corresponding to <1% false discovery rate (FDR) (Figure 5.2.f; Appendix E). Binning cells into high (>60%; n=273), intermediate (10-60%; n=228), and low (<10%; n=313) heteroplasmy for the pathogenic allele highlighted distinct chromatin features near NR2F2, TRMT5, and the SENP5/ NCBP2-AS2 loci (Figure 5.2.g-i). Notably, genes near these loci have been broadly linked to mitochondria biology, including in mitochondrial pathology (Wu et al., 2015) and function (Zunino et al., 2007), respiratory chain deficiencies (Powell et al., 2015), and cell signaling under hypoxic conditions (Kugeratski et al., 2019). The accessibility profiles at other loci were virtually indistinguishable (Figure E.2.b,c), suggesting that the observed variations in Figure 5.2.g-i may be a consequence of disease allele heteroplasmy. Furthermore, we identified transcription factors (TFs) whose activity may be associated with the mutation by scoring TF binding sites (from ChIP-seq; Appendix E) whose accessibility was correlated with pathogenic heteroplasmy. In particular, MEF2A and MEF2C were strongly anti-correlated with pathogenic heteroplasmy. Notably, the transcription factor MEF2 is a target of mitochondrial apoptotic caspases, supporting a model where pathogenic allele heteroplasmy may regulate nuclear factor activity, suggesting mechanisms of coordination between both genomes (Brusco & Haas (2015); Figure E.2.d,e). While any individual cis- or trans- alteration requires further experimental investigation, these analyses demonstrate the potential to study the altered cellular circuits resulting from pathogenic mtDNA variants in a heteroplasmy-dependent manner.

Notably, a second highly heteroplasmic mutation, 8202T>C (bulk heteroplasmy 34%) was the most correlated mutation with the 8344A>G variant across the single cells (Figure 5.2.j). Using MITOMAP (Lott et al., 2013), we annotated the non-synonymous variant (phenylalanine to serine) as a "probably damaging" mutation in the mitochondrially encoded gene cytochrome C oxidase II (MT-CO2). 456 of our 818 high-quality GM11906 cells were positive for both mutations (>5% heteroplasmy), whereas the remaining cells showed 0% heteroplasmy for either both or 8202 alone, but not 8344 alone (Figure 5.2.k). Of the 5,230 mtDNA paired-end reads that covered both variants from the double-positive population, 99.6% exclusively contained either both mutated or wildtype alleles (Figure 5.2.l). The co-occurrence of both mutations on the same haplotype and the presence of 8344A>G+/8202T>C- cells suggests the evolution of at least two subclonal populations, each spanning the complete spectrum from low to very high 8344A>G heteroplasmy (Figure 5.2.k,m), demonstrating how the mtscATAC-seq approach can enhance our understanding of variation and clonal dynamics in the context of mitochondrial disease.

5.2.3 Inference of confident mutations for clonal lineage tracing

To facilitate clonal tracing of human cells based on reliable mtDNA variation, we developed the Mitochondrial Genome Analysis Toolkit (`mgatk`; Figure 5.3.a; Appendix E), as a computational pipeline to identify clonal substructure in complex populations profiled using mtscATAC-seq. Recent variant callers developed for single-cell genotyping were designed to distinguish amplicon error from true mutations or account for allelic dropout (Zafar et al., 2016), neither of which predominantly confound heteroplasmy estimates from mtscATAC-seq (Appendix E). Instead, `mgatk` focuses specifically on clonal mtDNA variant calling in single cells, by leveraging the high mtDNA copy number, near-uniform coverage across the mtDNA genome (Figure 5.1.g), and an overall high per-cell coverage in mtscATAC-seq. Because our focus is on clonal variants, `mgatk` not only estimates the heteroplasmy for every possible mitochondrial variant in individual cells (~50,000), but then prioritizes individual mutations based on aggregate properties from experimental batches. Specifically, `mgatk` identifies high-confidence clonal mutations by aggregating signal across cells, leveraging between-cell variability and quantifiable strand bias (Figure 5.3.a; Appendix E). Thus, rather than calling variants in individual cells, `mgatk` leverages the high-throughput nature of our data to identify between-cell properties to distinguish signal from noise. The resulting mutations are then used as a feature set for downstream analyses, such as the inference of clonal families.

We validated `mgatk` by identifying anticipated clonal substructure in the 855 TF1 cells (>50x mean mitochondrial genome coverage) profiled in the mixture experiment (Figure 5.1). Because these cells were expanded from a population of 30 individually flow cytometry sorted TF1 cells, we expected observing multiple sub-clones. We identified 48 reliable mtDNA variants by bivariate filtering of variants with a relatively high variance mean ratio

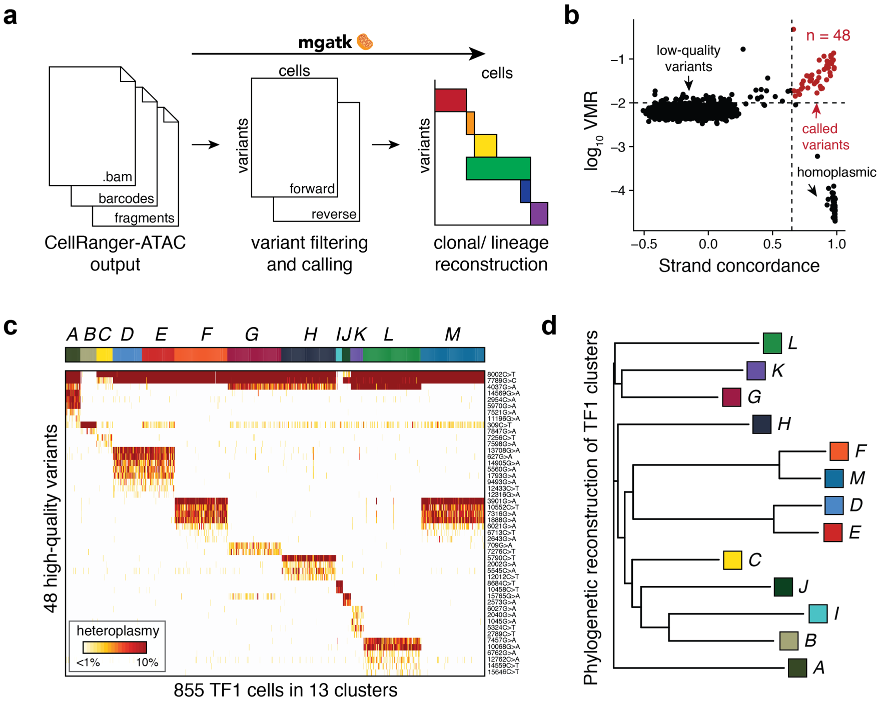

Figure 5.3: Identification of high-confidence variants and subclonal structure in TF1 cells. **(a)** Schematic of `mgatk` workflow. **(b)** Identification of high-confidence variants from high strand concordance in paired-end sequencing data and high variance mean ratio (VMR). **(c)** Unsupervised clustering of TF1 cells using 48 high-quality variants into 13 population clusters. Each column is a cell. Rows show detected mutation. Heatmap color indicates percent heteroplasmy. **(d)** Phylogenetic reconstruction of clonal TF1 groups. The tree was constructed using neighbor joining; each tip represents a cell cluster from **(c)**.

(VMR) and concordant heteroplasmy from both strands (Figure 5.3.b; Appendix E). Using these 48 variants as features, we determined 13 clonal cell subsets using a shared nearest neighbor clustering approach, with most cells carrying multiple cluster-distinct variants (Figure 5.3.c; Appendix E). Variants called by other approaches lacked sensitivity compared to `mgatk` (Figure E.3.a,b), and variants called only by these other methods had substantial strand bias (Figure E.3.c; Appendix E). The 48 high-confidence variants not only allowed us to reconstruct a putative phylogenetic tree for the identified TF1 subclones (Figure 5.3.d), as we previously showed with low throughput methods6, but to do so at a throughput that can be readily scaled up to many thousands of cells per experiment.

Though `mgatk` was optimized for mtscATAC-seq data, its unsupervised application performed comparably well to our previous supervised identification of multiple hematopoietic colony specific variants from 935 cells profiled by SMART-seq2 (from Chapter 1; Figure E.3.d-h; Appendix E). Furthermore, variants identified by `mgatk` substantially outperformed other unsupervised variant calling approaches in discerning cells that shared a clonal origin (Figure E.3.g,h; Appendix E). However, as SMART-seq2 and other scRNA-seq methods detect a substantial number of false-positive variants, corroboration by mtDNA sequencing is still highly recommended (see Chapter 1); conversely, mtscATAC-seq captures DNA directly, thus minimizing potential artifacts. Overall, `mgatk` analysis of mtscATAC-seq data provides the most robust and high-throughput means to identify high-quality mtDNA variants associated with cell states by a single-cell genomic assay.

5.2.4 Linking cell state to fate in hematopoietic differentiation

The multi-modal output of mtscATAC-seq simultaneously informs us of features of cell state and clonal relatedness, allowing us to better study complex human differentiation processes, where genetic barcoding is not possible, and high throughput is required. To illustrate this potential, we focused on a case study in hematopoiesis, a process fueled by possibly $10,000 - 100,000$s of distinct hematopoietic stem/progenitor cells (HSPCs; Lee-Six et al. (2018)), potentially requiring the sampling of large cell numbers to capture clonal spectra. Furthermore, previous reports suggest the presence of functional heterogeneity and differentiation (lineage) bias within the early HSPC pool (Rodriguez-Fraticelli et al., 2018; Jacobsen & Nerlov, 2019), though in most instances we lack ways to simultaneously link HSPC cell state to downstream fate of recently derived differentiating daughter cells (Weinreb et al., 2020), especially in humans.

To examine this, we first benchmarked mtscATAC-seq in an *in vitro* model of human hematopoiesis, where clonal contributions could be anticipated. We cultured ~500 or ~800 CD34+ HSPCs in progenitor expansion media, before induction of monocytic or erythroid differentiation with erythropoietin (EPO), stem cell factor (SCF), and interleukin-3 (IL-3). Over the course of 20 days we profiled cells from two independent cultures (two and three timepoints for the 500 and 800 cell input, respectively), yielding 18,964 high quality mtscATAC-seq cell profiles (Figure 5.4.a; Appendix E), with a mean 24,333 unique nuclear fragments per cell, 49.0% of which were in accessibility peaks, and a mean 73.6x mtDNA coverage per cell. Dimensionality reduction (Granja et al., 2019), transcription-factor motif scoring (Schep et al., 2017), and inference of pseudotime trajectories highlighted differentiation continuums from HSPCs to either erythroid or monocytic populations, consistent with our experimental expectations (Figure 5.4.b,c; Figure E.4.a-d; Appendix E). These findings verify that mtscATAC-seq can reconstruct continuous cell state transitions comparable to previous scATAC-seq studies (Granja et al., 2019; Satpathy et al., 2019; Buenrostro et al., 2018).

Application of `mgatk` identified 179 and 308 high-confidence, heteroplasmic variants in the 500 cell and 800 cell input cultures, respectively, which were enriched for transitions (95.0 and 95.1%; Figure 5.4.d; Appendix E), consistent with previous findings. In both cultures, there were substantial shifts in heteroplasmy, including significantly wider distribution of allele frequency fold changes than expected if the HSPCs underwent differentiation uniformly (Figure 5.4.e,f; Kolmogorov–Smirnov p<2.2x10-16). Along our sequential sampling experiment, the heteroplasmy change in the 800-cell input culture from the second sampling (day 14 / day 8) largely explained the third (day 20 / day 14; Figure 5.4.g), suggesting that clonal contributions largely did not diverge further during continued differentiation. However, our sequential clonal tracing captures complexities in these temporal cell state transitions. For example, we observe patterns suggestive of variable clone proliferation dynamics, such as cells that expanded earlier (3712G>A) or later (14322A>G) in the culture system (Figure 5.4.h). Analysis of 18

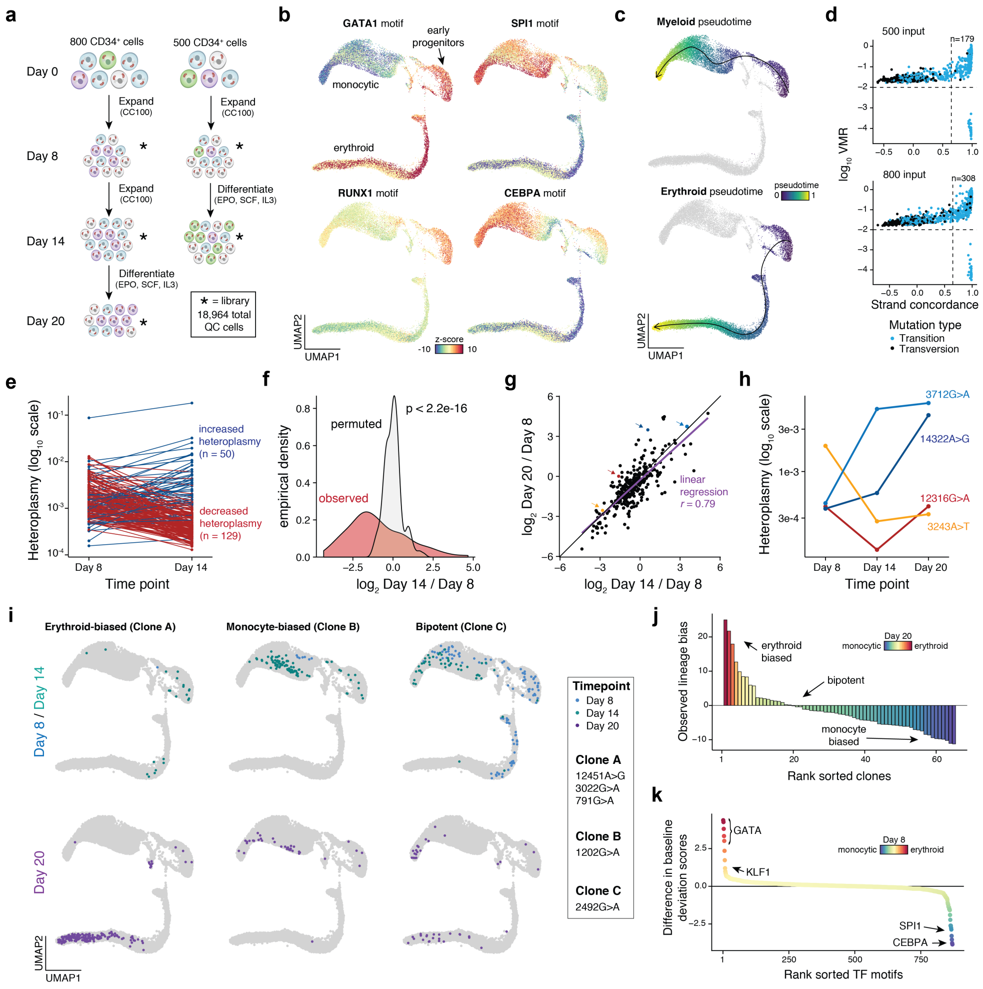

Figure 5.4: Clonal lineage tracing across accessible chromatin landscapes in an *in vitro* model of hematopoiesis. **(a)** Schematic of experimental design. Approximately 800 or 500 CD34+ HSPCs were derived from the same donor, expanded, and differentiated in two independent cultures over the course of 20 days as shown. Stars represent timepoints/ populations of cells that were profiled via mtscATAC-seq. **(b)** Two dimensional embedding of all quality controlled cells using UMAP. Single-cell transcription factor motif deviation scores for indicated factors are shown in color for all cells. **(c)** Pseudotime trajectories for monocytic and erythroid trajectories are depicted. **(d)** Identification of high confidence variants derived from both cultures. The number of variants passing both thresholds (dotted lines) is indicated. **(e)** Changes in heteroplasmy for 179 variants identified from the 500 input culture from day 8 to day 14. Values represent the mean over all single-cells in the library. **(f)** Increased variability in heteroplasmy shifts for the 500 cell input culture. P-value is reported from a Kolmogorov–Smirnov test comparing the observed and permuted distributions log fold-changes. **(g)** Comparison of heteroplasmy shifts for the 800 cell input culture. Linear regression indicates that most of the variability in heteroplasmy changes at the late time point (day 20, y-axis) can be explained by the intermediate time point (day 14, x-axis). Colored dots are highlighted in the next panel. **(h)** Heteroplasmy trajectories for four selected mutations from **(g)**. Values represent the mean over all single-cells in the library for the indicated time point. **(i)** Three examples of clonal populations marked by indicated mutations identified in the 800 cell input culture that result in erythroid, monocytic, or bipotent differentiated cell outcomes. **(j)** Systematic identification of clonal outcomes using the late time point (day 20). Y-axis depicts the difference between z-score in erythroid and monocytic bias of a single clone. **(k)** Differences in transcription factor motif activity comparing erythroid-biased and monocytic biased clones at the earliest sampled time point (day 8).

shared mutations between the two cultures suggested that proliferation capacity was independent of the specific mutation at least for these mutations, as their heteroplasmy fold-changes were not correlated between the two experiments (Figure E.4.e,f).

Interestingly, we observed six "confirmed" pathogenic mutations between the two cultures, including 12316G>A and 3243A>T (Figure 5.4.h), both of which alter normal mitochondrial tRNA function, possibly explaining their observed decreased population frequencies over the course of the culture. Each of these six mutations occurs at a maximum of 0.1% allele frequency in the bulk population, but exceed 30% heteroplasmy in some individual cells (Figure E.4.g), confirming that our approach enables the detection and study of deleterious somatic mtDNA variants in cells of otherwise healthy individuals.

Combining the mtDNA mutation and clonal status with the cells' chromatin profiles, we inferred properties and possible fates of HSPCs in our cultures, distinguishing bi-potent progenitors from those biased in favor of an erythroid vs. monocytic fate. We used a community detection algorithm to partition the cells from the two cultures to 167 clonal groups by mtDNA mutations (Figure E.4.h,i; Appendix E), with most cells carrying at least one high-quality somatic mtDNA mutation (Figure E.4.j). We then examined the states of the cells in each clone, to identify HSPCs from day 8 in clones with biased (enriched) membership of monocytic or erythroid cells on day 20 (Figure 5.4.i). Specifically, of the 65 clonal populations with at least 10 cells at day 20 we observed in the 800 input culture (Figure 5.4.j; Appendix E), 9 were erythroid-biased and 22 were monocytic-biased (z-score >5; Figure 5.4.j).

To further leverage our data association of cell state and fate, we examined the chromatin features of HSPCs in biased clones and in bi-potent ones. Indeed, well characterized erythroid (GATA1 and KLF1) or monocytic transcriptional regulator motifs (SPI1 and CEBPA) were more accessible in day 8 cell clones that preferentially gave rise to daughter cells of erythroid or monocytic lineage by day 20, respectively (Figure 5.4.k). However, when restricting this analysis towards day 8 cells within the early progenitor cluster (cluster 8; Figure E.4.c), this association diminishes, though our power to detect such lineage biasing features (if present and causal for such observations) may be limited given the number of cells profiled at this stage (n=257). Overall, these results suggest that applying our framework for clonal inferences in an *in vivo* human context could facilitate systematic studies that were previously limited to model organisms or gene therapy trials (Sun et al., 2014; Scala & Aiuti, 2019).

5.2.5 Clonal heterogeneity in chronic lymphocytic leukemia

Finally, we applied mtscATAC-seq to cells obtained directly *in vivo* from patients with putatively clonal malignancies. We profiled peripheral blood mononuclear cells (PBMCs) from two patients with chronic lymphocytic leukemia (CLL), which is conventionally characterized as a monoclonal B-cell malignancy (Figure 5.5.a). Single-cell B-cell receptor sequencing by *ex vivo* 5' scRNA-Seq (Appendix E) confirmed a predominantly monoclonal population of leukemic cells in both patients (Figure 5.5.b). Based on our previous work, we hypothesized that somatic mtDNA mutations may arise during tumorigenesis, which mark and enable tracking of genetic subclones that may further aid to resolve intra-tumor heterogeneity6. We collected 23,467 high quality mtscATAC-seq profiles (mean 55.5x mtDNA coverage; 11,423 unique nuclear fragments per cell and 70.8% in peaks), and applied `mgatk` to CD19+ and predominantly leukemic cells to reveal 43 mutations and 15 putative subclones across the two patients (Figure 5.5.c; Figure E.5.a,b). This marked genetic diversity in a perceived highly clonal malignancy reinforces the effectiveness of our high-throughput approach to identify rare subclonal structure, including a cluster marked by the 12067C>T mutation present in 0.4% of the entire leukemic cell population (Figure 5.5.c).

To better understand the functional consequences of this subclonal structure, we related the mtDNA clones with both their chromatin profiles and receptor clonotypes, leveraging the mtDNA coverage from 5' scRNA-seq (Figure E.5.c,d) to relate to variants identified from mtscATAC-seq. Interestingly, leukemic cells with the 14858G>A mtDNA mutation did not carry the predominant BCR clonotype, presenting a distinct sub-clonal population showing various differentially-expressed genes (Figure 5.5.b,d; Figure E.5.e; Appendix E). Moreover, all cells in Patient 1 were positive for trisomy 12, a common cytogenetic abnormality in CLL (Roos-Weil et al., 2018), suggesting that the copy number alteration preceded the somatic mtDNA diversity detected (Figure 5.5.e). Performing a per-peak association with our putative subclones, we observed hundreds of loci associated

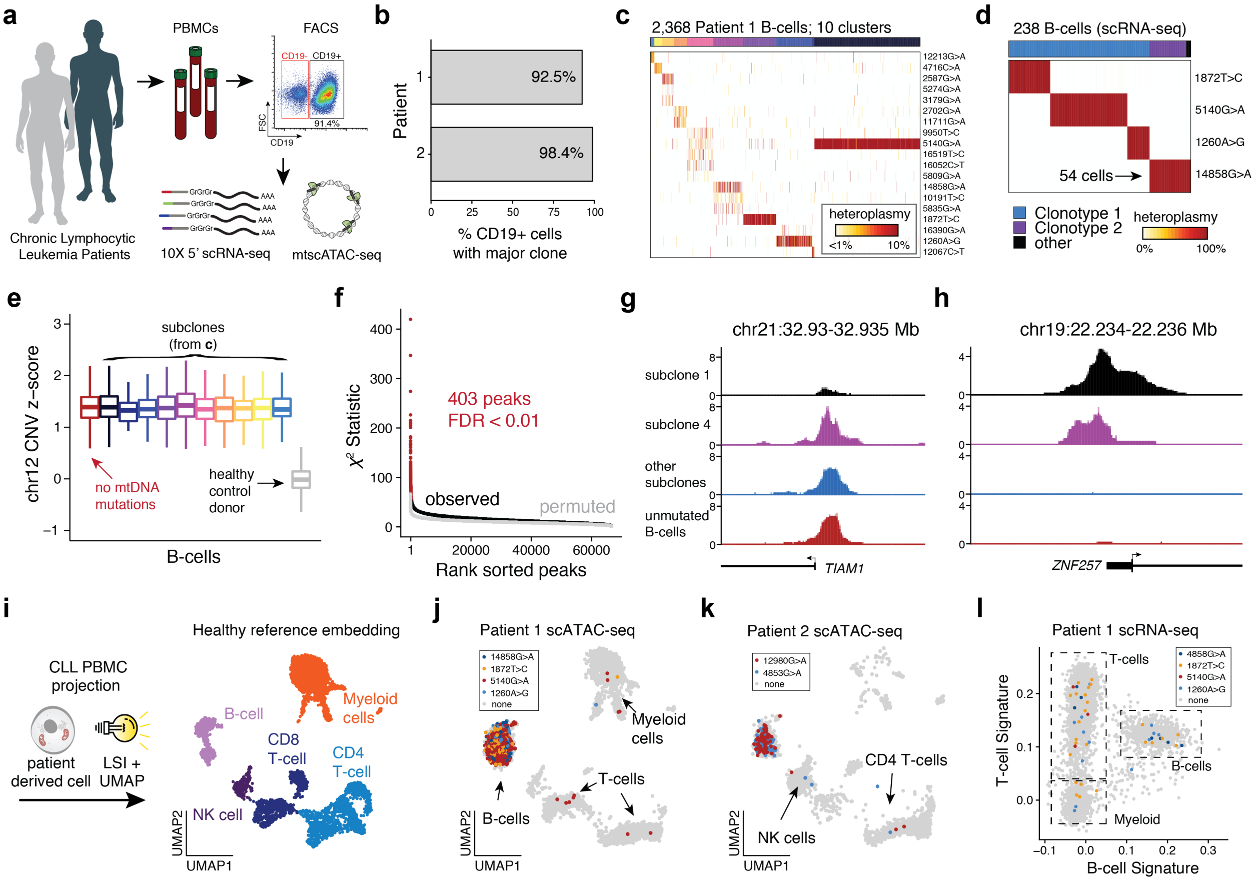

Figure 5.5: Clonal and functional heterogeneity in chronic lymphocytic leukemia resolved by somatic mtDNA mutations. **(a)** Schematic of experimental design. Populations of peripheral blood mononuclear cells (PBMCs) from two CLL patients were separated by FACS or magnetic bead enrichment and profiled with mtscATAC-seq and 10x 5' scRNA-seq. **(b)** Fraction of CD19+ cells with major B cell receptor (BCR) clonotype as determined from V(D)J receptor sequencing. **(c)** Inference of subclonal structure from somatic mtDNA mutations for patient 1. Cells (columns) are clustered based on mitochondrial genotypes (rows). Colors at the top of the heatmap represent clusters or putative subclones. Color bar, heteroplasmy (% allele frequency). **(d)** Clonotype receptors (columns) associated with somatic mtDNA mutations (rows) from patient 1. Colors at the top of the heatmap represent BCR clonotypes. Color bar, heteroplasmy (% allele frequency). **(e)** Estimated copy number of chromosome 12 across putative subclones for patient 1. **(f)** Sub-clone associations with accessible chromatin. Red dots denote peaks associated at a false-discovery rate of <0.01. **(g)** Examples of subclone-associated differential accessibility peaks near the TIAM1 and **(h)** ZNF257 promoters. **(i)** Schematic of scATAC projection framework using latent semantic indexing (LSI) and UMAP. A healthy PBMC reference embedding with indicated cell types is shown. **(j)** Projection of cells collected from Patient 1 and **(k)** Patient 2. Colors indicate cells positive for indicated somatic mtDNA mutations. Non-B-cells are highlighted. **(l)** Gene signature plots of PBMCs from single-cell RNA-seq for Patient 1 corroborating mtDNA mutations in non-B-cells.

with subclonal structure in these tumors (Figure 5.5.f; Figure E.5.f), including promoters of the ZNF257 and TIAM1 genes, the latter of which had previously been associated with chemoresistance in CLL and colorectal cancer (Izumi et al. (2019); Hofbauer et al. (2014); Figure 5.5.g,h). These results provide a broad basis for how paired chromatin accessibility and mtDNA genotyping can resolve epigenetic differences in malignant subpopulations at single-cell resolution.

Among the identified variants from mgatk, six mutations (four in patients 1, two in patient 2) attained homoplasmy in a subset of cells and were markedly enriched in the CD19+ population (Figure E.5.g,h). Notably, the same variants were also identified in non-B cells, including T lymphocytes, natural killer (NK), and myeloid cells (Figure 5.5.i-l; Figure E.5.i,j). These results point to the involvement of an early hematopoietic progenitor cell with residual multi-lineage capacity in the pathogenesis of CLL, as suggested by previous reports (Alizadeh & Majeti, 2011), but that could now be demonstrated *in vivo* in patient samples with the use of mtscATAC-seq/ mgatk. These results could further be corroborated in the scRNA-seq data of patient 2 upon integration of calling somatic mutations in nuclear genes (i.e. chr4:109,084,804A>C "LEF1" and chr19:36,394,730G>A "HCST"; identified by exome sequencing) (Figure E.5.i,j). Taken together, our results demonstrate the wide applicability of our mtscATAC-seq/ mgatk platform enabling the retrospective inference of cellular population dynamics in healthy and disease states.

5.3 Discussion

Here, we develop and validate our high-throughput platform for measuring mtDNA mutation heteroplasmy and concomitant accessible chromatin states in thousands of single-cells per reaction. Notably, we verify data standards (Figure 5.1.), chart the cis- and trans- effects of pathogenic mutations (Figure 5.2.), and infer subclonal population structure (Figure 5.3.), all from a single experiment. By leveraging somatic mtDNA variation in more complex settings, our results further indicate the potential of natural genetic mtDNA barcodes to inform cellular dynamics (Figure 5.4.) and clonal heterogeneity within malignant cells *in vivo* (Figure 5.5.). Furthermore, our platform provides an intrinsic coupling of these mutations to cell state and function due to concomitant accessible chromatin readouts. Unlike high-throughput scRNA-seq approaches that suffer from uneven coverage of mitochondrial RNA and a high false positive error rate, our improved technical platform and computational identification of variants enables robust inferences in complex settings, readily extending the scope of single-cell genomic applications. While our demonstration of mtscATAC-seq focused on the popular 10x Genomics Chromium system, we expect its adaptation to alternative scATAC-seq workflows.

In addition to pathogenic mitochondrial variants, such as 8344A>G, our high-throughput platform should facilitate the examination of functional mtDNA mutations in these relatively common disease settings (Stewart & Chinnery, 2015). We note that as 46 out of 90 "confirmed" MITOMAP-predicted pathogenic mtDNA mutations alter tRNA function (Lott et al., 2013), heteroplasmy estimation for most of these causal variants require a DNA-based assay for robust detection and analysis of single-cell variability rather than a poly-A based RNA-based technique. Furthermore, alterations in mtDNA have been associated with a variety of complex human diseases, including Alzheimer's Disease (Corral-Debrinski et al., 1994), Parkinson's Disease (Bender et al., 2006), cardiomyopathies (Lee & Han, 2017), pediatric cancers (Triska et al., 2019), and various other malignancies. More generally, the accumulation of somatic mtDNA mutations may contribute to aging phenotypes (Stewart & Chinnery, 2015). As our approach facilitates rapid genotyping and concomitant chromatin profiles in thousands of cells, potential molecular consequences of mtDNA variants may now be dissected using our platform (Figure 5.2), which is not otherwise possible using bulk approaches due to the unappreciated diversity of somatic variants with possible distinct effect sizes present in single cells of healthy tissues (Kang et al., 2016).

Despite the relatively small size of the mitochondrial genome, the prevalence of somatic mutations, though not necessarily present in every cell, is expected to enable inferences of clonal contributions and cellular population dynamics of complex human tissues at any moment and over time *in vivo* (see Chapters 1 and 2). In contrast to other high-throughput single-cell somatic mutation detection technologies that typically require a priori knowledge of specific variants called from mRNA transcripts (Nam et al., 2019), our approach enables de novo discovery of variants to inform the inference of subclonal structure in primary human cells. Though not all variants correlated with a data-driven population (*e.g.* Figure E.5.b), we expect that additional improvements in variant calling, community detection methods, and heteroplasmy-specific distance functions will further aid to resolve cellular hierarchies in greater detail. Furthermore, our analyses in the context of CLL provides a vignette of integrating nuclear point mutations, copy number alterations, immune receptor rearrangements, and mtDNA variation to further resolve clonal structure and functional heterogeneity. The advances presented here now enable new avenues to study how cellular dynamics plays a role in human health and disease.

5.4 Acknowledgements

We are grateful to Erik Bao, Jacob Ulirsch, Evgenij Fiskin, and members of the Sankaran and Regev labs for useful discussion. We acknowledge support from the Broad Institute and the Whitehead Institute Flow Cytometry core facilities. This research was supported by National Institutes of Health grants F31 CA232670 (C.A.L.), P01 CA206978 (C.J.W. and G.G.), U10 CA180861 (C.J.W.), R01 DK103794 (V.G.S.), and R33 HL120791 (V.G.S.), a gift from the Lodish Family to Boston Children's Hospital (V.G.S.), the New York Stem Cell Foundation (NYSCF, V.G.S.), and the Howard Hughes Medical Institute (A.R.). S.G. is supported by funding from the Kay Kendall Leukaemia Fund. G.G. is supported by the Paul C Zamecnick chair. C.J.W. is a scholar of the Leukemia and Lymphoma Society. F.C. and J.D.B were supported by the Allen Distinguished Investigator Program. V.G.S. is a NYSCF-Robertson Investigator. We are grateful to patients who made this work possible.

5.5 Author Contributions

C.A.L. and L.S.L. conceived and designed the project with guidance from A.R and V.G.S. C.A.L. developed the software and led data analysis. L.S.L. and C.M. developed the mtscATAC-seq experimental protocol. L.S.L led, designed, and performed experiments with assistance from C.M., W.L., and E.C. S.G. processed CLL patient samples with L.S.L. T.Z. performed the in situ genotyping experiments. Z.C. and J.M.V. analyzed data. D.R. and G.G. aided in the exome sequencing. F.C., J.D.B., M.J.A., C.J.W., A.R., and V.G.S. each supervised various aspects of this work. A.R. and V.G.S. provided project oversight and acquired funding. C.A.L., L.S.L., V.G.S and A.R. wrote the manuscript with input from all authors.

5.6 Code and data availability

Review access to processed data is available at GEO accession GSE142745 (with access token ofchomkepxsvrqd). Software and documentation for mitochondrial variant calling via `mgatk` is available at http://github.com/caleblareau/mgatk.

”You miss 100% of the shots that you don’t take” - Wayne Gretsky

Michael Scott, *The Office*

6 Conclusion

In the 63 years since the conception of the Waddington landscape, our understanding of the molecular actors governing cell fate decisions and transitions has greatly expanded. In particular, the identification of master regulator transcription factors underlying cell lineages, particularly in the hematopoietic system, has annotated the ”guy-wires” underlying cellular differentiation. Over the past decade, the capabilities of single-cell sequencing has further clarified this model where stem and progenitor cells do not occupy discrete cell states but instead exist along continuous trajectories– akin to pathways along the Waddington landscape (Buenrostro et al., 2018; Jacobsen & Nerlov, 2019). While the picture of this model of cellular differentiation has become increasingly clear with new technologies, we still have a limited understanding of how individual cell behaviors are dictated in humans *in vivo*. In particular, methods that couple cell state and cell fate in single cells have been largely inaccessible but should enable greater resolution in the hierarchical processes governing cell fate, differentiation, and composition of complex human tissue. As such, the innovations presented in this work were designed to develop scalable technologies to enable approaches to answer these fundamental questions.

Prior to the work presented in this dissertation, a primary method for inferring lineages of human tissue relied on single-cell whole-genome sequencing (scWGS), requiring ~$1, 000 / cell and severely affected by error rates in amplification of nucleic acids (Lodato et al., 2015), key limitations of these techniques. Further, these scWGS approaches do not provide measures of cell state, limiting inferences about cellular composition of human tissue. In Chapter 1, I introduced the concept of lineage tracing in human tissue with somatic mitochondrial mutations. Co-opting existing single-cell assays, we showed that both the plate-based Smart-seq2 and the microfluidic Fluidigm C1 assays enabled single-cell mitochondrial genotyping (in addition to concomitant transcriptomic and epigenomic readouts). Ultimately, these approaches profile batches of up to 96 cells per experiment at a cost of ~$15 / cell (including sequencing). In Chapters 3 and 5, I introduce droplet microfluidic approaches to perform these single-cell genomics assays (specifically single-cell ATAC-seq) at a substantially greater throughput (~5,000 cells / reaction) and significantly lower cost ($0.1 / cell). Specifically, Chapter 3 introduces a scATAC-seq platform on via the BioRad ddSEQ platform while Chapter utilizes the 10x Genomics Chromium controller for the application of mtscATAC-seq. As technologies in this space continue to rapidly advance, I expect that additional layers of genomic information may be ascertained per cell utilizing similar principles of droplet microfluidics.

Ultimately, these technical innovations make applications of clonal lineage tracing substantially more tractable in human tissue. Forecasting trends in the field to potential findings over the next two decades, I predict the following. First, I anticipate that multi-modal measurements in single-cells will become routine wherein cellstate may be simultaneously ascertained from a compendium of cell surface markers, whole transcriptome abundances, genome-wide accessible chromatin, and intracellular protein state. These, coupled with lineage information (from immune cell clonotype receptors and/or somatic mutations) will enable the *de novo* inference of key molecular effectors and trajectories assumed by stem and progenitor cells that comprise human tissues and will lead to the identification of therapeutic targets to combat a variety of complex diseases. Second, I predict that these multimodal and clonal measurements will provide direct instruction on the transition states in our adaptive immune system that facilitates immunity to viruses and other highly contagious pathogens. Such innovations will be essential in defending our expanding population against new pathogens, such as SARS-Cov-19. Finally, I predict that as a consequence of this work, a routine diagnostic blood test will be facilitated by single-cell technology to infer pre-malignant hematologic states and enable preventative measures against blood cancers and heart disease.

In summary, the goal of this dissertation was to continue the development of scalable experimental and computational solutions to infer attributes of clonal composition and cell states in complex human tissue. By innovating and understanding new approaches for lineage tracing in human tissues (Chapters 1,2) and high-resolution (*i.e.* single-cell) measures of cell state (Chapter 3,4), my work addresses these limitations from two different angles. In Chapter 5, I introduce a concomitant high-throughput clonal lineage (via mtDNA mutations) and cell state (via chromatin accessibility) with scalable throughput and demonstrate its utility in the human hematopoietic system in a variety of applications. While many biological questions remain unanswered, the technical advances forwarded by this dissertation directly enable previously inconceivable experimental designs to chart cell states and composition in complex human tissues. Importantly, lessons learned from the continued investigation of human biology with these approaches stand to enable new avenues of therapeutic intervention.

A

Supplemental material for Chapter 1

A.1 Experimental model, subject, and method details

A.1.1 TF1 Cell Culture

TF1 cells (ATCC) were maintained in Roswell Park Memorial Institute Medium (RPMI) 1640, 10% fetal bovine serum (FBS), 2mM L-Glutamine and 2ng/ml recombinant human Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF) (Peprotech) and incubated at 37°C and 5% CO2. 293T cells (ATCC) were maintained in Dulbecco's Modified Eagle Medium-High Glucose (DMEM), 10% fetal bovine serum (FBS), and 2mM L-Glutamine and incubated at 37°C and 5% CO2.

A.1.2 Primary Cell Culture and methylcellulose colony assays

CD34+ hematopoietic stem and progenitor cells were obtained from the Fred Hutchinson Hematopoietic Cell Processing and Repository (Seattle, USA) and were cultured in StemSpan II with 1x CC100 (Stemcell Technologies) at 37°C and 5% CO2. For methylcellulose colony assays, 500 cells per ml were plated in MethoCult H4034 Optimum (Stemcell Technologies) according to the manufacturer's instructions. Individual colonies were picked at day 10 or 12 after plating for single cell sorting.

A.1.3 Human colorectal cancer specimen

Primary untreated colorectal tumor and adjacent non-neoplastic tissue were surgically resected from a 75-year-old male patient with pathologically diagnosed colorectal adenocarcinoma at Massachusetts General Hospital. Written informed consent for tissue collection was provided in compliance with IRB regulations (IRB compliance protocol number 02-240. Broad Institute ORSP project number ORSP-1702).

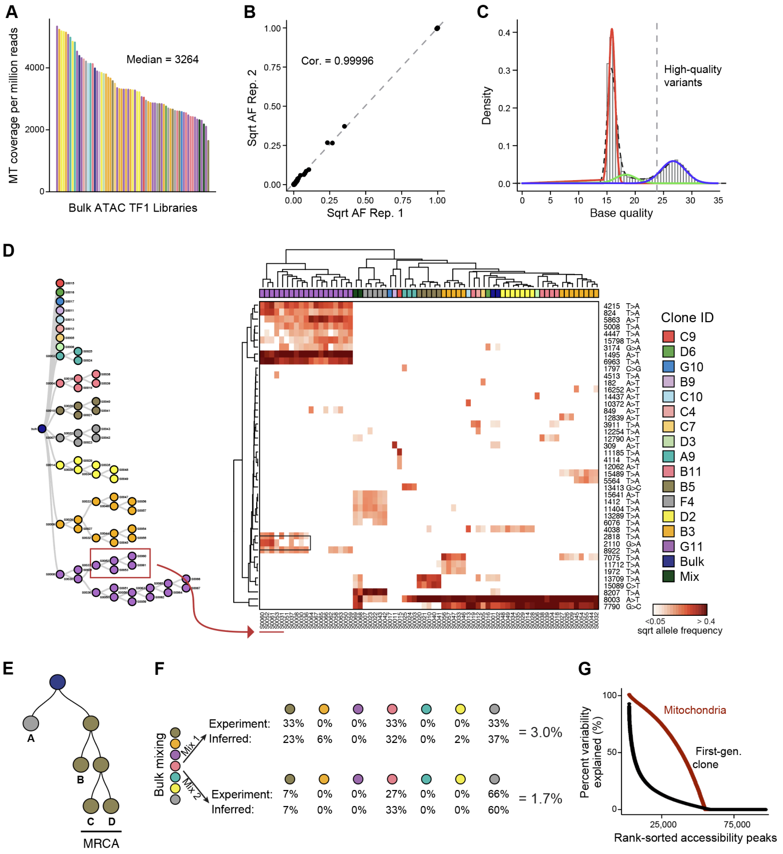

Figure A.1: Detection of mitochondrial mutations with ATAC-Seq. (A) Coverage of mitochondrial genome by bulk ATAC-Seq. The mitochondrial genome coverage per million reads (y axis) of each TF1 bulk ATAC-seq sample (x axis), sorted by coverage and colored by parent clone as in Figure 1C. **(B)** mtDNA mutations are consistently detected across replicate sequencing runs. Heteroplasmy (square root of allele frequency) for each high-confidence mutation (x, y axis) in two technical replicates of the bulk TF1 sample. Pearson correlation coefficient between the replicates is indicated. **(C)** Gaussian mixture model fit over per-base pair, per-allele base qualities. Shown is the distribution of per-base pair, per-allele base qualities scores (x axis), fit with three Gaussian curves (colors) representing three mixture components: blue: high-confidence variants. Vertical dashed line: threshold for 99% probability of belonging to the distribution of high confidence variants. **(D)** Left: known lineage of TF1 clones annotated with sample IDs. Right: Hierarchical clustering of bulk TF1 clones by high confidence mtDNA variants. Shown are the samples (columns) labeled by clone (color code as in Figure 1.1.C, sample IDs are annotated at the bottom of the heatmap) and ordered by hierarchical clustering (dendrogram, top) based on the square root of the allele frequency (color bar) of high-confidence variants (rows) identified in **(C)**. Box indicates a subclone-specific mutation as highlighted in Figure 1.1.D (right). The square root transformation shows lower-frequency variants with more intensity. The color bar is shown with a square root transformation that maps to an allele frequency range of 0.0025-0.2. Position of each mutation and the base pair change is shown. **(E)** Most recent common ancestor (MRCA) analysis to quantify lineage reconstruction accuracy. Schematic showing hypothetical clones where colors represent arbitrary clonal populations. Trios are analyzed to determine the pair that has the MCRA, including between-clone (*e.g.*, A, C, D) and a within-clone (*e.g.*, B, C, D) example. **(F)** Deconvolution of synthetic samples. For each of two mixture experiments shown are the true proportions ("Experiment") and inferred proportions ("Inferred") for each clone in the mixture, as well as the average deviation. **(G)** Variance component model. Variance explained by the sample structure (y axis, %) for each chromatin accessibility peak (points, rank ordered by variance explained), by the mitochondrial genotypes (red) and the clone ID (black).

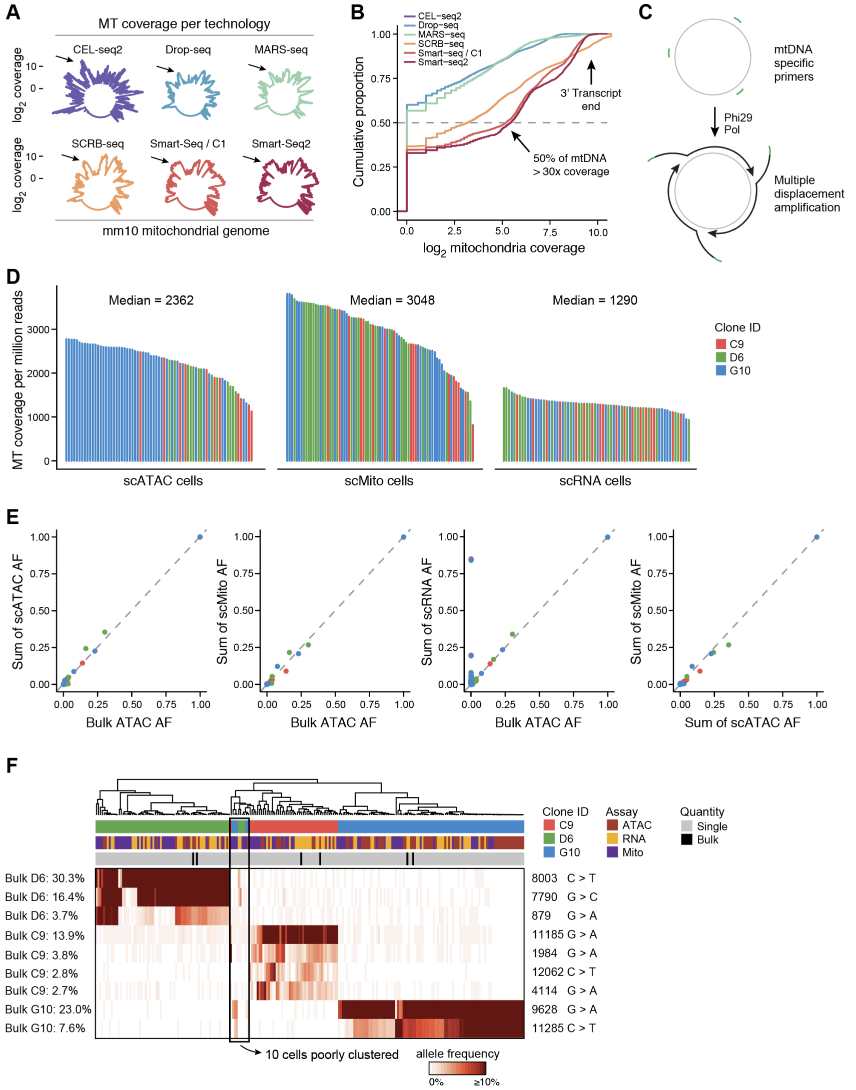

Figure A.2: Assessment of mitochondrial mutations by single cell genomics. (A,B) Coverage of the mitochondrial genome by six different scRNA-seq methods applied to mESCs. **(A)** Log2(coverage) along the mm10 mitochondrial genome for each method. Arrows: a gene uniformly covered by full-length scRNA-seq (SMART-seq methods) but showing, as expected, increased coverage of the 3' end of the transcribed gene in all other methods. **(B)** Cumulative density plot of the mean base pair coverage for each method. Grey dashed line: median coverage. Bottom arrow: SMART-seq approaches cover 50% of bases at 30x or greater. Top arrow: CEL-seq2 and SCRB-seq cover 3' transcript ends more deeply. **(C)** scMito-seq. Mitochondrial sequence specific primers are used for replication of circular mtDNA using the Phi29 polymerase. (D, E) Performance of scATAC-, scRNA-, and scMito-seq. **(D)** Coverage of the mitochondrial genome per million sequence reads (y axis) for cells (bars) from three primary clones (color as in Figure 1C) in each of the three methods. The median cell coverage per million reads is noted. **(E)** Allele frequencies as ascertained by the sum of reads from single cells from each method (y axis) compared to bulk ATAC-seq (x axis) for the same three clones as **(D)**. **(F)** Clones identified by genotype-based clustering across methods. Hierarchical clustering of all TF1 mitochondrial genotyping profiles (columns), including bulk (black) and single cells (grey) from independent single cell assays (purple, yellow, maroon), across the three TF1 clones assayed (red, green, blue as in Figure A.1.C). Color bar: Heteroplasmy frequency (%).

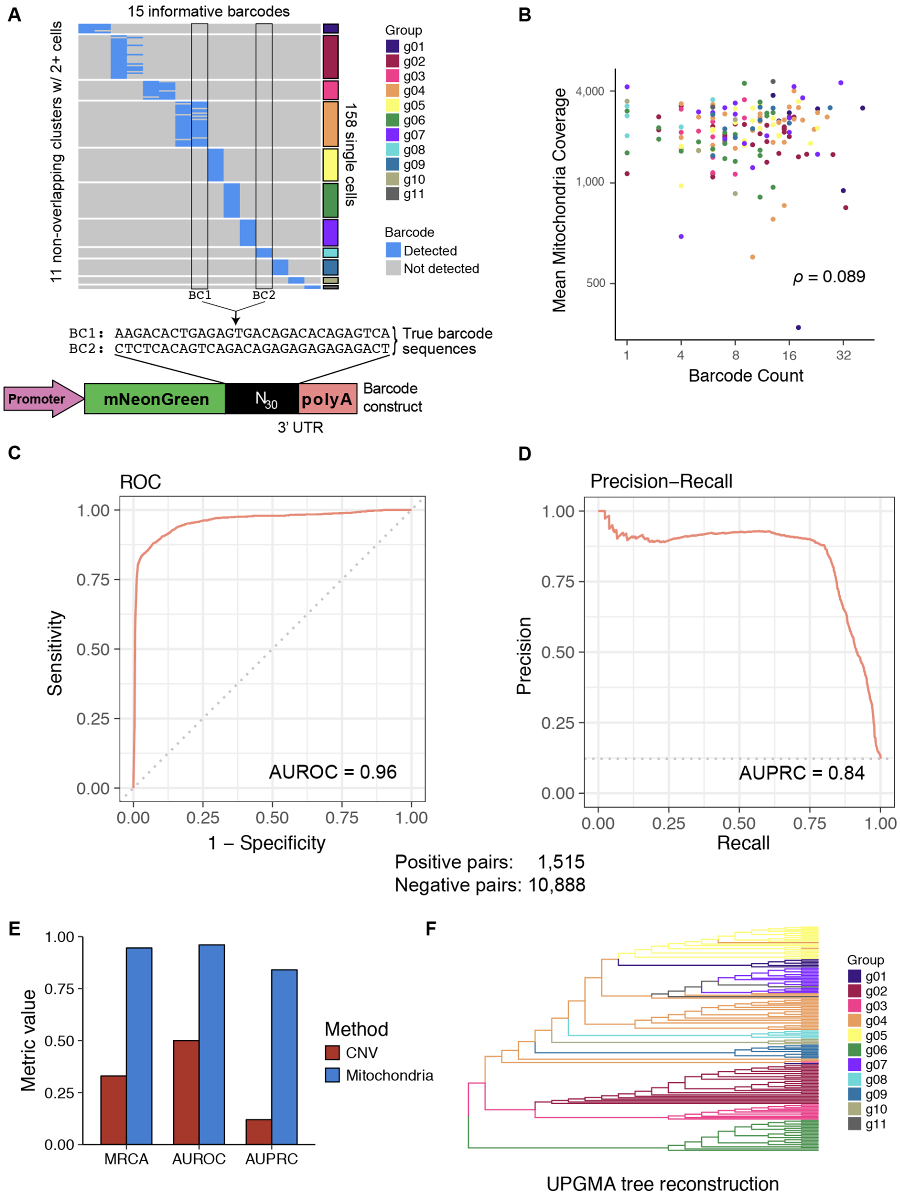

Figure A.3: mtDNA based clone assignment of single cells agrees with lentiviral barcode assignment. (A) Lentiviral barcodes. 15 informative lentiviral barcodes (columns) were used to classify 158 cells to 11 barcode clusters (rows) of at least two cells per cluster. Two 30-mer barcodes are highlighted at the bottom with a scheme of the lentiviral construct. Groups g01-04 are cells that contain two distinct barcodes (multiplicity of infection > 1). **(B)** Low correlation (Spearman ρ = 0.089) between barcode and mitochondrial coverage. Per-cell (dots) mitochondrial coverage (y axis) and lentiviral barcode coverage (x axis). Colors: barcode clones as in in (A). **(C, D)** Concordance between barcode and mtDNA clones. Receiver operating characteristic (ROC) and precision-recall (PR) curves using the Pearson correlation distance as a metric for pairs of cells sharing barcodes. Area under the ROC (AUROC) and PR (AUPRC) are denoted. **(E)** The same metrics (MRCA, AUROC, AUPRC) for mitochondrial and CNV-based distance predicting the same barcode identity in this experiment. **(F)** Visualization of the scRNA-seq data, colored by barcode as in **(A)**, using the UPGMA algorithm.

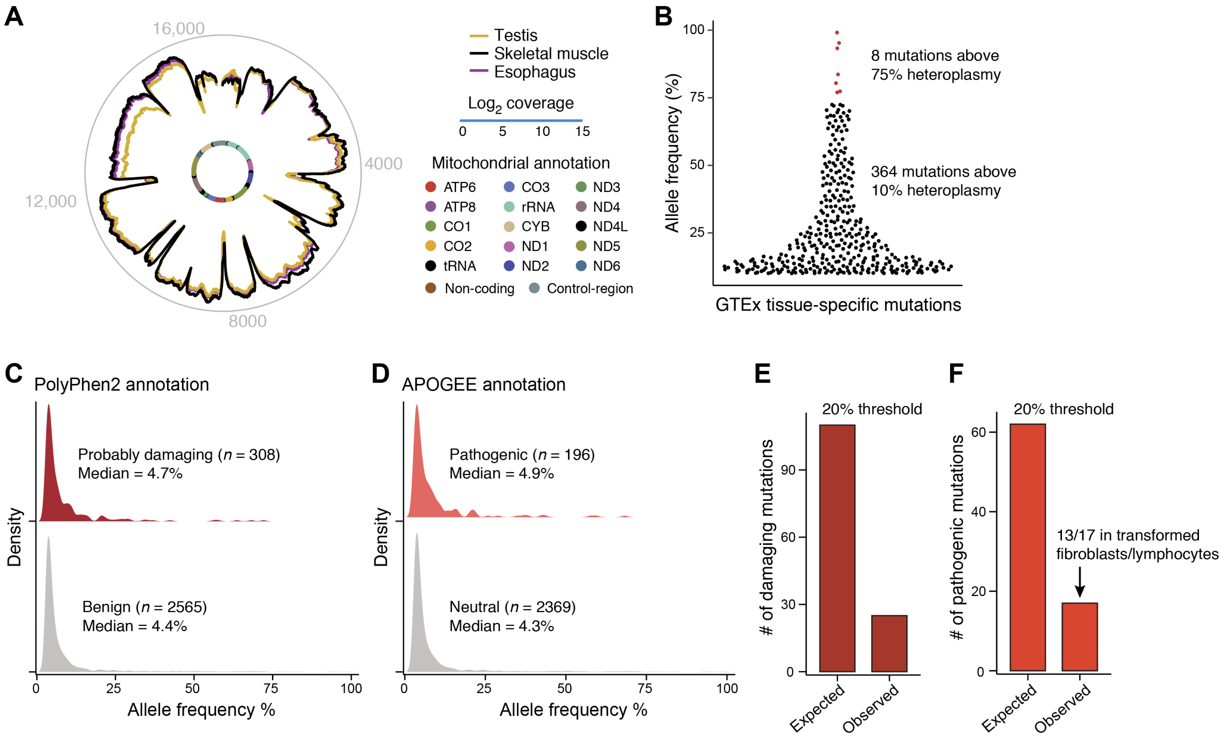

Figure A.4: Detection of heteroplasmic mitochondrial mutations across human tissues. (A) Mitochondrial genome coverage for three tissues (additional to those in Figure 1.4.C). Inner circle: mitochondrial genome annotation; middle circular tracks: mean coverage for testis (orange), skeletal muscle (black), and esophagus (purple); outer grey circle: coordinates of the mitochondrial genome. **(B)** Tissue specific mutations. Beeswarm plot shows the allele frequency (y axis, %) of 372 tissue-specific mutations with a heteroplasmy >10%. Dots: mutation in a tissue from a specific donor. Red: eight mutations with above 75% heteroplasmy. **(C-F)** Reduced number of protein damaging mutations than expected. **(C, D)** Empirical distributions of tissue-specific allele frequencies (x axis, %) for variants annotated as **(C)** protein-damaging (red) or benign (grey) by PolyPhen2, or **(D)** pathogenic (red) or neutral (grey) from APOGEE. Median heteroplasmy is noted and similar across all annotations (between 4-5%). **(E, F)** The number of damaging **(E)** and pathogenic **(F)** mutations (y axis) expected and observed at the tails of the distributions (>20% heteroplasmy). The number of expected mutations are calculated as the product of the number of mutations and the marginal proportions in each category. Many of the pathogenic mutations with higher heteroplasmy were found in transformed fibroblasts/ lymphocytes.

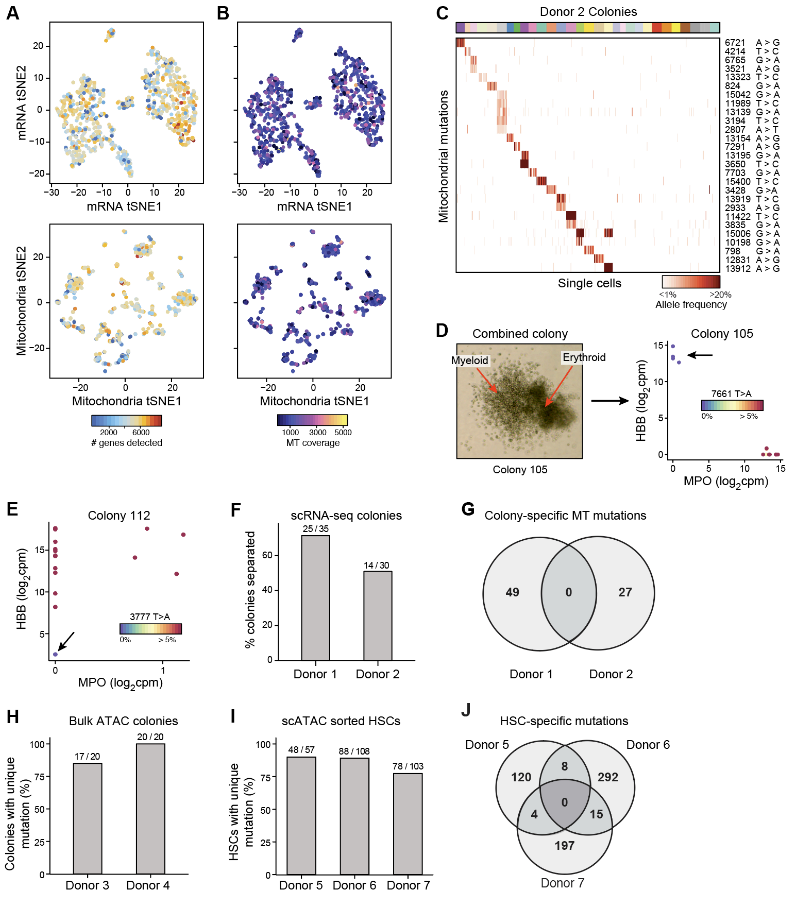

Figure A.5: Mitochondrial mutations in primary hematopoietic cells. **(A, B)** Cell relations based on expression profiles or mitochondrial genotype. tSNE plots computed on expression profiles (top) and mitochondrial genotypes (bottom) colored by **(A)** the number of genes detected (min. 5 counts) per cell, related to Figure 1.5, or **(B)** the fold coverage of the mitochondrial genome per cell, related to Figure 1.5. **(C)** Colony-specific mitochondrial mutations for donor 2. Shown are the allele frequencies of mutations (rows) that are found by supervised analysis as specific to the cells (columns) in each colony (sorted by colony membership; colored bar on top). Position of each mutation and the base pair change is shown. Color bar: Heteroplasmy frequency (%). **(D)** Mixed colonies. Left: Image of colony 105, a mixture of two hematopoietic colonies as confirmed by imaging, gene expression data, and mtDNA genotypes. Right: Scatter plots of the expression levels for a myeloid (MPO, x axis) and erythroid (HBB, y axis) for each cell (dot) in the colony, colored by the allele frequency (color bar) of a heteroplasmic mutation identified only in the myeloid cells. **(E)** Identification of potential contaminant cell in colony 112 based on expression and mtDNA genotype. Scatter plots as in (D) for the cells in colony 112. Arrow: cell lacking the mitochondrial mutation identified in all other cells of this colony, also lacks HBB expression. **(F)** Percentage of individual colonies separated based on mitochondrial mutations (y axis) for donor 1 and donor 2 for the scRNA-seq colony experiment in Figures 1.5.H and A.5.C. **(G)** Colony-specific mutations for donor 1 and donor 2 identified in 1.5.H and A.5. Care non-overlapping. **(H)** Mitochondrial mutations identified through bulk ATAC-seq in primary hematopoietic colonies derived from individual CD34+ HSPCs separate 85% and 100% of those colonies in each of two donors. **(I)** Sorted phenotypic HSCs (CD34+CD38-CD45RA-CD90+) assayed with scATAC-seq for three additional donors show unique mutations in >75% of cells. **(J)** Mutations that distinguish individual HSCs are mostly non-overlapping between donors.

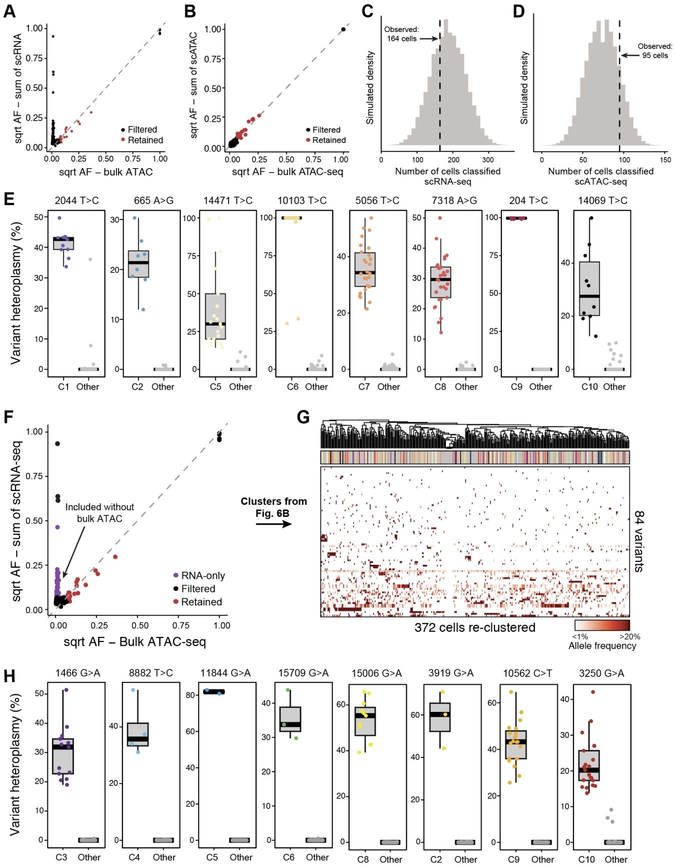

Figure A.6: Mitochondrial mutations identify clonal contributions in polyclonal mixtures of human cells. (A) Allele frequencies for retained mutations agree between scRNA-seq and bulk ATAC-seq. Allele frequencies determined by the sum of single cells from scRNA-seq (y axis) and bulk ATAC-seq (x axis). Black – filtered; red – retained. **(B)** Concordance of allele frequencies between single cell and bulk ATAC-seq. Variant allele frequencies determined by the sum of single cells from scATAC-seq (y axis) and bulk ATAC-seq (y axis), which were retained for (red) or filtered from (black) further analysis. **(C, D)** Number of cells classified by clustering by mitochondrial genotypes. Distribution of the number of cells clustered successfully by mitochondrial genotypes across simulations using cell input from **(C)** scRNA-seq (compare to Figure 1.6.B) or **(D)** scATAC-seq (compare to Figure 1.6.C). Dotted line: observed number of classified cells. **(E)** Selected cluster-specific mutations (compare to Figure 1.6.B). Box plots show the distribution of heteroplasmy (%, y axis) of 8 selected cluster-specific mutations in individual cells for each of 8 clusters, in the specific cluster for the mutation, and in the cells in all other clusters. Dots: individual cells. Dark bar indicates the median single-cell heteroplasmy. **(F, G)** Inclusion of scRNA-seq-specific mutations hampers successful clustering of cells. **(F)** Variant allele frequencies determined by the sum of single cells from scRNA-seq (y axis) and bulk ATAC-seq (x axis). Red: RNA-seq specific mutations retained in the analysis in **(G)** but not in Figure 1.6.B. **(G)** Hierarchical clustering of cells from Figure 1.6.B but when also including the RNA-only mutations from **(F)**. Shown are the allele frequencies of retained heteroplasmic mutations (rows) from scRNA-seq across cells (columns), where cells are sorted by unsupervised clustering. The color bar shown above the cells is the classification inferred from Figure 1.6.B, demonstrating the utility of the addition of the bulk sample for high confidence-variant filtering and exclusion of artefactual variants. **(H)** Cluster specific mutations (compare to Figure 1.6.C). Boxplots for eight selected cluster-specific mutations from each of eight clusters derived from the scATAC-seq experiment. Individual cells are denoted by dots and colored by their cluster membership in the unsupervised analysis.

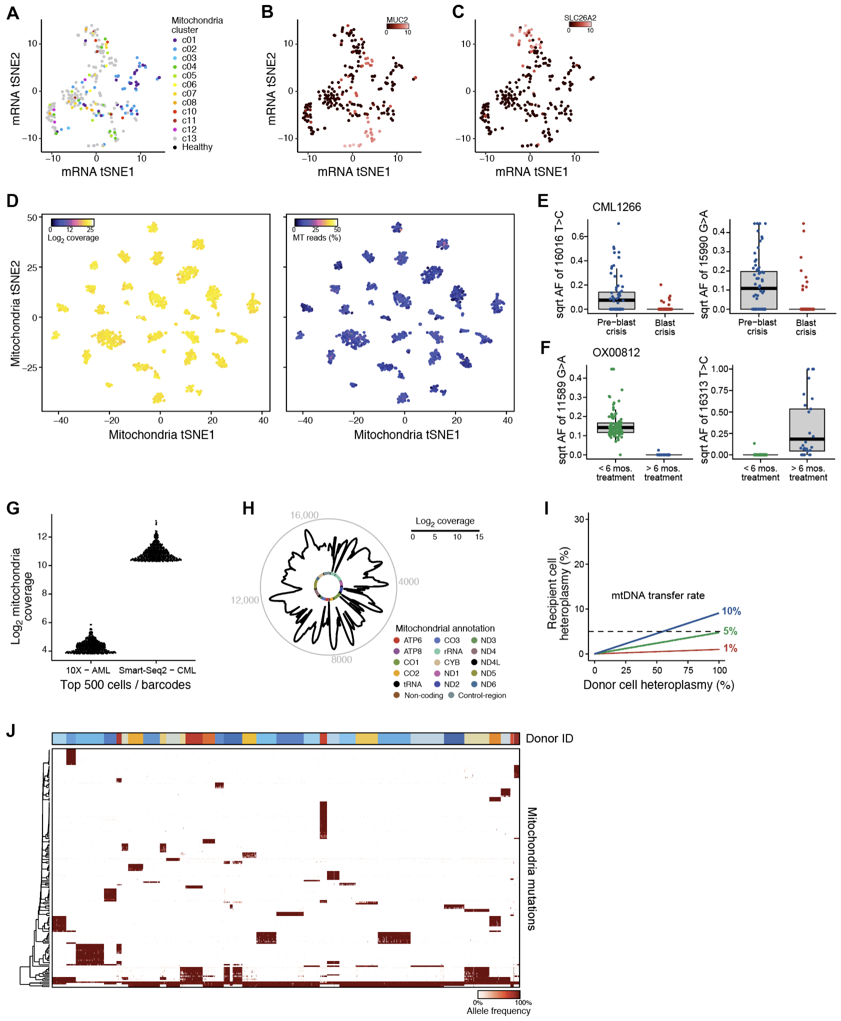

Figure A.7: Application of mitochondrial mutation tracking in human cancer *in vivo*. **(A)** tSNE of clones identified from mitochondrial mutations in Figure 1.7.B. The same coordinates are used to show **(B)** MUC2 expression and **(C)** SLC26A2 expression. Color bar: log2 counts per million. **(D)** Separation of donors by mitochondrial genotype does not reflect coverage. tSNE plots of 2,145 single cells from 31 donors computed on mitochondrial genotypes (as in Figure 1.7.G), with each cell colored by total coverage (left) or the proportion of mitochondrial reads mapping to the mitochondrial transcriptome (right). **(E, F)** Changes in observed allele frequencies at different stage of disease. Box plots show the distribution of allelic frequencies of a specific mutation at different timepoints of disease/ sampling as indicated in Figures 1.7.H,I. Dots are individual single cells; dark bar represents median heteroplasmy. **(G, H)** Reduced mitochondrial coverage by 3' droplet based scRNA-seq. **(G)** The mitochondrial transcriptome coverage (y axis) for the top 500 barcodes and cells (dots) from the 10x Chromium Single Cell 3′ scRNA-Seq (left) and SMART-seq2 (right) datasets, respectively. **(H)** Aggregate mitochondrial transcriptome across cells in the 10x Chromium Single Cell 3′ scRNA-seq dataset. Rounded edges: 3' ends of transcripts, which are relatively well-covered (compare to Figure 1.2.E). **(I)** mtDNA transfer. Heteroplasmy in donor cell (x axis) vs. recipient cell (y axis) from simulations assuming different rates (1, 5 and 10%; colored lines) of horizontal mtDNA transfer from donor to recipient cell and fixed mtDNA content per cell. Dashed line: 5% heteroplasmy in the recipient cell. **(J** Near homoplasmy mutations. Heatmap of the allele frequency (color bar, %) of each of 164 mitochondrial mutations (rows) with near-homoplasmy in one or more of the 2,145 single cells (columns) from 31 donors, sorted by donor annotations (color code on top, as in Figure 1.7.G).

A.1.4 Lentiviral barcoding of TF1 cells

TF1 cells were infected with a modified Perturb-seq lentiviral construct (Dixit et al., 2016) expressing a mNeonGreen gene carrying a 30bp random nucleotide sequence in its untranslated region (Figure A.3.A). For production of lentiviruses, 293T cells were transfected with the appropriate viral packaging and genomic vectors (pVSV-G and pDelta8.9) using FuGene 6 reagent (Promega) according to the manufacturer's protocol. The medium was changed the day after transfection to RPMI 1640 supplemented with 10% FBS, L-Glutamine and Penicillin/ Streptomycin. After 24h, this medium was collected and filtered using an 0.22-μm filter immediately before infection of TF1 cells. The cells were mixed with viral supernatant in the presence of 8 μg/ml polybrene (Millipore) in a 6-well plate at a density of ~300,000 cells per well. The cells were spun at 2,000 r.p.m. for 90 min at 22 °C and left in viral supernatant overnight. The medium was replaced the morning after infection. Twenty-five barcoded mNeonGreen+ cells were sorted at day 3 post infection and expanded for 11 days before processing using a combination of bulk ATAC-seq and scRNA-seq.

A.1.5 Single cell sorting

Single cells were sorted into 96 well plates using the Sony SH800 sorter with a 100μm chip at the Broad Institute Flow Cytometry Facility. Sytox Blue (ThermoFisher) was used for live/ dead cell discrimination. For scRNA-seq, plates were spun immediately after sorting and frozen on dry ice and stored at -80C until further processing.

A.1.6 Bulk ATAC-seq

For ATAC-seq library preparations 5,000-10,000 cells were washed in PBS, pelleted by centrifugation and lysed and tagmented in 1x TD buffer, 2.5μl Tn5 (Illumina), 0.1% NP40, 0.3x PBS in a 50μl reaction volume as described (Corces et al., 2017). Samples were incubated at 37°C for 30min at 300rpm. Tagmented DNA was purified using the MinElute PCR kit (Qiagen). The complete eluate underwent PCR, as follows. After initial extension, 5 cycles of pre-amplification using indexed primers and NEBNext® High-Fidelity 2X PCR Master Mix

(NEB) were conducted, before the number of additional cycles was assessed by quantitative PCR using SYBR Green. Typically, 5-8 additional cycles were run. The final library was purified using a MinElute PCR kit (Qiagen) and quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

A.1.7 Single cell ATAC-seq

The C1 Fluidigm platform using C1 single cell Auto Prep IFC for Open App and Open App Reagent Kit were used for the preparation of single cell ATAC-seq libraries as previously described (Buenrostro et al., 2015). Briefly, cells were washed and loaded at 350 cells/µl. Successful cell capture was monitored using a bright-field Nikon microscope and was typically >85%. Lysis and tagmentation reaction and 8 cycles of PCR were run on chip, followed by 13 cycles off chip using custom index primers and NEBNext® High-Fidelity 2X PCR Master Mix (NEB). Individual libraries were pooled and purified using the MinElute PCR kit (Qiagen) and quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

A.1.8 Bulk RNA-seq

Cells were lysed in RLT or TCL lysis buffer (Qiagen) supplemented with beta-mercaptoethanol and RNA was isolated using a RNeasy Micro kit (Qiagen) according to the manufacturer's instructions. An on-column DNase digestion was performed before RNA was quantified using a Qubit RNA HS Assay kit (Invitrogen). 1-10ng of RNA were used as input to a modified SMART-seq2 (Picelli et al., 2014) protocol and after reverse transcription, 8 cycles of PCR were used to amplify transcriptome library. Quality of whole transcriptome libraries was validated using a High Sensitivity DNA Chip run on a Bioanalyzer 2100 system (Agilent), followed by library preparation using the Nextera XT kit (Illumina) and custom index primers according to the manufacturer's instructions. Final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity

DNA chip run on a Bioanalyzer 2100 system (Agilent).

A.1.9 Single cell RNA-seq

Single cells were sorted into 5ul TCL lysis buffer (Qiagen) supplemented with 1% beta-Mercaptoethanol. RNA isolation, reverse transcription and PCR were conducted as described using a modified SMART-seq2 protocol (Picelli et al., 2014). Quality control and library preparation were conducted as described above.

A.1.10 Single cell Mito-seq

Single cells were sorted in to 500l TCL lysis buffer (Qiagen) supplemented with 1% beta-mercaptoethanol. DNA was isolated with AMPure XP beads (Beckman Coulter) and the REPLI g Mitochondrial DNA kit (Qiagen) was used for amplification at 33C for 8h in a 16.5ul reaction volume. Amplified DNA was cleaned up with AMPure XP beads (Beckman Coulter), quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and library preparation was performed using the Nextera XT kit (Illumina) using custom index primers according to the manufacturer's instructions.

A.1.11 Processing of human colorectal cancer and adjacent healthy tissues

Fresh tissue was collected into RPMI 1640 medium supplemented with 2% human serum (Sigma), cut into 1 mm2 pieces, and enzymatically digested for 20min at 37°C using the Human Tumor Dissociation Kit (Miltenyi Biotec) in the presence of 10μM ROCK inhibitor Y-2763 (Sigma). Cell suspension was passed through 70μm cell strainers and centrifuged for 7min at 450g at 4°C. Supernatant was removed and cells were subject to ACK Lysing Buffer (Life Technologies) for 2min on ice, centrifuged for 7min at 450 g at 4C, and resuspended in RPMI 1640 supplemented with 2% human serum (Sigma). The single cell suspension was stained with Zombie Violet in PBS (Invitrogen) for 10min on ice and subsequently with antibodies against human CD326, CD45, and CD235a (Biolegend) in RPMI 1640 medium supplemented with 1% human serum in the presence of 10μM

Y-2763 for 15 min on ice. Zombie Violet- CD235a- CD45- CD326+ cells were bulk sorted into 1.5ml Eppendorf tube containing 1x TD buffer, 2.5µl Tn5 (Illumina), 0.1% NP40, 0.3x PBS in a 50µl reaction volume for ATAC-seq as described above. Using the identical gating scheme, single cells were sorted into Eppendorf twin-tec PCR plates containing 10µl TCL lysis buffer (Qiagen) supplemented with 1% beta-Mercaptoethanol and processed for scRNAseq as described above.

A.1.12 Sequencing

All libraries were sequenced using Nextseq High Output Cartridge kits and a Nextseq 500 sequencer (Illumina). Libraries were sequenced paired-end (2x 38 or 2x 75 cycles).

A.2 Bioinformatics methods for bulk sequencing

A.2.1 Data processing and read alignment

For each sequencing library generated in this study, libraries were sequenced on an Illumina NextSeq 500 and demultiplexed using the bcl2fastq program. For each library, raw .fastq reads were aligned using either Bowtie2 version 2.3.3 (Langmead & Salzberg, 2012) or STAR version 2.5.1b (Dobin et al., 2013) to the hg19 reference genome. For the mESC scRNA-seq coverage comparison (Figure 1.2.A), reads from the published dataset (Ziegenhain et al., 2017) were aligned to the mm10 reference genome.

RNA-seq and scRNA-seq transcript counts were computed using STAR's "–quantModes GeneCounts" flag using the Gencode 19 release .gtf file.

For the published droplet based scRNA-Seq (10X Genomics) AML dataset (Zheng et al., 2017b), processed `.bam` files (aligned to GRCh37) were downloaded from the public downloads page on the `10x` website.

Raw .fastq files for public RNA-seq and scRNA-seq data were downloaded from the Gene Expression Omnibus (GEO), European Nucleotide Archieve (ENA), the database of Genotypes and Phenotypes (dbGaP) resources or European Genome-Phenome Archive (EGA), as follows: GSE75790 (mESC scRNA-seq); PRJEB20143 (SIDR scDNA/RNA-seq); phs000424.v7.p2 (GTEx); T lyphocytes (EGAS00001002072; EGAS00001002430).

A.2.2 Mitochondrial genotyping

For each sequencing sample, per-base, per-allele counts were determined using a custom Python script that imported aligned `.bam` files using the pysam module (https://github.com/pysam-developers/pysam). Raw reads were filtered such that they had an alignment quality of 30 and were uniquely mapping to only the mitochondrial genome. The mean base-quality score was computed per-base, per-allele for each sample for quality control. At a given mitochondrial genome position x, the allele frequency (AF) of a base b was computed using the number of reads R supporting that particular base at position x:

$$AF_{x,b} = \frac{R_b}{\sum_{b \in A,C,G,T} R_b}$$

where $\sum_{b \in A,C,G,T} R_b$ is the coverage of a given position x.

A.2.3 Variant quality control and filtering

To remove variants whose inferred heteroplasmy may reflect sequencing errors, we examined the distribution of per-base, per-allele base-quality scores, noting a clear pattern of high quality and low-quality variants (Figure A.1.C). To determine high quality variants, we fit a mixture of three Gaussian distributions (Figure A.1.C, labeled by different colors), and filtered such that only alleles that had >99% probability of belonging to the blue (largest mean BQ) Gaussian were retained. This conservative empirical threshold for a BQ cutoff was determined to be 23.8 based on this mixture model approach (Figure A.1.C, vertical dotted line).

As one poorly quantified position allele would affect the estimates for all other alleles at the specific position, we filtered positions that contained one or more alleles with a BQ less than the empirical threshold unless the allele had a non-significant (*i.e.*, less than 1 in 500) effect on heteroplasmy. In total, we called 44 high-quality variants across our TF1 (sub-)clones (Figure A.1.D) that were present at a minimum of 2.5% heteroplasmy in at least one sample. Throughout the study, we observed a preponderance of C>T, T>C, G>A, and A>G mutations (transitions), consistent with previous reports (Ju et al., 2014). Of note, we used bulk ATAC-seq to nominate high-quality variants across three other hematopoietic cell lines (GM12878, K562, and Jurkat) and observed 29-64 heteroplasmic mutations per line, suggesting our inferences in Figures 1.1-1.3 would generalize to other cell lines.

A.2.4 Mitochondrial distance matrix

As input to the variance components models (Figure 1.1.G), we computed a mitochondrial relatedness matrix $K_{mito} = 1 - D$, where D is a symmetric, pairwise distance matrix whose elements encode the distance between pairs of cells or clones based on the differences in their respective allele frequencies. We define D for pairs of observations i, j over high-quality variants $x \in X$ using the matrix of allele frequencies (AF) and coverage frequencies (C), such that only variants sufficiently well-covered (minimum number of reads at the position > 100) are included. Explicitly, we define the mitochondrial distance between observations i, j using the distance $d_{i,j}$ as follows:

$$d_{i,j} = \frac{\sum_x \sqrt{\left|AF_{x,i} - AF_{x,j}\right|} * (1_{C_{x,i} > 100} * 1_{C_{x,j} > 100})}{\sum_x (1_{C_{x,i} > 100} * 1_{C_{x,j} > 100})}$$

where 1 is the indicator function. Intuitively, this representation of mitochondrial distance simultaneously accounts for variation in rare heteroplasmy (through the square root transformation) and only compares pairs of cells by their high-confidence variants. We note that the square root transformation yields a one-to-one mapping of allele frequencies and provides relative weight to variants whose allele frequencies are very close to zero.

For the bulk ATAC-seq of TF1 (sub-)clones analyzed in Figure 1.1, all quality-controlled variants passed the coverage requirement; however, the additional indicator functions for coverage were necessary for subsequent single cells experiments.

For the hierarchical clustering of the TF1 lineage cells, we used a modified mitochondrial distance metric computed from the Pearson correlation distance. Intuitively, this metric is less dependent on the absolute values of the variant heteroplasmy. We note that while an ideal tree reconstruction algorithm would facilitate the inclusion of internal nodes, we found no such algorithms readily available, as most tree reconstruction approaches do not allow for internal observations. Further, we did not pursue the development of such approaches here.

A.2.5 Variance components model

To determine the proportion of the variance of chromatin accessibility that could be explained by the mitochondrial lineage in each peak, we performed a variance decomposition using a random effects model (Figure 1.1.H). Briefly, the chromatin accessibility counts measured from ATAC-seq for 91,607 accessibility peaks were summed, centered, and scaled for each sample. We then estimated for each peak the proportion of variance explained due to the random variance component (σ_e^2) and due to the variance component from the sample-sample structure inferred by the mitochondrial genotype (σ_m^2), using average information restricted maximum likelihood (AIREML). Explicitly, our model for the variance of chromatin accessibility account for an individual peak is:

$$\text{Peak Accessibility} \sim N(0,\ \sigma_m^2 \mathbf{K}_{\textbf{mito}} +\ \sigma_e^2 \mathbf{I})$$

and the proportion of the variance explained by the mitochondrial structure then is the ratio of σ_m^2 over the total variation:

$$\frac{\sigma_m^2}{\sigma_m^2 + \sigma_e^2}$$

The proportion of the variance explained by the mtDNA mutation substructure is shown for each peak in Figure A.1.G alongside an analogous calculation, where the substructure is only defined by a binary indicator of clonal membership for pairs of samples.

A.2.6 Most Common Recent Ancestor (MRCA) analysis

To determine our ability to accurately reconstruct the experimental lineage in Figure 1.1 by mitochondrial mutations, we determined the proportion of correctly identified Most-Recent Common Ancestors (MRCA) for trios of (sub-)clones, similar to an approach recently reported by Biezuner et al. (2016). For any given set of three samples in the predicted tree (*e.g.* A, C, and D; in Figure A.1.E), three possible arrangements are possible: (1) A and C share an MRCA compared to D; (2) C and D share an MRCA compared to A; or (3) A and D share an MRCA compared to C. Given the true experimental lineage tree (in this example, arrangement 2), we determined whether or not our reconstructed lineage correctly identified the MRCA. Thus, by chance, a random tree reconstruction would be 33% accurate. Here, we distinguish comparisons within-clone (*e.g.*, B,C,D in Figure A.1.E) or between clones (*e.g.* A,C,D) and demonstrate that our tree reconstruction significantly outperforms what is expected by chance in both settings.

A.2.7 Clonal mixture deconvolution (TF1 clones)

To demonstrate that clonal mixtures can be deconvoluted, we mixed our second-generation clones in known proportions and inferred these proportions from the mitochondrial genotype of the mixture. For two known mixture fractions (Figure A.1.F), we genotyped each mixed sample with bulk ATAC-seq and then used the second-generation allele frequencies to infer each mixture, by fitting a support vector regression model to estimate the mixing proportions, in a manner analogous to CIBERSORT (Newman et al., 2015). As shown in Figure A.1.F, the average deviation of the inferred and true mixing proportions are 1.7% and 3.0%, demonstrating that *a priori* defined genotypes can be used to approximate the contributions of complex mixtures.

A.2.8 Comparison of scRNA-seq methods

To compare mitochondrial coverage with different scRNA-seq methods, we downloaded a dataset of 583 scRNA-seq profiles from mouse embryonic stem cell (mESC) (Ziegenhain et al., 2017). Reads were aligned to the mm10 reference genome using STAR. Per-base pair coverage estimates were computed for each single cell using reads uniquely mapping to the mitochondrial genome.

To verify that heteroplasmic variants were expressed at a comparable frequency as these heteroplasmies in DNA, we downloaded 38 high-quality profiles, where both mitochondrial genome and transcriptome were available (Han et al., 2018b). Reads from mtDNA and RNA were aligned as described above to the hg19 reference genome, using Bowtie2 and STAR, respectively, and heteroplasmic allele frequencies were plotted for variants with at least 50 reads covering the locus in both RNA and DNA both with a minimum BQ score of 20 in the same cell.

A.2.9 Comparison of scRNA-Seq, scATAC-Seq and scMito-Seq (TF1 clones)

To compare given single cell profiling methods to the corresponding bulk method or to other single cell and bulk methods, we summed all raw allele counts for high-quality cells (minimum of 100X mitochondrial genome coverage). We performed such comparisons for nine characterized, clone-specific heteroplamsic variants (Figure A.2.F) and for variants identified as RNA-specific (Figure 1.2.D). We further plotted the allele frequency comparing the two technologies for heteroplasmic variants, revealing concordance across all the technologies (Figure A.2.E).

A.2.10 Validation of clonal mutations in single cells using lentiviral barcoding

To detect barcodes in TF1 scRNA-seq libraries, we appended a 221 base pair “chromosome” to the standard STAR hg19 reference genome where the 30bp random sequence was soft-masked. Custom Python scripts determined reads uniquely aligning to the lentiviral construct that overlapped the random 30bp barcode. From the 20 mutations nominated in Figure 1.3.C, a cell-cell distance metric was computed from the Pearson correlation of the square root of the heteroplasmy matrix. This metric was similarly used for the MRCA analysis as described for Figure 1.1. For each pair of cells, we used the group designation from the lentiviral barcode assignment as a binary classifier and the mitochondrial distance metric as a diagnostic metric of cell-cell similarly to compute receiver operating characteristics.

A.2.11 CNV calling for lentivirally barcoded TF1 cells

Copy number variation (CNV) was determined using the InferCNV tool run using the default settings (Patel et al., 2014). We modified the main script to return the cell-cell distance matrix computed before performing the default hierarchical clustering. This cell-cell distance matrix (computed over the CNV bins) was used as input to our MRCA computation.

A.3 GTEx analyses

Raw .fastq files were downloaded from dbGAP as noted above for nearly 10,000 samples sequenced on Illumina Hi-Seq with 75bp paired-end reads. We retained 8,820 samples belonging to one of 49 tissues that had at least 25 total samples, from individuals with at least 10 tissues, and with mean mitochondrial genome coverage of 1000x. We define a "tissue specific mutation" (Figures 1.4.D,F,G) for a given mitochondrial variant if the variant is present at least at 3% heteroplasmy (or more where indicated) in an individual tissue but no more than 0.5% (within our margin of error for bulk RNA-seq) in any of the other tissues for a specific donor. We removed mutations that occurred within a given tissue in more than 10 individuals to exclude the possibility of tissue-specific mitochondrial RNA-editing events. While the noise in the RNA-seq assay inherently leads to more false positives and less certainty in the heteroplasmy estimation, our procedure of comparing heteroplamsic values against other tissues within a donor provides a conservative means toward identifying putative somatic mutations that arose during development or homeostasis.

To compute the expected number of pathogenic and damaging mutations (Figures 1.4.E,F), we multiplied the number of loci that were observed above a defined heteroplasmy threshold (*e.g.*, 20%) by the rate at which damaging or pathogenic mutations occur in the mitochondrial genome.

Dimensionality reduction using mRNA expression profiles or mitochondrial genotypes We performed a t-stochastic neighbor embedding (t-SNE) of the cells by either their expression or mitochondrial genotype profiles (Figures 1.5 and A.5). First, we identified a set of 935 high quality scRNA-seq profiles that (1) have at least 500 genes detected, (2) had a total count of at least 2,000 across expressed genes, and (3) had a mean mitochondrial genome coverage of at least 100x. For dimensionality reduction by expression profiles, we first batch-corrected a log counts-per-million matrix of gene expression values using sva (Leek et al., 2012) and used the top 10 principal components for our t-SNE. For the dimensionality reduction by mitochondrial genotype profiles, we used all variants with a mean BQ score of 25 present at a heteroplasmy of at least 0.5% in our population of cells and similarly computed t-SNE coordinates using the top 10 principal components of the heteroplasmy matrix. We observed no significant batch effect in the mitochondrial allele frequencies.

Supervised identification of colony and cell-specific mutations in hematopoietic cells To identify mutations that effectively separate individual colonies in donors 1 and 2 (Figures 1.5 and A.5), we searched for mutations present at a minimum of 80% of cells within a colony, at a minimum heteroplasmy of 5%, but are not present at greater than 5% heteroplasmy in more than two cells from all the other colonies together.

To identify mutations that separate individual bulk ATAC colonies (donor 3 and 4), we searched for mutations that were present at a heteroplasmy > 5% in a particular colony but absent (< 0.5% heteroplasmy) in all other colonies.

To identify cell-specific mutations in FACS-sorted HSCs (donors 5, 6, and 7), we searched for mutations that were present at > 5% heteroplasmy for a particular cell, but otherwise absent (< 0.5%) in all other cells for a specific donor.

A.4 Single-cell bioinformatics analyses

A.4.1 Separation of clonal mixtures of CD34+ HSPCs

For the analysis of CD34+ HSPCs, we identified variants that had a mean BQ score of at least 20 for both the sum of single cells and the bulk ATAC-seq and were detected in bulk at a heteroplasmy of at least 0.5%. This identified 14 for scRNA- (Figure A.6.A) and 16 high quality variants for scATAC-seq (Figure A.6.B).

Using these variants and cells passing filter (minimum average mitochondrial genome coverage of 100x), we performed a fuzzy k-medoids clustering and assigned a cell to a cluster if it had an assignment probability greater than 95% and left it unassigned otherwise. We identified 9 clusters for scATAC-seq and 10 for scRNA-seq that corresponded directly to one or more mutations (Figures A.6.B,C). While other cells showed evidence of mutations, these occurred at lower heteroplasmy values than the frequencies for cells assigned to the group (Figures A.6E,F).

A.4.2 Simulated density of assignment

To verify that our probabilistic cluster was within the range of expectation, we performed a simulation study by parameterizing attributions of our mixing experiment (Figures A.6C,D). Specifically, for each of the 30 input CD34+ cells, we simulated a proportion of the specific cell in the final population $p_i,\ i \in 1, \dots, 30$, using a Beta distribution:

$$p_i \sim \mathit{Beta}(1,\ 29)$$

In expectation, the proportion in the terminal cell populations would be 1/30, consistent with the expectation of the draw from the Beta distribution. From this vector of population proportions p, we simulate the number of cells N sampled from our single-cell sampling using a multinomial distribution:

$$n \sim Multinomial(N, p)$$

where $N = 372$ and 148 for the scRNA-Seq and scATAC-Seq, respectively. Thus, n_i represents the number of cells that were derived from a single original cell i. Next, we simulated whether cell i contained a mutation that could be detected and clustered in a group of cells ($r = 1$). This was achieved using a Bernoulli draw for each cell:

$$r_i = Bern(q)$$

where q was estimated to be 0.5 based on our analyses in Figure 1.5 for scRNA-seq. Finally, the total number of cells clustered (c, the unit shown on Figures A.6.C,D) is computed from the following: $c = \sum_{i=1}^{30} r_i * n_i$

For both scATAC- and scRNA-seq, we computed c over 10,000 simulations each. Our observed number of cells clustered in Figures A.6.C,D fell comfortably within the 95% coverage interval for both scATAC- and scRNA-Seq (Figures A.6.C,D).

A.4.3 Analysis of colorectal cancer data

Bulk ATAC-seq and scRNA-seq libraries were aligned using bowtie2 and STAR as described above. We identified variants that had a mean BQ score of at least 20 for both the sum of single cells and the bulk ATAC-seq and were detected in bulk at a heteroplasmy of at least 0.5%, yielding 12 high-quality variants. Clusters were defined using a similar procedure as described in the previous section. With the exception of 15044 G>A, the highest heteroplasmy in the bulk healthy samples was 0.0009. In total, 12 high-confidence clusters were identified with at least 2 cells. A t-SNE mapping of cells was rendered for the mRNA profiles as described above (Figures 1.7.D-F and A.7.A-C).

A.4.4 Dimensionality reduction of CML scRNA-seq data

To address spurious variants in scRNA-seq in the absence of a bulk DNA guide (Figure A.6.G), we hypothesized that using a more stringent measure of quality, base alignment quality (BAQ) (Li, 2011), could facilitate the identification of fewer higher quality variants. Indeed, we identified 242 high-quality variants that had a minimum BAQ score greater than 20 with a mean heteroplasmy of 0.5% in the population of high quality cells (minimum mean mitochondrial genome coverage of 100x).

We performed a t-SNE on the first 25 principal components from the z-score normalized heteroplasmy matrix using default parameters (perplexity = 30). We used a Mann-Whitney U-Test to identify variants that co-varied with annotated patient sub-phenotypes at a significance of $p < 10^{-3}$ within a given donor.

A.4.5 Analysis of CML scRNA-seq data

Clustering of the scRNA-seq data for donor CML656 was performed using SC3 (Kiselev et al., 2017) on processed expression values available through GEO accession GSE76312, with default parameters for clusters of size 2, 3, and 4. The data form the 29 cells in cluster 1 were re-processed using STAR (Dobin et al., 2013) using parameters noted above, followed by differential expression testing using limma-voom (Law et al., 2014). The lowest non-zero allele frequency of 4824 T>C for a cell in cluster 1 was 4%, providing a clear basis for determining cells that were 4824 T>C + (that is, any cell with a non-zero allele frequency for 4824 T>C were considered 4824 T>C +). In total, 14 cells in cluster 1 were negative for the mutation whereas 15 were positive, which served to define categories for differential gene expression within cluster 1 cells.

A.4.6 Analysis of T lymphocyte scRNA-seq data

Raw .fastq files were downloaded from the European Genome-phenome Archive. Meta data associated with each cell was further downloaded with the raw sequencing data, and included a definition of clones based on TCR sequences inferred by TRACER (Stubbington et al., 2016). In instances where we observed heterogeneity in mitochondrial mutations within a clonal marker (*e.g.* Figure 1.6.F), we verified that TCR annotations were supported by > 100 reads as reported in the meta data.

A.4.7 Processing the AML scRNA-Seq dataset

For the AML datasets previously generated by `10x` Genomics (Zheng et al., 2017b), cells from two patients (AML027 and AML035) were analyzed for mitochondrial genotypes. Aligned and processed `.bam` files were downloaded from the `10x` website and further processed using custom Python scripts. Cell barcodes associated with at least 200 reads uniquely aligning to the mitochondrial genome were considered for downstream analysis. Barcodes were further filtered by requiring coverage by at least one read at two specific variants at mtDNA positions 3010 and 9698. We note that we did not observe a barcode that contained a read to support both alternate alleles (3010G>A and 9698T>C). We determined that 4 out of 1,077 cells were derived from the recipient (1.7.M), a higher estimate than in the previously reported analysis performed with nuclear genome variants (reported exactly 0%) (Zheng et al., 2017b), though these four cells were not included in the published analysis as they did not pass the author's barcode/ transcriptome filters. We did not observe a well-covered set of variants separating the donor/ recipient pair in the AML027 dataset, and did not further analyze it for mutations but only for determining well-covered barcodes (Figure A.7.G,H).

B

Supplemental material for Chapter 2

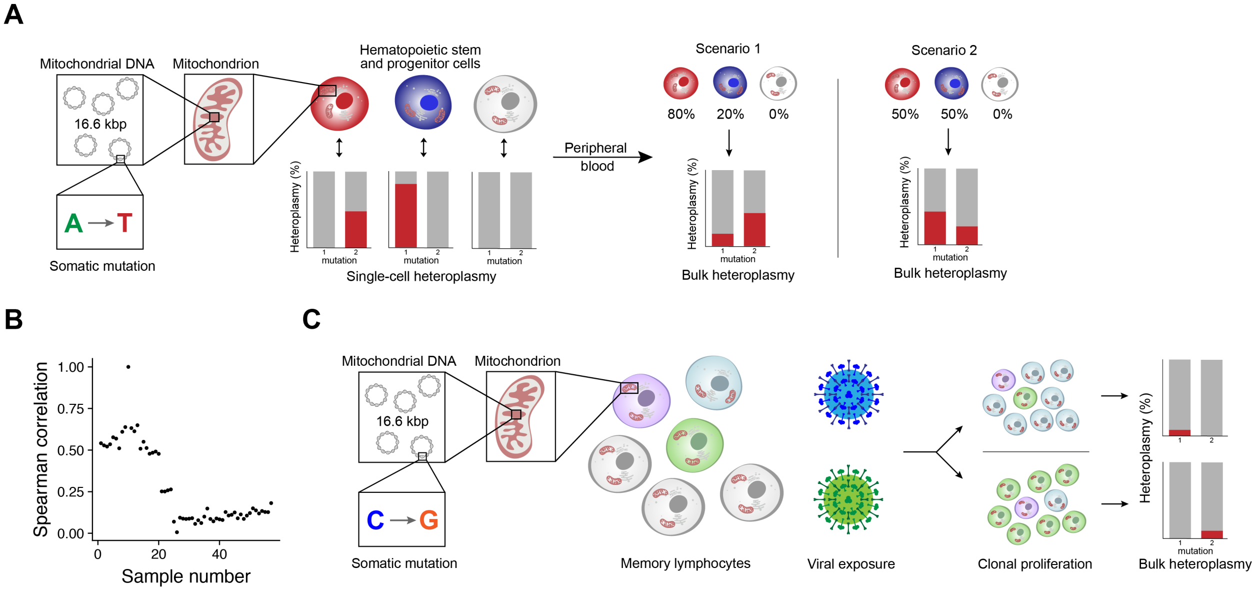

Figure B.1: Theoretical basis for heteroplasmy variation *in vivo*. **(A)** Schematic illustrating how the varying contributions of progenitor cells, carrying specific somatic mtDNA mutations at indicated allele frequencies, may affect heteroplasmy levels in bulk population level measurements of peripheral blood. **(B)** Spearman correlation of 57 time points (ordered by relative time of sampling) across time points sampled. Correlation value is measured with the tenth sample. Compare to Figure 2.1.C. **(C)** Schematic illustrating how the clonal expansion of antigen-specific lymphocytes carrying clone-specific somatic mtDNA mutations may lead to fluctuations in heteroplasmy levels in bulk population level measurements of peripheral blood.

B.1 Data acquisition

Raw sequencing reads were downloaded from Gene Expression Omnibus (GEO) accessions GSE33029 and GSE111405 for data related to Figure 2.1. For Figure 2.2, raw sequenced reads were obtained from accessions GSE85853 and GSE111015.

B.2 Bioinformatics methods

Alignment to the hg19 reference genome was performed using appropriate tools for RNA-seq: STAR (Dobin et al., 2013), ATAC-seq: bowtie2 (Langmead & Salzberg, 2012), and whole-genome bisulfite sequencing: bismark (Krueger & Andrews, 2011). Reads aligning to the mtDNA genome were extracted using samtools Li

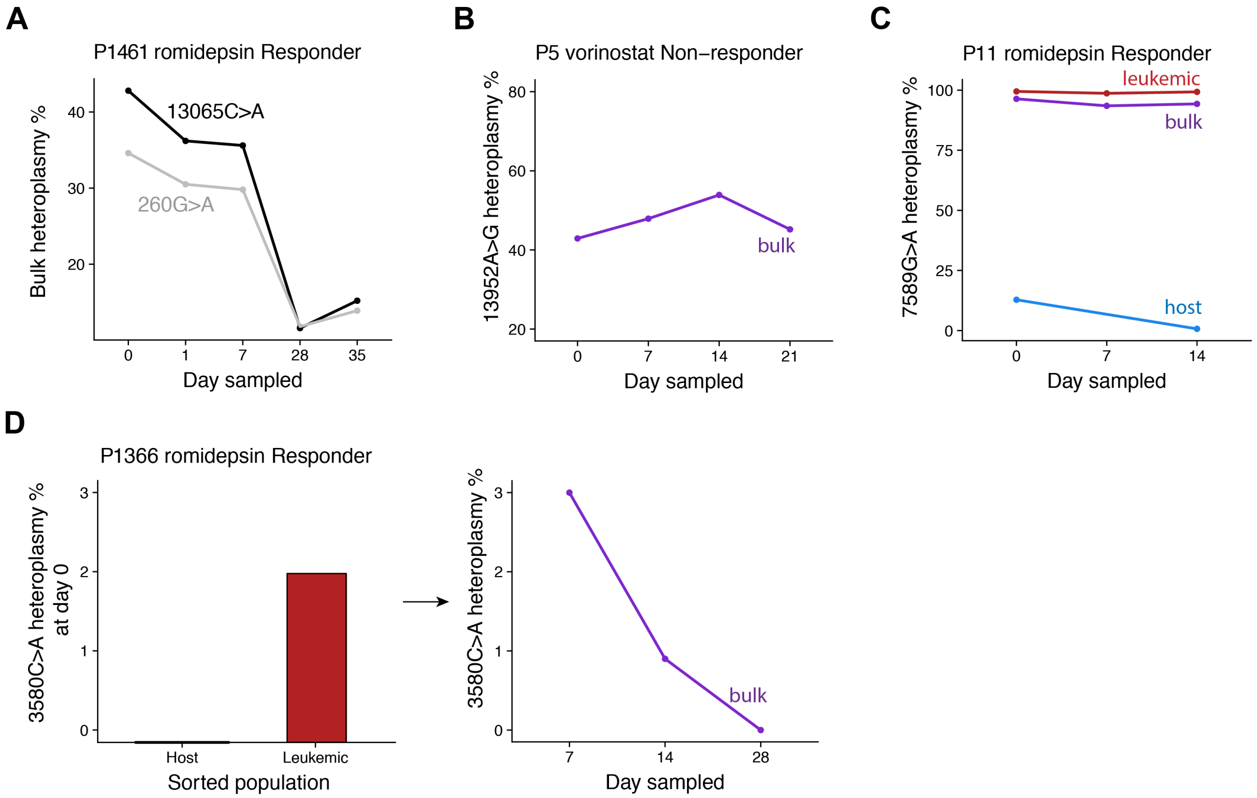

Figure B.2: Supporting evidence for mtDNA mutation dynamics in response to therapeutic treatment cases of CTCL *in vivo*. **(A)** Examples of two mutations losing heteroplasmy over 5 weeks of sampling from a responder treated with romidepsin. **(B)** Example of a heteroplasmic mutation that persists at steady states across three weeks in a non-responder to vorinostat treatment. **(C)** Heteroplasmy of 7586G>A mutation as measured in bulk peripheral blood and enriched leukemic or host cells over two weeks of treatment in patient P11, who responded to romidepsin therapy. Note loss of heteroplasmy in the host cells, but stable levels in the leukemic population, suggesting persistence of leukemic cells carrying the 7586G>A allele. **(D)** Heteroplasmy of the 3580C>A allele present in enriched leukemic cells, but absent in host cells at day 0 of treatment (left). Loss of the 3580C>A allele in bulk peripheral blood at the indicated time points following start of treatment, suggesting therapy-sensitivity of leukemic cells carrying the respective allele (right). We note that other populations (*e.g.* sorted after day 0) were not available.

(2011), and PCR-duplicated reads were removed using Picard tools. Per-sample, per-mutation heteroplasmy abundances were estimated using our previously reported pipeline. All depicted mutations were selected on the basis of supervised analyses. Mutations in RNA-seq were specifically filtered against a set of purported RNA-editing events as we have previously described in Chapter 1. All meta-data (*e.g.* sample, timepoint) was curated from the GEO accessions that contained the raw high-throughput sequencing data.

C
Supplemental material for Chapter 3

C.1 Biological methods

C.1.1 Cell lines

GM12878 (Coriell Institute for Medical Research) human lymphoblastoid cells were maintained in RPMI 1640 medium modified to include 2 mM L-glutamine (ATCC), 15% FBS (ATCC) and 1% Penicillin Streptomycin (Pen/Strep) (ATCC). K562 (ATCC) human chronic myelogenous leukemia cells were maintained in Iscove's Modified Dulbecco's Medium (IMDM) (ATCC) supplemented with 10% FBS and 1% Pen/Strep. NIH/3T3 (ATCC) mouse embryonic fibroblast cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM) (ATCC) supplemented with 10% Calf Bovine Serum and 1% Pen/Strep. All cell lines were maintained at 37°C and 5% CO2 at recommended density and were harvested at mid-log phase for all experiments. All suspension cells were harvested using standard cell culture procedure, and adherent cells were detached using TrypLE Express Enzyme (Gibco). After harvesting, cells were washed twice with ice cold 1x PBS (Gibco) supplemented with 0.1% BSA (MilliporeSigma). Cells were then filtered with a 35 µm cell strainer (Corning) and cell viability and concentration were measured with trypan blue on the TC20 Automated Cell Counter (Bio-Rad). Cell viability was greater than 90% for all samples. See the Life Sciences Reporting Summary for more information.

C.1.2 Mouse tissues

Flash frozen adult mouse whole brain tissue was purchased from BrainBits (SKU: C57AWB). Nuclei isolation was performed using the Omni-ATAC protocol for isolation of nuclei from frozen tissues (Corces et al., 2017). Nuclei permeability and concentration were measured with trypan blue on the TC20 Automated Cell Counter. For all samples, over 95% of the nuclei were permeable to trypan blue, meaning that the nuclei isolation was successful.

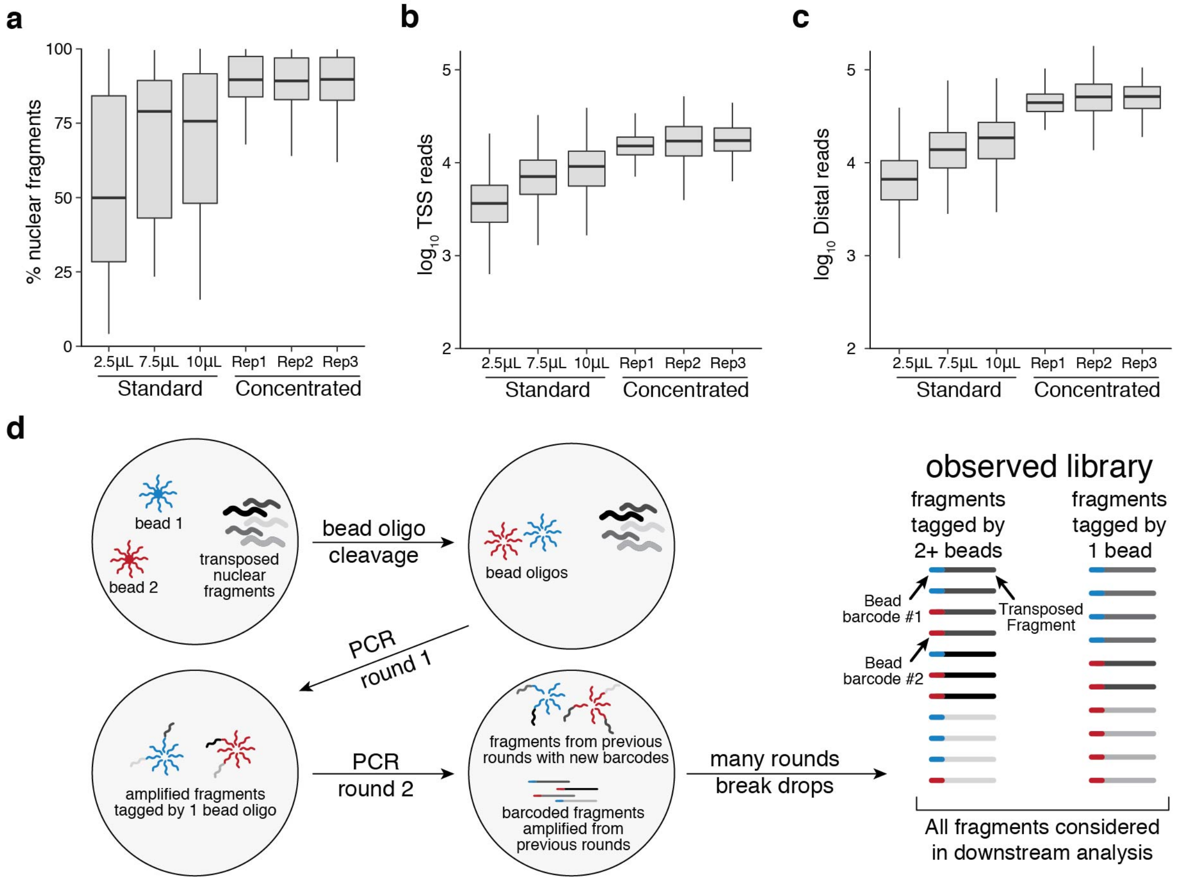

Figure C.1: Optimization of Tn5 transposition for dscATAC-seq. **(a)** Fraction of reads mapping to the nuclear genome for each of the Tn5 concentrations. The remaining reads map to the mitochondrial genome. Different volumes (2.5-10 µL) of the standard commercial Tn5 (TDE1) are compared against 3 replicates of a custom Tn5 concentration (2.5 µL) optimized for dscATAC-seq for K562 cells. **(b)** Number of unique reads mapping near transcription start sites (TSS) or **(c)** distal regulatory elements for the same Tn5 conditions. Center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. All three panels **(a-c)** show the top 500 cells sorted by library size. **(d)** Schematic of biochemical process leading to multiple fragments becoming tagged by multiple bead barcodes in the same droplet.

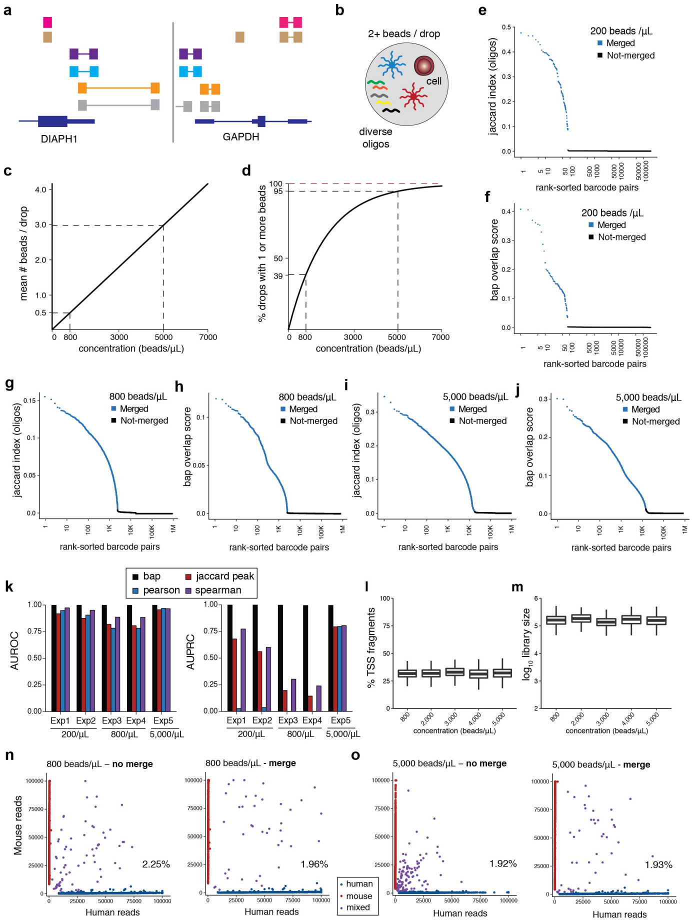

Figure C.2: Validation of bead merging computational approach. **(a)** Browser shot of paired-end reads near the DIAPH1 and GAPDH loci. Reads are colored by bead barcode sequence. **(b)** Schematic of verification experiment where a library of random oligonucleotides was encapsulated into droplets together with Tn5 transposed cells and barcoded beads. The schematic shows a droplet containing a library of random oligos, a cell and two beads with different barcode sequences. **(c)** The expected number of beads per drop as a function of bead concentration. Inference of this line was determined by a maximum likelihood estimation for a double-truncated Poisson distribution. **(d)** Percent of drops with one or more beads as a function of bead concentration. Values are estimated using the probability density function of a Poisson distribution parameterized by the mean number of beads per drop from **(c)**. **(e)** Jaccard index overlap metric for pairs of bead barcodes loaded at a concentration of 200 beads/µL. For each pair of bead barcodes observed, the Jaccard index was computed over the observed random oligonucleotide sequences. **(f)** The bap overlap score computed from the dscATAC-seq data (agnostic to oligonucleotides) from the same experiment. In each panel, pairs of bead barcodes nominated for merging are highlighted in blue. Merged pairs were determined by computing a "knee" inflection point. The same two panels are shown in **(g-j)** but for increased bead concentration: **(g,h)** 800 beads/µL; **(i,j)** 5,000 beads/µL. **(k)** (left panel) Area under the receiver operating curve (AUROC) values for true positive bead merges nominated from the random oligonucleotide sequences. Four metrics are compared, including our novel computational approach, termed bap. Various bead concentrations per experimental condition are shown below the x-axis. (right panel) The same conditions and metrics but showing the area under the precision-recall curve (AUPRC). **(l)** %TSS enrichment scores for the same pool of cells processed at different bead concentrations. **(m)** Per-cell library complexities across a range of tested bead concentrations, the same as in panel **(l)**. Both panels **(l,m)** show the top 500 cells sorted by library size. **(n)** Species mixing plots and collision rates (text) for the same experiment (800 beads/µL) with and without bead merging. **(o)** The same plots as in **(n)** but at a bead concentration of 5,000 beads/µL.

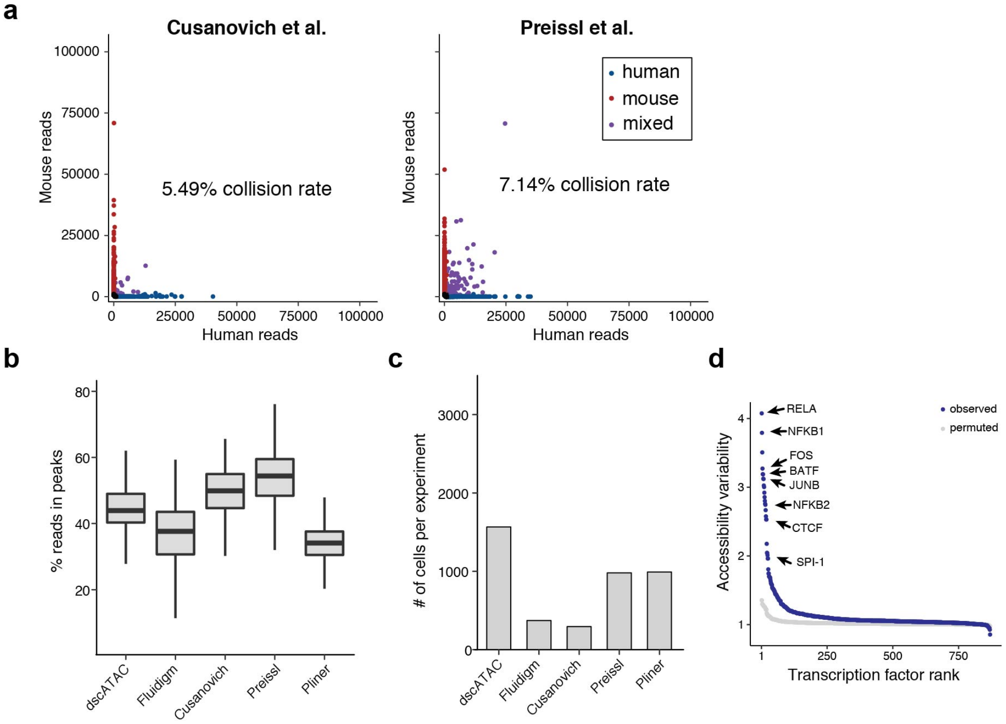

Figure C.3: Additional quality controls of dscATAC-seq. (a) Species mixing plots and estimated collision rates for existing scATAC-seq methods. **(b)** Fraction of reads in peaks for the comparison in Figure 3.1.f. The chromatin accessibility peak set was obtained from ENCODE DNase-seq data for GM12878 and thus agnostic to the datasets compared here. Center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. **(c)** Number of cells (GM12878 only) compared in panel **(b)** and Figure 3.1.f. **(d)** Rank sorted variability across transcription factor motifs within the GM12878 dscATAC-seq profiles.

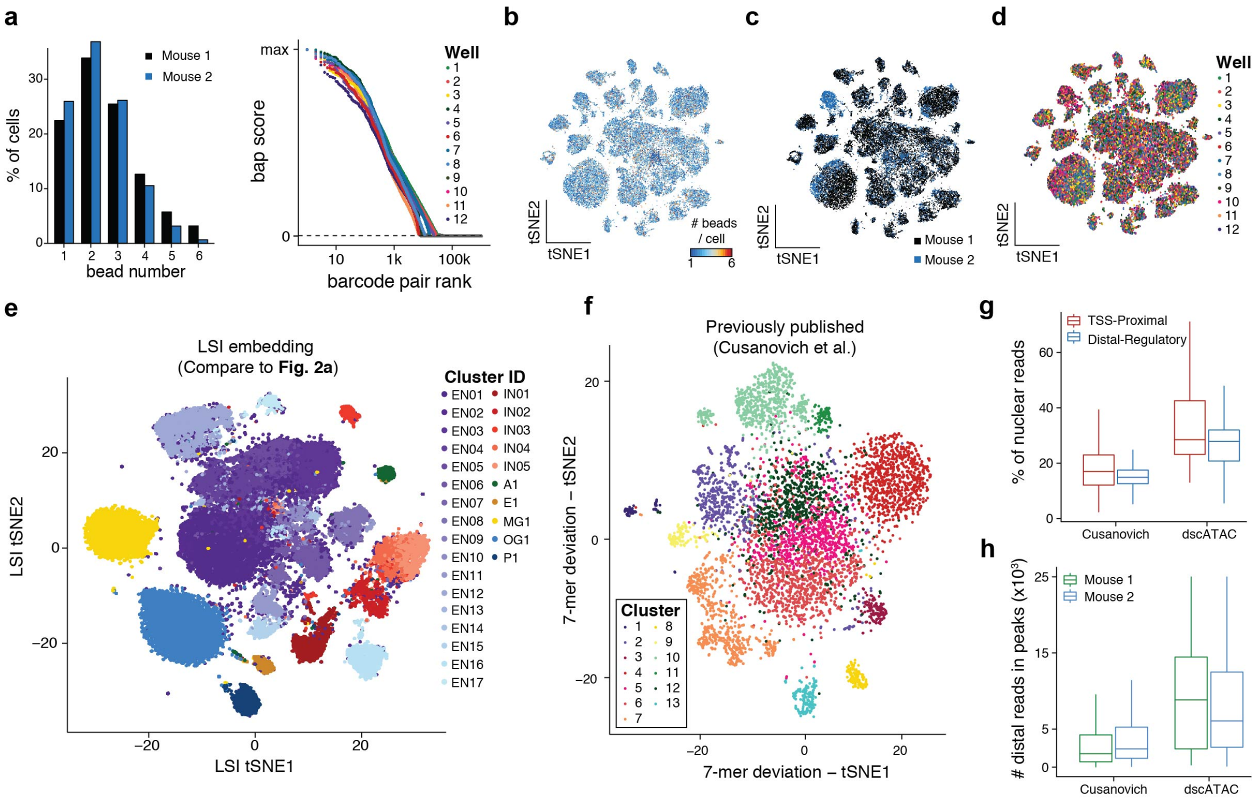

Figure C.4: Quality control information for the dscATAC-seq mouse brain and comparison with existing data. (a) Distribution of number of beads per cell identified across the two mice (bead input concentration = 5,000 beads/µL) for high-quality cells that pass quality controls. The corresponding bead merging curves are shown to the right for the twelve libraries. **(b)** Mouse brain cells in the t-SNE from Figure 3.2.a colored by number of bead barcodes detected per cell. The same coordinates are shown for **(c)** mouse donor, and **(d)** experimental well. **(e)** *de novo* embedding using latent semantic indexing (LSI). Colors match annotations from Figure 3.2.a. All plots show the same (n=46,653) cells shown in Figure 3.2.a. **(f)** t-SNE of previously published sciATAC-seq data for mouse brain (Cusanovich et al., 2018) using the same 7-mer method (Louvain, t-SNE; compare to Figure 3.2.a; n=5,744 cells). **(g)** Comparison of the percentage of reads mapping to the nuclear genome (separated into TSS-proximal or distal chromatin accessibility peaks) between whole mouse brain data generated using dscATAC-seq or a recently optimized sciATAC-seq method (Cusanovich et al., 2018). Center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. (h) Raw total number of reads mapping to distal chromatin accessibility peaks (see blue from panel **(g)** between dscATAC-seq and the sciATAC-seq method described in **(g)**). Boxplots summarize thousands of cells for each comparison. Center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range.

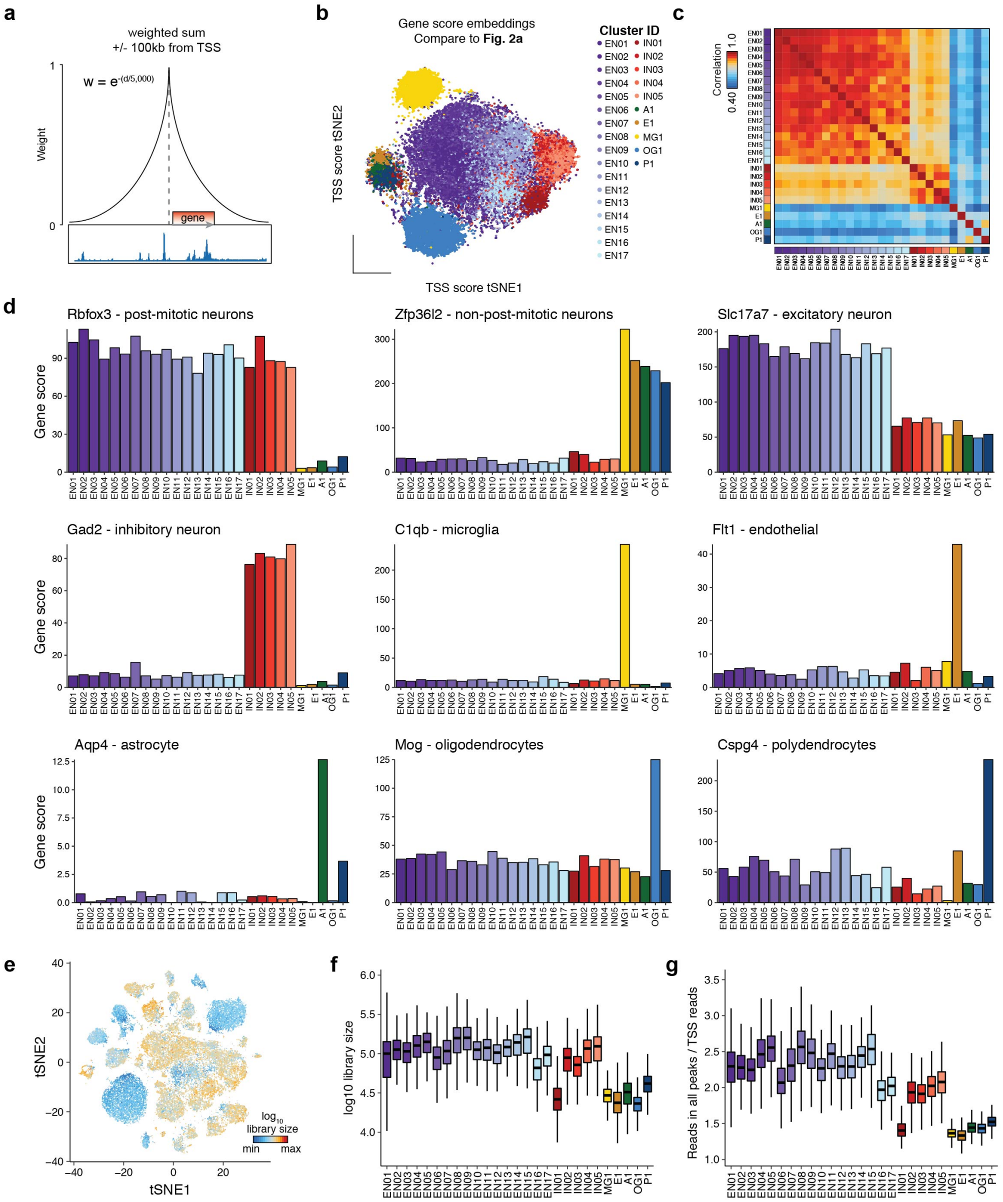

Figure C.5: Chromatin accessibility scores for validation of cell clusters from mouse brain. **(a)** Schematic demonstrating the approach used to define chromatin accessibility scores surrounding gene promoters. **(b)** t-SNE of cells by promoter region chromatin accessibility scores for all genes. The same colors and cells (n=46,653) used in Figure 3.2.a are shown here. **(c)** Hierarchical clustering of chromatin accessibility scores calculated as shown in **(a)** for each cluster derived from the mouse brain dscATAC-seq dataset using Pearson correlation. 27 clusters from Figure 3.2.a are depicted. **(d)** Representative chromatin accessibility scores for known marker genes defining cell types in the mouse brain, plots are titled by the marker gene and defined cell type. **(e)** Mouse brain cells in the t-SNE from Figure 3.2.a colored by per-cell log10 library complexity (n=46,653 cells). **(f)** Per-cell log10 library complexity for each cluster derived from the mouse brain dscATAC-seq dataset. Center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. **(g)** Per-cell ratio of total reads in peaks to TSS reads per cluster.

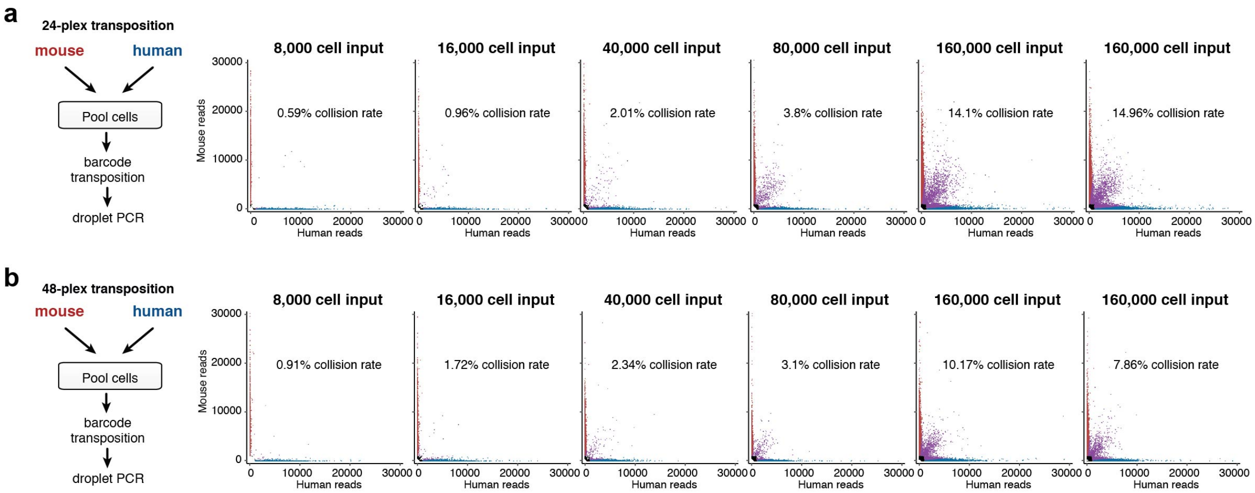

Figure C.6: Species mixing analysis of dsciATAC-seq. (a-b) Species mixing analysis for human (K562) and mouse (3T3) cell mix generated using **(a)** 24 or **(b)** 48 Tn5 transposase barcodes. For each panel a schematic of the experimental procedure is included (left), and primary results from a cell titration plotting total mouse or human nuclear fragments (right). In these plots points are labeled as either low quality (black), mouse (red), human (blue) or mixed (purple).

C.1.3 Human peripheral blood and bone marrow cells

Cryopreserved human bone marrow (BM) mononuclear cells, isolated BM CD34+ stem/progenitor cells, peripheral blood mononuclear cells (PBMC), and isolated peripheral blood CD4+, CD8+, CD14+, CD19+ and CD56+ cells were purchased from Allcells. Cells were quickly thawed in a 37°C water bath, rinsed with culture medium (IMDM medium supplemented with 10% FBS and 1% Pen/Strep) and then treated with 0.2 U/μL DNase I (Thermo Fisher Scientific) in 10 mL of culture medium at 37°C for 30 min. After DNase I treatment, cells were washed with medium once and then twice with ice cold 1x PBS + 0.1% BSA. Cells were then filtered with a 35 μm cell strainer (Corning) and cell viability and concentration were measured with trypan blue on the TC20 Automated Cell Counter (Bio-Rad). Cell viability was greater than 80% for all samples.

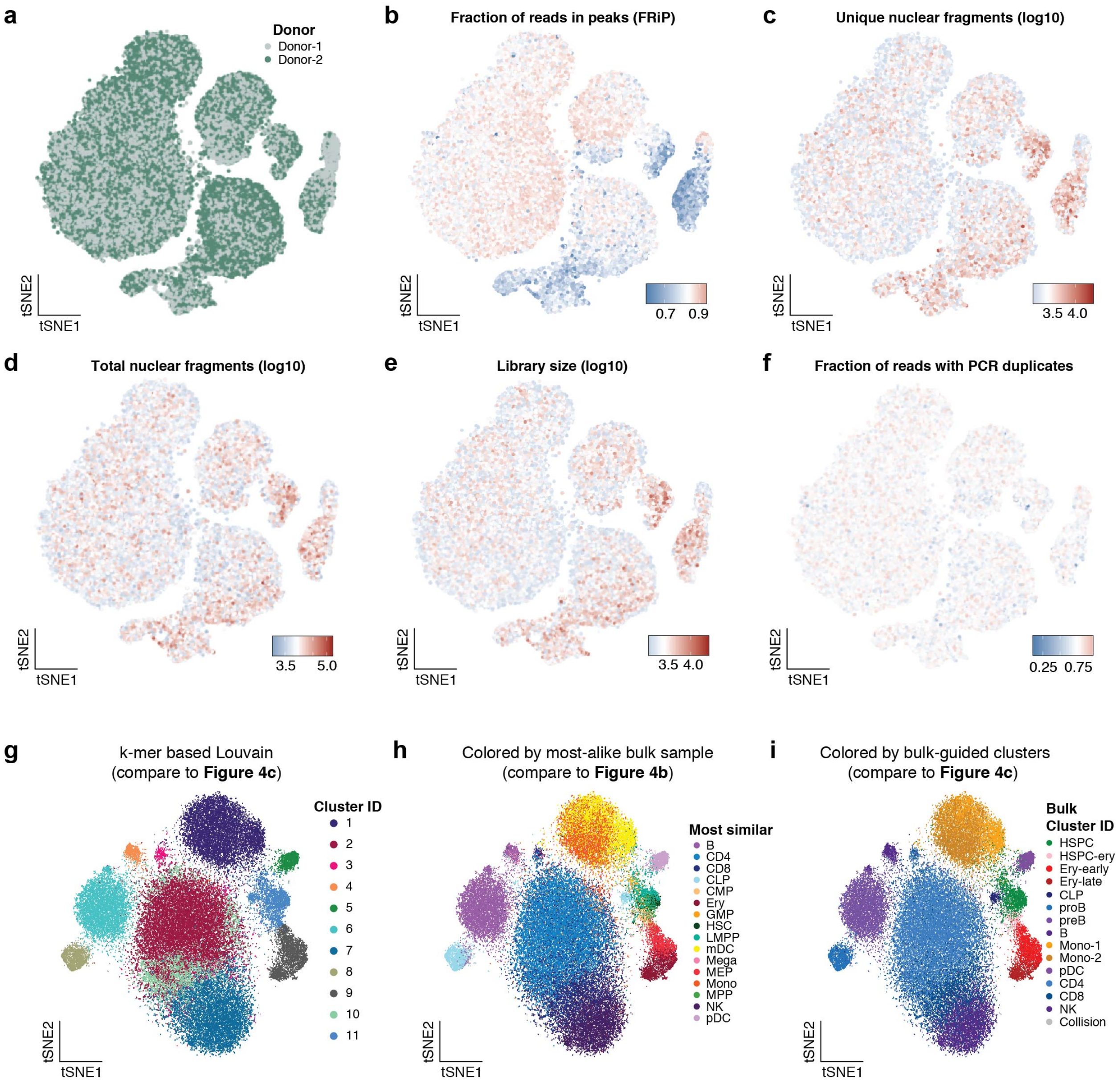

Figure C.7: Quality control analysis of human bone marrow dsciATAC-seq data. **(a-f)** Single-cell data derived from BMMCs colored by their **(a)** donor, **(b)** fraction of reads in peaks (FRiP), **(c)** log10 unique nuclear fragments, **(d)** log10 total aligned nuclear fragments, **(e)** log10 library size, and **(f)** fraction of reads with PCR duplicates. **(g)** *de novo* embedding and clustering of the human BMMC data using the 7-mer k-mer strategy. Colors represent Louvain clustering from the principal components of the 7-mer deviations. **(h,i)** Same coordinates as **(g)** but colored according to annotations defined in Figure 3.4.b,c, respectively. All panels show n=60,495 cells.

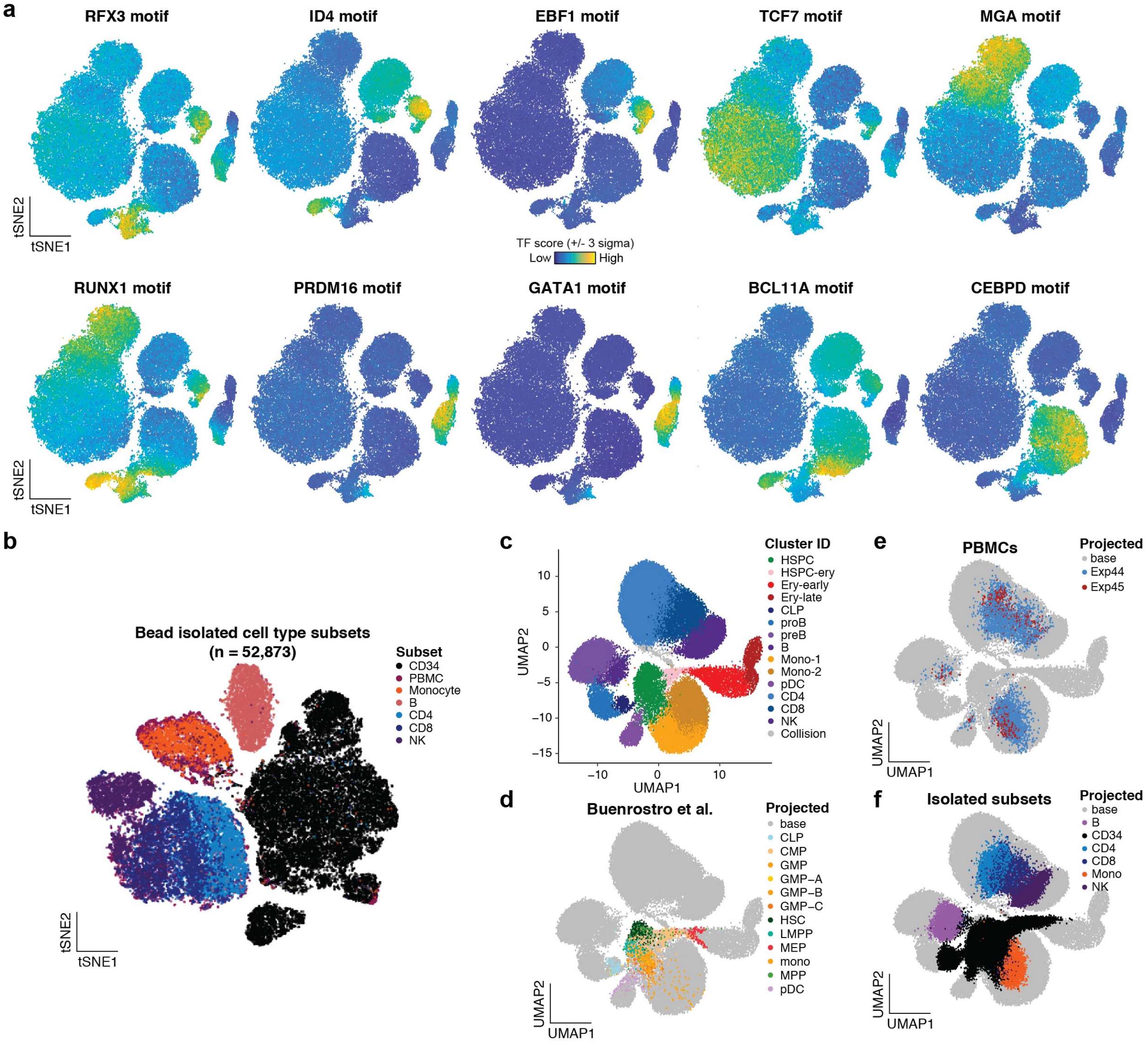

Figure C.8: Cell types identified in the human bone marrow dsciATAC-seq data. (a) Selected transcription factor deviation motifs shown for resting cells (n=60,495 cells) profiled using dsciATAC-seq. **(b)** Embedded cells from isolated subtypes profiled using the standard dscATAC-seq platform (n=52,873 cells). **(c)** UMAP embedding of single-cell data colored by clusters identified (compare to Figure 3.4.c). **(d-f)** Projection of additional single-cell data onto UMAP coordinates of the dsciATAC-seq bone marrow data, projecting **(d)** sorted progenitor subsets (Buenrostro et al., 2018), **(e)** peripheral blood mononuclear cells (PBMCs) or **(f)** isolated subsets (shown individually in **(b)**).

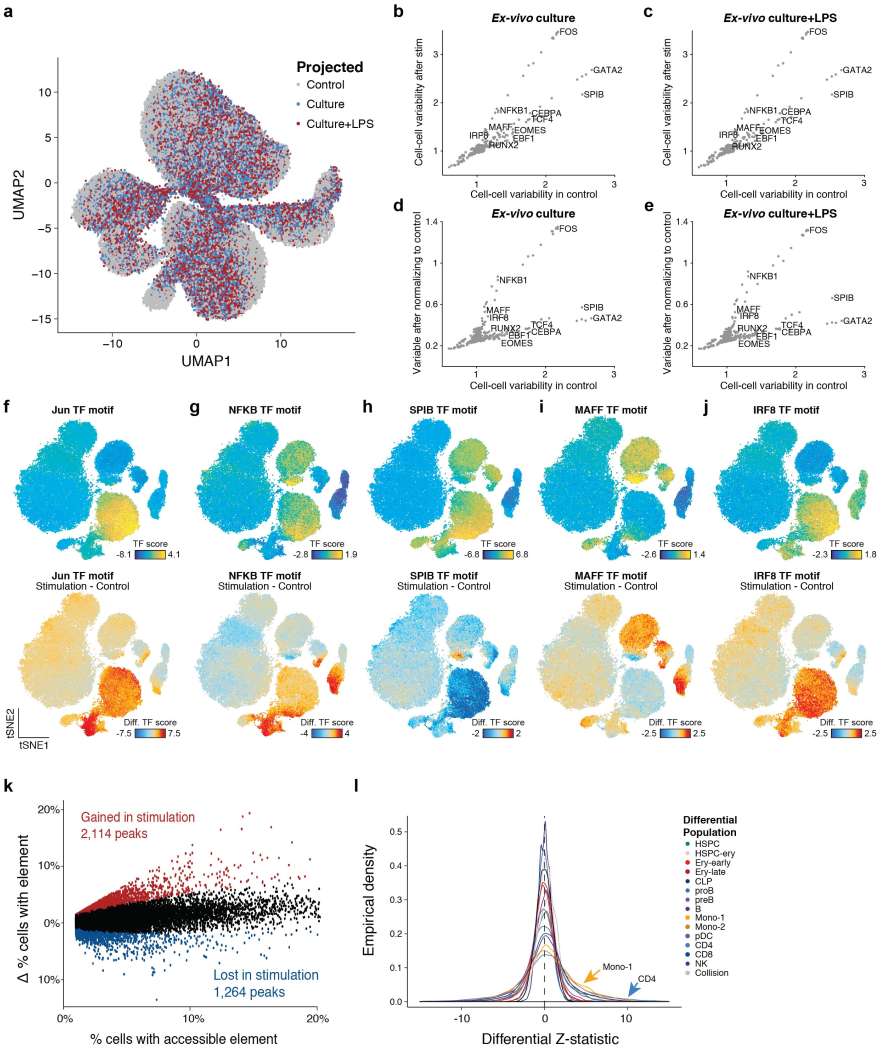

Figure C.9: Stimulation of human bone marrow derived cells. (a) Stimulated BMMC (n=75,968) cells projected onto the UMAP coordinates defined by the non-stimulated control cells (n=60,495 cells). **(b,c)** Cell-cell TF score variability for the stimulation and control cells showing **(b)** *ex vivo* culture and **(c)** *ex vivo* culture and LPS stimulation, only unique TF motifs are highlighted. **(d,e)** Cell-cell TF score variability for the control cells and variability of stimulation after normalizing to the control TF variability for **(d)** *ex vivo* culture and **(e)** *ex vivo* culture and LPS stimulation conditions, only unique TF motifs are highlighted. **(f-j)** Depictions of transcription factor deviation scores in resting cells (top) compared to the differential after stimulation (bottom) for selected motifs. A total of n=60,495 cells are plotted. **(k)** Sample summary of differential peak analysis for the Mono-1 cluster. Each dot represents a chromatin accessibility peak found in at least 1% of cells. The overall % of cells with element are shown on the x-axis whereas the y-axis depicts the difference in the % of cells with the element accessible (stimulated - resting). Peaks found significantly different at a 1% FDR (two-sided binomial test; Benjamini Hochberg corrected) are colored in red and blue. **(l)** Overall summary statistics per-population from differential peak analysis showing the Z-statistic from the two-sided permutation test for differential accessibility. Each colored curve represents the overall Z-statistics for all peaks in the specified cluster.

C.1.4 Human bone marrow mononuclear cells stimulations

BM mononuclear cells were quickly thawed in a 37°C water bath, rinsed with culture medium (RPMI 1640 medium supplemented with 15% FBS and 1% Pen/Strep) and then treated with 0.2 U/μL DNase I in 10 mL of culture medium at 37°C for 30 min. After DNase I treatment, cells were washed with medium once, filtered with a 35 μm cell strainer and cell viability and concentration were measured with trypan blue on the TC20 Automated Cell Counter. Cell viability was greater than 90% for all samples. Cells were plated at a concentration of 1 x 10^6 cell/mL, rested at 37°C and 5% CO2 for 1 h and then either incubated in serum containing media (RPMI 1640 medium supplemented with 15% FBS and 1% Pen/Strep) at 37°C and 5% CO2 for 6 h (*ex vivo* culture) or treated with 20 ng/mL Lipopolysaccharide (LPS) (tlrl-3pelps, Invivogen) for 6 h (LPS stimulation). After stimulation, cells were washed twice with ice cold 1x PBS + 0.1% BSA and cell viability and concentration were measured with trypan blue on the TC20 Automated Cell Counter. As a control, we processed cells immediately after counting, without any incubation.

C.1.5 Cell lysis and tagmentation

For a detailed description of tagmentation protocols and buffer formulations refer to the SureCell ATAC-Seq Library Prep Kit User Guide (17004620, Bio-Rad). Harvested cells and tagmentation related buffers were chilled on ice. For cell lines, a protocol based on Omni-ATAC was followed(Corces et al., 2017). Briefly, washed and pelleted cells were lysed with the Omni-ATAC lysis buffer containing 0.1% NP-40, 0.1% Tween-20, 0.01% Digitonin, 10 mM NaCl, 3 mM MgCl2, and 10 mM Tris-HCl pH7.4 for 3 min on ice. The lysis buffer was diluted with ATAC-Tween buffer that only contains 0.1% Tween-20 as a detergent. Cells were collected and resuspended in OMNI Tagmentation Mix. This mix is formulated with ATAC Tagmentation Buffer and ATAC Tagmentation Enzyme, both of which are included in the SureCell ATAC-Seq Library Prep Kit (17004620, Bio-Rad). The OMNI Tagmentation Mix was buffered with 1X PBS supplemented with 0.1% BSA. Cells were mixed and agitated on a ThermoMixer (5382000023, Eppendorf) for 30 min at 37°C. Tagmented cells were kept on ice prior to encapsulation.

For PBMCs and BM mononuclear cells, lysis was performed simultaneously with tagmentation. Washed and pelleted cells were resuspended in Whole Cell Tagmentation Mix containing 0.1% Tween-20, 0.01% Digitonin, 1X PBS supplemented with 0.1% BSA, ATAC Tagmentation Buffer and ATAC Tagmentation Enzyme. Cells were tagmented using a thermal protocol and maintained thereafter as described in the Omni-ATAC protocol described above.

For mouse tissues, nuclei were washed with ATAC-Tween buffer containing 0.1% Tween-20, 10 mM NaCl, 3 mM MgCl2, and 10 mM Tris-HCl pH7.4 prior to the whole cell protocol described above.

C.2 Methods for dscATAC-seq/dsciATAC-seq

C.2.1 Optimized Tn5 concentration

To test if the concentrated Tn5 (part of the SureCell ATAC-Seq Library Prep Kit, 17004620, Bio-Rad) performed better than the standard commercial Tn5 enzyme (TDE1, 15027865, Illumina), we prepared dscATAC-seq libraries for K562 cells using different amounts of TDE1 and our new concentrated Tn5. K562 cells were prepared and lysed as described in the Omni-ATAC protocol described above. Cells were then resuspended in OMNI Tagmentation Mix containing ATAC Tagmentation Buffer and either 1) different amounts of TDE1 (2.5, 7.5 or 10μL in a 50μL reaction, see Figure 3.1.b) or 2) the concentrated Tn5 (2.5μL in a 50μL reaction, 3 replicates, see Figure 3.1.b). Cells were mixed and agitated on a ThermoMixer for 30 min at 37°C. Tagmented cells were kept on ice prior to encapsulation and libraries were prepared using our standard method as described below. The top 500 cells based on library complexity are shown for all comparisons.

C.2.2 Droplet library preparation and sequencing

For a detailed protocol and complete formulations, refer to the SureCell ATAC-Seq Library Prep Kit User Guide (17004620, Bio-Rad). Tagmented cells or nuclei were loaded onto a ddSEQ Single-Cell Isolator (12004336, Bio-

Rad). Single-cell ATAC-seq libraries were prepared using the SureCell ATAC-Seq Library Prep Kit (17004620, Bio-Rad) and SureCell ddSEQ Index Kit (12009360, Bio-Rad). Bead barcoding and sample indexing were performed in a C1000 Touch™ Thermal cycler with a 96-Deep Well Reaction Module (1851197, Bio-Rad): 37°C for 30 min, 85°C for 10 min, 72°C for 5 min, 98°C for 30 sec, 8 cycles of 98°C for 10 sec, 55°C for 30 sec, and 72°C for 60 sec, and a single 72°C extension for 5 min to finish. Emulsions were broken and products cleaned up using Ampure XP beads (A63880, Beckman Coulter). Barcoded amplicons were further amplified using a C1000 Touch™ Thermal cycler with a 96-Deep Well Reaction Module: 98°C for 30 sec, 6-9 cycles (cycle number depending on the cell input, Section 4 Table 3 of the User Guide) of 98°C for 10 sec, 55°C for 30 sec, and 72°C for 60 sec, and a single 72°C extension for 5 min to finish. PCR products were purified using Ampure XP beads and quantified on an Agilent Bioanalyzer (G2939BA, Agilent) using the High-Sensitivity DNA kit (5067-4626, Agilent). Libraries were loaded at 1.5 pM on a NextSeq 550 (SY-415-1002, Illumina) using the NextSeq High Output Kit (150 cycles; 20024907, Illumina) and sequencing was performed using the following read protocol: Read 1 118 cycles, i7 index read 8 cycles, and Read 2 40 cycles. A custom sequencing primer is required for Read 1 (16005986, Bio-Rad; included in the kit).

C.2.3 Assembly of indexed Tn5 transposome complexes

To generate indexed Tn5 transposome complexes, we modified the Illumina Nextera Read 1 Adapter to contain a 6 nt barcode (96 distinct barcodes). Each indexed oligo was mixed with the Illumina Nextera Read 2 Adapter and annealed to a 15 nt mosaic end complementary oligonucleotide (5' phosphorylated and 3' Dideoxy-C). All oligonucleotides were HPLC purified (IDT). For the annealing reaction, oligonucleotides were mixed at a 1:1:2 molar ratio (Read 1: Read 2: complementary mosaic end) at 100 μM final concentration in 50mM NaCl. The mixture was incubated at 85°C, ramped down to 20°C at a rate of -1°C/min, and then 20°C for 2 additional minutes. After being diluted 1:1 in glycerol, the annealed oligonucleotide mixture was then mixed 1:1 with 14.8 μM purified Tn5. The Tn5/oligonucleotide mixture was incubated for 30 min at room temperature and then kept at -20°C prior to the tagmentation reactions.

C.2.4 Species mixing controls

Human and mouse cell lines were processed and lysed using the Omni-ATAC-seq protocol as described above. For the 24-plex control experiment in Figure 3.3 and C.6, K562 and NIH/3T3 cells were mixed at a 1:1 ratio and tagmented with Tn5 loaded with indexed oligonucleotides 1-3, 13-15, 25-27, 37-39, 49-51, 61-63, 73-75, 85-87 in 50 µL reactions (10 µL of indexed Tn5 per reaction) with 25,000 cells each. Cell line tagmentation buffer components and reaction conditions were the same as described above. After the tagmentation reaction, all cells were pooled, washed with tagmentation buffer without Tn5 and processed using our standard protocol for droplet library preparation and sequencing. Different cell numbers were used as input, as indicated in Figure 3.3.

For the 48-plex control experiment in Figure C.6, K562 and NIH/3T3 cells were mixed at a 1:1 ratio and tagmented with Tn5 loaded with indexed oligonucleotides 1-6, 13-18, 25-30, 37-42, 49-54, 61-66, 73-78, 85-90 in 50 µL reactions (10 µL of indexed Tn5 per reaction) with 25,000 cells each. Cell line tagmentation buffer components and reaction conditions were the same as described above. After the tagmentation reaction, all cells were pooled, washed with tagmentation buffer without Tn5 and processed using our standard protocol for droplet library preparation and sequencing. Different cell numbers were used as input, as indicated in Figure C.6.

C.2.5 Human BM mononuclear cells stimulations

BM-MNCs from 2 donors were stimulated and washed as described above. For the experiment in Figures 4 and 5, BM-MNCs were tagmented with Tn5 loaded with indexed oligonucleotides 1-96 in 20 µL reactions (4 µL of indexed Tn5 per reaction) with 8,000 cells each (Control, *ex vivo* culture and LPS stimulation as described above). BM-MNC tagmentation buffer components and reaction conditions were the same as described above. After the tagmentation reaction, all cells were pooled, washed with tagmentation buffer without Tn5 and processed using our standard protocol for droplet library preparation and sequencing. Pooled cells were split into 16 different samples for droplet library preparation, with varying cells inputs (20,000, 40,000 or 80,000 cells). After sequencing, data from all 16 samples were merged for the analyses. Sequencing data for the dsciATAC-seq experiments were processed with bap as described below using the "–tn5-aware" flag that inhibits cell merging across different Tn5 barcodes.

C.3 Bioinformatics methods for sequencing data analysis

C.3.1 Raw read processing

Per-read bead barcodes were parsed and trimmed using UMI-TOOLs (Smith et al., 2017), and the remaining read fragments were aligned using BWA (http://bio-bwa.sourceforge.net/) on the Illumina BaseSpace online application. Constitutive elements of the bead barcodes were assigned to the closest known sequence allowing for up to 1 mismatch per 6-mer or 7-mer (mean >99% parsing efficiency across experiments). For the dsciATAC-seq experiments, bead barcodes were parsed using a custom python script aware of the 96 possible Tn5 barcodes. All experiments were aligned to the hg19 or mm10 reference genomes (or a combined reference genome in the case of species mixing experiments).

To identify systematic biases (*i.e.* reads aligning to an inordinately large number of barcodes), barcode-aware deduplicate reads, and perform bead merging (see below), we developed a software suite called the bead-based ATAC-seq processing (bap) tool. This software uses as input a `.bam` file for a given experiment with a bead barcode identifier indicated by a SAM tag. We generalized this pre-processing pipeline to handle other datasets (Fluidigm C1, sciATAC-seq) to enable consistent comparisons across various technologies (Figure 3.1).

C.3.2 Identification of multiple beads per droplet

An integral part of the technique described herein relies on the robust identification of pairs of bead barcodes that share exact insertions at a rate that exceeds what may be expected by chance. We note that our procedure readily enables multiple beads per droplet (Figure C.2). First, highly abundant barcodes are detected in the experiment wherein each unique barcode sequence is quantified among nuclear-mapping reads, and our knee calling algorithm establishes a per-experiment bead threshold. Next, sequencing reads assigned to a bead barcode passing filter are de-duplicated using the insert positions of the paired-end reads (as previously implemented in Picard tools).

After initial deduplication, we further remove paired-end reads that map to more than 6 bead barcodes, reasoning that these represent a technical confounder. Next, for each pair of bead barcodes passing the bead filtering step, we compute the Jaccard index over the insertion positions of reads, providing a measure of how similar the Tn5 insertions are between any pair of bead barcodes. From these pairwise Jaccard index statistics, we perform a second knee call to determine pairs likely to have originated from the same droplet (Figure C.2.d). Finally, to assign droplet-level barcodes, we then loop over the original bead barcodes in order of their original nuclear read abundance. For a given bead barcode, if it is paired with any other bead barcodes that passed the pairwise knee, those bead barcodes are "merged" into one droplet barcode. This iteration repeats until all bead barcodes have been assigned to precisely one droplet barcode. To facilitate comparisons without droplet merging (*e.g.* Figure C.2.j,k), our pipeline facilitates the "–one-to-one" flag, which maps one bead barcode onto one droplet barcode; this option was employed primarily to process other scATAC-seq datasets that would not have beads that would require merging. Additional details regarding this procedure and comparisons in Figure C.2.g are discussed in the final paragraphs of this document.

C.3.3 Species mixing analysis

We carried out the same quantification procedure for all species mixing datasets analyzed in this work. Namely, reads were mapped to a hybrid hg19-mm10 reference genome using BWA. Cells were identified using the `bap` knee calling described above. The output of this pipeline yields the number of unique nuclear reads mapping to the mouse and human genomes, which were compared per-cell. We further excluded cells with less than 1,000 reads mapping to either the human or mouse genomes and identified collisions as those that had less than a 10x enrichment over the minor genome. The overall collision rate is reported as the number of annotated collision cells over the total number of cells compared (mouse + human + collisions).

C.3.4 Peak calling

For each scATAC-seq experimental sample, chromatin accessible summits were called using MACS2 callpeak with custom parameters previously described (Corces et al., 2016). To generate a non-overlapping set of peaks per analysis, we first extended summits of each experiment to 500 bp windows (+/- 250 bp). We combined these 500 bp peaks, ranked them by their summit significance value, and retained specific non-overlapping peaks based on this ordering. We further removed peaks that overlapped the ENCODE blacklist and a custom mitochondrial blacklist generated by aligning a synthetic mtDNA genome to the nuclear genome.

C.3.5 Library complexity estimation

Per-cell library complexities were estimated using the Lander-Waterman equation (Lander & Waterman, 1988) using a custom R function translated from a previously established Java function implemented in Picard tools. Per-cell counts of total number of mapped nuclear reads passing quality filters and the number of unique nuclear reads served as inputs. The library complexity thus represents a metric that estimates the total number of unique nuclear reads given by the cell independent of sequencing depth.

C.3.6 Comparison to public datasets

To benchmark the dscATAC-seq platform against existing datasets, we downloaded raw sequencing data (.fastq format) for GM12878 cells via three different combinatorial indexing scATAC-seq methods19,24,25 and 384 cells processed with the Fluidigm C1 (Buenrostro et al., 2015) from GEO. All dataset were processed using the same pipeline, which included BWA alignment and downstream processing with `bap` using the "–one-to-one" flag that skips bead merging. We note that in all three combinatorial indexing scATAC-seq experiments, GM12878 cells were mixed with mouse cells. As such, we compared only annotated human cells (>9:1 ratio of human to mouse reads) from these experiments for downstream analysis.

To determine the correlation between single-cell ATAC-seq experiments, we used a merged peak set comprising of 175,581 combined DNase-seq hypersensitivity peaks from GM12878 and K562 made available through the ENCODE Project. The sum of single cells (agnostic to cell ID) were compared against bulk Dnase-seq profiles generated from ENCODE and Omni-ATAC (Corces et al., 2017). To score the fraction of reads in peaks across single cell experiments, we used only the GM12878 DNase-seq peak set (124,321 peaks) to ensure that peak selection did not bias our quantification and comparison of technologies.

C.3.7 Validation of multiple beads per droplet inference

To validate our ability to merge cells marked by multiple droplet beads, we introduced a diverse library of random oligonucleotides (14 nucleotides random region) to our microfluidic reaction (Figure C.2). Human PBMCs were processed with this library of random oligos at bead concentrations of 200, 800, and 5,000 beads/μL, spanning the ranges used for the data presented in this work. The random oligonucleotides were spiked in to the cells at a final concentration of 5 nM after the tagmentation reaction, and samples were processed and sequenced using our standard protocol (described above). Among pairs of beads merged, the average number of oligos observed per bead ranged from 792-1,979 per experiment.

We reasoned that bead barcodes sharing a noticeable overlap of these oligos (Figure C.2.a,b) would be barcodes from two beads contained in the same droplet. We identified reads containing our random oligo by first identifying the 15 bp constant sequence and subsequently parsing the 14 bases downstream of the constant sequence. For each experiment, we called a knee on the bead barcode pairwise Jaccard indices (per observed 14 base oligonucleotide) and computed the overlap of random sequences observed (Figure C.2.c) for barcodes passing the nuclear read knee. For pairs of bead barcodes passing the oligo overlap knee, we annotated these as true positives.

Next, we computed our `bap` metric pairwise for each bead barcode using the overlap of pairs of inserts over each fragment (or paired-end read). This produces a metric for all pairs of bead barcodes with at least 500 unique nuclear reads observed per barcode (Figure C.2.d). Using the true-positives defined from the random oligos data and a continuous overlap metric from `bap`, we computed precision-recall and receiver operating curves (mean area under the receiver-operating curve (AUROC) = 1.000 and mean area under the precision recall curve

(AUPRC) = 0.997 (Figure C.2.d). We further compared other possible metrics for bead merging, including Pearson and Spearman correlation and a Jaccard index over reads in peaks, finding that our approach was the most robust and specific (Figure C.2.g). We note that the library of random oligonucleotides provides a completely orthogonal measure of bead overlap compared to the nuclear DNA fragments used in the `bap` algorithm.

C.4 Theory of beads and droplet concentrations

In this setting, we are interested in estimating the number of beads per droplet at variable bead concentrations using observed data. Given that our observed data does not yield any droplets with zero beads (cells not captured) nor can any measurement be relied on with greater than 6 beads (physical limit for bead; observed values likely reflect merged droplets), the observed number of beads per droplet is modeled by a double-truncated Poisson distribution. The probability density function of a double-truncated Poisson distribution for a single observation can be written as follows:

$$\Pr\left(Y_i = y_i \,|c_1 \leq y_i \ \leq c_2\right) = \frac{\lambda^{y_i}}{y_i\,!\ \sum_{k=c_1}^{c_2} \lambda^{y_i} /k!}$$

Here, c_1 is our lower bound (1 in our case) of the empirical data and c_2 is the upper bound (in our case 6) for observed numbers of beads / droplet y. Let $i \ \in 1, 2, \ldots, n$. Then, we observe n cells and y_i denotes the number of beads per drop for cell i. The log likelihood (l) of observing a value can thus be computed as follows:

$$l\left(\lambda|y\right) = \sum_{i=1}^{n} y_i \ \log(\lambda) - \ \sum_{i=1}^{n} \log\left(y_i\,!\right) - \log\left(\sum_{k=c_1}^{c_2} \frac{\lambda^k}{k!}\right)$$

Here, a closed form solution of λ (parameter of the Poisson distribution indicating the mean number of beads per cell) is impossible. Thus, we estimate the value using the `optim()` function in R, providing the maximum likelihood estimate (MLE).

Given the MLE estimate for λ, from plugging into the Poisson PDF, we can trivially compute:

$$p = \exp\{-\lambda\}$$

where p is the proportion of droplets with 0 beads. We can then approximate the number of droplets with a barcode as $1 - p$. Empirical values of λ were determined using GM12878 and mouse brain data at different bead concentrations (800 and 5,000 beads/μL) and were found to be robust across the various datasets analyzed throughout.

C.5 Bioinformatics methods for single-cell data

C.5.1 *de novo* K-mer clustering

Here, we computed bias-corrected deviation z-scores for K k-mers and a set of S samples (dscATAC-seq cells) with P peaks computed via the chromVAR methodology. Here, our implementation utilizes a binarized matrix M (dimension P by K) where $m_{i,k}$ is 1 if k-mer k is present in peak i and 0 otherwise based on the reference genome annotation. For all applications, we used k = 7, resulting in K = 8,192 ($4^7/2$) 7-mers. We note that the division by 2 is to account for reverse-complement k-mers that would be identical as both strands of the reference genome are considered when building M. Using the matrix of fragment counts in peaks X (dimension P by S), where $x_{i,j}$ represents the number of fragments from peak i in sample j, we produce a deviation score matrix Z of dimension S samples (rows) and K 7-mers (columns).

The matrix Z is computed using an expectation of peak accessibility based on technical confounders present in assays (differential PCR amplification or variable Tn5 tagmentation conditions). This is achieved by generating 50 background peaks intrinsic to the set of epigenetic data examined. The full details describing the computation of Z have been previously described in the chromVAR (Schep et al., 2017) manuscript. Finally, as any of the 8,192 7-mers are highly correlated, we then use the top principal components of the matrix Z as input for downstream processes, including the Louvain clustering and t-SNE embedding.

C.5.2 Cell type specific promoter region chromatin accessibility scores and regulatory region analysis in mouse brain

To define cluster-specific regulatory elements and promoter region chromatin accessibility scores, we defined pseudo-bulk cell types by aggregating the counts per cell over each of the annotated cluster definitions. First, the peak x cell type counts matrix (*X*) was count-per-million (CPM) normalized, and peaks with an overall mean CPM > 1 were retained. This filtered peak x cell type matrix was then z-score transformed. Explicitly, for cell type *j* and peak *i*, our transformed statistic was:

$$z_{i,j} = \frac{x_{i,j} - mean(x_{i,*})}{sd(x_{i,*})}$$

We identified 135,737 cell type-specific chromatin accessibility peaks with a $z_{i,j} > 3$ in at least one cell type (some value j), which were assigned to clusters based on the maximum z-score value ($\text{argmax}_j\, z_{i,j}$). Peaks were separated and clustered based on the population with the maximum value in Figure 3.2.e. An identical procedure was used for the promoter region accessibility scores x cell type matrix starting with the annotated set of 310 marker genes from a previous scRNA-seq analysis of mouse brain (Saunders et al., 2018), resulting in 262 genes where the $z_{i,j} > 3$ criterion was met for the promoter gene scores (Figure 3.2.d).

C.5.3 Promoter region chromatin accessibility scores

To annotate our *de novo* clusters from the whole mouse brain, we computed per-cluster promoter region chromatin accessibility scores representing a weighted-sum of chromatin accessibility around the transcription start site (TSS) of each gene in our reference data. Specifically, for gene *g* and cluster *i*, we define a chromatin accessibility score g_i from the following:

$$g_i = \sum_{j \in J} x_{i,j} * e^{-d_j/k}$$

Here, $x_{i,j}$ represents the counts-per-million normalized chromatin accessibility count for cluster i and chromatin accessibility peak j. Accessibility peaks used per gene J were restricted to those within 100,000 bp of a corresponding TSS, and d_j represents the distance (in base pairs) between the TSS and the center of peak j. The scaling constant, k, was fixed to 5,000 for all chromatin accessibility score computations.

C.5.4 Mouse brain cluster annotation

To annotate the dscATAC-seq mouse brain clusters in a data-driven manner based on the molecular signature of the distinct cell types in the brain, we used a resource containing scRNA-seq data for 690,000 individual cells sampled from 9 regions of the adult mouse brain (Saunders et al., 2018), which identified 565 subclusters within the broad classes of cell types in the brain. The list of cell types includes neurons, astrocytes, microglia, oligodendrocytes, polydendrocytes, and components of the vasculature. We note that many of these subclusters are from analysis of specific brain regions and further re-clustering within broadly defined clusters, leading to a large number of clusters. We use this data resource to 1) assign each one of our clusters to one of the broad cell classes identified in their study and 2) further refine the annotation by identifying which gene expression signature (within the 565 subclusters) provides an optimal match to each one of our dscATAC-seq clusters. To do this, we first obtained the union of the `class_marker` and `type_marker` genes identified in the scRNA-seq study (total of 310 unique genes). We then calculated the Spearman correlation coefficient between the per-cluster promoter region chromatin accessibility scores (27 clusters) and the aggregated scRNA-seq signal per cluster (565 clusters) at those 310 marker genes (Saunders et al., 2018). We then employed the Gale-Shipley algorithm to assign an optimal matching of scRNA-seq clusters to our scATAC-seq clusters. Here, the Gale-Shipley algorithm assigns pairs that maximize the global utility of the matches, noting our utility function was Spearman correlation. To classify the 27 dscATAC-seq clusters, we used the broad class assignment of the most correlated scRNA-seq cluster, except for the ”Neuron” class, which was further divided into Excitatory and Inhibitory neurons based on the annotation of Slc17a7 or Gad1 respectively. We then performed the same computational approach using another scRNA-seq dataset with 262 clusters (Zeisel et al., 2018) to validate the robustness of our approach. When displaying the overall correlation structure (Figure 3.2.c), we restricted the scRNA-seq clusters to those that had one or more class matches to the scATAC-seq data (500 out of 565 clusters).

C.5.5 Bulk-guided clustering

Bulk-guided clustering of single cells (Figure 3.4) was performed as previously described (Buenrostro et al., 2018). Briefly, a matched peakset (k=156,311 peaks) was used for both BMMCs dsciATAC-seq (n=136,463 single cells), and bulk ATAC-seq profiles previously generated for sorted hematopoietic cell populations for 16 cell types (Corces et al., 2016). PCA was first run on quantile-normalized bulk ATAC-seq data generating principal components (PCs) capturing variation across cell types. Single cells were then projected in the space of these bulk-trained PCs by multiplying the scATAC-seq reads in peaks matrix with the peaks x PC loading coefficients matrix to yield a matrix of single-cell projection scores (cells x PCs). The derived single-cell scores were then scaled and centered, and the corresponding single-cell data visualized using t-SNE. Predicted labels for single cells were obtained by correlating projected single cell scores with bulk PC scores, and choosing the most-correlated bulk cell type based on Pearson correlation coefficient. To define clusters for the control (unstimulated) BMMC dataset (Figure 3.4.c), Louvain clustering was performed using the igraph package where the 20 nearest neighbors per cell were used to build the embedding.

C.5.6 Single-cell classification

To assign most-alike clusters generated form the 15 clusters of the control (unstimulated) BMMC dataset (Figure 3.4.c) to additional datasets (Figure 3.4.e,f), the medoids of each per-cluster principal component were determined over all cells assigned from the Louvain clustering at baseline. Next, for each cell from a new dataset (*i.e.* FACS-sorted populations and stimulation-response cells), we assigned to the cell a reference cluster based on the minimum Euclidean distance between each cell's principal components and the medoids of the clusters.

C.5.7 Analysis of differential TF motifs

To compute differential TF scores in normal and stimulation conditions, we determine the 20-nearest stimulus condition neighbors for each single-cell in the resting condition using the bulk-guided PC scores and a Pearson correlation distance metric. To calculate differential TF motifs, we subtract the mean of the 20 stimulus cells by the TF score for each cell in the normal condition. Last, to suppress noise in the comparison, we smooth the differential TFs by taking the mean of the 20 nearest neighbors in the control condition. Again, the nearest neighbors are calculated using the bulk-guided PC scores, with Pearson correlation as a distance metric.

C.5.8 Differential peak identification in bone marrow stimulation

We devised a permutation test that accessed whether the proportion of cells with an accessibility element was differential between the stimulated and resting conditions, controlling for overall differences in accessibility (using measures at promoters). First, we filtered our consensus peak set such that the given peak was accessible in at least 1% of cells irrespective of stimulation or resting. Then, for an individual regulatory element i, we determined the proportion of cells in the resting p_r and the stimulated p_s conditions that observed one or more fragments overlapping the accessibility peak. Next, we computed the proportion of all promoters annotated in our dataset for both resting (p'_r) and stimulated (p'_s). Our observed differential statistic thus is given by:

$$\frac{p_s}{p'_s} - \frac{p_r}{p'_r}$$

To determine statistical significance, we permuted the stimulation and resting labels 1,000 times to generate a permuted distribution. We observed the corresponding z-statistic (Figure C.9.l) to be centered with a largely-Gaussian distribution. After converting these Z-statistics to p-values using a standard normal distribution, we computed a per-cluster false discovery rate (FDR) and established a significance threshold of 1% uniformly across clusters. We further computed an effect size of the difference between stimulated and resting, given simply by $p_s - p_r$. We summarized the differential association in Figure 3.5.g where the red bars ($FDR < 10^{-5}$) and pink bar ($FDR < 10^{-2}$) represent the statistical significance of the change in chromatin accessibility for each cell type cluster.

C.5.9 Overlap with fine-mapped GWAS SNPs

To identify regulatory regions affected by our stimulation conditions that may be relevant for human disease, we overlapped differential peaks identified per cell type with single nucleotide polymorphisms (SNPs) identified through genetic fine-mapping studies of 21 immune traits as previously described (Farh et al., 2015). Specifically, we downloaded the per-SNP meta-data available online (http://pubs.broadinstitute.org/pubs/finemapping/dataportal.php) and intersected differentially-accessible peaks with annotated positions of fine-mapped variants with a posterior probability > 0.3 computed by PICS (Farh et al., 2015) across all reported traits.

C.6 Supplemental note about computational bead merging with bap

C.6.1 Premise

The graphic shown in Figure C.2.a shows a simplified data example for six selected bead barcodes that were inferred to contribute to a total of three cells. For all six barcodes, one or more fragments was observed at each of the GAPDH and DIAPH1 promoters. However, upon closer examination, certain barcode pairs (*e.g.* pink/brown; orange/grey) annotate reads that share the same exact Tn5 insertion position. This demonstration highlights the value of considering the exact Tn5 insertion position for each fragment, rather than reads across a peak, when considering potential statistics to identify and merge these bead pairs. The computational approach used for bap does this.

The biochemical basis for this approach is characterized in Figure C.1.d. In brief, after the oligonucleotides (containing bead barcodes) are cleaved from the physical bead, they act as primers and anneal and amplify the cellular DNA fragments via PCR, therefore tagging each DNA fragment with a bead barcode. During successive rounds of PCR, oligonucleotides containing different bead barcodes may PCR amplify the same fragment. As a result, when the full library is sequenced, the same individual DNA fragment will be associated with multiple barcodes. Our approach seeks to define co-occurrence of bead barcodes from these DNA fragments.

C.6.2 Statistic

Explicitly, for arbitrary bead barcodes a and b, the fragment universe $U_{a \cup b}$ is all unique fragments (defined based on the genomic coordinates of the Tn5 insertion sites) tagged by the two bead barcodes. Further, U_a and U_b are the fragments in the universe of each fragment for the barcodes individually. All of these Tn5 insertion universes are defined by first removing highly abundant fragments (*i.e.* observed in >6 bead barcodes) as these likely represent a technical artifact. The `bap` statistic, s, for a and b is defined as follows:

$$s_{a,b} = \frac{|U_a \cap U_b|}{|U_{a \cup b}|}$$

C.6.3 Comparison to other statistical methods

To compare the efficacy of our chosen approach (`bap`) to other plausible approaches (that do not resolve the exact position of the Tn5 insertion), we considered three additional metrics shown in Figure C.2.k. We define each metric below:

- `bap`: Pairwise (between bead barcodes) jaccard index computed over Tn5 insertion sites after removing highly abundant fragments (*i.e.* observed in >6 bead barcodes). This is represented by the $s_{a,b}$ statistic described above.

- Pearson: Pairwise Pearson correlation metric of the reads by barcodes matrix.

- Spearman: Pairwise Spearman correlation metric of the reads by barcodes matrix.

- jaccard peak: Pairwise jaccard index value computed over fragments observed within chromatin accessibility peaks.

From these metrics, the corresponding AUROC and AUPRC values were computed for each experiment using the same set of true-positive bead barcode merges. The true-positive set is defined by the orthogonal measure that includes droplet PCR with a high-diversity oligonucleotide as a template alongside the transposition fragments within cells. These metrics are reported in the barplots shown in Figure C.2.k.

Finally, we distinguish between the metric that is computed pairwise (used to compute AUROC and AUPRC statistics) and the defined threshold value on the bap statistic that results in bead merges from our algorithm. This data-driven, dynamic threshold for bead merges per-library is determined using the "knee-calling" algorithm described in the following section.

C.6.4 Potential drawbacks to this approach

As open chromatin fragments are cell type specific and are also the basis for our bead merging algorithm, theoretically, there may be a tendency for 'false-positive' merges (*i.e.* two bead barcodes tag two unique cells but are merged into one cell) to be enriched across cells of the same cell type. Notably, in our analyses of these data, we have found no empirical evidence of this occurring or confounding our results. However, we encourage others to acknowledge this possibility, particularly as this approach is used in new ways. Notably, if false positives are a concern additional approaches may be used, such as indexing single-cells using combinatorial Tn5 and/or the introduction of random oligonucleotides into the droplet mix.

Additionally, we note that all of our data settings in which we evaluated the bead merging algorithm were diploid genomes. Highly variable ploidy could impact the accuracy of bead merges, particularly for model organisms with a chromosomal copy number significantly greater than two. Further, the computational efficiency of our approach is on the order of $O\left(n^2\right)$, where n is the number of bead barcodes. Larger datasets with significantly greater diversity of bead barcodes may hinder our approach for bead merging.

C.6.5 Dynamic threshold determination or "knee-calling"

To identify high-quality cells, knee calling is performed by first generating a Gaussian kernel density estimate (KDE) of the log10 transformed unique nuclear read counts per bead barcode. The KDE is then used to create a density distribution of 10,000 evenly spaced values between the minimum (c_{min}) and maximum (c_{max}) log10 transformed unique nuclear read counts (we denote this vector as c). Local minima in this density distribution were identified and used as potential inflection points below which barcodes were filtered from further analysis.

To choose a specific local minimum for thresholding (c_{thres}), we picked the smallest value (among the vector of possible minima) that satisfied the following criteria:

$$\log_{10}(c_{max}) - \log_{10}(c_{thres}) > d$$

and

$$\frac{\sum_{i \in I} 1(\log_{10}(c_i) > \log_{10}(c_{thres}))}{|I|} > 0.20$$

In other words, criterion (1) enforces that the difference between the number of reads for the most abundant bead and the threshold is larger than some value d ($d = 0.5$). We observed this criterion to be useful in settings where the top few (~10) beads contained a disproportionately high number of reads. For this second criterion, we nominate a set of barcodes I such that $\forall\ i \in I, c_i >$ mode (c), where $|I|$ represents the total number of barcodes in this set. As we observed many more droplets with no cell, the mode of the vector c most likely represents a value associated with a bead barcode that is not associated with a cell. Thus, this second criterion enforces that at least 20% of these plausible barcodes that are more abundant than the mode pass the knee detection. In a scenario where no data-driven parameter can be determined, our algorithm fixes the fragment threshold to a default value of 500. We note that for all libraries presented in this manuscript, our approach successfully determined a data-driven threshold for each individual library. Further, we note that this conceptually takes a conservative approach and in almost all analyses, more stringent filtering criteria are applied.

C.6.6 Jaccard index for bead merging

Here, we note that this pairwise computation attempts to find a relatively small number of true bead pairs that originated from the same droplet. To achieve this, the same procedure and criterion using a Gaussian KDE is further applied to the `bap` Jaccard index score (s) for pairs of bead barcodes to identify beads originating from the same droplet. Notably, the log10 transformation is not applied for the scores. Additionally, we define the threshold difference to be $d = 0.05$, and the set of barcode pairs I is all non-zero pairs (which is consistent with the definition above using the mode since 0 is empirically almost always the mode observed per-library). In a scenario where no data-driven parameter can be determined, our algorithm fixes the fragment threshold to a default value of 0.005. We note that for all libraries presented in this manuscript, our approach successfully determined a data-driven threshold for each individual library.

D

Supplemental material for Chapter 4

D.1 Biological and experimental methods

D.1.1 Loading and visualizing bead loading in droplets

We used the 10x Chromium Controller Training Kit (PN-12024, PN-120238) to generate GEMs following manufacturer's instructions. The GEMs were carefully collected without disrupting the emulsion. After GEM formation, 10 μL of GEMs from each 10x channel was immediately loaded onto Countess Cell Counting Chamber Slides (C10228, Thermofisher) for visualization. We captured 10 bright field images under an Olympus IX70 microscope, and beads per droplet were counted based on manual inspection of images. To quantify the proportion of barcodes affected by multiple beads (barcode multiplets), we used the following equation:

$$\%\text{multiplets} = \frac{\sum_{b=2}^{4} bn_b}{\sum_{b=1}^{4} bn_b} \text{ x } 100$$

where b is the number of beads present in a given droplet and n_b is the number of droplets with b beads. Here, the expression is capped at 4 as droplets with 4+ beads could not be reliably quantified. Thus, in these instances, the value of barcodes per droplet were conservatively assigned a count of 4. For the Zheng et al. (2017b) data, we used the following abundances from previous imaging data: 15% of droplets had 0 beads; 80% of droplets had 1 bead; and 5% of droplets had 2 beads. As neither the raw data nor the quantification values have been published, these values were approximated from an examination of a plot previously reported.

D.1.2 Profiling PBMCs using 10x scATAC-seq

For 10x scATAC-seq experiments with PBMCs (PB003F, AllCells), frozen cells were quickly thawed in a 37°C water bath for about 30s and transferred to a 15 mL tube. 5 mL of pre-warmed RPMI 1640 (ATCC, 30-2001) supplemented with 10% Fetal Bovine Serum (FBS) were added to the sample drop by drop. The cells were pelleted by spinning at 300g for 5min at room temperature. The supernatant was removed, and cells were washed with 1 mL PBS. The cells were then pelleted again, resuspended in 1 mL PBS, and used for 10x ATAC v1.0 pro-

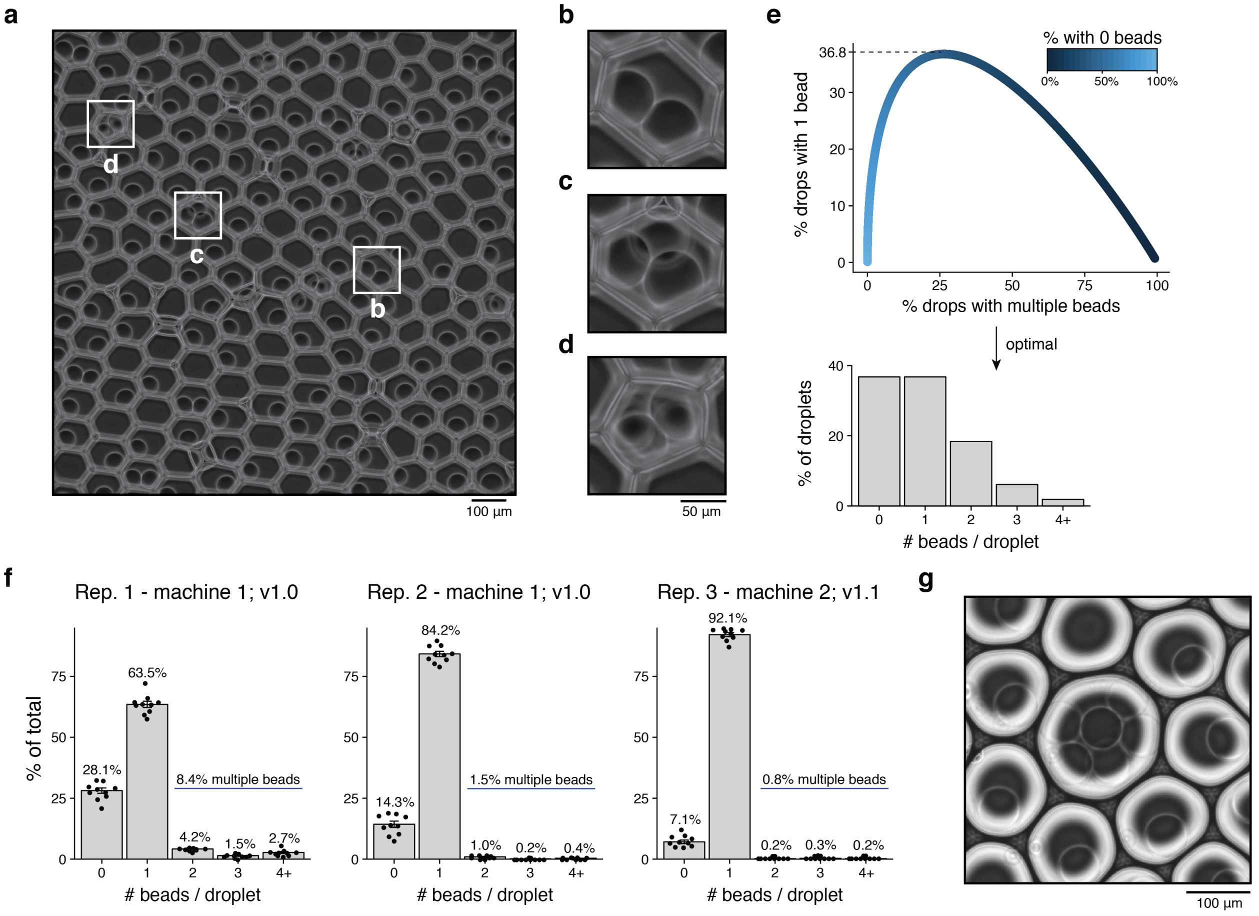

Figure D.1: Details of barcode multiplet quantification via imaging. (a) Alternative field of view. Boxes highlight individual droplets shown in subsequent panels. The image is representative of a total of 30 fields of view taken from 3 independent experiments. **(b-d)** Examples of 2, 3, and 4+ beads per droplet, respectively. **(e)** Theoretical support for optimal bead loading under Poisson distribution assumptions. The dotted line (top) represents the theoretical maximum for 1 bead loaded into droplets, and the full distribution at this point is shown in the bar graph. **(f)** Quantification of beads per droplet for each replicate. Above each panel, the machine and the version of the chip used for the training kit is indicated. Error bars represent standard error of mean over n=10 independent fields of view for each of the three experimental replicates (n=30 total). **(g)** Example of presumed merged droplet containing multiple (~6) beads. The selected droplet was one of ~10 droplets that was likely the consequence of merging taken from a total of 30 fields of view taken from 3 independent experiments.

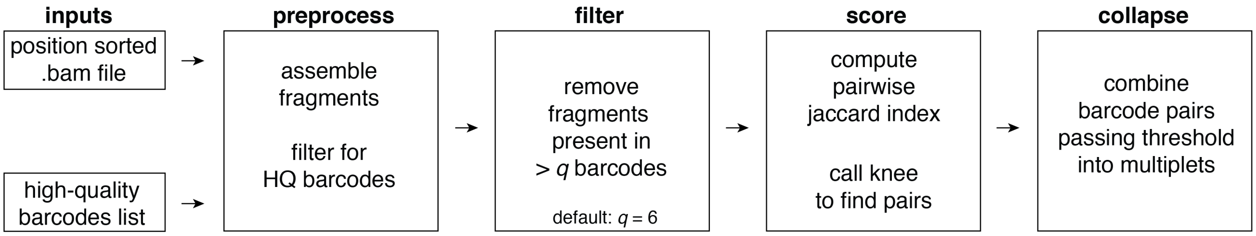
inputs
position sorted .bam file
high-quality barcodes list
preprocess
assemble fragments
filter for HQ barcodes
filter
remove fragments present in > q barcodes
default: q = 6
score
compute pairwise jaccard index
call knee to find pairs
collapse
combine barcode pairs passing threshold into multiplets

Figure D.2: Summary of the `bap` **workflow for** `10x` **scATAC-seq data.** An overview of the inputs and computational workflow for the application of `bap` to `10x` scATAC-seq data.

tocol following manufacturer's instructions. The corresponding library was sequenced on an Illumina NextSeq 500.

D.2 Bioinformatics and data analysis methods

D.2.1 Data preprocessing

Raw sequencing data was processed with Cell Ranger ATAC version 1.0.0. Reads were aligned to the hg19 reference genome available on the `10x` Genomics website. Processed `10x` PBMC datasets were downloaded from https://www.10xgenomics.com/resources/datasets/ from the version 1.1 PBMC 5k scATAC-seq dataset. The requisite input files for `bap` included the `.bam` file and the high-quality barcodes file. Additional annotations from Louvain clustering and t-SNE coordinates were also downloaded for downstream visualization and analyses. For the comparison of the chip technologies (Figure 4.3.g), we again downloaded the PBMC 5k scATAC-seq datasets from the "Chromium Next GEM ATAC Demonstration."

D.2.2 Processing `10x` scATAC-seq data with `bap`

In order to facilitate the processing of `10x` scATAC-seq data with `bap`, no major substantive changes were required for the underlying barcode multiplet identification algorithm that has been previously outlined in Chapter 3. However, additional command-line options were added, including the –barcode-whitelist flag, which im-

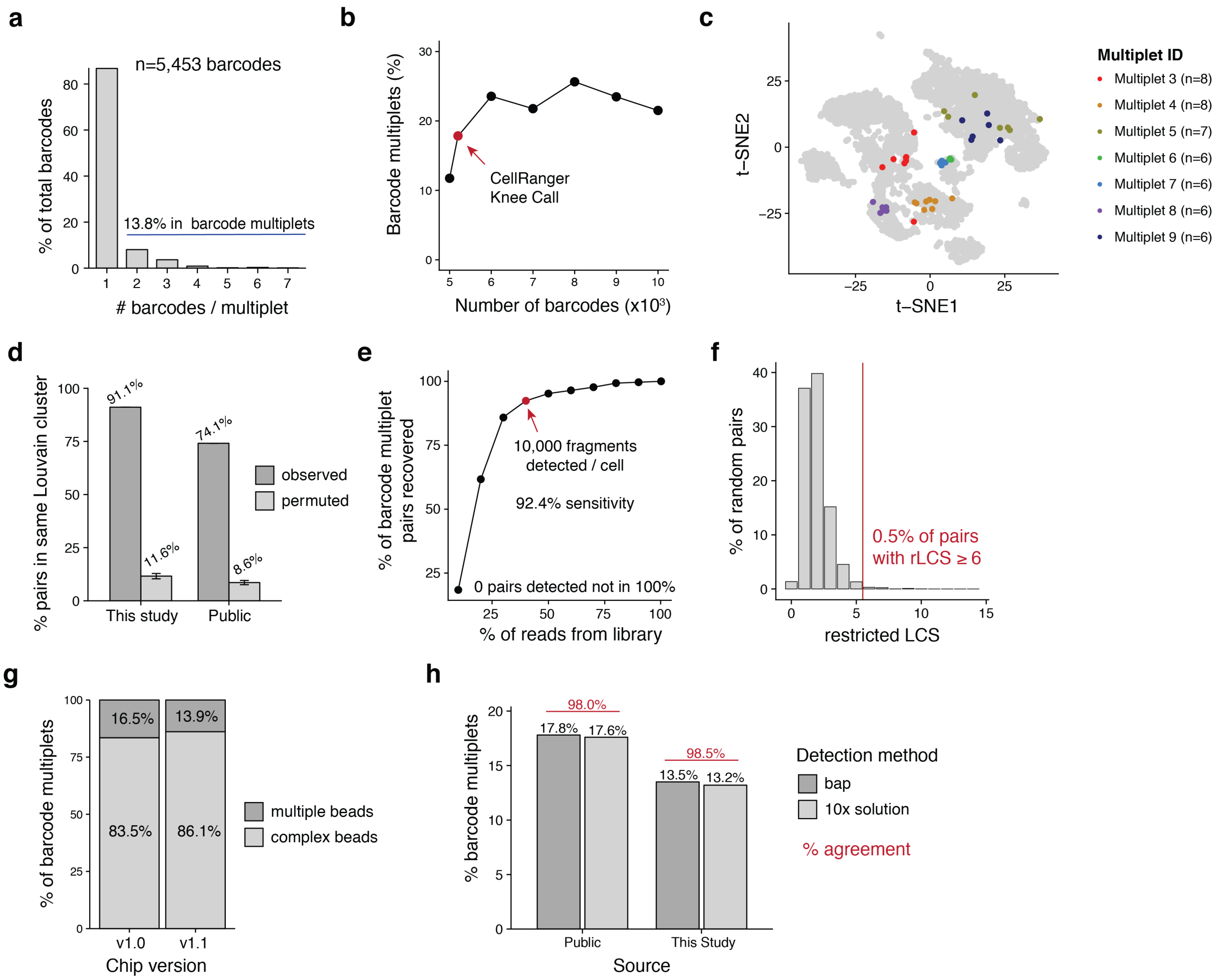

Figure D.3: Supporting information for barcode multiplets learned from scATAC-seq data. (a) Quantification of barcodes affected by barcode multiplets for the PBMC dataset generated with this work ("This Study"). **(b)** Percentage of barcode multiplets identified for different numbers of input barcodes. **(c)** Visualization of seven additional barcode multiplets from the Public dataset. **(d)** Proportion of bead pairs occurring in the same chromatin accessibility-defined Louvain cluster compared to a permuted background. Error bars represent standard error of mean over n=100 independent permutations per each dataset (two independent experimental replicates). **(e)** Downsampling analysis of the dataset generated in this work ("This Study"). Barcode multiplets were examined at downsampled intervals from 10%-90% by units of 10%. The highlighted sample represents 40% downsampling and corresponds to a median 10,000 fragments detected per barcode. At all downsampled thresholds, we detected 0 pairs that were not present in the 100% sample. **(f)** Distribution of the restricted longest common subsequence (rLCS) for 1,000,000 randomly sampled barcode pairs in the 10x barcode universe. A threshold at 6 is drawn for use in other analyses. **(g)** Breakdown of types of barcode multiplets from the Next-gem comparison data. **(h)** Comparison of methods to detect barcode multiplets. The rates of barcode multiplets detected by each solution is shown in black. The % agreement between the two methods (per barcode) is shown in red.

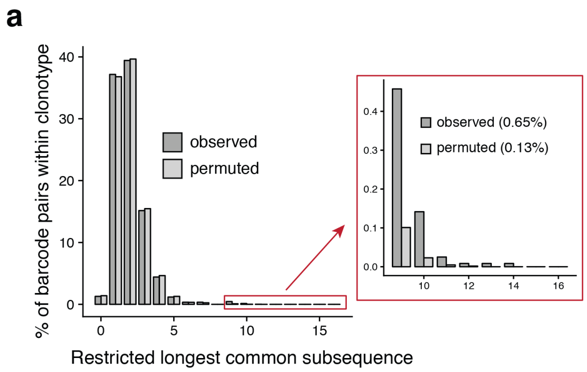

Figure D.4: Evidence of barcode multiplets from rLCS of BCR clones. (a) Observed and permuted (within clonotype) restricted longest common subsequence for the BCR clone dataset. The inset shows a zoom for rLCS ≥9 and the percent of barcodes depicted in the panel.

ports the error-corrected, quality-controlled barcodes identified as "cells" by CellRanger, enabling analysis of the filtered output from the default `10x` pipeline. This functionality augments the default process in `bap` where abundant barcodes are identified via quantification and knee-calling in terms of total reads observed per barcode. Versions 0.5.9+ of `bap` facilitate full analysis and merging of barcode multiplets with `10x` scATAC-seq data.

D.2.3 In silico mixing experiment

Using two different public PBMC 5k datasets, we sought to determine a putative false positive rate for the application of `bap` to `10x` scATAC-seq data. Here, we denoted the PBMC-5k "Public" dataset as Channel 1 and the PBMC-5k from the NextGEM beads as Channel 2. We modified the CB tags (which contains the error-corrected barcodes) in the `.bam` files for each channel to ensure that each barcode for each experiment was uniquely identifiable. These modified bam files were subsequently merged. Next, the same modification to the barcodes was made, and the two high-quality barcodes files were combined into a single file. We then executed `bap` using the default parameters with this merged `.bam` and merged barcode list file. Using a single threshold determined by the knee call, we identified pairs of barcodes originating from the same or different channels as summarized in Figure D.2.c-e. The top 500,000 barcode pairs were plotted in rank order for each of these three plots, and the same single threshold was visualized in all three panels.

D.2.4 Assigning bead barcodes to multiplets

The identification of multiplets follows the same strategy previously described in Chapter 3. In brief, a per-barcode pair summary statistic (modified jaccard index) is computed using the one base pair location of Tn5 insertions. We emphasize that this statistic has been validated using an orthogonal oligonucleotide library as we have previously described in Chapter 3. From this distribution of millions of barcode pairs, we computationally infer an inflection point threshold T (similar to a ”knee-call” used by CellRanger to identify true cell barcodes). To derive multiplets, we iteratively consider the barcode pairs (*e.g.* b_1 and b_2) with the highest remaining overlap score and append any additional barcodes whose overlap value with either b_1 or b_2 exceeds T. For example, if the statistic between b_1 and b_3 exceeds T, then b_1, b_2, and b_3 are assigned to one multiplet. This process continues until all barcodes are assigned a multiplet that had an overlap score exceeding T. All remaining barcodes are assigned as singlets. To facilitate processing of the `10x` scATAC-seq data, we modified the command line interface and internal data structures of `bap`, but the conceptual basis and execution is the same as previously described3.

D.2.5 Classifying and quantifying complex beads

To determine multiplets driven by putative bead barcode synthesis errors, we considered all pairs of barcodes within an annotated multiplet and computed the restricted longest common subsequence (rLCS) between them. Explicitly, the rLCS is the largest consecutive number of characters that match between two strings without shifting the strings. We note the necessity of defining a distance metric (rLCS) that is distinguished from the longest common subsequence (LCS) as our metric does not allow insertions or deletions when performing the string matching. Additionally, rLCS is distinguished from the Hamming distance as the matching characters must all occur in a continuous unit (which is not enforced by Hamming).

To determine an appropriate threshold to classify multiplets as having originated from multiple beads or a single heterogeneous bead, we established a null distribution of the rLCS shown in Figure D.3.f. To achieve this, 1,000,000 random draws of barcode pairs were determined and the rLCS was computed. We selected an rLCS threshold of 6 as pairs with an rLCS ≥6 represented less than 0.5% of the data, which was used to classify multiplets from the real data (Figure 4.3.f). To determine whether the number of fragments was similarly captured between barcodes contained in multiplets, we computed the pairwise percent difference of the log2 unique fragments ("passed_filter" in the CellRanger-ATAC .csv file). The per-multiplet average of the mean pairwise percent difference is plotted in the boxplots in Figure 4.3.g, and we used a two-sided Kolmogorov–Smirnov test to verify that the droplets containing multiple beads had a more even ratio of reads compared to multiplets driven by bead heterogeneity.

To quantify the percent of beads that had heterogeneity, the numerator was the number of multiplets identified with an rLCS ≥ 6 (from Figure 4.3.f). The denominator was the total number of barcodes analyzed while 1) still counting all barcodes in perceived bead multiplets but 2) collapsing the heterogenous barcode multiplets to only 1 barcode. For example, in the "This Study" dataset, the total number of barcodes passing the CellRanger knee was 5,453. Of these, 4,732 barcodes were from singlets, 121 barcodes were associated with multiplet beads per droplet (and thus not complex), and 600 barcodes were associated with 253 complex beads. The complex bead rate can be computed as follows:

$$complex\ bead\ rate = \frac{\#\ complex\ beads}{\#\ singlet\ beads + \#\ beads\ in\ bead\ multiplets + \#\ complex\ beads}$$

For our example of the "This Study" dataset:

$$\frac{253}{4732 + 121 + 253} = 4.95\%$$

D.2.6 Chi-square test for cluster / multiplet

To test for association between barcode multiplets and cluster identification, we performed a chi-square test for independence. For the n Louvian clusters identified by CellRanger, we assembled a 2xn contingency table, tabulating barcodes into corresponding entries in the contingency table. The two rows specified whether each bead barcode was predicted to occur in a multiplet or not as identified by `bap`. P-values were computed using the chi-squared statistic with $n - 1$ degrees of freedom.

D.2.7 Evaluation of barcode multiplets with different numbers of variable input barcodes

To test the abundance of barcode multiplets with different numbers of considered barcodes, we executed `bap` with 5,000-10,000 barcodes at intervals of 1,000 barcodes (6 additional executions) in addition to the 5,205 found by CellRanger's knee call. Each barcode set was nominated based on the ranking of fragments in peaks, the same metric used by CellRanger to determine an optimal threshold. Our results (Figure D.3.b) show that the inferred cutoff underestimates the barcode multiplets in the Public data, consistent with our imaging results. We interpret this plot to show that barcode multiplets often occur near the inflection point (consistent with these barcodes having fewer reads due to the fractionated data). However, this rate flattens when additional barcodes added do not represent multiplets but other ambient fragments that cannot be associated with a highly-observed barcode.

D.2.8 Enrichment for barcode multiplet pairs in the same cluster

For each barcode multiplet identified by `bap`, we considered all possible pairwise combinations of constitutive barcodes. For example, multiplets consisting of precisely two bead barcodes had one pair whereas multiplets consisting of four barcodes contained six barcode pairs (all combinations; choose two). For these pairs, we computed the proportion that occurred in the same Louvain cluster produced by the default CellRanger execution.

A background rate was generated by performing 100 permutations of the full dataset where cluster labels were permuted.

D.2.9 Downsampling analyses

To evaluate the stability of the bap statistic as a function of coverage, we downsampled the dataset generated here ("This Study") at intervals of 10% and reran bap on the resulting downsampled .bam files. Here, we used the full set of high-quality barcodes determined from the CellRanger execution on the full dataset. Moreover, we determined the set of identified barcode pairs from the full dataset as a 'true positive' set of pairs to compare the downsampled results. Figure D.3.e shows the results of this downsampling, including the 40% subsample (that corresponded to a median 10,132 fragments per barcode) that achieved >90% sensitivity in detecting the set of barcode pairs from the full data. Critically, in each of the 9 downsampled executions of bap, no barcode pairs were identified that were not present in the full dataset.

D.3 Comparison of our approach with 10x solution

After contacting 10x support, we obtained the "clean_barcode_multiplets_1.0.py" script, which identifies barcode multiplets in single-cell ATAC-seq data. We executed this code and evaluated the output for the two scATAC-seq datasets closely analyzed in this work ("Public" and "This Study"). While the procedure used to identify multiplets similarly utilizes shared Tn5 insertions, the treatment of multiplets once detected is different from bap. Specifically, for each multiplet, the barcode with the most unique fragments is retained and the other barcodes are filtered out. Further, 10x refers only to the barcodes that are filtered out as 'multiplets', rather than counting the most prevalent barcode as part of a barcode multiplet as we've done throughout this manuscript. For comparison purposes, we used our definition of barcode multiplet (as stated in the abstract) and reported the rates from each tool (see script in Code Availability for the exact procedure). Finally, to compute the concordance between the two methods, we assigned each barcode whether or not it was part of a barcode multiplet from both sources and report the percentage of barcodes that had a matching annotation across the detection methods.

D.4 Clonotype analyses with 10x 5' kit

D.4.1 Estimation of multiplet-adjust BCR / TCR clonotype abundances

In order to estimate the number of cells contributing to each clonotype (defined by a unique BCR or TCR sequence), we downloaded the per-barcode clone identification files (BCR: vdj_v1_hs_nsclc_b_all_contig_annotations.csv; TCR: vdj_v1_hs_nsclc_t_clonotypes.csv) from the 10x CellRanger output for the public NSCLC tumor dataset. Here, each barcode is assigned a clonotype group when detected with high confidence in the CellRanger pipeline. To simulate the occurrence of barcode multiplets, we executed the following simulation procedure.

For each barcode i with a total of n barcodes in the experiment (all assigned a clonotype), we simulate a corresponding multiplet value m_i which defines the barcode multiplicity; *i.e.* the number of unique barcodes that overall co-occur with barcode i inside a theoretical droplet. We performed our simulation by specifying the following probability distribution function:

$$P(m_i = 1) = 0.93;\ P(m_i = 2) = 0.05;\ P(m_i = 3) = 0.01 P(m_i = 4) = 0.005;\ P(m_i = 5) = 0.005$$

Importantly, the values defined in the probability distribution function are grounded in the empirical estimates from bap across our two datasets but likely represent conservative estimates assuming a similar distribution of barcode multiplets from scATAC-seq holds in this assay. In other words, $P(m_i = 1) \ = \ 0.93$ is likely overestimated and $P(m_i > \ 5) = 0$ is underestimated, and from this parameterization, the expected rate of barcode multiplets is 15.8%. Here, we denote the set of values m_i as M (of length n). To account for k clonotypes with exactly one barcode that could only be generated from a barcode singlet, we define a new set $M\prime$ such that $M\prime \cup K = M$ where $|K| \ = \ k$ and $\forall m_i \ \in \ K, \ m_i \ \ = \ 1$. Thus, the elements of $M\prime$ represent the barcode multiplicities for clonotypes annotated with two or more cells.

To estimate the multiplet-adjusted cell number per clonotype, we iteratively sample from the set M' until we have observed sufficient barcode numbers to explain the original clonotype abundances, akin to observing droplets with variable barcode abundances. More precisely, for a given clonotype j comprised of c_j barcodes (from the raw CellRanger output), we seek to compute the multiplet-adjusted number of cells c_j'. To achieve this, we sample from M' until the sum meets or exceeds c_j. c_j' then is the number of draws corresponding to the number of multiplet-aware droplets needed to explain the clonotype abundance and can be interpreted as the number of cells present in the clone under the simulation setting. As an example, suppose $c_j = 4$, representing a clone of four barcodes. If we sample a 4 or 5 from M', then $c_j' = 1$, meaning that one droplet explains the clone in this scenario. Last, the new per-clonotype abundances in the library are then represented by the union of K with the set of all c_j. These multiplet-adjusted abundances were computed over 100 iterations, and the numbers reported in the main text represent the mean over these simulations. We note that an *R* script that achieves this approach is available in the repository noted in Code Availability.

We define the "clone false discovery rate" as the proportion of clonotypes with at least 2 cells that then becomes explained by a barcode multiplet (*i.e.* $c_j' = 1;\ c_j > 1$) under our simulation setting. The numbers reported in the main text represent means for each of the BCR and TCR clones over the 100 simulations. Finally, we note that while this simulation assumes that the multiplet rates inferred for scATAC-seq are transferable to scRNA-seq, alternative approaches, such as estimating the complex bead rate from scRNA-seq directly, are likely unreliable without a sensitive multiplet detection approach as presented with `bap`. Ultimately, our simulation results provide an anchor to interpret the potential shift in clonotype abundance from the lens of our barcode multiplet artifact. However, additional experiments and analytical tools are needed to accurately determine clonotype abundance.

D.4.2 Determination of multiplet-driven clonotypes

In scATAC-seq data, barcode multiplets were identified using our approach previously described. However, no such approach exists for scRNA-seq. Thus, to identify potential multiplets, we were required to consider potential multiplets defined only by barcode similarity, which would be reflective of synthesis errors resulting in a bead with heterogeneous barcodes (Figure 4.1.a). To determine these potential multiplets, we considered all pairs of barcodes within an annotated clonotype and computed the restricted longest common subsequence (rLCS) between them. Analysis of the distribution of pairs (Figure D.4.a) within clonotype labels revealed was used to identify the clones shown in Figure 4.4. When computing a permuted distribution (Figure D.4.a), labels of clonotypes were shuffled such that random barcode pairs were considered.

E

Supplemental material for Chapter 5

E.1 Biological methods

E.1.1 Cell lines and cell culture

TF1 cells (ATCC) were maintained in Roswell Park Memorial Institute Medium (RPMI) 1640, 10% fetal bovine serum (FBS), 2mM L-Glutamine and 2ng/ml recombinant human Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF) (Peprotech) and incubated at 37C and 5% CO2. GM11906 cells (Corriell) were maintained in Roswell Park Memorial Institute Medium (RPMI) 1640, 15% fetal bovine serum (FBS) and 2mM L-Glutamine and incubated at 37C and 5% CO2.

E.1.2 Primary cells and cell culture

CD34+ hematopoietic stem and progenitor cells were obtained from the Fred Hutchinson Hematopoietic Cell Processing and Repository (Seattle, USA). The CD34+ samples were de-identified and approval for use of these samples for research purposes was provided by the Institutional Review Board and Biosafety Committees at Boston Children's Hospital. CD34+ cells were thawed and cultured in StemSpan II with 1x CC100 (StemCell Technologies, Inc.) at 37C and 5% CO2. At indicated time points, these cells were seeded in media supporting the differentiation into monocytic and erythroid cells. Briefly, cells were cultured at a density of $10^5 - 10^6$ cells per milliliter (ml) in IMDM supplemented with 2% human AB plasma, 3% human AB serum, 1% penicillin/streptomycin, 3 IU/ml heparin, 10 mg/ml insulin, 200 mg/ml holo-transferrin, 1 IU erythropoietin (Epo), 10 ng/ml stem cell factor (SCF) and 1 ng/ml IL-3 and incubated at 37C and 5% CO2.

E.1.3 Chronic lymphocytic leukemia samples

Cryopreserved peripheral blood mononuclear cells from chronic lymphocytic leukemia patients consented on institutional review board approved protocols were obtained from AllCells or from Adrian Wiestner at the National Institute of Health. Cryopreserved cells were thawed by serial dilution in RPMI with 10% fetal bovine

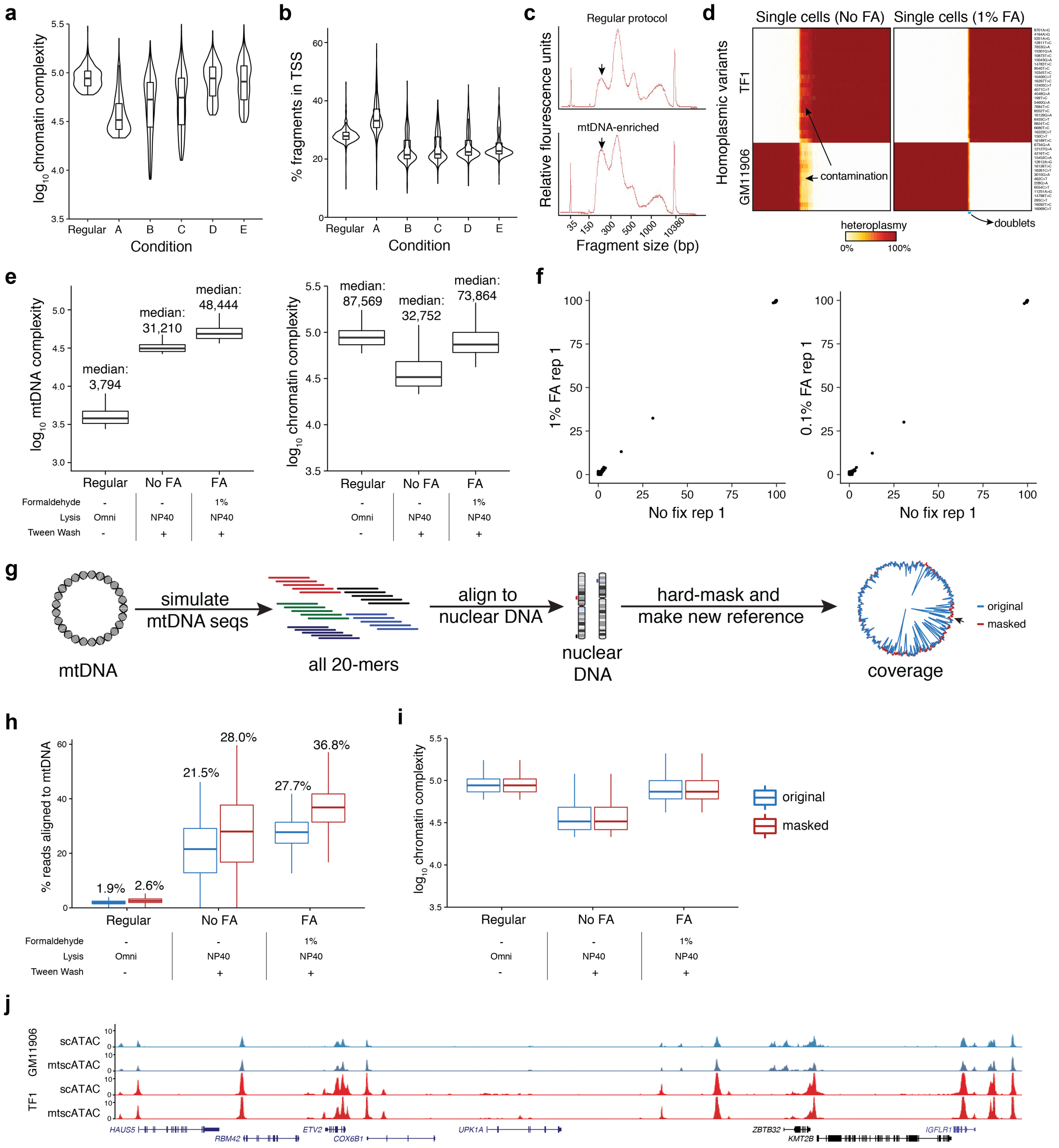

Figure E.1: Additional validation of biotechnical and computational basis for single-cell mtDNA genotyping. **(a)** Comparison of chromatin library complexity (estimated number of unique fragments) across screened lysis conditions as shown in Figure 5.1. **(b)** The same variable lysis conditions showing the TSS rate per cell. **(c)** BioAnalyzer traces of mtscATAC-seq library fragment size distribution for regular conditions and mtDNA-enriched conditions. **(d)** Heteroplasmy heatmap of single cells (columns) for 43 private homoplasmic mutations (rows) in the TF1 or GM11906 cell lines with (left) and without (right) FA treatment. Color bar, heteroplasmy (% allele frequency). **(e)** Comparison of mtDNA fragment complexity and chromatin complexity between the original/ regular 10x scATAC protocol and modified lysis conditions with and without formaldehyde (FA) treatment. **(f)** Heteroplasmy of sum of single-cell ATAC-seq libraries with variable FA treatment. **(g)** Schematic, method, and results of improving mtDNA genome coverage via hard-masking the reference genome. **(h)** Comparison of % reads mapping to mtDNA and **(i)** chromatin complexity with (red) and without (blue) the hard masking. Boxplots: center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. **(j)** Accessible chromatin landscapes aggregated from single cells near the ETV2 locus for both cell lines as assayed via regular scATAC-seq and mtscATAC-seq.

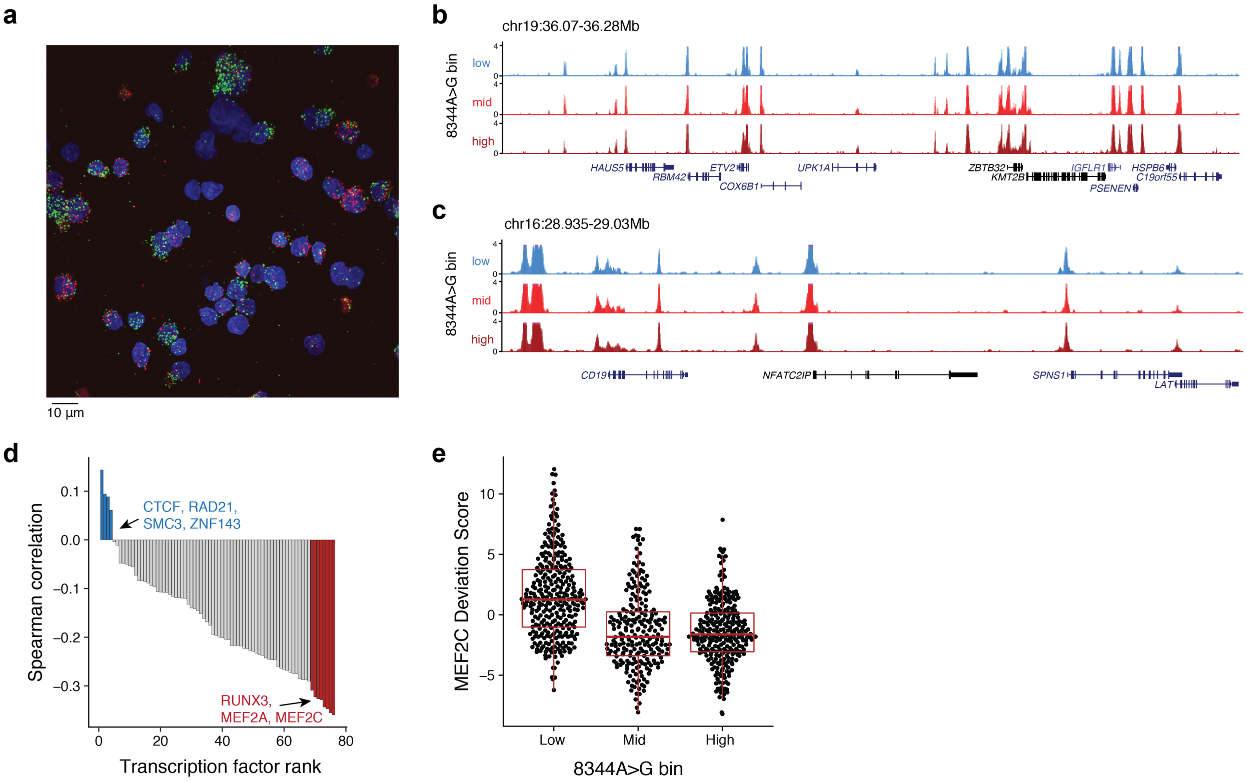

Figure E.2: Further inferences in analysis of the GM11906 (MERRF) lymphoblastoid cell line. **(a)** Alternative field of view for GM11906 in situ genotyping imaging experiment. Pseudo bulk accessibility track plots are shown for the **(b)** ETV2 and **(c)** CD19 loci. Pseudo-bulk groups represent 0-10% (low), 10-60% (mid), and 60-100% (high) m.8344A>G heteroplasmy. **(d)** Spearman correlation of heteroplasmy against the ChIP-seq deviation scores computed via chromVAR. Each bar is a single transcription factor with selected factors highlighted. **(e)** Depiction of MEF2C deviation scores from chromVAR for m.8344A>G heteroplasmy bins, corresponding to 0-10% (Low), 10-60% (Mid), and 60-100% (High). Boxplots: center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range.

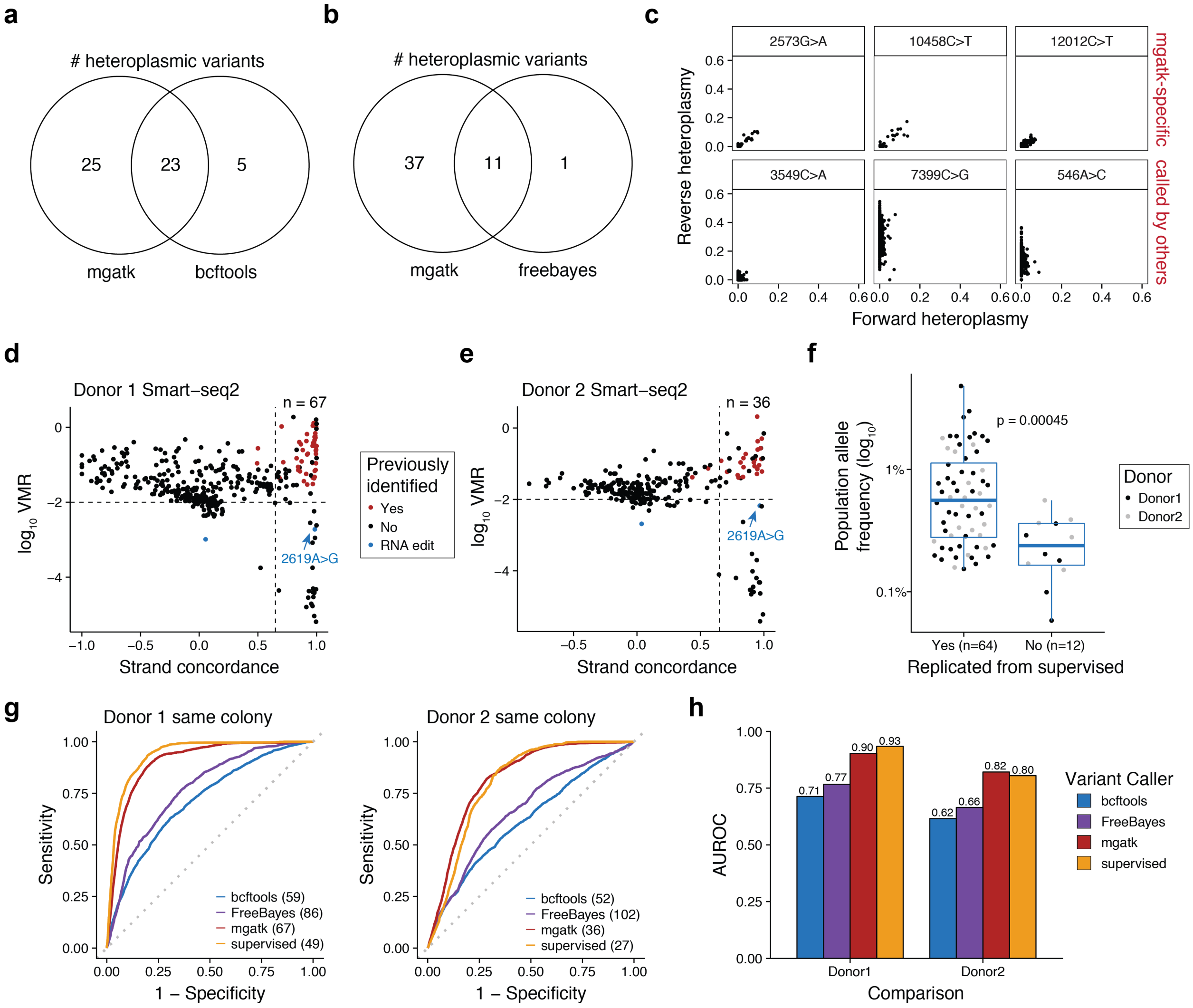

Figure E.3: Validation of somatic mtDNA mutation calling via mgatk. **(a)** Venn diagrams depicting comparisons of heteroplasmic mutations identified by mgatk, samtools/bcftools, and **(b)** FreeBayes. **(c)** Comparison of heteroplasmy estimated from reads aligned to either strand. The top row are three variants called specifically by mgatk; 3549C>A was identified only by FreeBayes. 7399C>G and 546A>C were called specifically by bcftools. **(d)** Identification of 67 and **(e)** 36 heteroplasmic variants from previously published SMART-seq2 hematopoietic colony data. Blue variants represent known RNA-editing events. **(f)** Comparison of population heteroplasmy values for variants replicated by mgatk from a previous supervised approach. Boxplots: center line, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range. **(g)** Concordance between discerning cells sharing a clonal origin based on colony-specific mtDNA mutations and their unsupervised identification using indicated algorithms (mgatk, bcftools, FreeBayes) and previously described supervised approach (see Chapter 1). Receiver operating characteristic (ROC) using the per cell pair mtDNA similarity metric to identify pairs of cells sharing a clonal origin based on sets of mtDNA variants. The number of variants in each set is also depicted. **(h)** Area under the ROC (AUROC) is denoted for each donor group and indicated variant caller as depicted in **(g)**.

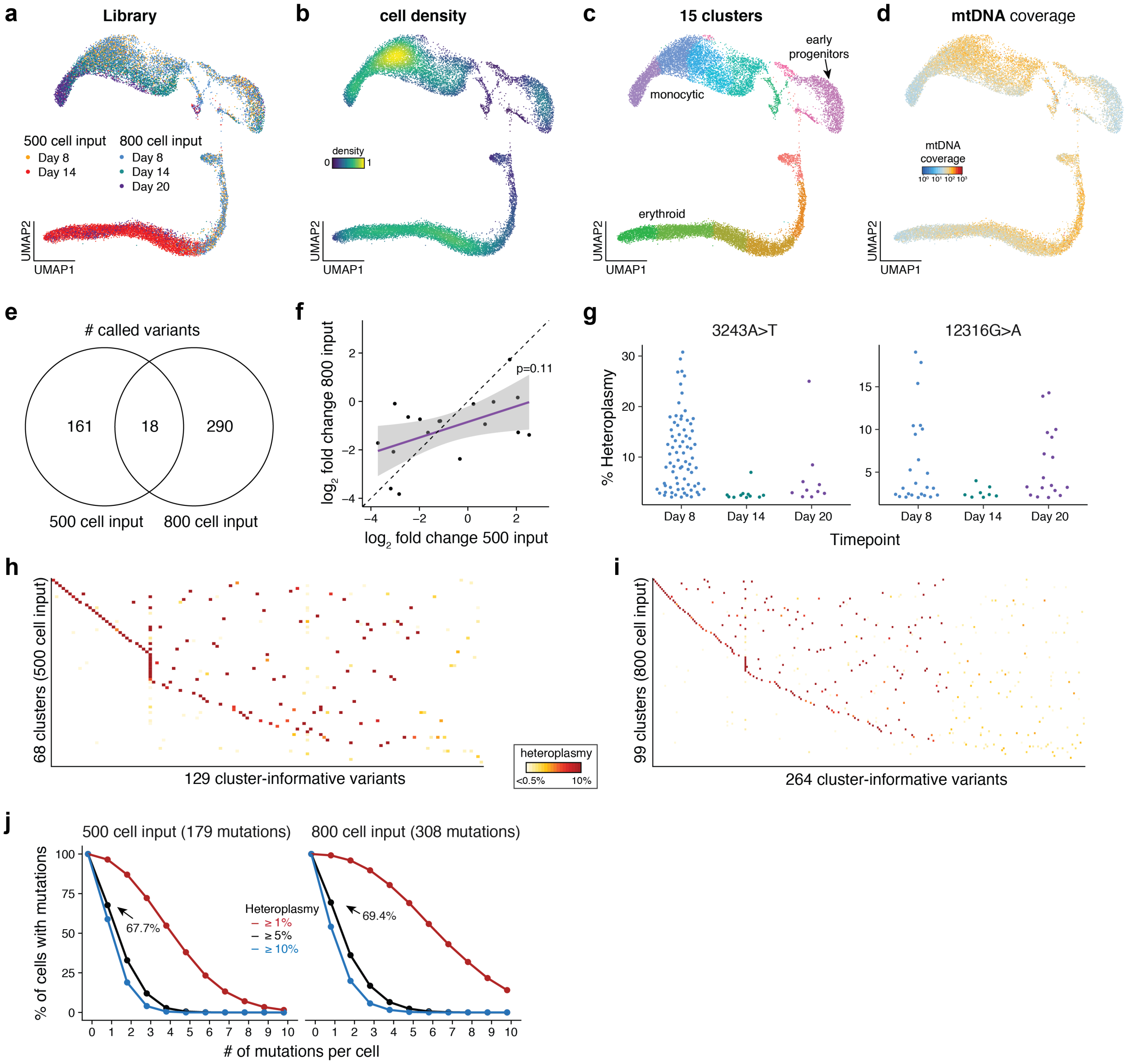

Figure E.4: Supporting information for clonal lineage tracing across accessible chromatin landscapes in an *in vitro* model of hematopoiesis. (a) Depiction of single-cell UMAP embedding showing the original distribution of cells for each library/ time point, **(b)** relative cell density, **(c)** Louvain cluster, and **(d)** mitochondrial DNA coverage per single cell. **(e)** Overlap of variants called for each of the two datasets. **(f)** Comparison of log2 fold change in heteroplasmy from day 14 to day 8 for 18 overlapping variants. The p-value shown is for the beta 1 coefficient of the depicted linear regression model. **(g)** Known pathogenic mtDNA mutations detected from a healthy donor. Each dot is a cell separated by the sampled library. All cells with a heteroplasmy of at least 2% are shown. **(h)** Depiction of unsupervised clustering of groups of cells based on shared somatic mtDNA mutations (y-axis) with corresponding individual mtDNA mutations (x-axis) associated with each cluster for the 500 cell input and **(i)** 800 cell input culture. Color bar, heteroplasmy (% allele frequency). **(j)** Fraction of cells (y-axis) carrying number of somatic mtDNA variants (x-axis) above indicated thresholds (≥1%, ≥5%, ≥10% heteroplasmy; red, black, and blue lines, respectively) for indicated cultures.

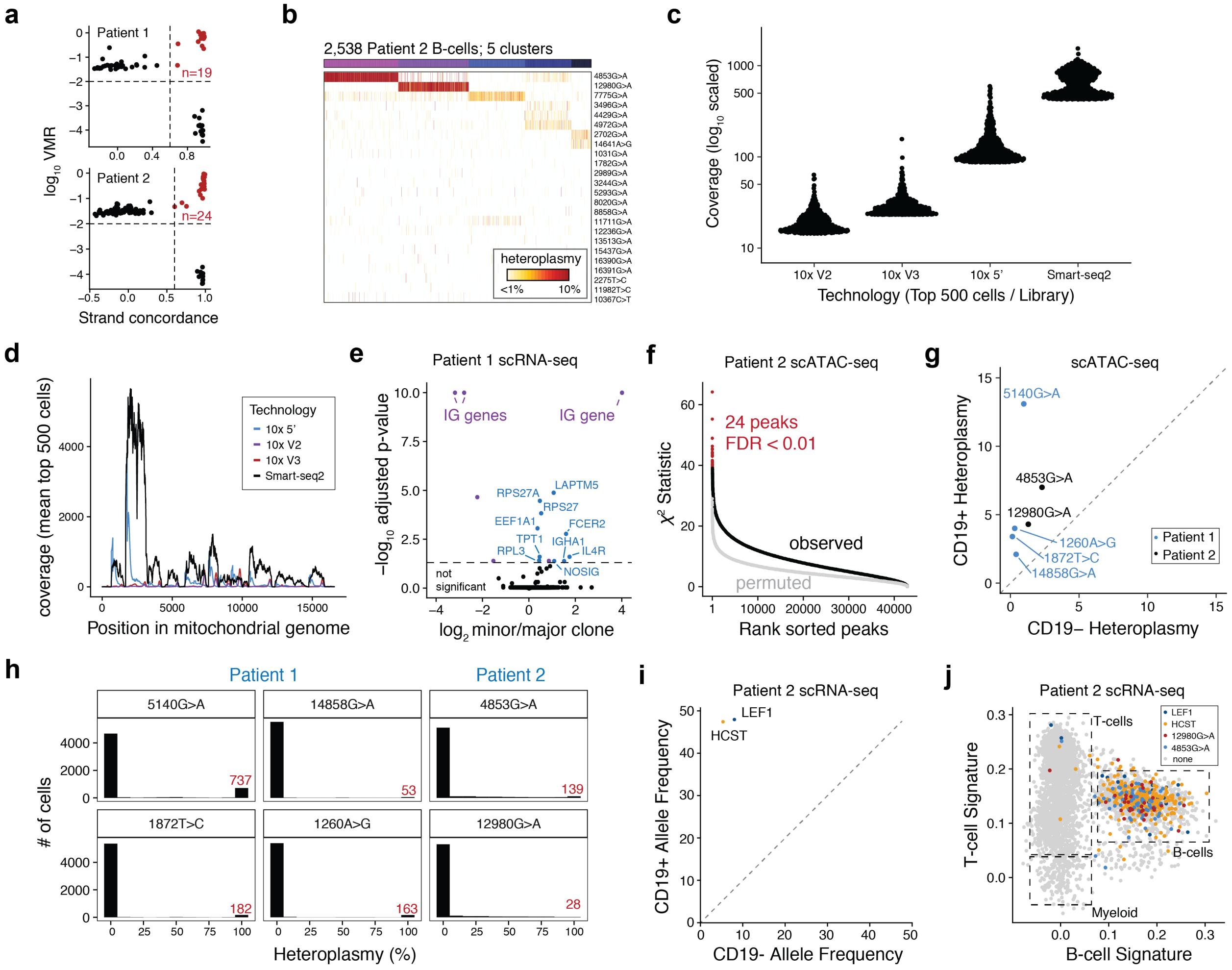

Figure E.5: Additional details for clonal and functional heterogeneity in chronic lymphocytic leukemia revealed by somatic mtDNA mutations. (a) Identification of high-confidence variants for Patient 1 (top) and Patient 2 (bottom). The number of variants is indicated. **(b)** Inference of subclonal structure from somatic mtDNA mutations for patient 2. Cells (columns) are clustered based on mitochondrial genotypes (rows). Colors at the top of the heatmap represent clusters or putative subclones. Color bar, heteroplasmy (% allele frequency). **(c)** Dot plots showing the mitochondrial genome coverage (log10; y-axis) for the top 500 cells per technology for four indicated scRNA-seq technologies. **(d)** The mean per-position mitochondrial genome coverage for the same 500 cells as in **(c)**. **(e)** Volcano plot showing differential gene expression analysis from major and minor clonotypes defined by BCR sequence. Immunoglobulin (IG) genes are shown in purple; all other genes with an FDR < 0.05 are shown in blue. **(f)** Histograms showing the distribution of heteroplasmy across the profiled population of cells for six selected variants, four from Patient 1 (left) and two from Patient 2 (right). The number of variants in the top heteroplasmy bin (>90%) are shown in red. **(g)** Heteroplasmy from the sum of single-cells in the CD19+ and CD19- mtscATAC-seq experiments for indicated mutations and patients. **(h)** Results for per-peak chi-squared association with subclonal group. Each dot is a peak rank-sorted by the chi-squared statistic. **(i)** Allele frequency from the sum of single cells from the 5' CD19+ and CD19- scRNA-seq libraries for two indicated variants - chr4:109,084,804A>C ("LEF1") and chr19:36,394,730G>A ("HSCT"). **(j)** Corroboration of T cells based on gene expression signatures and carrying indicated somatic nuclear and mtDNA mutations (patient 2).

serum. B lymphocytes were isolated using the negative selection Mojosort Human Pan B Cell Isolation Kit (Biolegend, 480082) and CD19 negative immune cells were isolated from a separate aliquot using the positive selection Mojosort Human CD19 selection Kit (Biolegend, 480106).

E.1.4 Flow cytometry analysis and sorting

For flow cytometry analysis and sorting cells were washed in FACS buffer (1% FBS in PBS) before antibody staining. For the CLL patient derived PBMC staining a FITC-conjugated CD19 antibody (HIB19, 302206, Biolegend) was used at 1:50 dilution. For live/ dead cell discrimination Sytox Blue was used according to the manufacturer's instructions (Thermo Fisher, S34857). FACS analysis was conducted on a BD Bioscience Fortessa flow cytometer at the Whitehead Institute Flow Cytometry core. Data were analyzed using FlowJo software v10.4.2. Cell sorting was conducted using the Sony SH800 sorter with a 100 µm chip at the Broad Institute Flow Cytometry Facility. Sytox Blue (ThermoFisher) was used for live/ dead cell discrimination.

E.2 Genomics methods

E.2.1 Single cell ATAC-seq (C1 Fluidigm)

The C1 Fluidigm platform using C1 single cell Auto Prep IFC for Open App and Open App Reagent Kit were used for the preparation of single cell ATAC-seq libraries as previously described20. Briefly, cells were washed and loaded at 350 cells/ml. Successful cell capture was monitored using a bright-field Nikon microscope and was >85%. Lysis and tagmentation reaction and 8 cycles of PCR were run on chip, followed by 13 cycles off chip using custom index primers and NEBNext High-Fidelity 2X PCR Master Mix (NEB). Individual libraries were pooled and purified using the MinElute PCR kit (QIAGEN) and quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

E.2.2 Single cell ATAC-seq and mtscATAC-seq

ScATAC-seq libraries were generated using the 10x Chromium Controller and the Chromium Single Cell ATAC Library and Gel Bead Kit (1000111) according to the manufacturer's instructions (CG000169-Rev C; CG000168-Rev B) or as detailed below with respect to the modifications enabling increased mtDNA yield and genome coverage. 1.5ml - 2ml DNA LoBind tubes (Eppendorf) were used to wash cells in PBS and downstream processing steps. After washing cells were fixed in 0.1 or 1% formaldehyde (FA; ThermoFisher 28906) in PBS for 10 min at RT, quenched with glycine solution to a final concentration of 0.125M before washing cells twice in PBS via centrifugation at 400g, 5 min, 4C. Cells were subsequently treated with lysis buffer (10mM Tris-HCL pH 7.4, 10mM NaCl, 3mM MgCl2, 0.1% NP40, 1% BSA) for 3 min for primary hematopoietic cells and 5 min for cell lines on ice, followed by adding 1ml of chilled wash buffer and inversion (10mM Tris-HCL pH 7.4, 10mM NaCl, 3mM MgCl2, 1% BSA) before centrifugation at 500g, 5 min, 4C. The supernatant was discarded and cells were diluted in 1x Diluted Nuclei buffer (10x Genomics) before counting using Trypan Blue and a Countess II FL Automated Cell Counter. If large cell clumps were observed a 40μm Flowmi cell strainer was used prior to processing cells according to the Chromium Single Cell ATAC Solution user guide with no additional modifications. Briefly, after tagmentation, the cells were loaded on a Chromium controller Single-Cell Instrument to generate single-cell Gel Bead-In-Emulsions (GEMs) followed by linear PCR as described in the protocol. Additional incubation (30 min to 12h) at 60C to further facilitate decrosslinking prior to the first 72C elongation step did not improve results (data not shown) and we suggest using the PCR conditions specified in the 10x scATAC-seq protocol. After breaking the GEMs, the barcoded tagmented DNA was purified and further amplified to enable sample indexing and enrichment of scATAC-seq libraries. The final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

E.2.3 Single cell RNA-seq

ScRNA-seq libraries were generated using the `10x` Chromium Controller and the Chromium Single Cell 5′ Library Construction Kit and human B cell and T cell V(D)J enrichment kit according to the manufacturer's instructions. Briefly, the suspended cells were loaded on a Chromium controller Single-Cell Instrument to generate single-cell Gel Bead-In-Emulsions (GEMs) followed by reverse transcription and sample indexing using a C1000 Touch Thermal cycler with 96-Deep Well Reaction Module (BioRad). After breaking the GEMs, the barcoded cDNA was purified and amplified, followed by fragmenting, A-tailing and ligation with adaptors. Finally, PCR amplification was performed to enable sample indexing and enrichment of scRNA-Seq libraries. For T cell and B cell receptor sequencing, target enrichment from cDNA was conducted according to the manufacturer's instructions. The final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

E.2.4 Sequencing

All libraries were sequenced using Nextseq High Output Cartridge kits and a Nextseq 500 sequencer (Illumina). `10x` scATAC-seq libraries were sequenced paired end (2 x 72 cycles). `10x` 5' scRNA-seq libraries were sequenced as recommended by the manufacturer.

E.2.5 Processing scATAC-seq data

Raw sequencing data was demuliplexed using CellRanger-ATAC mkfastq. Raw sequencing reads for all libraries were aligned to the hg19 reference genome using CellRanger-ATAC count. The raw output of the CellRanger-ATAC count execution, including the barcodes passing knee and the position-sorted `.bam` file served as inputs into the command-line interface of `mgatk`, which produces a PCR-deduplicated, per-cell, per-strand count of all alleles at all positions in the reference mitochondrial genome. To minimize the impact of barcode multiplets (see Chapter 4), we placed stringent thresholds on the mean mtDNA coverage per-barcode, which also provided greater confidence in the downstream heteroplasmy analyses.

E.3 Masked reference genome and NUMT comparison

To effectively assign putative multi-mapping reads to the mtDNA, we modified the existing CellRanger-ATAC reference genome by hard-masking nuclear mitochondrial DNA segments (NUMT). These regions were detected by simulating reads of length 20 from the reference mtDNA genome and encoding 1 base "errors" via the ART program60. Simulated reads were then aligned to the reference genome (with the mitochondrial chromosome excluded). As these reads were simulated to originate from the mtDNA genome but aligned to the nuclear genome, we hard masked these regions using bedtools (Quinlan & Hall, 2010). Comparisons of data from Figure 5.1 were performed by re-aligning the same datasets to the reference genome with and without masking. Complete documentation to reproduce the masking and modification of the CellRanger-ATAC reference genome are available as part of the `mgatk` wiki (https://github.com/caleblareau/mgatk/wiki).

To estimate the number of accessible NUMT fragments that would be assigned to mtDNA, we considered two different approaches. First, we used a public GM12878 dataset from 10x Genomics that was aligned to the standard hg19 reference and counted the number of fragments per cell overlapping our NUMT blacklisted regions, which resulted in a mean 1.4 and median 1.0 fragments per cell. Second, we used a compendium of DNase accessible peaks from 164 distinct samples from the ENCODE and Roadmap Consortia, and estimated that these samples contained a mean 22.6 peaks overlapping our NUMT blacklist. Next, using the GM12878 peakset and the same scATAC-seq dataset, we determined that a mean 4.1% of the GM12878 DNase peaks were detected over all cells. The product of these two numbers (22.6*0.041=0.93 fragments/cell) provides an alternative estimate for the number of accessible chromatin fragments overlapping NUMTs (~1 fragment) that were blacklisted. As our mtscATAC-seq assay generates ~$5,000 - 10,000$ mtDNA fragments, we conclude that our blacklist approach yields negligible NUMT contamination.

E.4 Comparison of experimental conditions

For all comparisons shown in the boxplots and violin plots, the top 1,000 cells/barcodes based on chromatin library complexity were plotted. The top 1,000 number was chosen to ensure the selection of real cells rather than barcode multiplets (see Chapter 4. For the overall coverage comparison (Figure 5.1.g), the top 2,000 cells based on nuclear complexity were averaged (to represent the 2,000 cells loaded). Cells were assigned TF1, doublet, or GM11906 using the sum of alleles at homoplasmic mitochondrial SNP loci (Figure E.1.d) using a 99% threshold for assignment to either major cell-type for our final protocol. We assigned barcodes as cell doublets (Figure 5.1.d,e) when this 99% threshold was not met for the major celltype. For both mtDNA and chromatin complexity estimation (Figure E.1.e), we used the number of unique and duplicate fragments as part of the CellRanger-ATAC (chromatin) and `mgatk` (mitochondria) output as inputs into the Lander-Waterman equation (Lander & Waterman, 1988), which estimates the total number of unique molecules present given these two measurements. Complexity measures were computed per barcode passing the knee filter from the default CellRanger-ATAC execution. To verify that cell type-specific accessible peaks were retained in mtscATAC-seq, we determined 77,704 peaks present in either the TF1 or GM11906 cell lines using the regular `10x` scATAC-seq conditions. These were determined from assigning barcodes to either cell line using mtDNA SNPs and calling peaks on the aggregate bulk population as previously described10. We repeated this peak calling procedure with our mtscATAC-seq data, identifying 72,887 peaks that overlapped the 77,704 peaks (93.8%).

E.5 Mitochondrial pathogenic variants

We queried MITOMAP(Lott et al., 2013) version `r102` and filtered for "Confirmed" pathogenic base-substitution variants. 46 variants were annotated to alter tRNA function whereas 42 were annotated to alter protein coding sequences in one or more protein-coding genes. Two additional variants were annotated to alter rRNA function.

E.5.1 In situ detection of mitochondrial heteroplasmy

All solutions below were prepared in 1x phosphate buffered saline (PBS), and incubations were carried out at RT unless otherwise specified. Two million GM11906 cells were fixed with 2 mL 1% paraformaldehyde for 10 min and quenched by adding 666 µL 1M Tris-HCl pH 8 and incubating for 5 min. Cells were then permeabilized with 0.5% Triton-X 100 for 20 min and embedded in 4% acrylamide gels63. The mitochondrial target sequence (on the antisense strand) was made accessible for hybridization by enzymatic removal of the sense strand64,65: restriction digest with 0.5 U/µL XbaI at 37C for 1 h, followed by adding 0.2 U/µL lambda exonuclease (both New England Biolabs) at 37C for 30 min. The oligonucleotide probe sequences against the wildtype and mutant alleles were pooled at 100 nM each in 2x SSC and 20% formamide, hybridized to the cell gels at 37C overnight, and circularized with 6U/µL T4 ligase (Enzymatics) for 2 hours. Rolling circle amplification, crosslinking, and *in situ* sequencing were performed as previously described (Lee et al., 2015). The cell gel was stained with DAPI (Thermo Fisher) and imaged on a Nikon Eclipse Ti microscope with a Yokogawa CSU-W1 confocal scanner unit and an Andor Zyla 4.2 Plus camera using a Nikon Plan Apo 60X/1.40 objective. Z-stack images spanning 24 µm at 0.4 µm intervals were acquired in the following channels: 405nm excitation with a 452/45 emission filter; 488nm excitation with a 525/50 emission filter; 561nm excitation with a 579/34 emission filter.

E.5.2 Image processing and heteroplasmy quantification

Each image stack was transformed into 2D by taking the maximum intensity projection across z-planes. Individual nuclei boundaries were defined by performing watershed segmentation on the DAPI staining. Wild-type and mutant probes were detected using a local maxima finder and uniquely assigned to individual cells based on spatial proximity. Probes that could not be unambiguously assigned to a cell were excluded from heteroplasmy and coverage measurements.

E.5.3 Epigenomic correlates with pathogenic heteroplasmy

To identify chromatin accessibility features associated with pathogenic heteroplasmy in the GM11906 cell line, we considered two approaches that complemented our estimation of heteroplasmy at the single-cell level. First, to assess cis-associations, we computed single-cell gene scores as previously described10,11 and computed per-gene associations with heteroplasmy using Spearman correlation (Figure 5.2.f). To establish a background distribution, we permuted heteroplasmy per-cell and recomputed the per-gene association statistic. We reported the number of gene scores correlated with heteroplasmy if the magnitude of the Spearman correlation exceeded 0.2. However, we note that a 1% false positive rate from the permutation testing would be a threshold of 0.087, resulting in 752 positively and 1,992 negatively correlated gene scores. We reported the more conservative results after examination of the accessible chromatin tracks where loci exceeding a magnitude 0.2 correlation revealed more robust peak differences. Second, to assess trans-associations, we downloaded a compendium of 78 high-quality ChIP-seq peak sets from lymphoblastoid cell lines from the ENCODE project (ENCODE Project Consortium, 2012). Per single-cell deviation scores were computed for these factors using chromVAR (Schep et al., 2017).

E.6 Variant calling and evaluation

To best identify informative clonal mutations from our mtscATAC-seq assay, we first considered existing variant calling approaches. Notably, algorithms designed for genotyping typically utilize a Bayesian framework to determine the empirical probability of a certain non-reference allele being truly observed at a particular location. In this setting, the ploidy of the genome is often parameterized in the model, and the allele frequency directly influences the confidence of detecting the mutation. As mtDNA copy number per cell is variable and informative clonal mutations may occur at very low allele frequencies, we found these existing approaches to be unsuitable for our mtscATAC-seq assay. Therefore, we developed a variant calling framework to identify high-confidence heteroplasmic mutations in a manner that 1) is largely independent of the mean allele frequency; 2) is robust to variability in genome ploidy of a cell; and 3) utilizes the features intrinsic to the high-throughput single-cell mtDNA data, including near-uniform deep coverage, minimal dropout per-cell, and thousands of single-cells per experiment. Our resulting variant calling framework, mgatk, achieves these goals.

Analyses of mtscATAC-seq from this manuscript revealed that certain positions with substantial heteroplasmy across biological diverse sources was primarily driven by sequencing error. These "recurrently-mutated" loci were due in part to several low-complexity stretches in the mitochondrial genome. However, by further evaluation of these variants, we determined that the erroneous heteroplasmy was primarily driven by one strand, reflective of a photobleaching effect from the sequencing machine. Hence, we devised the per-variant "strand concordance" value to capture the agreement of heteroplasmy between the strands, which is defined as the Pearson correlation between allele counts for all cells that have at least one count observed for the specific alternate allele. We note that for most variants, retention of all cells results in most observations being (0,0) for the strand correlation, inflating the statistics globally, making it less useful for discriminating variants. Additionally, while our approach works for mtscATAC-seq and full-length scRNA-seq methods (*e.g.* SMART-seq2), our approach is not appropriate for 3' scRNA-seq methods.

To compare our proposed variant calling approach to other tools, we analyzed the 855 TF1 single cells (Figure 5.3.) profiled in this manuscript. First, our execution of monovar (Zafar et al., 2016) failed as the genotype likelihood model is a function of a factorial of the max depth, which cannot be stored for the extremely deep coverage that results from our protocol. We then evaluated samtools/bcftools (Li, 2011) and FreeBayes (Garrison & Marth, 2012), treating each of the 855 cells as individual samples. To compare to mgatk (Figure E.3.a,b), the resulting .vcf files from each of these tools were filtered to remove clear homoplasmic variants and that had a variant quality ≥100. While our analyses indicated mgatk had greater sensitivity in resolving heteroplasmic variants informative for subclonal structure, relaxing this variant quality threshold did not improve detection of these informative variants and instead resulted in far more variants with strand discordance (Figure E.3.c). Finally, we acknowledge that other variant calling tools, such as GATK, utilize a Fisher's exact test to flag variants with high strand discordance that can be removed in downstream processing. We found this approach to be unsuitable for this data due to our high copy-number, resulting in extremely-small p-values for all variants, including those that clearly correlated with subclonal structure.

E.6.1 Evaluation of mgatk with SMART-seq2 data

To further benchmark our variant calling algorithm, we reanalyzed 895 high-quality cells from poly-clonal hematopoietic cells carrying somatic mtDNA mutations identified from SMART-seq2 scRNA-seq data (see Chapter 1). Previously aligned .bam files were re-processed with mgatk for each donor, and variant calling mirror the parameters established in the TF1 example (*i.e.* strand concordance ≥ 0.65; -log10(VMR) ≥ 2; see Figure E.3.g,h). From these samples, we had previously identified 78 variants showing subclonal structure using a supervised approach (*i.e.* the per-cell colony annotations were used in the identification of the variants). This set of 78 variants represents a "silver standard" as variants showed disproportionate heteroplasmy in a particular clone based on a Mann-Whitney U-test previously described (see Chapter 1).

Overall, mgatk identified 103 variants across the two donors. This set replicated 64 of the 76 (84.2%) previously identified sub-clonal variants. The variants that were not replicated were rarer in the population of cells (p=0.00045; Wilcoxen Rank-Sum Test; Figure E.3.f). While we generally believe the mgatk variant calling approach to be sensitive to low-frequency variants, we note that this supervised variant calling procedure (when clonal annotations are known) is theoretically better-powered to detect low-frequency mutations. However, we note that one previously-identified variant, 4214T>C, had only non-zero heteroplasmy on one strand, strongly suggestive of an artifactual variant that was nonetheless identified by our previous supervised approach.

To evaluate the efficacy of variant identification approaches for inferring clones, we tested their ability to correctly classify true-positive pairs of cells that were derived from the same clone (see Chapter 1). We computed per cell pair mtDNA similarity metric (the negation of our previous mitochondria distance; see Chapter 1), using mutations identified by three unsupervised approaches (mgatk, bcftools, and FreeBayes), as well as our previous supervised approach for each donor. Area under the receiver operating curve (AUROC, Figure E.3.g,h) were computed and can be interpreted as the efficacy of classifying pairs of cells from the same clone based on sets of mtDNA variants.

E.6.2 TF1 analyses

To identify putative subclones, we used the heteroplasmy matrix (capped at 10% as shown in Figure 5.3.c) as input into Principal Component Analysis. Next, we used the top 10 PCs as inputs into the FindNeighbors/ FindClusters functions from Seurat68 with default hyperparameters for these functions (k.param = 20; resolution = 0.8). In principle, this approach identifies communities of cells whose overall mutations are similar, and subclones are identified using a modularity optimization. Finally, we performed tree reconstruction using neighbor-joining on the distance between the average heteroplasmy of cells per clone using the phangorn R package (Schliep, 2011).

E.7 *In vitro* analyses

For each mtscATAC-seq library, cells were processed using CellRanger-ATAC with default settings, including the '–force-cells 6000' flag. Each library was further filtered such that cells had minimum 25% fragments in accessibility peaks, 1000 unique nuclear fragments, and 20x mtDNA coverage. Cutoffs were determined from examination of the density of each parameter. Somatic mtDNA mutations were identified using default thresholds from `mgatk`.

Clustering and embedding using Uniform Manifold Approximation and Projection70 (UMAP) were performed on the top 30 reduced dimensions from Latent Semantic Indexing (LSI) as previously described for the chromatin accessibility features (Cusanovich et al., 2018). Annotation of cell states were determined using transcription factor motif scoring via chromVAR(Schep et al., 2017) with default parameters, noting that the background peak selection was performed using all libraries merged. Pseudotime trajectories were defined using a semi-supervised approach from LSI and embedding as previously described (Granja et al., 2019).

To determine cell clones, we used the mutations by cells matrix per culture (capped at 10%) as input to the FindNeighbors/ FindClusters functions from Seurat with hyperparameters k.param = 10; resolution = 2, which yielded good separation of the rare cell clones. Clone-specific mutations were shown for all mutations exceeding

0.5% mean heteroplasmy in cell clones (Figure 5.4.i,j). We defined erythroid and monocytic cells in the day 20 library as those that exceeded a 0.5 pseudotime score along the specific axes (from Figure 5.4.c) and retained 65 clones from the 800 cell culture that had at least 10 total cells that were differentiated. To compute the lineage bias z-score (Figure 5.4.j), we computed the fraction of monocytic/erythroid labels in a cell clone and permuted these labels 100 times over the day 20 library. Finally, to infer putative lineage-priming chromatin accessibility, we identified 9 erythroid-biased and 22 monocytic-biased clones (z-score >5 from Figure 5.4.j) and computed the mean transcription factor deviation scores from the day 8 cells belonging to each clone. The difference in means between the erythroid and monocytic-biased clones is plotted (Figure 5.4.k).

E.8 Chronic lymphocytic leukemia scATAC analyses

For each mtscATAC-seq library, cells were processed using CellRanger-ATAC with default settings, including the '–force-cells 6000' flag. Each library was further filtered such that cells had minimum 50% fragments in accessibility peaks, 1000 unique nuclear fragments, and 20x mtDNA coverage. Somatic mtDNA mutations were identified using `mgatk` with the default parameters for the CD19 positive cells profiled with mtscATAC-seq (Figure 5.5.a). Putative sub-clones were identified using the mutations for patient 1 (n=19) and patient 2 (n=24) separately using the FindNeighbors/ FindClusters functions from Seurat where the heteroplasmy matrix was capped at 10%. We used the default parameters for patient 1 (k.param = 20; resolution = 0.8 Figure 5.5.c) and modified parameters for patient 2 (k.param = 50; resolution = 1.5; Figure E.5.b) to effectively identify subclones. For visualization of cell x mutation heatmaps, subsets of cells from patient 1 (2,368/5,631; Figure 5.5.c) and patient 2 (2,538/5,865; Figure E.5.b) were visualized as the remaining cells had 0% heteroplasmy at all called mutations.

To determine copy number alterations (Figure 5.5.e), we first constructed overlapping 10Mb bins genome-wide using a step size of 2Mb. Next, we overlapped the .fragments.tsv file from the `10x` CellRanger-ATAC output with these bins to compute a bin x cell matrix for both the CLL samples as well as a healthy control PBMC sample. Next, we computed a per-cell, per-bin z-score of the number of fragments after normalizing each cell to a consistent sequencing depth. The chromosome 12 z-score (Figure 5.5.e) represents the per-cell mean of the z-scores from the bins mapping to chromosome 12.

To identify chromatin accessibility peaks associated with mtDNA mutation-derived subclones, we performed a series of χ^2 association tests. After binarizing the chromatin accessibility count per-peak, per-cell, a contingency table of dimension $n * 2$ was assembled, where n is the number of subclones per tumor. The resulting chi-squared statistics were associated with p-values using $n - 1$ degrees of freedom, and correction for multiple testing was performed using the Benjamini–Hochberg procedure. To further visualize a null association statistics, we permuted the subclone annotations per peak to visualize a null distribution of the chi-square statistics (see gray from Figure 5.5.f; Figure E.5.f). The TIAM1 and ZNF257 loci were selected based on strong association (both in the top 10 most-associated peaks) and proximity to annotated transcription start sites.

To identify non-B-cells with mtDNA mutations, we first embedded a healthy PBMC 5k sample from the 10x Genomics public dataset using LSI and UMAP as previously described (Granja et al., 2019). Using the LSI components the projection capability of UMAP, we projected CD19 negative cells from both CLL donors onto the reduced dimension space (Figure 5.5.j,k). Cells were annotated as positive for specific mtDNA mutations if the heteroplasmy exceeded 20% (corresponding to at least 4 unique molecules containing the alternate allele; Figure 5.5.j,k).

E.8.1 Exome sequencing

Enriched CLL cells and in vitro expanded CD3+ T lymphocytes to serve as a germline control were subjected to whole exome sequencing using the clinical somatic exome workflow through the Broad Institute Genomics Platform. The exome product targets 35.1 Mb with a total bait size of 38.9 Mb and are optimized to cover the following: 99% of ClinVar variants; complete Mitochondrial genome; full ACMG59 gene list; Online Mendelian Inheritance in Man (OMIM) putative gene sequences; Catalogue of Somatic Mutations in Cancer (COSMIC) variants; Internal 'ONCO Panel' and additional key promoters and other motifs that have been identified as potential cancer hot spots. Automated library preparation occurs as follows. Samples were plated at a concentration of 2 ng/µl and volume of 50 µl (total 100ng input) into fresh matrix tubes allowing positive barcode tracking throughout the process. Library Construction: Samples were sheared to yield ~180 bp size distribution. cfDNA samples do not proceed through this step. Kapa Hyperprep kits were used to construct libraries in a process optimized for somatic samples, involving end repair, adapter ligation with forked adaptors containing unique molecular indexes and addition of P5 and P7 sample barcodes via PCR. After SPRI purification libraries were quantified with Pico Green. Libraries were normalized and equimolar pooling was performed to prepare multiplexed sets for hybridization. Sample pools were then split and hybridized in up to 8 separate reaction wells to accommodate volumes. Automated capture was performed, followed by PCR of the enriched DNA and SPRI purification. Post-capture QC: Multiplex pools were quantified with Pico Green and DNA fragment size was estimated using Bioanalyzer electrophoresis. Sequencing: Final libraries were quantitated by qPCR and loaded across the appropriate number of Illumina flow cell lanes to achieve the target coverage. Completed exomes contained >= 85% of target bases covered at >= 50x depth and ranged from 130-160x mean coverage of the targeted region. Both tumor and normal samples were processed and used for variant identification.

E.8.2 CLL scRNA-seq analyses

5' scRNA-seq libraries, including VDJ sequencing, were processed using default parameters with CellRanger 3.1.0. Mitochondrial genotyping was conducted using `mgatk` with the "–umi-barcode" tag specifying the SAM tag from the CellRanger `.bam` output marking the error-corrected UMI barcode. Cell-type specific signatures (Figure 5.5.k; Figure E.5.j) were computed using Seurat's AddModuleScore68 where gene bins were computed on a control set of healthy PBMCs. Cell-type specific genes were determined from the Immune Cell Atlas (available here: https://github.com/caleblareau/immune_cell_signature_genes). Two nuclear variants, chr4:109,084,804A>C ("LEF1"; p.S112A) and chr19:36,394,730G>A ("HSCT"; p.A56T), encoded missense mutations that were detected using whole-exome sequencing and somatic mutation calling. These mutations were covered by the 5' scRNA-seq libraries, enabling single-cell examination (Figure E.5.j). Cells were annotated as positive for mtDNA mutations (Figure 5.5.l; Figure 5.5.j) when supported by at least two distinct UMIs.

References

Abate, A. R., Chen, C.-H., Agresti, J. J., & Weitz, D. A. (2009). Beating poisson encapsulation statistics using close-packed ordering. *Lab Chip*, 9(18), 2628–2631.

Alizadeh, A. A. & Majeti, R. (2011). Surprise! HSC are aberrant in chronic lymphocytic leukemia. *Cancer Cell*, 20(2), 135–136.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L., & Steemers, F. J. (2014). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.*, 46(12), 1343–1349.

Bar-Yaacov, D., Avital, G., Levin, L., Richards, A. L., Hachen, N., Rebolledo Jaramillo, B., Nekrutenko, A., Zarivach, R., & Mishmar, D. (2013). Rna-dna differences in human mitochondria restore ancestral form of 16s ribosomal rna. *Genome Res*, 23(11), 1789–96.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-A. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., & Nolan, G. P. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030), 687–696.

Bender, A., Krishnan, K. J., Morris, C. M., Taylor, G. A., Reeve, A. K., Perry, R. H., Jaros, E., Hersheson, J. S., Betts, J., Klopstock, T., Taylor, R. W., & Turnbull, D. M. (2006). High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and parkinson disease. *Nat. Genet.*, 38(5), 515–517.

Biasco, L., Pellin, D., Scala, S., Dionisio, F., Basso-Ricci, L., Leonardelli, L., Scaramuzza, S., Baricordi, C., Ferrua, F., Cicalese, M. P., Giannelli, S., Neduva, V., Dow, D. J., Schmidt, M., Von Kalle, C., Roncarolo, M. G., Ciceri, F., Vicard, P., Wit, E., Di Serio, C., Naldini, L., & Aiuti, A. (2016). In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and Steady-State reconstitution phases. *Cell Stem Cell*, 19(1), 107–119.

Biezuner, T., Spiro, A., Raz, O., Amir, S., Milo, L., Adar, R., Chapal-Ilani, N., Berman, V., Fried, Y., Ainbinder, E., Cohen, G., Barr, H. M., Halaban, R., & Shapiro, E. (2016). A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res*, 26(11), 1588–1599.

Bodenmiller, B., Zunder, E. R., Finck, R., Chen, T. J., Savig, E. S., Bruggner, R. V., Simonds, E. F., Bendall, S. C., Sachs, K., Krutzik, P. O., & Nolan, G. P. (2012). Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.*, 30(9), 858–867.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., & Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2), 311–322.

Brusco, J. & Haas, K. (2015). Interactions between mitochondria and the transcription factor myocyte enhancer factor 2 (MEF2) regulate neuronal structural and functional plasticity and metaplasticity. *J. Physiol.*, 593(16), 3471–3481.

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., Majeti, R., Chang, H. Y., & Greenleaf, W. J. (2018). Integrated Single-Cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6), 1535–1548.e16.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12), 1213–1218.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561), 486–490.

Caicedo, A., Fritz, V., Brondello, J. M., Ayala, M., Dennemont, I., Abdellaoui, N., de Fraipont, F., Moisan, A., Prouteau, C. A., Boukhaddaoui, H., Jorgensen, C., & Vignais, M. L. (2015). Mitoception as a new tool to assess the effects of mesenchymal stem/stromal cell mitochondria on cancer cell metabolism and function. *Sci Rep*, 5, 9073.

Calo, E. & Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, 49(5), 825–837.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409), 1380–1385.

Chen, R., Xia, L., Tu, K., Duan, M., Kukurba, K., Li-Pook-Than, J., Xie, D., & Snyder, M. (2018a). Longitudinal personal DNA methylome dynamics in a human with a chronic condition. *Nat. Med.*, 24(12), 1930–1939.

Chen, X., Miragaia, R. J., Natarajan, K. N., & Teichmann, S. A. (2018b). A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.*, 9(1), 5345.

Consortium, G., Aguet, F., Brown, A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., Nguyen, D. Y., & Noble, M. S. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675), 204–213.

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., & Chang, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, 48(10), 1193–1203.

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., Montine, T. J., Greenleaf, W. J., & Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*, 14(10), 959–962.

Corral-Debrinski, M., Horton, T., Lott, M. T., Shoffner, J. M., McKee, A. C., Beal, M. F., Graham, B. H., & Wallace, D. C. (1994). Marked changes in mitochondrial DNA deletion levels in alzheimer brains. *Genomics*, 23(2), 471–476.

Cui, J.-H., Lin, K.-R., Yuan, S.-H., Jin, Y.-B., Chen, X.-P., Su, X.-K., Jiang, J., Pan, Y.-M., Mao, S.-L., Mao, X.-F., & Luo, W. (2018). TCR repertoire as a novel indicator for immune monitoring and prognosis assessment of patients with cervical cancer. *Front. Immunol.*, 9, 2729.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237), 910–914.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., & Shendure, J. (2018). A Single-Cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5), 1309–1324.e18.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., & Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol*, 29(12), 1120–7.

Dames, S., Chou, L.-S., Xiao, Y., Wayman, T., Stocks, J., Singleton, M., Eilbeck, K., & Mao, R. (2013). The development of next-generation sequencing assays for the mitochondrial genome and 108 nuclear genes associated with mitochondrial disorders. *J. Mol. Diagn.*, 15(4), 526–534.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting molecular circuits with scalable Single-Cell RNA profiling of pooled genetic screens. *Cell*, 167(7), 1853–1866.e17.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21.

Elliott, H. R., Samuels, D. C., Eden, J. A., Relton, C. L., & Chinnery, P. F. (2008). Pathogenic mitochondrial DNA mutations are common in the general population. *Am. J. Hum. Genet.*, 83(2), 254–260.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.

Espín-Palazón, R., Stachura, D. L., Campbell, C. A., García-Moreno, D., Del Cid, N., Kim, A. D., Candel, S., Meseguer, J., Mulero, V., & Traver, D. (2014). Proinflammatory signaling regulates hematopoietic stem cell emergence. *Cell*, 159(5), 1070–1085.

Essers, M. A. G., Offner, S., Blanco-Bose, W. E., Waibler, Z., Kalinke, U., Duchosal, M. A., & Trumpp, A. (2009). IFNalpha activates dormant haematopoietic stem cells in vivo. *Nature*, 458(7240), 904–908.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., & Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539), 337–343.

Gao, H., Sun, B., Fu, H., Chi, X., Wang, F., Qi, X., Hu, J., & Shao, S. (2016). Pdia6 promotes the proliferation of hela cells through activating the wnt/beta-catenin signaling pathway. *Oncotarget*, 7(33), 53289–53298.

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J. J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., & Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414), 91–100.

Giladi, A. & Amit, I. (2018). Single-cell genomics: A stepping stone for future immunology discoveries. *Cell*, 172(1-2), 14–21.

Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P. S., Povinelli, B. J., Booth, C. A. G., Sopp, P., Norfo, R., Rodriguez-Meira, A., Ashley, N., Jamieson, L., Vyas, P., Anderson, K., Segerstolpe, A., Qian, H., Olsson-Stromberg, U., Mustjoki, S., Sandberg, R., Jacobsen, S. E. W., & Mead, A. J. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med*, 23(6), 692–702.

Granja, J. M., Klemm, S., McGinnis, L. M., Kathiria, A. S., Mezger, A., Corces, M. R., Parks, B., Gars, E., Liedtke, M., Zheng, G. X. Y., Chang, H. Y., Majeti, R., & Greenleaf, W. J. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.*, 37(12), 1458–1465.

Griessinger, E., Moschoi, R., Biondani, G., & Peyron, J. F. (2017). Mitochondrial transfer in the leukemia microenvironment. *Trends Cancer*, 3(12), 828–839.

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., Kang, B., Liu, Z., Jin, L., Xing, R., Gao, R., Zhang, L., Dong, M., Hu, X., Ren, X., Kirchhoff, D., Roider, H. G., Yan, T., & Zhang, Z. (2018). Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med*, 24(7), 978–985.

Han, K. Y., Kim, K. T., Joung, J. G., Son, D. S., Kim, Y. J., Jo, A., Jeon, H. J., Moon, H. S., Yoo, C. E., Chung, W., Eum, H. H., Kim, S., Kim, H. K., Lee, J. E., Ahn, M. J., Lee, H. O., Park, D., & Park, W. Y. (2018a). Sidr: simultaneous isolation and parallel sequencing of genomic dna and total rna from single cells. *Genome Res*, 28(1), 75–87.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G.-C., Chen, M., & Guo, G. (2018b). Mapping the mouse cell atlas by Microwell-Seq. *Cell*, 173(5), 1307.

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, 9(1), 75.

Hatori, M., Gill, S., Mure, L. S., Goulding, M., O'Leary, D. D. M., & Panda, S. (2014). Lhx1 maintains synchrony among circadian oscillator neurons of the SCN. *Elife*, 3, e03357.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., & Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, 6(4), 283–289.

Hofbauer, S. W., Krenn, P. W., Ganghammer, S., Asslaber, D., Pichler, U., Oberascher, K., Henschler, R., Wallner, M., Kerschbaum, H., Greil, R., & Hartmann, T. N. (2014). Tiam1/Rac1 signals contribute to the proliferation and chemoresistance, but not motility, of chronic lymphocytic leukemia cells.

Izumi, D., Toden, S., Ureta, E., Ishimoto, T., Baba, H., & Goel, A. (2019). TIAM1 promotes chemoresistance and tumor invasiveness in colorectal cancer.

Jacobsen, S. E. W. & Nerlov, C. (2019). Haematopoiesis in the era of advanced single-cell technologies.

Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T. M., Childs, R., Levens, D., & Zhao, K. (2015). Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, 528(7580), 142–146.

Ju, Y. S., Alexandrov, L. B., Gerstung, M., Martincorena, I., Nik-Zainal, S., Ramakrishna, M., Davies, H. R., Papaemmanuil, E., Gundem, G., Shlien, A., Bolli, N., Behjati, S., Tarpey, P. S., Nangalia, J., Massie, C. E., Butler, A. P., Teague, J. W., Vassiliou, G. S., Green, A. R., Du, M. Q., Unnikrishnan, A., Pimanda, J. E., Teh, B. T., Munshi, N., Greaves, M., Vyas, P., El-Naggar, A. K., Santarius, T., Collins, V. P., Grundy, R., Taylor, J. A., Hayes, D. N., Malkin, D., Group, I. B. C., Group, I. C. M. D., Group, I. P. C., Foster, C. S., Warren, A. Y., Whitaker, H. C., Brewer, D., Eeles, R., Cooper, C., Neal, D., Visakorpi, T., Isaacs, W. B., Bova, G. S., Flanagan, A. M., Futreal, P. A., Lynch, A. G., Chinnery, P. F., McDermott, U., Stratton, M. R., & Campbell, P. J. (2014). Origins and functional consequences of somatic mitochondrial dna mutations in human cancer. *Elife*, 3.

Ju, Y. S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L. B., Rahbari, R., Wedge, D. C., Davies, H. R., Ramakrishna, M., Fullam, A., Martin, S., Alder, C., Patel, N., Gamble, S., O'Meara, S., Giri, D. D., Sauer, T., Pinder, S. E., Purdie, C. A., Borg, A., Stunnenberg, H., van de Vijver, M., Tan, B. K., Caldas, C., Tutt, A., Ueno, N. T., van 't Veer, L. J., Martens, J. W., Sotiriou, C., Knappskog, S., Span, P. N., Lakhani, S. R., Eyfjord, J. E., Borresen-Dale, A. L., Richardson, A., Thompson, A. M., Viari, A., Hurles, M. E., Nik-Zainal, S., Campbell, P. J., & Stratton, M. R. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647), 714–718.

Kadkhodaei, B., Ito, T., Joodmardi, E., Mattsson, B., Rouillard, C., Carta, M., Muramatsu, S.-I., Sumi-Ichinose, C., Nomura, T., Metzger, D., Chambon, P., Lindqvist, E., Larsson, N.-G., Olson, L., Björklund, A., Ichinose, H., & Perlmann, T. (2009). Nurr1 is required for maintenance of maturing and adult midbrain dopamine neurons. *J. Neurosci.*, 29(50), 15923–15932.

Kang, E., Wang, X., Tippner-Hedges, R., Ma, H., Folmes, C. D. L., Gutierrez, N. M., Lee, Y., Van Dyken, C., Ahmed, R., Li, Y., Koski, A., Hayama, T., Luo, S., Harding, C. O., Amato, P., Jensen, J., Battaglia, D., Lee, D., Wu, D., Terzic, A., Wolf, D. P., Huang, T., & Mitalipov, S. (2016). Age-Related accumulation of somatic mitochondrial DNA mutations in Adult-Derived human iPSCs. *Cell Stem Cell*, 18(5), 625–636.

Kelsey, G., Stegle, O., & Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. *Science*, 358(6359), 69–75.

Kester, L. & van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell*.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., & Hemberg, M. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nat Methods*, 14(5), 483–486.

Klein, A. M. & Macosko, E. (2017). InDrops and drop-seq technologies for single-cell sequencing. *Lab Chip*, 17(15), 2540–2541.

Krueger, F. & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11), 1571–1572.

Kugeratski, F. G., Atkinson, S. J., Neilson, L. J., Lilla, S., Knight, J. R. P., Serneels, J., Juin, A., Ismail, S., Bryant, D. M., Markert, E. K., Machesky, L. M., Mazzone, M., Sansom, O. J., & Zanivan, S. (2019). Hypoxic cancer–associated fibroblasts increase NCBP2-AS2/HIAR to promote endothelial sprouting through enhanced VEGF signaling. *Sci. Signal.*, 12(567), eaan8247.

Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3), 231–239.

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4), 357–359.

Laroni, A., Armentani, E., Kerlero de Rosbo, N., Ivaldi, F., Marcenaro, E., Sivori, S., Gandhi, R., Weiner, H. L., Moretta, A., Mancardi, G. L., & Uccelli, A. (2016). Dysregulation of regulatory CD56(bright) NK cells/t cells interactions in multiple sclerosis. *J. Autoimmun.*, 72, 8–18.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2), R29.

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., Turczyk, B. M., Yang, J. L., Lee, H. S., Aach, J., Zhang, K., & Church, G. M. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3), 442–458.

Lee, S. R. & Han, J. (2017). Mitochondrial mutations in cardiac disorders. *Adv. Exp. Med. Biol.*, 982, 81–111.

Lee-Six, H., Øbro, N. F., Shepherd, M. S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R. J., Huntly, B. J. P., Martincorena, I., Anderson, E., O'Neill, L., Stratton, M. R., Laurenti, E., Green, A. R., Kent, D. G., & Campbell, P. J. (2018). Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724), 473–478.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6), 882–3.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993.

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D'Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J., & Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256), 94–98.

Lopez, R. D., Waller, E. K., Lu, P. H., & Negrin, R. S. (2001). CD58/LFA-3 and IL-12 provided by activated monocytes are critical in the in vitro expansion of CD56+ T cells. *Cancer Immunol. Immunother.*, 49(12), 629–640.

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., Procaccio, V., & Wallace, D. C. (2013). mtDNA variation and analysis using mitomap and mitomaster. *Curr. Protoc. Bioinformatics*, 44, 1.23.1–26.

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., participants in the 1st Human Cell Atlas Jamboree, & Marioni, J. C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, 20(1), 63.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., & Voet, T. (2015). G&t-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods*, 12(6), 519–22.

Marlein, C. R., Zaitseva, L., Piddock, R. E., Robinson, S. D., Edwards, D. R., Shafat, M. S., Zhou, Z., Lawes, M., Bowles, K. M., & Rushworth, S. A. (2017). Nadph oxidase-2 derived superoxide drives mitochondrial transfer from bone marrow stromal cells to leukemic blasts. *Blood*, 130(14), 1649–1660.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., & Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190–1195.

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet detection in Single-Cell RNA sequencing data using artificial nearest neighbors. *Cell Syst*, 8(4), 329–337.e4.

McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., & Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298), aaf7907.

Meneses, A. (2015). Serotonin, neural markers, and memory. *Front. Pharmacol.*, 6, 143.

Mishra, P. & Chan, D. C. (2014). Mitochondrial dynamics and inheritance during cell division, development and disease. *Nature reviews Molecular cell biology*, 15(10), 634–646.

Morris, J., Na, Y.-J., Zhu, H., Lee, J.-H., Giang, H., Ulyanova, A. V., Baltuch, G. H., Brem, S., Chen, H. I., Kung, D. K., Lucas, T. H., O'Rourke, D. M., Wolf, J. A., Grady, M. S., Sul, J.-Y., Kim, J., & Eberwine, J. (2017). Pervasive within-mitochondrion Single-Nucleotide variant heteroplasmy as revealed by Single-Mitochondrion sequencing. *Cell Rep.*, 21(10), 2706–2713.

Moschoi, R., Imbert, V., Nebout, M., Chiche, J., Mary, D., Prebet, T., Saland, E., Castellano, R., Pouyet, L., Collette, Y., Vey, N., Chabannon, C., Recher, C., Sarry, J. E., Alcor, D., Peyron, J. F., & Griessinger, E. (2016). Protective mitochondrial transfer from bone marrow stromal cells to acute myeloid leukemic cells during chemotherapy. *Blood*, 128(2), 253–64.

Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J., & Adey, A. C. (2018). Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.*, 36(5), 428–431.

Nam, A. S., Kim, K.-T., Chaligne, R., Izzo, F., Ang, C., Taylor, J., Myers, R. M., Abu-Zeinah, G., Brand, R., Omans, N. D., Alonso, A., Sheridan, C., Mariani, M., Dai, X., Harrington, E., Pastore, A., Cubillos-Ruiz, J. R., Tam, W., Hoffman, R., Rabadan, R., Scandura, J. M., Abdel-Wahab, O., Smibert, P., & Landau, D. A. (2019). Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature*, 571(7765), 355–360.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*, 12(5), 453–7.

Ni, T., Wei, G., Shen, T., Han, M., Lian, Y., Fu, H., Luo, Y., Yang, Y., Liu, J., Wakabayashi, Y., Li, Z., Finkel, T., Xu, H., & Zhu, J. (2015). Mitorca-seq reveals unbalanced cytocine to thymine transition in polg mutant mice. *Sci Rep*, 5, 12049.

Orkin, S. H. & Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4), 631–644.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suva, M. L., Regev, A., & Bernstein, B. E. (2014). Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190), 1396–401.

Pei, W., Feyerabend, T. B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., Chen, W., Sauer, S., Wolf, S., Höfer, T., & Rodewald, H.-R. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, 548(7668), 456–460.

Picelli, S., Björklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, 24(12), 2033–2040.

Plasschaert, L. W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A. M., & Jaffe, A. B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature*, 560(7718), 377–381.

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., & Trapnell, C. (2018). Cicero predicts cis-regulatory DNA interactions from Single-Cell chromatin accessibility data. *Mol. Cell*, 71(5), 858–871.e8.

Powell, C. A., Kopajtich, R., D'Souza, A. R., Rorbach, J., Kremer, L. S., Husain, R. A., Dallabona, C., Donnini, C., Alston, C. L., Griffin, H., Pyle, A., Chinnery, P. F., Strom, T. M., Meitinger, T., Rodenburg, R. J., Schottmann, G., Schuelke, M., Romain, N., Haller, R. G., Ferrero, I., Haack, T. B., Taylor, R. W., Prokisch, H., & Minczuk, M. (2015). TRMT5 mutations cause a defect in post-transcriptional modification of mitochondrial tRNA associated with multiple Respiratory-Chain deficiencies. *Am. J. Hum. Genet.*, 97(2), 319–328.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K., & Ren, B. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.*, 21(3), 432–439.

Qu, K., Zaba, L. C., Satpathy, A. T., Giresi, P. G., Li, R., Jin, Y., Armstrong, R., Jin, C., Schmitt, N., Rahbar, Z., Ueno, H., Greenleaf, W. J., Kim, Y. H., & Chang, H. Y. (2017). Chromatin accessibility landscape of cutaneous T cell lymphoma and dynamic response to HDAC inhibitors. *Cancer Cell*, 32(1), 27–41.e4.

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842.

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A., & Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.*, 36(5), 442–450.

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, 12(8), 529–541.

Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., Bodenmiller, B., Bruneau, B., Campbell, P., Carmichael, M., Carninci, P., Castelo-Soccio, L., Clatworthy, M., Clevers, H., Conrad, C., Eils, R., Freeman, J., Fugger, L., Goettgens, B., Graham, D., Greka, A., Hacohen, N., Haniffa, M., Helbig, I., Heuckeroth, R., Kathiresan, S., Kim, S., Klein, A., Knoppers, B., Kriegstein, A., Lander, E., Lee, J., Lein, E., Linnarsson, S., Macosko, E., MacParland, S., Majovski, R., Majumder, P., Marioni, J., McGilvray, I., Merad, M., Mhlanga, M., Naik, S., Nawijn, M., Nolan, G., Paten, B., Pe'er, D., Philippakis, A., Ponting, C., Quake, S., Rajagopal, J., Rajewsky, N., Reik, W., Rood, J., Saeb-Parsy, K., Schiller, H., Scott, S., Shalek, A., Shapiro, E., Shin, J., Skeldon, K., Stratton, M., Streicher, J., Stunnenberg, H., Tan, K., Taylor, D., Thorogood, A., Vallier, L., van Oudenaarden, A., Watt, F., Weicher, W., Weissman, J., Wells, A., Wold, B., Xavier, R., Zhuang, X., & Human Cell Atlas Organizing Committee (2018). The human cell atlas white paper. *eLife*.

Rendeiro, A. F., Krausgruber, T., Fortelny, N., Zhao, F., Penz, T., Farlik, M., Schuster, L. C., Nemc, A., Tasnády, S., Réti, M., Mátrai, Z., Alpar, D., Bödör, C., Schmidl, C., & Bock, C. (2020). Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib drug response in chronic lymphocytic leukemia. Nature Communications.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., & Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317–330.

Rodriguez-Fraticelli, A. E., Wolock, S. L., Weinreb, C. S., Panero, R., Patel, S. H., Jankovic, M., Sun, J., Calogero, R. A., Klein, A. M., & Camargo, F. D. (2018). Clonal analysis of lineage fate in native haematopoiesis. *Nature*, 553(7687), 212–216.

Roos-Weil, D., Nguyen-Khac, F., Chevret, S., Touzeau, C., Roux, C., Lejeune, J., Cosson, A., Mathis, S., Feugier, P., Leprêtre, S., Béné, M.-C., Baron, M., Raynaud, S., Struski, S., Eclache, V., Sutton, L., Lesty, C., Merle-Béral, H., Cymbalista, F., Ysebaert, L., Davi, F., Leblond, V., & on behalf of the FILO working group (2018). Mutational and cytogenetic analyses of 188 CLL patients with trisomy 12: A retrospective study from the french innovative leukemia organization (FILO) working group.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*, 14(5), R51.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, 33(11), 1165–1172.

Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., & others (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *BioRxiv*.

Saunders, A., Macosko, E. Z., Wysoker, A., Goldman, M., Krienen, F. M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., Goeva, A., Nemesh, J., Kamitaki, N., Brumbaugh, S., Kulp, D., & McCarroll, S. A. (2018). Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, 174(4), 1015–1030.e16.

Scala, S. & Aiuti, A. (2019). In vivo dynamics of human hematopoietic stem cells: novel concepts and future directions.

Scala, S., Basso-Ricci, L., Dionisio, F., Pellin, D., Giannelli, S., Salerio, F. A., Leonardelli, L., Cicalese, M. P., Ferrua, F., Aiuti, A., & Biasco, L. (2018). Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.*, 24(11), 1683–1690.

Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, 14(10), 975–978.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592–593.

Shema, E., Bernstein, B. E., & Buenrostro, J. D. (2019). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nature genetics*, 51(1), 19–25.

Shoffner, J. M. & Wallace, D. C. (1992). Mitochondrial genetics: principles and practice. *Am. J. Hum. Genet.*, 51(6), 1179–1186.

Simone, M. D., De Simone, M., Rossetti, G., & Pagani, M. (2018). Single cell T cell receptor sequencing: Techniques and future challenges.

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3), 491–499.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., & Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using crispr-cas9-induced genetic scars. *Nat Biotechnol*, 36(5), 469–473.

Spitz, F. & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9), 613–626.

Stewart, J. B. & Chinnery, P. F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.*, 16(9), 530–542.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, 3rd, W. M., Smibert, P., & Satija, R. (2018). Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.*, 19(1), 224.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, 3rd, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of Single-Cell data. *Cell*, 177(7), 1888–1902.e21.

Stubbington, M. J. T., Lonnberg, T., Proserpio, V., Clare, S., Speak, A. O., Dougan, G., & Teichmann, S. A. (2016). T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods*, 13(4), 329–332.

Sun, J., Ramos, A., Chapman, B., Johnnidis, J. B., Le, L., Ho, Y. J., Klein, A., Hofmann, O., & Camargo, F. D. (2014). Clonal dynamics of native haematopoiesis. *Nature*, 514(7522), 322–7.

Tao, L., Raz, O., Marx, Z., Biezuner, T., Amir, S., & Milo, L. (2017). A duplex mips-based biological-computational cell lineage discovery platform. *BioRxiv*.

Taylor, R. W., Barron, M. J., Borthwick, G. M., Gospel, A., Chinnery, P. F., Samuels, D. C., Taylor, G. A., Plusa, S. M., Needham, S. J., Greaves, L. C., Kirkwood, T. B., & Turnbull, D. M. (2003). Mitochondrial dna mutations in human colonic crypt stem cells. *J Clin Invest*, 112(9), 1351–60.

Teixeira, V. H., Nadarajan, P., Graham, T. A., Pipinikas, C. P., Brown, J. M., Falzon, M., Nye, E., Poulsom, R., Lawrence, D., Wright, N. A., McDonald, S., Giangreco, A., Simons, B. D., & Janes, S. M. (2013). Stochastic homeostasis in human airway epithelium is achieved by neutral competition of basal cell progenitors. *Elife*, 2, e00966.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., & Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414), 75–82.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, 2nd, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A.-C., Johannessen, C. M., Andreev, A. Y., Van Allen, E. M., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jané-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A., & Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282), 189–196.

Torralba, D., Baixauli, F., & Sanchez-Madrid, F. (2016). Mitochondria know no boundaries: Mechanisms and functions of intercellular mitochondrial transfer. *Front Cell Dev Biol*, 4, 107.

Triska, P., Kaneva, K., Merkurjev, D., Sohail, N., Falk, M. J., Triche, T. J., Biegel, J. A., & Gai, X. (2019). Landscape of germline and somatic mitochondrial DNA mutations in pediatric malignancies. *Cancer Res.*

Ullrich, B. & Südhof, T. C. (1995). Differential distributions of novel synaptotagmins: comparison to synapsins. *Neuropharmacology*, 34(11), 1371–1377.

Urban-Ciecko, J. & Barth, A. L. (2016). Somatostatin-expressing neurons in cortical networks. *Nat. Rev. Neurosci.*, 17(7), 401–409.

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., Hovestadt, V., Escalante, L. E., Shaw, M. L., Rodman, C., Gillespie, S. M., Dionne, D., Luo, C. C., Ravichandran, H., Mylvaganam, R., Mount, C., Onozato, M. L., Nahed, B. V., Wakimoto, H., Curry, W. T., Iafrate, A. J., Rivera, M. N., Frosch, M. P., Golub, T. R., Brastianos, P. K., Getz, G., Patel, A. P., Monje, M., Cahill, D. P., Rozenblatt-Rosen, O., Louis, D. N., Bernstein, B. E., Regev, A., & Suva, M. L. (2017). Decoupling genetics, lineages, and microenvironment in idh-mutant gliomas by single-cell rna-seq. *Science*, 355(6332).

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., McKinsey, G. L., Pattabiraman, K., Silberberg, S. N., Blow, M. J., Hansen, D. V., Nord, A. S., Akiyama, J. A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., Kaplan, T., Kriegstein, A. R., Rubin, E. M., Ovcharenko, I., Pennacchio, L. A., & Rubenstein, J. L. R. (2013). A high-resolution enhancer atlas of the developing telencephalon. *Cell*, 152(4), 895–908.

Waddington, C. H. (1957). *The strategy of the genes*. Routledge.

Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11), 1145–1160.

Wallace, D. C. & Chalkia, D. (2013). Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harb. Perspect. Biol.*, 5(11), a021220.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation.

Weintraub, H. & Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. *Science*, 193(4256), 848–856.

Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: Computational identification of cell doublets in Single-Cell transcriptomic data. *Cell Syst*, 8(4), 281–291.e9.

Woodworth, M. B., Girskis, K. M., & Walsh, C. A. (2017). Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat Rev Genet*, 18(4), 230–244.

Wu, S.-P., Kao, C.-Y., Wang, L., Creighton, C. J., Yang, J., Donti, T. R., Harmancey, R., Vasquez, H. G., Graham, B. H., Bellen, H. J., Taegtmeyer, H., Chang, C.-P., Tsai, M.-J., & Tsai, S. Y. (2015). Increased COUP-TFII expression in adult hearts induces mitochondrial dysfunction resulting in heart failure. *Nat. Commun.*, 6, 8245.

Xu, J., Nuno, K., Litzenburger, U. M., Qi, Y., Corces, M. R., Majeti, R., & Chang, H. Y. (2019). Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *Elife*, 8.

Yap, E.-L. & Greenberg, M. E. (2018). Activity-Regulated transcription: Bridging the gap between neural activity and behavior. *Neuron*, 100(2), 330–348.

Ye, K., Lu, J., Ma, F., Keinan, A., & Gu, Z. (2014). Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc Natl Acad Sci U S A*, 111(29), 10654–9.

Yu, V. W. C., Yusuf, R. Z., Oki, T., Wu, J., Saez, B., Wang, X., Cook, C., Baryawno, N., Ziller, M. J., Lee, E., Gu, H., Meissner, A., Lin, C. P., Kharchenko, P. V., & Scadden, D. T. (2016). Epigenetic memory underlies Cell-Autonomous heterogeneous behavior of hematopoietic stem cells.

Zafar, H., Wang, Y., Nakhleh, L., Navin, N., & Chen, K. (2016). Monovar: single-nucleotide variant detection in single cells. *Nat. Methods*, 13(6), 505–507.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., & Linnarsson, S. (2018). Molecular architecture of the mouse nervous system. *Cell*, 174(4), 999–1014.e22.

Zemmour, D., Zilionis, R., Kiner, E., Klein, A. M., Mathis, D., & Benoist, C. (2018). Single-cell gene expression reveals a landscape of regulatory t cell phenotypes shaped by the tcr. *Nat Immunol*, 19(3), 291–301.

Zheng, C., Zheng, L., Yoo, J. K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J. Y., Zhang, Q., Liu, Z., Dong, M., Hu, X., Ouyang, W., Peng, J., & Zhang, Z. (2017a). Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing. *Cell*, 169(7), 1342–1356 e16.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., & Bielas, J. H. (2017b). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8, 14049.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., & Enard, W. (2017). Comparative analysis of single-cell rna sequencing methods. *Mol Cell*, 65(4), 631–643 e4.

Zunino, R., Schauss, A., Rippstein, P., Andrade-Navarro, M., & McBride, H. M. (2007). The SUMO protease SENP5 is required to maintain mitochondrial morphology and function. *J. Cell Sci.*, 120(Pt 7), 1178–1188.