Analyze lineage commitment during human hematopoiesis using somatic mitochondrial mutations

Daphne van Ginneken

A thesis presented for the
Minor Research Project of the Master
Bioinformatics and Biocomplexity

Institut Curie (UMR168 Team Perie)
Utrecht University
July 18, 2023

Daily supervisor:	dr. Wenjie Sun
Supervisor host institute:	prof. dr. Leila Perie
Examiner:	prof. dr. Rob de Boer

Abstract

Lineage tracing is a crucial method to provide insights into the evolving landscape of progenitor potential during lineage commitment in human hematopoiesis. A deeper understanding of the hematopoietic process can have promising clinical implications for the development of immune therapies where the composition of the immune system needs to be altered. Numerous lineage tracing techniques which introduce genetic modifications have been employed in model systems. Approaches in humans usually require the detection of somatic mutations in nuclear or mitochondrial DNA at single-cell level. Here, we perform retrospective lineage tracing by detecting somatic mitochondrial mutations in single-cell sequencing data, and use them as natural genetic barcode to link genetic regulators of hematopoietic stem and progenitor cells (HSPCs) to cell fate in PBMCs. We employ previously published scATAC-seq data from two replicates of CD34+ HSPCs and two replicates of PBMCs from the same donor with a three months time interval, to identify mitochondrial variants. Lineages biased clones were quantified with a chi-squared test, and differentially expressed genes and enriched pathways were inferred. In contrast with previously published results, we observed clones with a significant bias towards the lymphoid, myeloid, and HSC self-renewal lineage. However, regulatory networks of these clones showed minimal overlap between the replicates, indicating uncertainty in the results. Despite this uncertainty, the observation of lineage biased clones provide an opportunistic perspective for the suitability of mitochondrial mutations for lineage tracing.

Layman summary

The hematopoietic system consists of the bone marrow and the blood cells and tissues it produces. Different paths of cell differentiation lead to different mature cells. We call these paths lineages, and the function of the mature cells we call 'fate'. It is important to understand how this system works, and why some cells follow a different lineage than other cells. By knowing exactly how this works, we can improve medical procedures such as stem cell transplantation and immune therapy.

Here, we try to follow a single cell and all its daughter cells by using mutations in the DNA of mitochondria in cells in the bone marrow and the blood. When cells divide, the mitochondria are distributed over the two daughter cells. The same mutation in two cells can mean that they belong to the same family, we call a 'clone'. That is why we search for mutations in the bone marrow, and in the blood a few months later, to see the different fates the bone marrow cells have developed in.

To search for these mutations in the mitochondria we use a method called single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq). scATAC-seq sequences the parts of the open part of the genome that can be transcribed. This is a very usefull approach to sequence the mitochondria, because the entire mitochondrial genome is open. Next to the mitochondria, this method also sequences the nuclear genome of each cell. We can use the genes on the open parts of the chromatin to identify what function and fate a cell has.

By combining these methods, we can link the genes on the open nuclear genome in the bone marrow cells, to the genes (and therefor cell fate) in the matured blood cells and see if there is a connection. If cells from a specific clone in the bone marrow all express more of gene 'A', and the daughter cells from this clone in the blood are mostly of cell fate 'B', then we can say that gene 'A' might cause cells to differentiate into type 'B'. And when someone has a disease and lacks cells in the blood of type 'B', we could manipulate the bone marrow to express more gene 'A'. This way we can change the composition of the blood cells.

We discovered some clones that mostly have cells in the lineages lymphoid, myeloid and hematopoietic stem cell (HSC) self-renewal lineage, we call these clones 'lineage biased clones'. The lymphoid lineage contains immune cells such as B cells, T cells and natural killer cells. The myeloid lineage makes red blood cells, granulocytes, monocytes and platelets. HSC can either differentiate into these lineages, or renew themselves.

These lineage biased clones contain specific genes for which they have more or less expression. But these genes changed a lot when you repeat the procedure. This means that we cannot be sure that these genes are really correlated to the cell fates, or that we found them by coincidence. But the fact that we found lineage biased clones means that we can use the mutations in mitochondria to follow the lineages in the hematopoietic system.

Acknowledgments

I would like to thank my daily supervisor Wenjie Sun for all the guidance, helpful feedback and interesting discussions. I would like to thank Leila Perie for allowing me to work in her team for the past 6 months, providing me with everything I needed, and challenging me to think deeper about various biological processes. I thank for everyone in Team Perie, and Institut Curie, for making my time in Paris a true pleasure. And finally, I thank Rob de Boer for introducing me to Leila Perie, and being my examiner during this internship.

Contents

Chapter 1

Introduction

1.1 Insights in hematopoiesis through lineage tracing

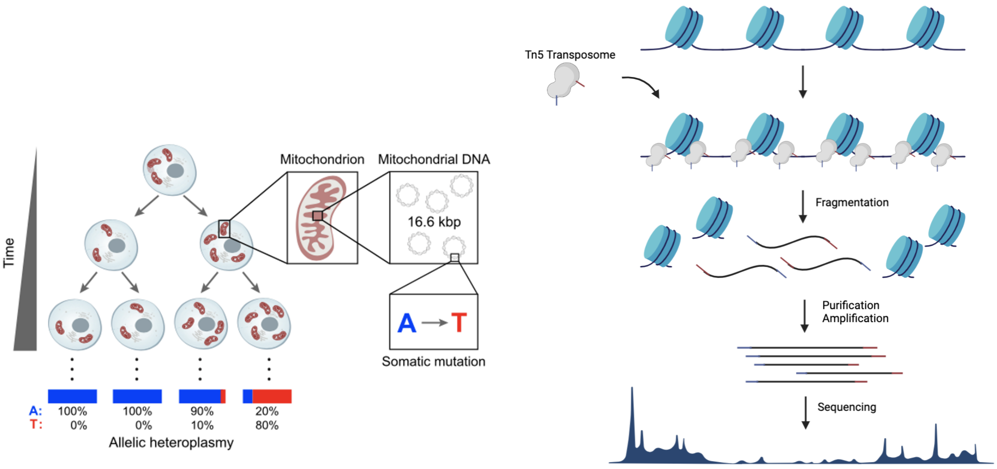
Lineage tracing refers to a set of techniques to track and trace the migration, proliferation and differentiation of a single cell. Accurate lineage tracing is a crucial goal in the complex process of hematopoiesis. During this highly regulated process, immature hematopoietic stem cells (HSCs) differentiate in progenitors of specialized blood cell lineages, and ultimately develop into mature immune cells, red blood cells and platelets. By utilizing lineage tracing techniques, valuable insights in the hematopoietic system in both pathological and physiological context can be attained. Moreover, this approach could provide information about the evolving landscape of progenitor potential during lineage commitment and cell fate decisions. Relevant findings potentially have clinical implications in fields such as HSC transplantations, ageing, immune-based therapy, tumor evolution and regenerative medicine.[1][2]

Some commonly employed lineage tracing techniques include the use of fluorescent proteins as lineage markers, retroviral or lentiviral labeling, and genetic barcoding. By introducing fluorescent proteins whose expressions are under the control of enzymes such as Cre recombinase, cellular clones can be visualized and tracked. Fluorescent labels can also be introduced by infecting cells with retroviral or lentiviral vectors. These vectors integrate in the host cell genome, and are inherited by daughters cells. Unique DNA sequences can be inserted and used as genetic barcodes by CRISPR/Cas9 genome editing. These barcodes can then be retrieved by high throughput sequencing. However, all these techniques introduce genetic tags or mutations, and are therefore not ethical to be applied in humans *in vivo*.[3][4]

1.2 Somatic mitochondrial mutations as genetic barcode

An alternative to genome altering lineage tracing approaches is retrospective lineage tracing, where somatic mutations are used as natural genetic barcodes [1]. A single cell has multiple mitochondria, which contain 100-1000 copies of mitochondrial DNA (mtDNA) that can acquire somatic mutations, causing various levels of allelic heteroplasmy (Figure 1.1a). Somatic mutations in the mtDNA with heteroplasmy levels of at least 5 percent can be stably propagated to daughter cells[5][6]. The mitochondrial genome has various characteristics that provide an advantage over its nuclear counterpart in distinguishing true mutations from background noise. The size of the circular human mitochondrial genome (approximately 16.6 kilobases) makes it small enough to be sequenced cost effective, and large enough to be a substantial target. In comparison to nuclear DNA (nDNA), mtDNA has a higher mutation rate (10-100 folds) and higher copy numbers (100-1000 per cell).[5][6]

Despite the advantages of mtDNA for lineage tracing, there are some considerations. For example, the inheritance patterns of mtDNA are more complicated than those of nDNA, making lineage inference a challenge [1]. Another potential limitation is the horizontal transfer of mitochondria (HMT) between cells. While HMT has been observed in various pathological conditions, triggered by stress responses, its extent and role in natural development and tissue homeostasis remains unclear [7]. Nonetheless, to significantly confound a lineage tracing analysis, HMT would require frequent occurrences. This scenario appears unlikely, as Ludwig *et al.*[5] found no evidence of such event in their data.



(a) Dynamics of mtDNA heteroplasmy. Derived from Ludwig *et al.*, 2019, Cell 176

(b) ATAC sequencing approach.

Figure 1.1: Schematic overview of mutation propagation in mitochondria and ATAC sequencing.

1.3 scATAC-seq for mitochondrial mutation detection

With the abundant availability of single-cell RNA sequencing (scRNA-seq) data, it may seem beneficial to use this data for detecting mtDNA mutations as clonal markers. However, factors such as RNA editing, transcription errors, technical artifacts, and limitations in mitochondrial sequencing depth pose challenges on the effectiveness of scRNA-seq for this purpose. As an alternative, single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) bypasses most of these obstacles. ATAC-seq identifies open accessible chromatin regions in the genome by fragmenting these regions with the use of hyperactive Tn5 transposase and adds adapters. These fragments can then be purified, amplified and sequenced (Figure 1.1b). The microfluidics-based single cell approach using 10X Genomics has become widely used. This system captures single transposed nucleus and adds unique barcodes to the DNA fragments. As the mitochondrial genome is not packaged into chromatin, it can be sequenced effectively with ATAC-seq [6].

Lareau *et al.*[8] introduced a modified version of the droplet-based 10X Genomics workflow, known as mitochondrial single-cell assay for transposase-accessible chromatin sequencing (mtscATAC-seq). This approach offers several improvements over scATAC-seq to enable higher and more uniform coverage of the mitochondrial genome. In contrast to scATAC-seq, mtscATAC-seq involves processing whole cells instead of isolated nuclei to retain a higher abundance of mitochondrial DNA. To minimize the potential mixing of mtDNA between cells, the mtscATAC-seq method incorporates modified cell lysis techniques and introduces a formaldehyde fixation step.

The utilization of scATAC-seq is not without its limitations. Differentially accessible regions currently lack power in cell-type annotation in comparison to differentially expressed genes in transcriptomics data [9]. 'Gene activity scores' can be computed based on distal and proximal accessible elements to a promoter region, enabling the inference of gene expression. Nevertheless, these gene scores do not not have the same golden standard signatures for unsupervised cell-type discovery that scRNA-seq has. The extensively available annotated scRNA-seq data can fill this gap by serving as a reference data to transfer cell labels. During this label transfer process, the query and reference datasets are projected into a shared lower-dimensional space defined by a canonical correlation structure. In this shared space, mutual nearest-neighbors are identified across the query and reference cells, which serve as anchors during data integration (Figure 1.2). However, certain limitations should be considered when employing this method of annotation. Firstly, dominant cell-types in the reference datasets may bias the cell-type prediction in the query data. Secondly, cell-types that are absent from the reference data will not be identified in the query data. The reference data should therefore be selected based on the expected cell-types and their abundance in the query data.[9][10]

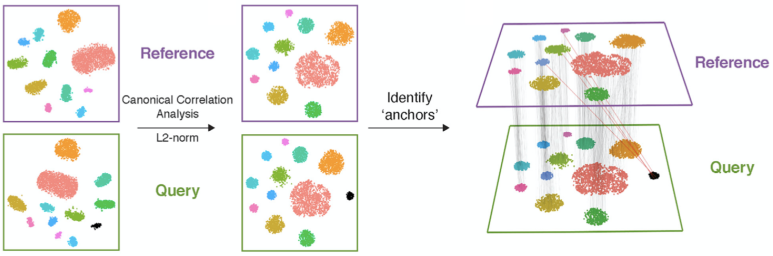


Figure 1.2: Data integration for cell-type label transfer. Query and reference data are projected into a shared lower dimensional space, in which anchors are identified. Derived from Stuart *et al.*, 2019, Cell 177 [10]

1.4 Lineage biased clones and genetic regulators in hematopoiesis

One of the key objectives in hematopoietic lineage tracing is to gain a deeper understanding of the regulatory mechanisms underlying cell fate decision. To study these mechanisms, it is necessary to identify clones with a distinct lineage bias. Lineage biased clones refer to cells with shared ancestry that are more prone to differentiate into specific cell fates. Through the identification of differentially expressed genes (DEGs) and enriched pathways between these lineage biased clones, genetic predictors of lineages commitment could be identified (Figure 1.3).[11][12]

Previous studies have successfully detected lineage biased clones both *in vitro* in humans and *in vivo* in mice models. For instance, Cosgrove *et al.*[13] discovered myeloid biased, erythroid biased and differentiation-inactive clones along with their metabolic pathways and DEGs in mice. They conducted lineage tracing *in vivo* using DRAG *in situ* barcoding [14], followed by scRNA-seq to recover the barcodes from hematopoietic stem and progenitor cells (HSPCs), and bulk sequencing of DNA after 47-67 weeks. If at least 75 percent of cells in a clone was myeloid or erythroid, the clone was assigned as lineage biased. Among the DEGs discovered were established markers of myeloid potential such as Mpo, Ctsg, Ms4a3 and Cpa3.

Lareau *et al.*[8] identified erythroid and monocyte biased clones in 20-day cultures of human CD34+ HSPCs. They performed lineage tracing using mtDNA mutations detected with mtscATAC-seq, and identified lineage biased clones based on the fraction of monocyte/erythroid cells in a clone. They permutated this cell-type distribution 100 times and computed a z-score for each clone. The clone was identified as lineage biased if the z-score between the observed and permutated fraction was at least 5. They identified various transcription factor motifs associated with these biased clones, including SPI1 and CEBPA for monocyte bias. This same study aimed to explore clonal tracing in human hematopoiesis *in vivo*. They profiled mtDNA mutations in mtscATAC-seq data obtained from CD34+ HSPCs and PBMCs collected 3 months later, and applied a Chi-squared test to investigate the association between clonal output and inferred cell states. In contrast to their *in vitro* study, no evidence of lineage biased clones was observed. To date, based on our existing knowledge, lineage biased clones have not been documented in human hematopoiesis *in vivo*.

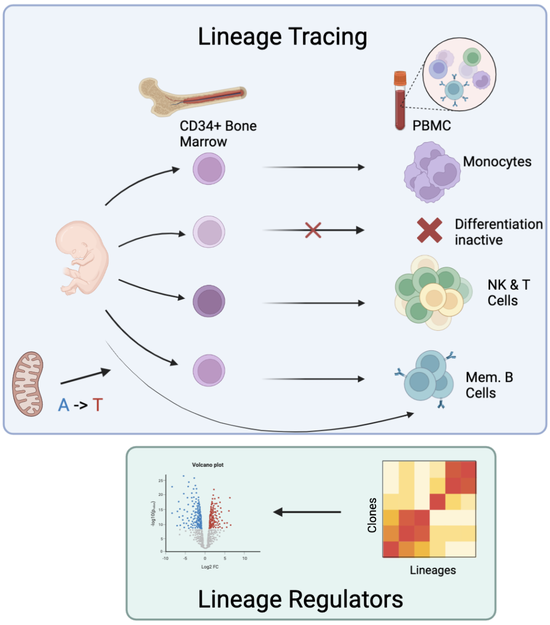


Figure 1.3: By tracing the lineage between CD34+ HSPCs and PBMCs, clones with a predisposition towards a certain cell fate could be identified. By analyzing DEGs of these clones, genetic regulators of lineage commitment could be discovered.

1.5 Research Questions

The primary objective of our study is to examine the utility of mtDNA mutations for *in vivo* lineage tracing in humans. This investigation poses challenges due to the absence of a definitive ground truth. Potential confounding variables, such as loss of mutations or mitochondrial transfer, may influence our findings. Nevertheless, the identification of substantial lineage biased clones with distinct genetic networks can mitigate these limitations and provide valuable insights.

Second, employing these methods we aim to investigate the presence of genetic regulators that can predict lineage commitment in HSPCs. The identification of such predictors would indicate the potential for manipulating HSPCs to modulate the cellular output of the bone marrow, leading to beneficial outcomes in diverse pathogenic processes. This could be promising for the development of therapeutic strategies that need to alter the composition of the immune system for clinical applications.

Chapter 2

Methods

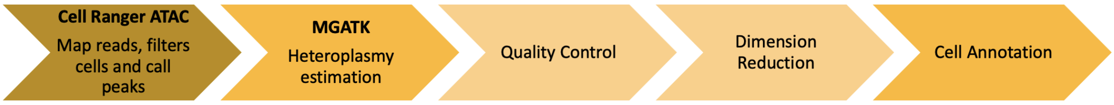
Instructions and code to reproduce this analysis can be found in the Github repository:
`https://github.com/TeamPerie/Report_Daphne`

2.1 Samples

We used the public human *in vivo* mtscATAC-seq data set from Lareau *et al.* [8] for our analysis. This data set contains two replicates of bone marrow-derived CD34+ HSPCs and two replicates of PBMCs with a 3 months time interval. Both samples were from the same healthy donor (male, 47 years old). All 10x mtscATAC-seq libraries were sequenced paired end and were aimed to contain 100 million reads and at least 20x coverage of the mitochondrial genome.

2.2 Preprocessing

All replicates were preprocessed separately to yield annotated high-quality cells.



2.2.1 Cell Ranger ATAC

Reads were mapped to the a modified version of the h38 human reference genome using Cell Ranger ATAC count version 2.1.0. The reference genome was modified to hard-mask regions on the nuclear genome that shared homology with the mitochondrial genome. This modification forces homologous reads to the mitochondrial genome during mapping, increasing mtDNA depth and capturing every mitochondrial variant. Cell Ranger ATAC count identifies peaks of accessible DNA, and calls cells based on the fragments overlapping these peaks. All default settings were used, but '–force-cells' was set to 6000 to keep a high cell count for an accurate heteroplasmy estimation.

2.2.2 MGATK

Filtered cells and peaks were then used to identify somatic mtDNA mutations using the mitochondrial genome analysis toolkit mgatk version 0.6.7[8]. In contrast to other variant callers, mgatk focuses on clonal mtDNA variants by combining signal across and between cells.

2.2.3 Quality control

An additional quality filtering was applied to keep cells with a minimum of 1000 unique fragments, 25 percent (CD34+ HSPCs) or 60 percent (PBMCs) fragments in peaks, and 20x mtDNA coverage. These cutoffs were selected based on the density, and are equal to the cutoff utilized by Lareau *et al.*[8]

2.2.4 Dimension reduction

The high quality chromatin data was further analyzed with the R package Signac[15] (R version 4.2.2, Signac version 1.9.0). Dimension reduction was performed based on latent semantic indexing (LSI). This a combination of term frequency-inverse document frequency (TF-IDF) normalization followed by a singular value decomposition (SVD) on the top variable features. LSI normalizes for sequencing depth across cells and across peaks and identifies relationships between peaks to reduce dimensions. Note that the first TF-IDF dimension often captures sequencing depth, thus technical variation, therefore we excluded the first dimension from the downstream analysis. The first 2:30 LSI dimensions were used to construct a shared nearest-neighbor graph, on which a Louvain clustering was performed with a resolution of 0.85. A UMAP based on the LSI dimensions was constructed to visualize the cells.

2.2.5 Cell annotation

Cells were annotated by transferring labels from reference scRNA-seq data. For the CD34+ bone marrow samples, the CITE-seq reference of human BMNC from the Seurat package was employed (Seurat version 4.3.0). For the PBMCs, existing 10X scRNA-seq v3 PBMC data was used as reference [10].

2.3 Identify mitochondrial variants

To identify mitochondrial variants, mtDNA mutation counts for the replicates were added together. Somatic mutations suitable for lineage tracing were identified based on thresholds. First, mutations must be detected in at least 5 cells, and maximum 1000 cells across the joined replicates. With these cutoffs, we exclude mutations that might be a result of sequencing error or potential germline mutations. Next, mutations must exhibit a strand concordance of at least 0.5, ensuring consistency between forward and reverse reads to eliminate technical noise. A variable mean ratio (VMR) threshold of at least 0.01 was set to assess the variation in allele frequency across cells. A low VMR indicates that most cells have this mutations with the same allele frequency (VAF), suggesting the mutation to be wild type and unsuitable for lineage tracing. Next, replicates were separated again to classify cells into clones. In order to classify a cell as mutant, the mutation must have a minimum VAF of 5 percent, and at least 2 forward and 2 reverse reads containing the mutation.

2.4 Calculate lineage bias

Clones from which the cells are more prone differentiate into a specific cell fate, and thus are lineage biased, are identified with a Chi-squared goodness of fit test. For each clone, we tested whether the distribution of cell-types among the mutant cells significantly differed from the distribution observed across all cells. The null hypothesis assumed no difference in the distribution, indicating the absence of lineage bias. The alternative hypothesis suggested the presence of lineage bias. For the CD34+ HSPCs, we compared the HSCs against all other cell-types to test for clones with a predisposition towards self-renewal or proliferation. For the PBMCs, we focused on comparing the lymphoid (activated- and memory B cells, gamma/delta-, naive CD4-, memory CD4- and cytotoxic CD8 T cells, and T regulatory cells) and myeloid (CD14- and CD16-monocytes) lineage. We excluded the NK cells and dendritic cells to diminish ambiguity in lineage assignment.

2.5 Analyze genetic mechanisms

In each replicate, we conducted a differential gene expression analysis within CD34+ hematopoietic stem and progenitor cells (HSPCs) between lineage biased clones, employing the Wilcoxon rank sum test. Genes exhibiting a significant Wilcoxon statistic, a minimum log2 fold change of 0.1, and a p-value below 0.05 were selected for subsequent KEGG pathway analysis [16].

Chapter 3

Results

3.1 mtscATAC-seq retrieved substantial mtDNA depth

We utilized mtscATAC-seq data of two CD34+ HSPC replicates and two PBMC replicates after a three months interval to perform *in vivo* retrospective lineage tracing. Cell Ranger ATAC analysis yielded approximately 6000 cells per sample, with an average of approximately 16,000 (BM) and 9000 (PBMC) fragments detected per cell. Subsequent filtering based on unique fragments, reads in the peak (FRIP), and mtDNA depth resulted in an average of 3800 cells with 167,000 peaks and an average mtDNA depth of 100 (BM), as well as 5110 cells with 140,000 peaks and an average mtDNA depth of 50 (PBMC) (Figure 3.1). Exact numbers can be found in appendix A.1.

3.2 Cell states annotated with scRNA-seq label transfer

Following quality control, cells were clustered and embedded in UMAP dimensions (Figure 3.2). As anticipated, the first LSI dimension exhibited a strong correlation with sequencing depth (Appendix Figure A.2). Cell-type annotation based on scRNA-seq label transfer revealed distinct subsets of T cells, B cells, monocytes and dendritic cells in the PBMC samples (Figure 3.2b). However, the identification of NK-cells proved to be less straightforward. Cluster 4 in PBMC1 and cluster 4 and 13 in PBMC2 are confidently annotated as NK cells. Additionally, cluster 5 and 8 for PBMC1, and cluster 0 for PBMC2 are also annotated as NK cells. But, this prediction carries uncertainty due to significant overlap observed with memory CD4 T cells. Various hematopoietic lineages were identified in the CD34+ HSPC samples (Figure 3.2a). However, these samples exhibited higher cellular heterogeneity within the clusters than the PBMC samples. In contrast to the PBMC samples, the bone marrow samples exhibit a more continuous landscape of cell identities rather than distinct clusters.

3.3 Mitochondrial variants identified as clonal markers

Threshold based mutation identification yielded a total of 2143 unique mutations as clonal markers, 433 of these were shared between CD34+ HSPCs and PBMCs (Figure 3.3). A strong correlation between clone sizes was observed among replicates, suggesting minimal sampling issues. Clone size correlation between bone-marrow and PBMCs was comparatively lower. In general, clones appeared larger in the PBMC samples than in the bone marrow samples.

When visualizing the CD34+ HSPCs and PBMCs in UMAPs and coloring the cells on allele frequency for specific mutations (2788C>A, 12868G>A, 3209A>G), we observe a visual lineage bias. Cells with high allele frequencies tended to cluster in specific locations locations within the UMAPs (Figure 3.4 bottom). Mutation 2788C>A seems to be biased towards T cells, 12868G>A seem to be biased towards cytotoxic CD8 T cells, and mutation 3209A>G seems to have a bias towards monocytes. These findings are contrasting with the results from Lareau *et al.*[8], where cells with higher allele frequencies appeared to be randomly distributed across their UMAP plots

(Figure 3.4 top). It should be noted that our analysis treated replicates separately, whereas Lareau *et al.* merged replicates, resulting in the presence of four UMAPs in our analysis compared to their two UMAPs.

3.4 Lineage biased clones detected

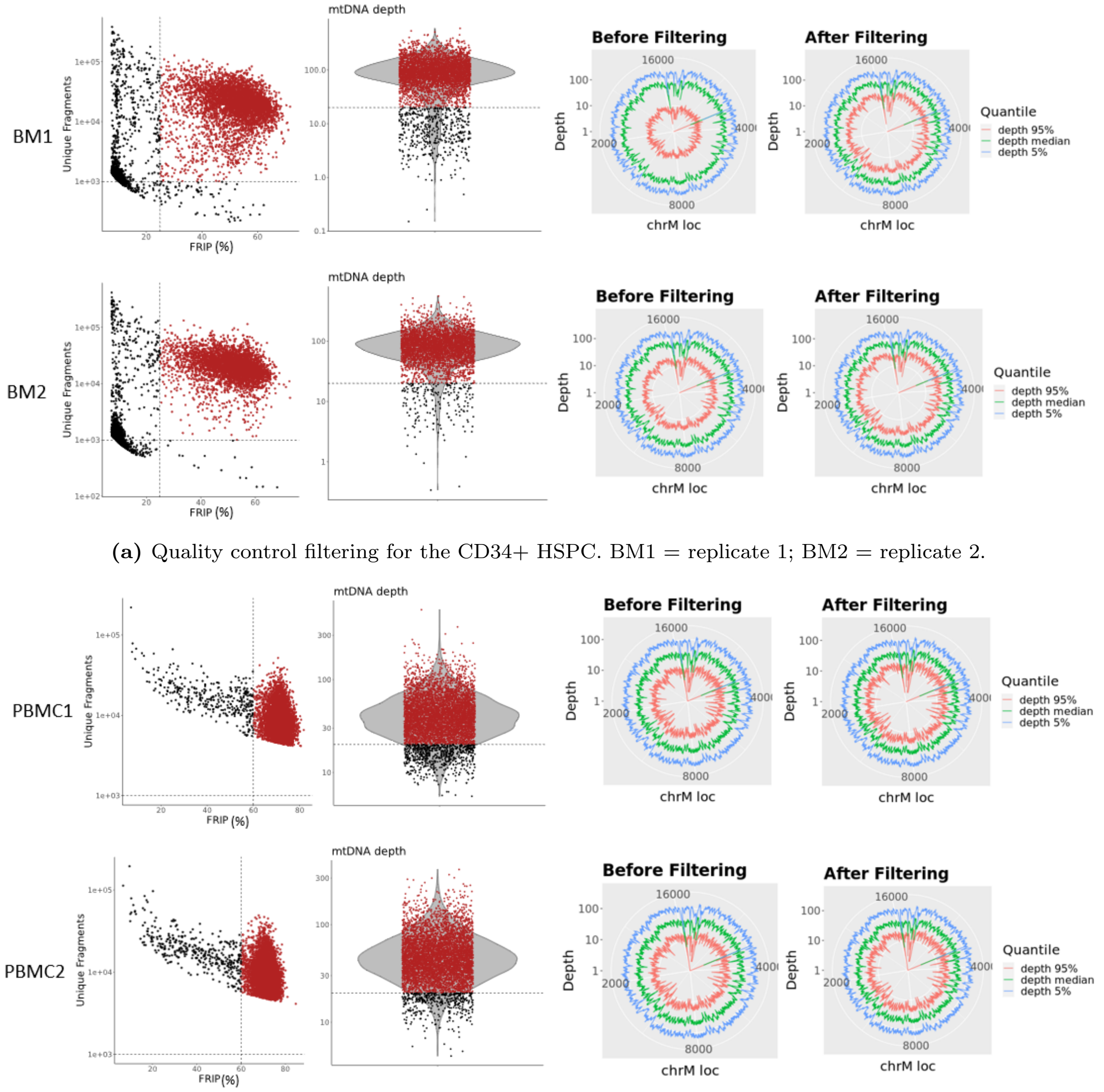
To test whether the clones have a lineages bias, cells were assigned to categories for comparison. In the CD34+ HSPCs, we compared cells annotated as HSC (818 in total) against all other cells (6617 in total) (Figure 3.5a). Employing a Chi-square test, 26 clones, with 271 cells in total, were detected with a bias towards HSCs (Figure 3.5b). No clones with a bias towards non-HSCs were detected. See Appendix Table A.2 and Figure A.3, A.4 and A.5, for details on all biased clones.

In the PBMC clones, we compared cells annotated in the lymphoid lineage (4179 in total) with the myeloid lineage (973 in total) and excluded NK cells and DCs (Figure 3.6a). We excluded NK cells due to the uncertain cell-type prediction (Figure 3.2), and DCs because they can arise from both the myeloid and the lymphoid lineage [17]. Utilizing the Chi-square test, 2 clones with 414 cells in total, were detected with a lymphoid bias, and 25 clones with 248 cells with a myeloid bias (Figure 3.6b). See Appendix Table A.3 and Figure A.6, A.7 and A.8, for details on all biased clones.

3.5 Potential regulatory networks observed

The lineage biased CD34+ HSPC clones were colored in the bone marrow UMAPs to visualize their heterogeneity in accessible chromatin features (Figure 3.7a). Consistent with expectations, we observed a higher proportion of cells from HSC-biased clones in the HSC region of the UMAP compared to other regions. However, some cells of the HSC biased clones were identified in other regions of the UMAP, indicating that these clones contain differentiation active cells. In order to investigate potential genetic predictors associated with hematopoietic stem cell (HSC) self-renewal or proliferation, we tested the clones for differentially expressed genes (DEGs) and enriched pathways (Figure 3.7b). 16,607 genes were detected in replicate 1, and 16,612 in replicate 2. Among these, 817 (replicate 1) and 888 (replicate 2) exhibited a high log2 fold change and significant p-value and where subsequently identified as differentially expressed. Between the replicates, 20 DEGs showed overlap. Subsequent KEGG pathway analysis of the DEGs revealed no shared pathways between the replicates, except for the regulation of actin cytoskeleton pathway observed in non-HSC biased clones.

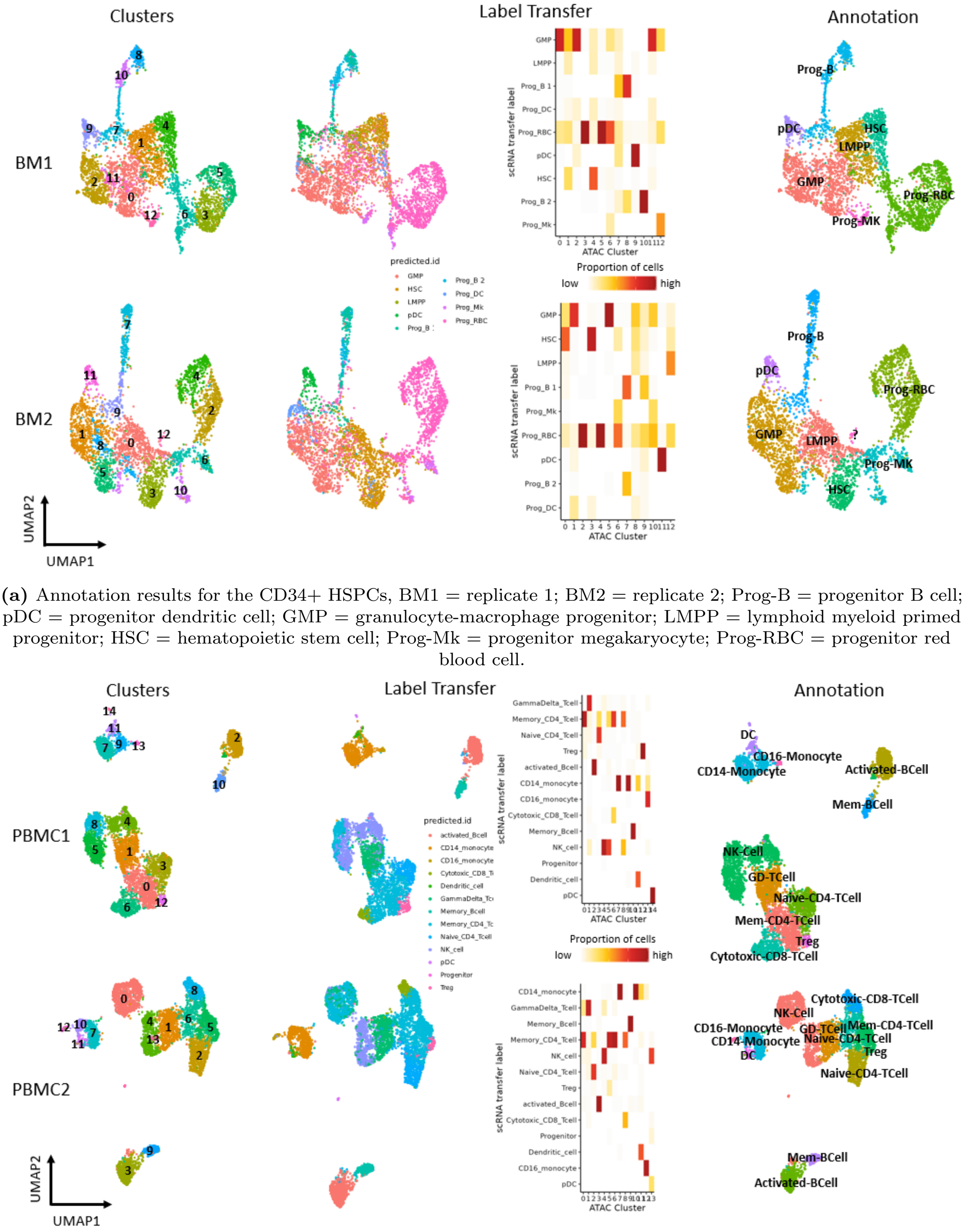
The lineage biased PBMC clones were traced back to the CD34+ bone marrow samples, and visualized in the bone marrow UMAPs (Figure 3.8a). No visual lineage bias is observed in these clones in the CD34+ HSPCs, indicating that lymphoid and myeloid biased progenitor have similar accessible chromatin features. Specifically, cells from myeloid-biased clones were also detected in the lymphoid progenitor area of the UMAP, while lymphoid-biased cells were also found in the granulocyte-macrophage progenitors (GMP) area. 16,938 genes were detected in replicate 1, and 16,526 in replicate 2 (Figure 3.8b). Among these, 734 (replicate 1) and 769 (replicate 2) exhibited a high log2 fold change and significant p-value and where subsequently identified as differentially expressed. Between the replicates, 9 DEGs showed overlap. The differentially expressed genes displayed a relatively low average log2 fold change in comparison to the results from the HSC biased clones. Further KEGG pathway analysis of the significant DEGs did not reveal any shared pathways between the replicates.



(a) Quality control filtering for the CD34+ HSPC. BM1 = replicate 1; BM2 = replicate 2.

(b) Quality control filtering for the PBMCs. PBMC1 = replicate 1; PBMC2 = replicate 2.

Figure 3.1: Quality control results of the CD34+ HSPCs and PBMCs. Cells are filtered based on the unique nuclear fragments, fraction of reads in peaks (FRIP) and average mtDNA sequencing depth. Circular plots show the coverage over the mitochondrial genome before and after filtering for the 5/50/95 percent quantile.



(a) Annotation results for the CD34+ HSPCs, BM1 = replicate 1; BM2 = replicate 2; Prog-B = progenitor B cell; pDC = progenitor dendritic cell; GMP = granulocyte-macrophage progenitor; LMPP = lymphoid myeloid primed progenitor; HSC = hematopoietic stem cell; Prog-Mk = progenitor megakaryocyte; Prog-RBC = progenitor red blood cell.

(b) Annotation results for the PBMCs. PBMC1 = replicate 1; PBMC2 = replicate 2; DC = dendritic cell; Mem-BCell = memory B cell; NK-Cell = natural killer cell; GD-TCell = gamma delta T cell; Mem-DC4-TCell = memory CD4 T cell, Treg = regulatory T cell.

Figure 3.2: UMAP visualizations of all quality controlled CD34+ HSPCs and PBMCs. Cells are colored based on the Louvain clustering (left), label transfer (middle), and final annotation (right). The heatmap depicts the proportion of cells in each clustered that has been predicted each cell-type.

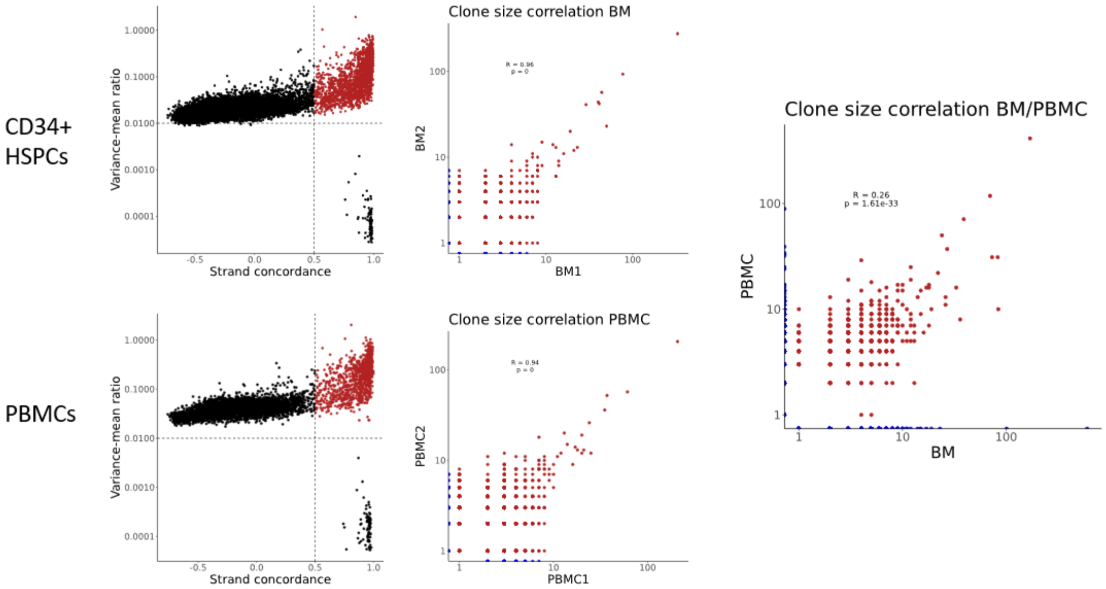


Figure 3.3: Identified clones and their sizes in CD34+ HSPCs (BM), PBMCs, and shared clones (right). R represent the Pearson correlation coefficient. Clones are colored blue when they are not present in one of the samples.

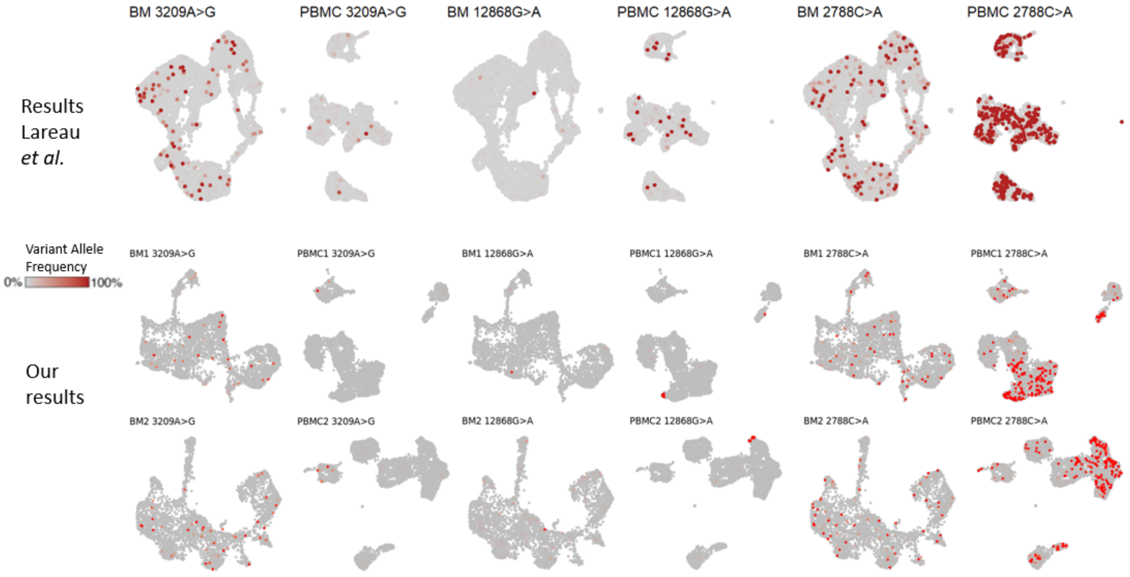
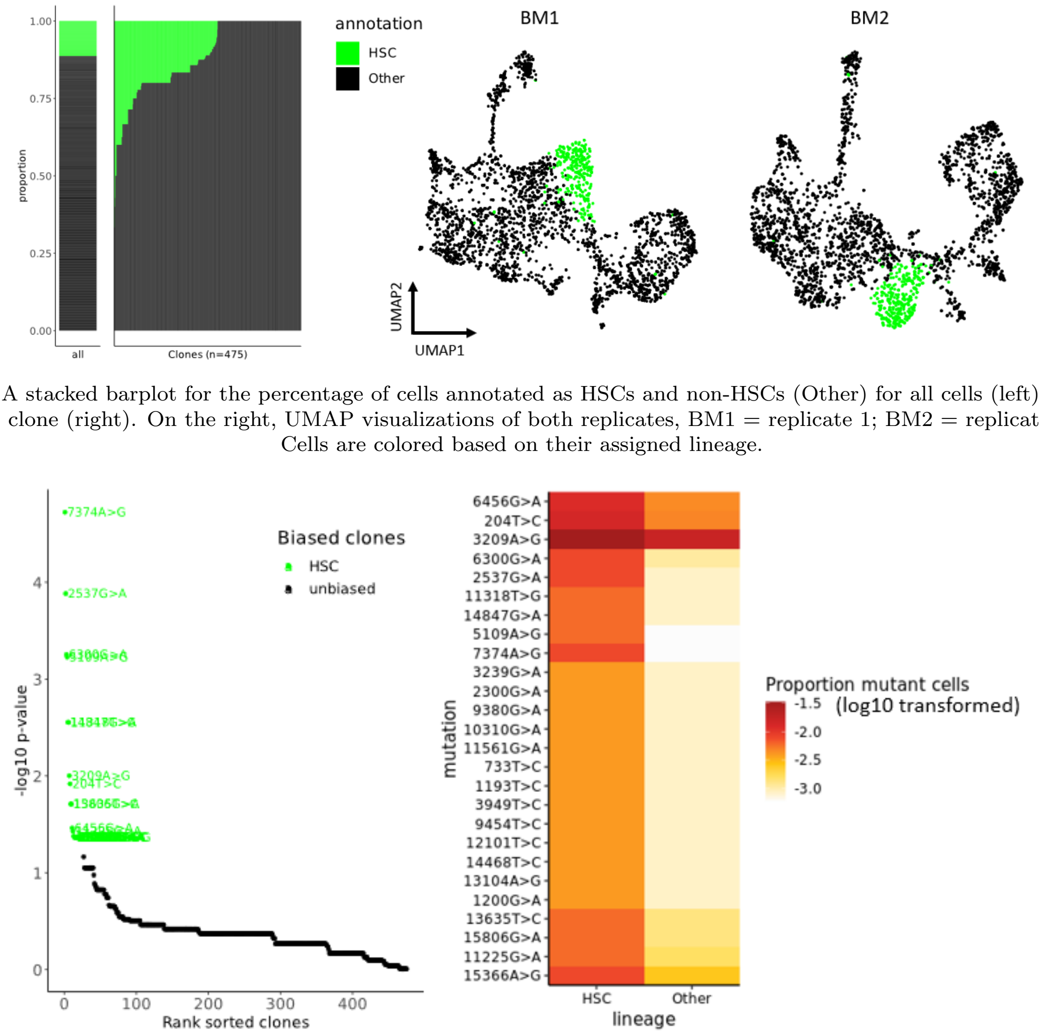


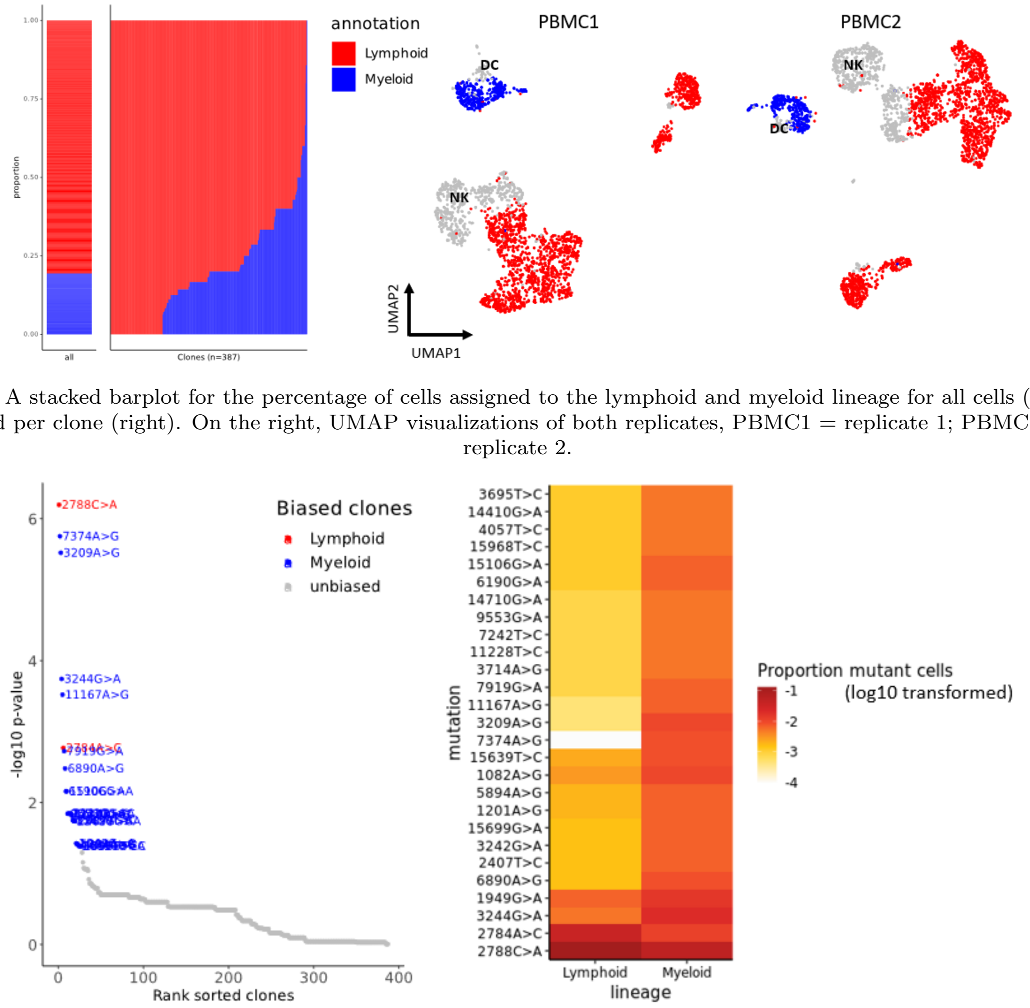
Figure 3.4: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for each mutation. Results reproduced with the code from Lareau *et al.*[8] (top), and our results (bottom).



(a) A stacked barplot for the percentage of cells annotated as HSCs and non-HSCs (Other) for all cells (left) and per clone (right). On the right, UMAP visualizations of both replicates, BM1 = replicate 1; BM2 = replicate 2. Cells are colored based on their assigned lineage.

(b) Left, the results of the chi-square test. Clones are colored based on their lineage bias for HSCs, or no observed lineage bias. Right, a heatmap of the log10 transformed proportion of cells per lineage in each lineage biased clones.

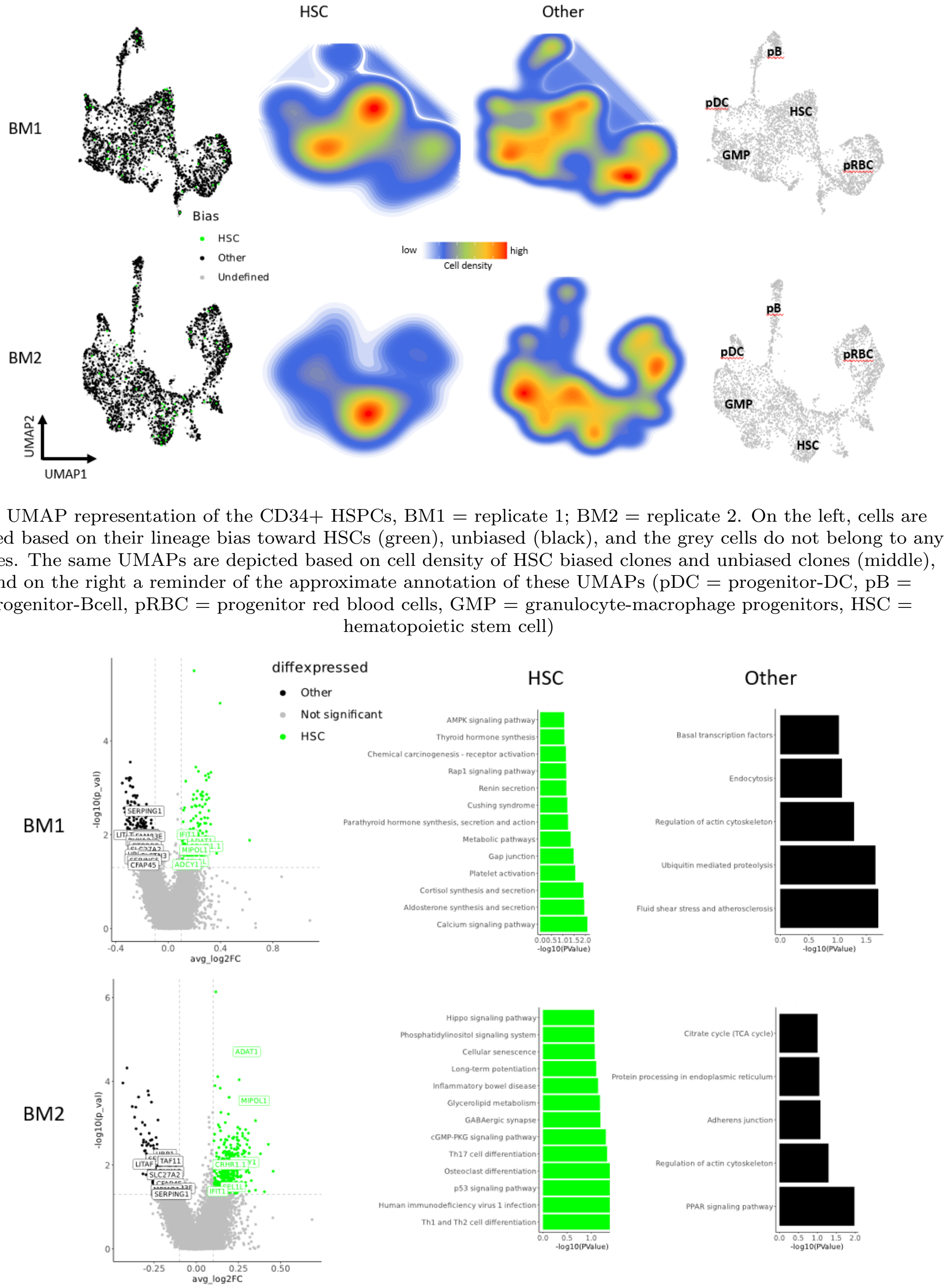
Figure 3.5: Results of the lineage bias test for CD34+ HSPC clones towards HSCs or non-HSCs (other).



(a) A stacked barplot for the percentage of cells assigned to the lymphoid and myeloid lineage for all cells (left) and per clone (right). On the right, UMAP visualizations of both replicates, PBMC1 = replicate 1; PBMC2 = replicate 2.

(b) Left, the results of the chi-square test. Clones are colored based on their lineage bias for lymphoid or myeloid, or no observed lineage bias. Right, a heatmap of the log10 transformed proportion of cells per lineage in each lineage biased clones.

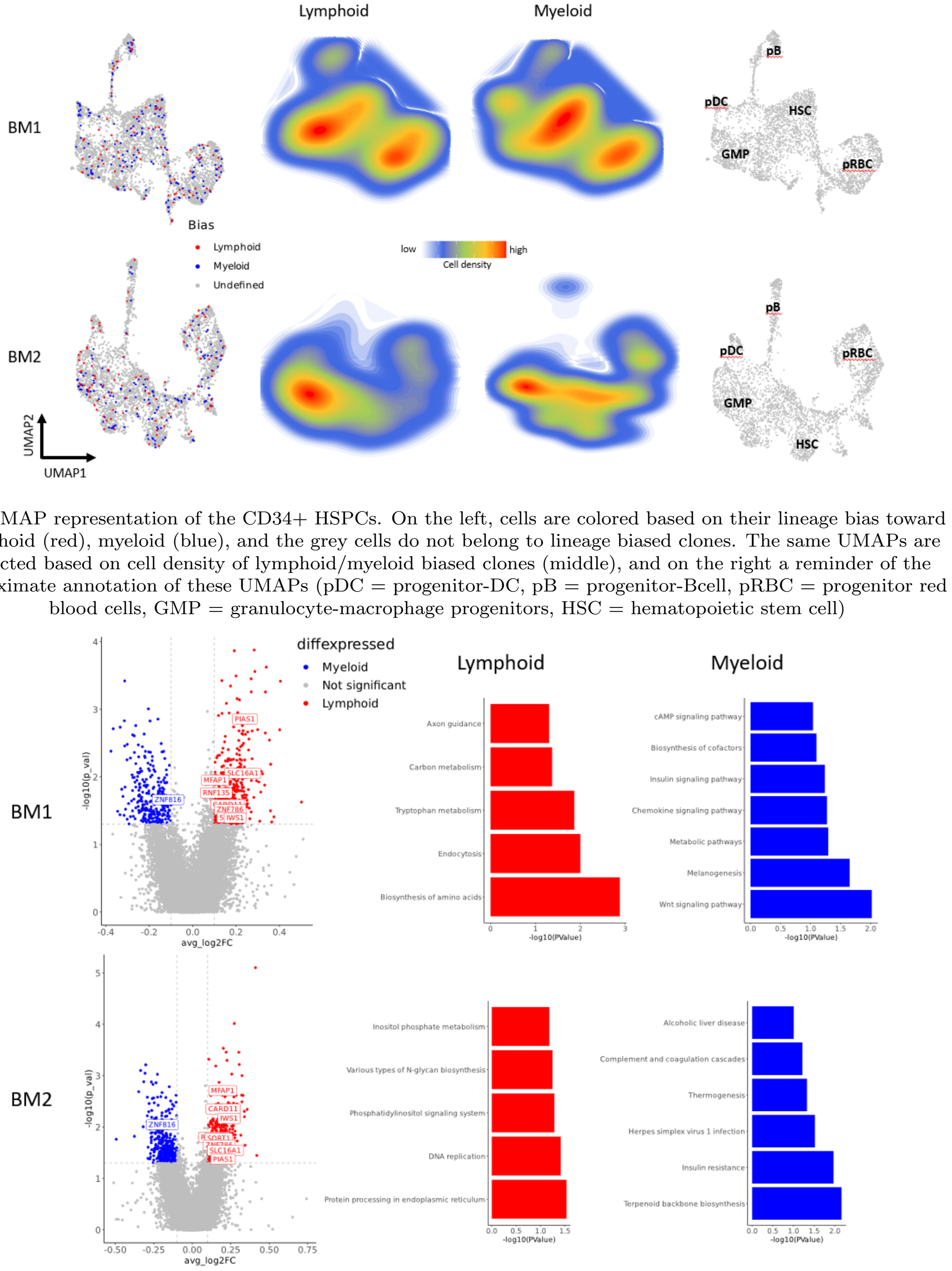
Figure 3.6: Results of the lineage bias test for PBMC clones towards lymphoid or myeloid



(a) UMAP representation of the CD34+ HSPCs, BM1 = replicate 1; BM2 = replicate 2. On the left, cells are colored based on their lineage bias toward HSCs (green), unbiased (black), and the grey cells do not belong to any clones. The same UMAPs are depicted based on cell density of HSC biased clones and unbiased clones (middle), and on the right a reminder of the approximate annotation of these UMAPs (pDC = progenitor-DC, pB = progenitor-Bcell, pRBC = progenitor red blood cells, GMP = granulocyte-macrophage progenitors, HSC = hematopoietic stem cell)

(b) Volcano plots showing differentially expressed genes between the HSC/Other biased clones. Genes are labeled when they have a significant p-value (below 0.05), high average log2 fold change (above 0.1) and are present in both replicates. Dashed line represents these cutoffs. Enriched KEGG pathways are displayed in the bar plots on the right.

Figure 3.7: Results of the regulatory network analysis for clones biased towards HSCs or non-HSCs (other) in the CD34+ HSPCs. Results are shown for both replicates (BM1 = replicate 1; BM2 = replicate 2)



(a) UMAP representation of the CD34+ HSPCs. On the left, cells are colored based on their lineage bias toward lymphoid (red), myeloid (blue), and the grey cells do not belong to lineage biased clones. The same UMAPs are depicted based on cell density of lymphoid/myeloid biased clones (middle), and on the right a reminder of the approximate annotation of these UMAPs (pDC = progenitor-DC, pB = progenitor-Bcell, pRBC = progenitor red blood cells, GMP = granulocyte-macrophage progenitors, HSC = hematopoietic stem cell)

(b) Volcano plots showing differentially expressed genes between the lymphoid/myeloid biased clones. Genes are labeled when they have a significant p-value (below 0.05), high average log2 fold change (above 0.1) and are present in both replicates. Dashed line represents these cutoffs. Enriched KEGG pathways are displayed in the bar plots on the right.

Figure 3.8: Results of the regulatory network analysis for clones biased towards lymphoid or myeloid in the CD34+ HSPCs. Results are shown for both replicates (BM1 = replicate 1; BM2 = replicate 2)

Chapter 4

Discussion and Conclusion

Here, we used somatic mitochondrial mutations to perform lineage tracing of the hematopoietic process in human *in vivo*. We employed single cell chromatin accessibility data to detect mitochondrial mutations in the CD34+ HSPCs and PBMCs of the same healthy donor with a 3 months time interval. Genetic regulators for lymphoid and myeloid progeny, and for HSC self-renewal, were detected using a chi-squared test for lineage bias and subsequent testing for differentially expressed genes (DEGs) and enriched pathways.

Despite adapting comparable methods as those utilized in the original publication of the data (Lareau *et al.*[8]), our results hold opposing views. Where Lareau *et al.* observe no lineage bias in the PBMC clones, we detect clones that are significantly biased. A small error while joining a matrix of allele frequencies with other metadata of the cells such as UMAP dimensions, without matching for cellular barcodes, explained the random distribution of mutant cells and absence of lineage bias in the results of Lareau *et al.* (Appendix Figure A.2). This observation highlights the importance of reproducibility.

Although we observe lineage biased clones and associated regulatory networks, these findings are not robust. The mutant cells from myeloid and lymphoid biased clones show comparable accessible chromatin features in the CD34+ HSPCs when displayed in the UMAP. This can explain the different DEGs and enriched pathways between replicates, as they might show significance due to coincidence. As the replicates displayed high similarity up until this point in the analysis, its divergent DEGs and pathways imply minor variations to have considerable consequences. Hence, the resulting regulatory networks are unreliable for predicting cell fate. The robustness of the results could be improved by setting more stringent thresholds for variant detection. The unstable regulatory networks can be explained by various additional factors, which we can divide into two categories.

First, cells could be assigned to the same clone even though they are from a different ancestor cell. This cause false negative results, where clones will not be detected as lineage biased. Cell-type annotation influences the accuracy of dividing cells into categories for the chi-squared test. The scRNA-seq label transfer shows a low confidence for various clusters, particularly in the continuous landscape of CD34+ HSPCs. This ambiguous annotation could cause distinct lineages to be categorized together, resulting in low fold changes and p-values for the DEGs. This limits the chi-squared test for lineage bias as it requires a priori category defining. The decision for comparing lymphoid with myeloid and HSC self-renewal with proliferation was because these are one of the earliest branching points in cell fate commitment. Still, resolution can be increased to compare more distinct cell subsets or progenitors. Additionally, a chi-squared test might not be the correct method for defining lineage bias towards HSC self-renewal, as many cells from a HSC biased clone still seem to differentiate into other progenitors and into the PBMC. Other factors that can generate false negative results are the possible horizontal transfer of mitochondria between cells and the potential rise of similar mutations by coincidence. With the size of the mitochondrial genome and with its high mutation rate, there could be a significant probability of two cells carrying the same mutations by coincidence instead of shared ancestry. Especially when analyzing long-lived cells such as naive T cells [18], this probability will increase. Another point of consideration is the time interval between CD34+ HSPC and PBMC sampling, and the longevity of certain cell-types. HSCs differentiate into mature blood cells in approximately 4-12 weeks, but lymphocytes might take even longer [19]. The 3 months time interval in our data should be able to sample daughter cells, or close descendants, of the clones sampled in the bone marrow. But some CD34+ HSPCs might not have differentiated into the PBMC yet, and some might already have multiple generations in the PBMC.

Secondly, a factor that could confound the observed lineage bias is the independence of cell fate from mitochondrial mutations. When using mitochondrial mutations as clonal marker to infer genetic predictors of lineage commitment, we assume these mutations to be independent from cell fate. This might not be the case, as mutations in certain mitochondrial genes could alter the function of the cell, and consequently its cell fate decisions.

Despite potential confounding factors, the generation of false positive results will be rare. Therefor, the observed lineage bias is most likely accurate and supports the proposition that the propagation of chromatin state enables long-term inheritance of cellular function[20][21]. Other promising results where the genetic regulators of clones biased towards HSC self-renewal and proliferation. These results showed higher robustness than for the lymphoid and myeloid biased clones, as more DEGs and pathways were overlapping between replicates, and the enrichment for certain pathways where in agreement with the biological hypothesis. The actin cytoskeleton pathway is a key regulator of migration of cells within the body, which aligns with it's enrichment in proliferating HSPCs [22]. The enrichment of the cellular senescence pathway in self-renewing HSCs is consistent with the function of these HSCs, as this pathway maintains tissue homeostasis and halts proliferation [23].

There are many uncertainties in the use somatic mitochondrial mutations as genetic barcode. Not much is known about the frequency of horizontal transfer, mitophagy and the distribution of mitochondria during mitosis. The impact and frequency of the confounding and limiting factors need to be studied in detail to draw conclusions from lineage tracing studies using mitochondrial mutations, and to trust the resulting genetic predictors. But, important steps have been made towards a better understanding of the strengths and limitations of this approach. In this study, lineage biased clones were observed in human *in vivo*. As far as we know, this has not been observed before and provides a promising perspective on the usability of mitochondrial mutations for lineage tracing.

References

[1] Cordes S, Wung C, and Dunbar CE. “Clonal tracking of haematopoietic cells: insights and clinical implications”. In: *BJH* 192 (2021), pp. 819–831. DOI: 10.1111/bjh.17175.

[2] Perie L and Duffy KR. “Retracing the in vivo haematopoietic tree using single-cell methods”. In: *FEBS Letters* 590 (2016), pp. 4068–4083. DOI: 10.1002/1873-3468.12299.

[3] Chen C, Liao Y, and Peng G. “Connecting past and present: single-cell lineage tracing”. In: *Protein Cell* 13.11 (2022), pp. 790–807. DOI: 10.1007/s13238-022-00913-7.

[4] Wagner DE and Klein AM. “Lineage tracing meets single-cell omics: opportunities and challenges”. In: *Nat Rev Genet* 21.7 (2020), pp. 410–427. DOI: 10.1038/s41576-020-0223-2.

[5] Ludwig LS, Lareau CA, Ulirsch JC, et al. “Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics”. In: *Cell* 176 (2019), pp. 1325–1339. DOI: 10.1016/j.cell.2019.01.022.

[6] Xu J et al. “Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA”. In: *eLife* 8 (2019), e45105. DOI: 10.7554/eLife.45105.

[7] Dong LF, Rohlena J, Zobalova R, et al. “Mitochondria on the move: Horizontal mitochondrial transfer in disease and health”. In: *JCB* 222.3 (2023), e202211044. DOI: 10.1083/jcb.202211044.

[8] Lareau CA, Ludwig LS, Muus C, et al. “Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling”. In: *Nature Biotechnology* 39.4 (2021), pp. 451–461. DOI: 10.1038/s41587-020-0645-6.

[9] Baek S and Lee I. “Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation”. In: *Comp. and Struct. Biotechn. Journal* 18 (2020), pp. 1429–1439. DOI: 10.1016/j.csbj.2020.06.012.

[10] Stuart T, Butler A, Hoffman P, et al. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177 (2019), pp. 1888–1902. DOI: 10.1016/j.cell.2019.05.031.

[11] Weinreb C, Rodriguez-Fraticelli A, Camargo FD, et al. “Lineage tracing on transcriptional landscapes links state to fate during differentiation”. In: *Science* 14.367 (2020). DOI: 10.1126/science.aaw3381.

[12] Buenrostro JD, Corces MR, Lareau CA, et al. “Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation”. In: *Cell* 173 (2018), pp. 1535–1548. DOI: 10.1016/j.cell.2018.03.074.

[13] Cosgrove J, Lyne A, Rodriguez I, et al. “Metabolically Primed Multipotent Hematopoietic Progenitors Fuel Innate Immunity”. In: *biorxiv* (2023). DOI: 10.1101/2023.01.24.525166.

[14] Urbanos J, Cosgrove J, Beltman JB, et al. “DRAG in situ barcoding reveals an increased number of HSPCs contributing to myelopoiesis with age”. In: *Nat Commun* 14.2184 (2023). DOI: 10.1038/s41467-023-37167-8.

[15] Stuart T, Srivastava A, Madad S, et al. “Single-cell chromatin state analysis with Signac”. In: *Nat Methods* 18.11 (2021), pp. 1333–1341. DOI: 10.1038/s41592-021-01282-5.

[16] Kanehisa M and Goto S. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic Acids Res* 28.1 (2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[17] McLellan AD and Kämpgen E. “Functions of myeloid and lymphoid dendritic cells”. In: *Immunology Letters* 72.2 (2000), pp. 101–105. DOI: 10.1016/S0165-2478(00)00167-X.

[18] den Braber I, Mugwagwa T, Vrieskoop N, et al. “Maintenance of Peripheral Naive T Cells Is Sustained by Thymus Output in Mice but Not Humans”. In: *Immunity* 36.2 (2012), pp. 288–297. DOI: 10.1016/j.immuni.2012.02.006.

[19] Upadhaya S, Sawai CM, Papalexi E, et al. “Kinetics of adult hematopoietic stem cell differentiation in vivo”. In: *J Exp Med* 215.11 (2018), pp. 2815–2832. DOI: 10.1084/jem.20180136.

[20] Bruno S, Williams RJ, and Del Vecchio D. “Epigenetic cell memory: The gene’s inner chromatin modification circuit”. In: *PLOS Comp. Biology* (2022). DOI: 10.1371/journal.pcbi.1009961.

[21] Ruckert T, Lareau CA, Mashreghi MF, et al. “Clonal expansion and epigenetic inheritance of long-lasting NK cell memory”. In: *Nat Imm* 23 (2022), pp. 1551–1563. DOI: 10.1038/s41590-022-01327-7.

[22] Kanaan Z et al. “The Actin-Cytoskeleton Pathway and Its Potential Role in Inflammatory Bowel Disease-Associated Human Colorectal Cancer”. In: *Genet Test Mol Biomarkers* 14.3 (2010), pp. 347–353. DOI: 10.1089/gtmb.2009.0197.

[23] Gorgoulis V et al. “Cellular Senescence: Defining a Path Forward”. In: *Cell* 179.4 (2019), pp. 813–827. DOI: 10.1016/j.cell.2019.10.005.

Appendix A

Appendix

A.1 Cell numbers

	Nr. cells after Cell Ranger	Nr. fragments per cell after Cell Ranger	Nr. cells after QC filtering	Nr. peaks after QC filtering	Average mtDNA depth after QC filtering
BM1	6032	16355	3897	171432	104
BM2	6044	15100	3689	163321	93
PBMC1	6000	8550	4994	139280	48
PBMC2	6000	9316	5226	141294	54

Table A.1: Exact numbers of cells, peaks and fragments after various preprocessing steps. BM1 = CD34+ HSPC replicate 1; BM2 = CD34+ HSPC replicate 2; PBMC1 = PBMC replicate 1; PBMC2 = PBMC replicate 2.

A.2 Correlation LSI dimensions and sequencing depth

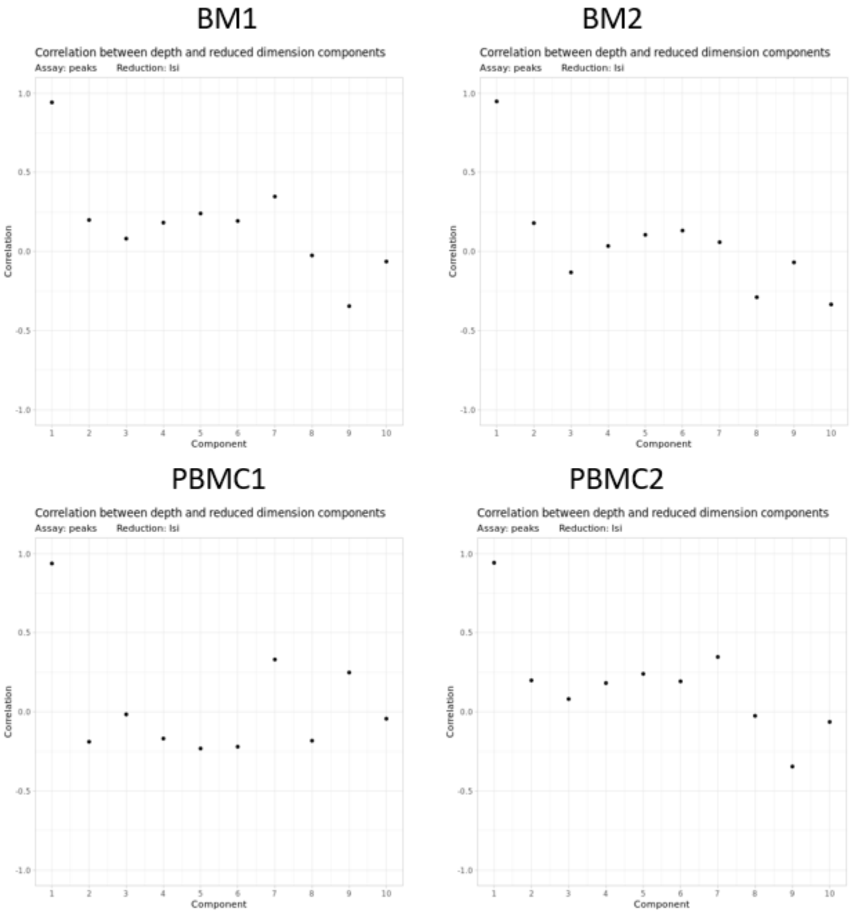
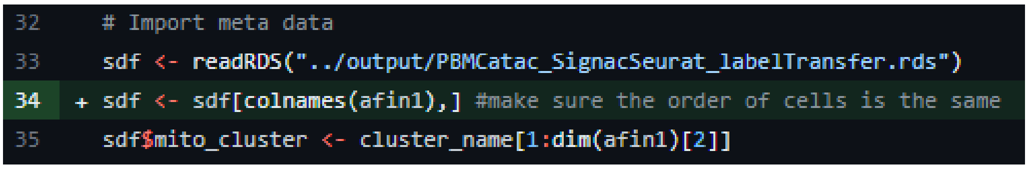


Figure A.1: Correlation between the sequencing depth and each LSI dimension for all samples. BM1 = CD34+ HSPC replicate 1; BM2 = CD34+ HSPC replicate 2; PBMC1 = PBMC replicate 1; PBMC2 = PBMC replicate 2.

A.3 Coding error Lareau *et al.*

After observing a coding error in the code of Lareau *et al.*[8], a pull request was created (`https://github.com/caleblareau/mtscATACpaper_reproducibility/pull/6`, Figure A.2).



```
32     # Import meta data
33     sdf <- readRDS("../output/PBMCatac_SignacSeurat_labelTransfer.rds")
34   + sdf <- sdf[colnames(afin1),] #make sure the order of cells is the same
35     sdf$mito_cluster <- cluster_name[1:dim(afin1)[2]]
```

Figure A.2: Line 34 was added to the code of Lareau *et al.*[8] to order cells in both dataframes in the same order.

A.4 Chi-Squared results

A.4.1 HSC self-renewal vs. proliferation bias

Mutation	Chisq stats	pvalue	p adjust	nr. cells	odds ratio
7374A>G	18.2969137166512	1.89013139053996e-05	0.008978	6	15.65
2537G>A	14.6314428400962	0.000130715351321805	0.062089	7	10.43
6300G>A	11.9142705617183	0.000557055465531023	0.264601	8	7.82
5109A>G	11.7951199893932	0.000593861841572743	0.282084	5	11.74
14847G>A	8.93570292128873	0.00279649828965589	1	6	7.82
11318T>G	8.93570292128873	0.00279649828965589	1	6	7.82
3209A>G	6.64346266680977	0.0099520316062046	1	84	1.98
204T>C	6.298052530843	0.012087068980956	1	26	2.88
15806G>A	5.45722422327308	0.0194875804105077	1	8	4.69
13635T>C	5.45722422327308	0.0194875804105077	1	8	4.69
6456G>A	4.46838511512794	0.0345276101699638	1	24	2.60
11225G>A	4.34030582931895	0.0372202371475907	1	9	3.91
15366A>G	4.14592427504435	0.0417348650410089	1	14	3.13
2300G>A	4.09343174084059	0.0430501503722248	1	5	5.21
3239G>A	4.09343174084059	0.0430501503722248	1	5	5.21
9380G>A	4.09343174084059	0.0430501503722248	1	5	5.21
10310G>A	4.09343174084059	0.0430501503722248	1	5	5.21
11561G>A	4.09343174084059	0.0430501503722248	1	5	5.21
733T>C	4.09343174084059	0.0430501503722248	1	5	5.21
1193T>C	4.09343174084059	0.0430501503722248	1	5	5.21
3949T>C	4.09343174084059	0.0430501503722248	1	5	5.21
9454T>C	4.09343174084059	0.0430501503722248	1	5	5.21
12101T>C	4.09343174084059	0.0430501503722248	1	5	5.21
14468T>C	4.09343174084059	0.0430501503722248	1	5	5.21
13104A>G	4.09343174084059	0.0430501503722248	1	5	5.21
1200G>A	4.09343174084059	0.0430501503722248	1	5	5.21

Table A.2: Results of the Chi-Squared test for lineage bias for HSC self-renewal or proliferation, only mutations with a p-value < 0.05 are reported. Odds ratio above 1 defines clones biased towards HSC self-renewal. P values were adjusted with a Bonferroni correction.

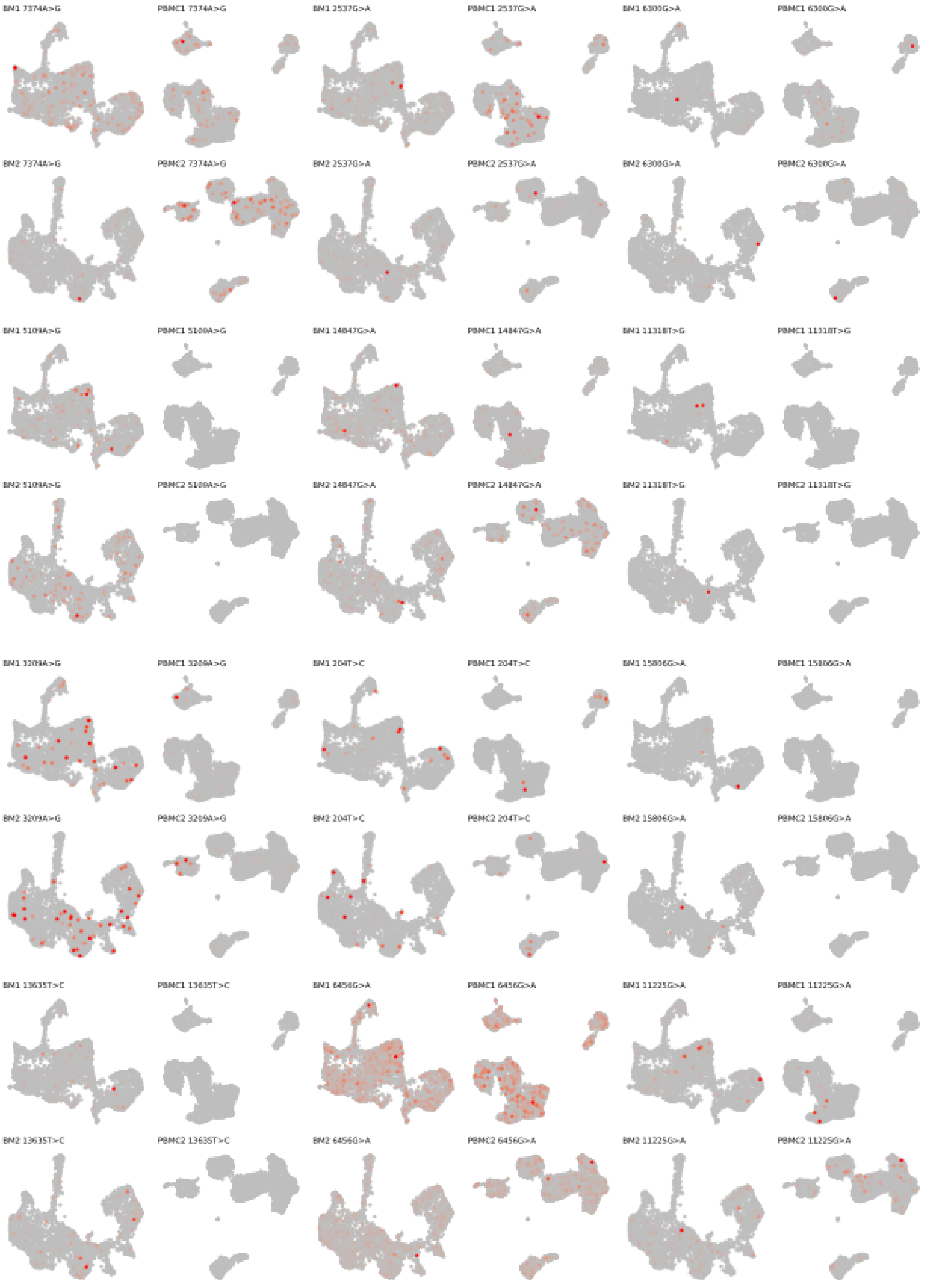


Figure A.3: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for the mutations from the lineage biased clones towards HSC self-renewal. (PART 1)

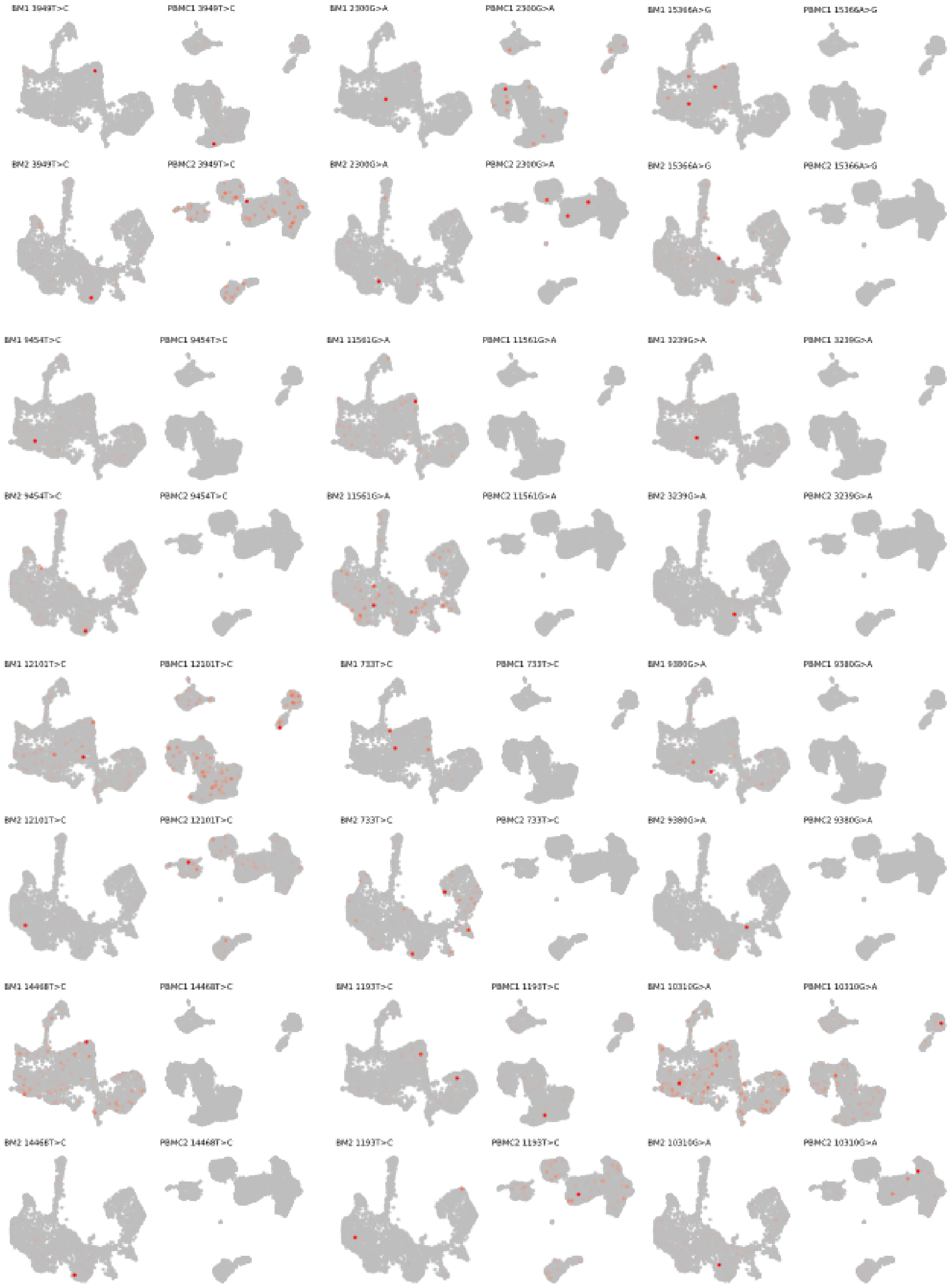


Figure A.4: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for the mutations from the lineage biased clones towards HSC self-renewal. (PART 2)

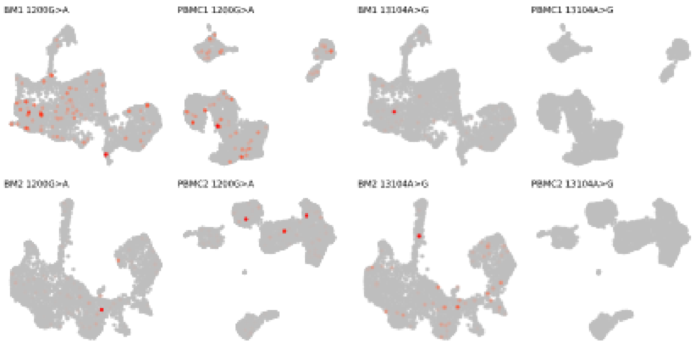


Figure A.5: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for the mutations from the lineage biased clones towards HSC self-renewal. (PART 3)

A.4.2 Lymphoid vs. Myeloid

Mutation	Chisq stats	pvalue	p adjust	nr. cells	odds ratio
2788C>A	24.7984191886274	6.36496478550977e-07	0.000246	380	2.46
7374A>G	22.8322259136213	1.76776583709284e-06	0.000684	5	0
3209A>G	21.8015736144636	3.02352054422938e-06	0.001170	7	0.03
3244G>A	14.0188421559237	0.000180987890606431	0.070042	26	0.25
11167A>G	13.056422329352	0.00030224622102532	0.116969	5	0.05
2784A>C	9.85426137463878	0.00169438814368686	0.655728	109	3.19
7919G>A	9.65651296937938	0.00188681430756684	0.730197	6	0.10
6890A>G	8.62943879320659	0.00330773460138613	1	9	0.17
6190G>A	7.29057449030523	0.00693173066224561	1	7	0.16
15106G>A	7.29057449030523	0.00693173066224561	1	7	0.16
9553G>A	5.99479230744131	0.0143481702365158	1	5	0.14
14710G>A	5.99479230744131	0.0143481702365158	1	5	0.14
7242T>C	5.99479230744131	0.0143481702365158	1	5	0.14
11228T>C	5.99479230744131	0.0143481702365158	1	5	0.14
3714A>G	5.99479230744131	0.0143481702365158	1	5	0.14
1082A>G	5.88623606150608	0.0152596822337696	1	14	0.29
15639T>C	5.64042941839096	0.0175509079237964	1	11	0.26
3242G>A	5.57086781179263	0.0182617078823791	1	8	0.21
15699G>A	5.57086781179263	0.0182617078823791	1	8	0.21
2407T>C	5.57086781179263	0.0182617078823791	1	8	0.21
1949G>A	4.32720022623751	0.0375078933510424	1	27	0.43
1201A>G	4.2819823336544	0.0385184004016032	1	9	0.27
5894A>G	4.2819823336544	0.0385184004016032	1	9	0.27
14410G>A	4.17815085884447	0.0409483552498958	1	6	0.21
3695T>C	4.17815085884447	0.0409483552498958	1	6	0.21
4057T>C	4.17815085884447	0.0409483552498958	1	6	0.21
15968T>C	4.17815085884447	0.0409483552498958	1	6	0.21

Table A.3: Results of the Chi-Squared test for lineage bias for lymphoid or myeloid, only mutations with a p-value < 0.05 are reported. Odds ratio above 1 defines clones biased towards lymphoid, below 1 towards myeloid. P values were adjusted with a Bonferroni correction.

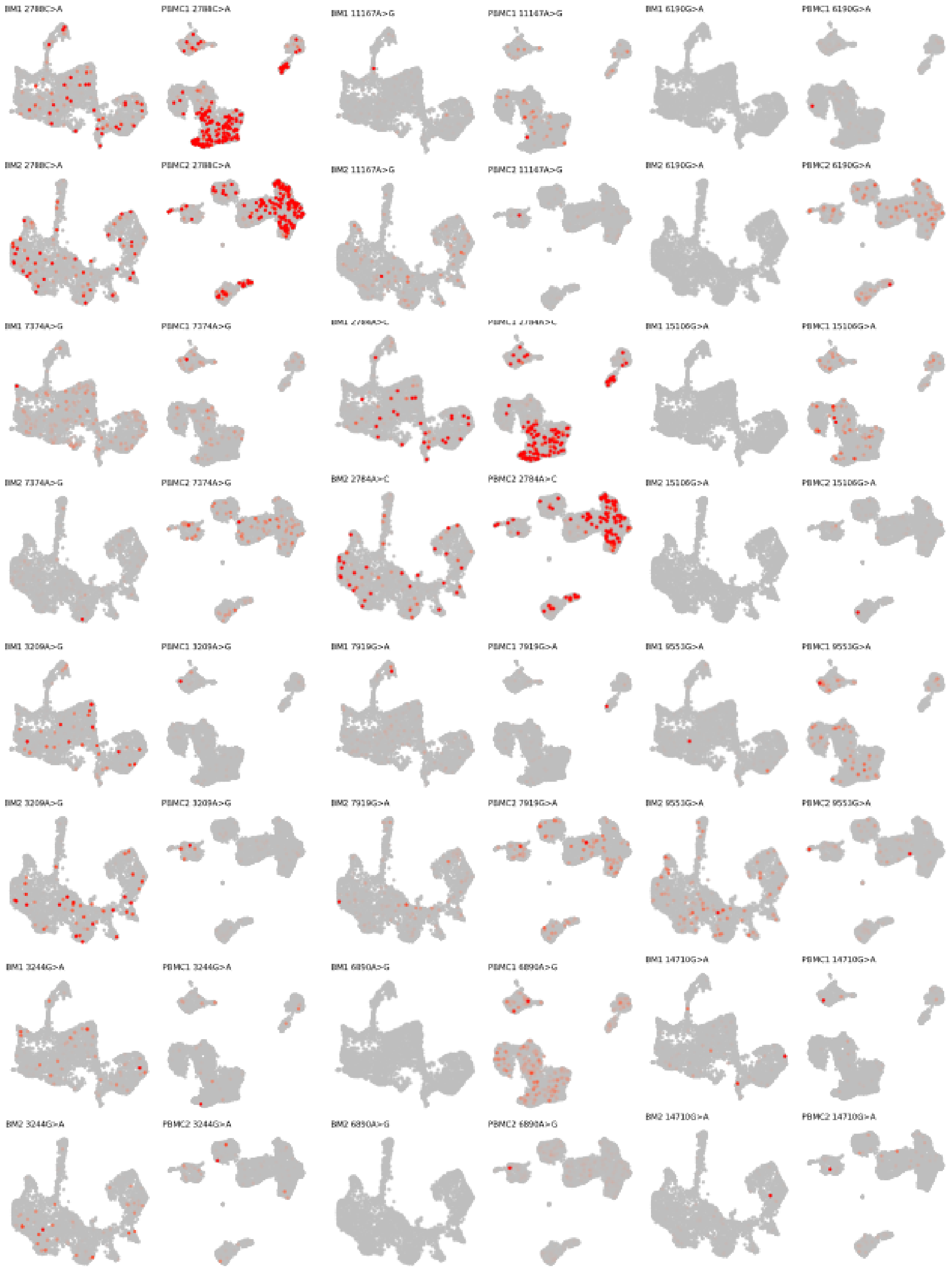


Figure A.6: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for the mutations from the lineage biased clones towards lymphoid or myeloid. (PART 1)

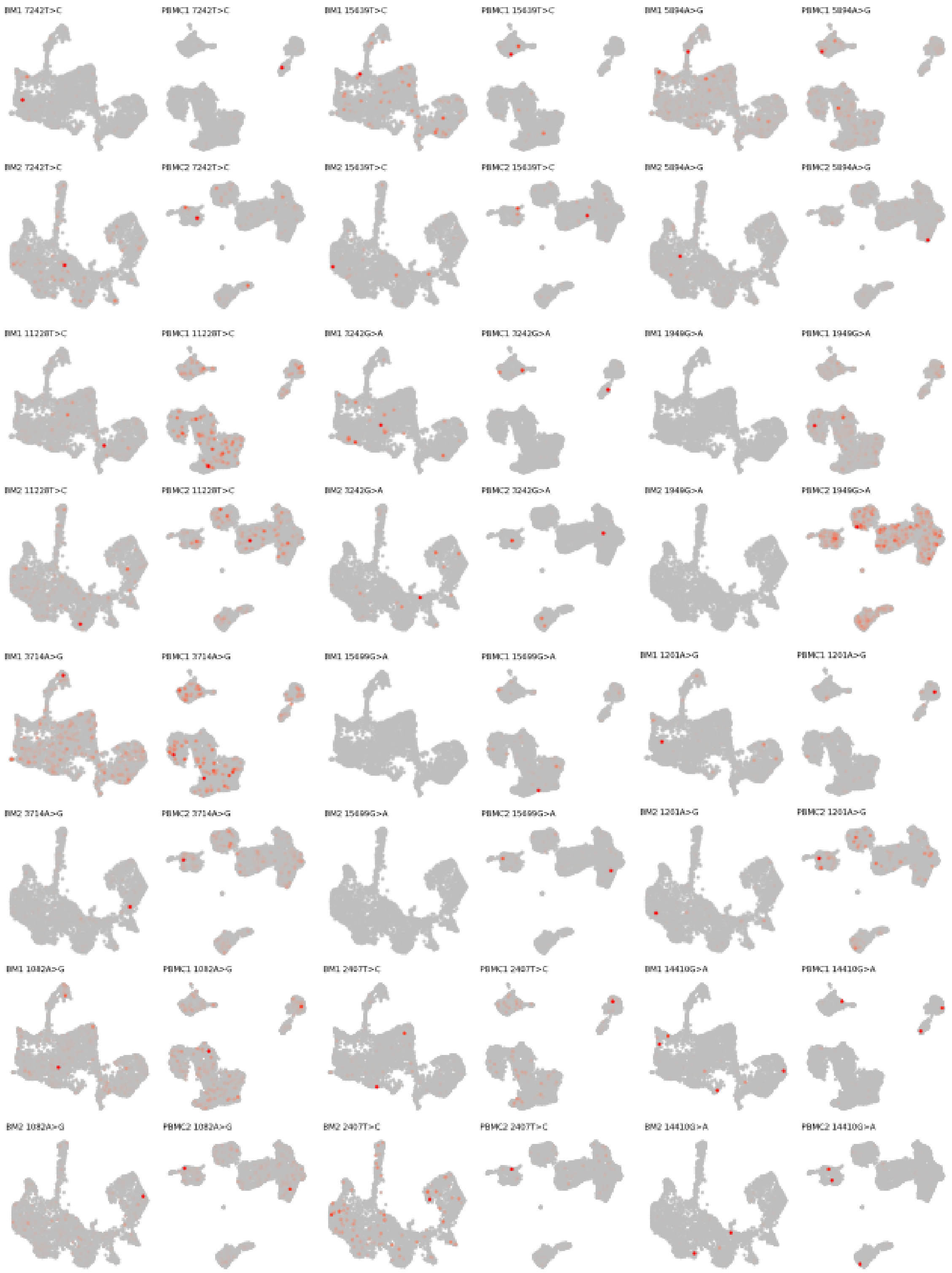


Figure A.7: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for the mutations from the lineage biased clones towards lymphoid or myeloid. (PART 2)

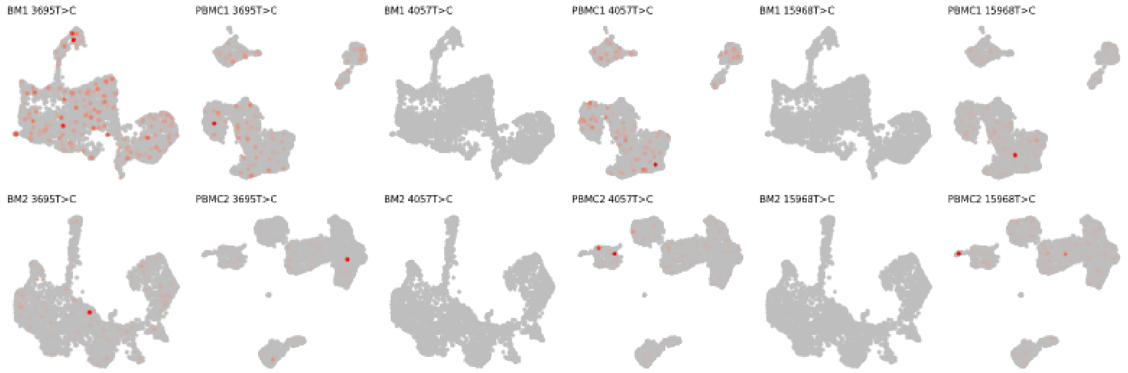


Figure A.8: CD34+ HSPCs and PBMCs embedding in the UMAP, cells are colored based on the allele frequency for the mutations from the lineage biased clones towards lymphoid or myeloid. (PART 3)