Probing Steady-state and Post-transplant Blood System Dynamics with Computational Analysis and Lineage-tracing

by

Michael K. Kuoch

S.B. Computer Science and Engineering, MIT, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by:	Michael K. Kuoch Department of Electrical Engineering and Computer Science January 19, 2024
Certified by:	Michael B. Yaffe Professor of Biology, Professor of Biological Engineering, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Probing Steady-state and Post-transplant Blood System Dynamics with Computational Analysis and Lineage-tracing

by

Michael K. Kuoch

Submitted to the Department of Electrical Engineering and Computer Science
on January 19, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Bone marrow transplants are an important tool in modern medicine due to their ability to treat a wide range of diseases, spanning both cancerous and non-cancerous conditions. We aim to study the blood system dynamics using sequencing data from paired pre-transplant and post-transplant samples and look for potential expression profiles that may be biased toward successful bone marrow engraftment. We find that some genes have increased expression in post-transplant samples compared to pre-transplant samples. We also discuss using clonal lineage tracing to track cell clones throughout the transplant process and present some preliminary analyses.

Thesis supervisor: Michael B. Yaffe
Title: Professor of Biology, Professor of Biological Engineering

Acknowledgments

I would like to thank Dr. Hojun Li for extensive mentorship and support throughout this thesis project and beyond. Without his guidance, none of the work presented here would be possible.

Contents

Title page 1

Abstract 3

Acknowledgments 5

List of Figures 9

List of Tables 11

1 Introduction 13
1.1 Bone Marrow Transplants . . . 13

2 Related Work 14

3 Preliminary Analysis of Bone Marrow Transplant Samples 16
3.1 Introduction . . . 16
3.2 Dataset and Methodology . . . 16
3.2.1 Dataset Overview . . . 16
3.2.2 Dataset Quality Control . . . 17
3.2.3 Dimensionality Reduction . . . 17
3.2.4 Differential Expression . . . 18
3.2.5 Clustering . . . 19
3.3 Results . . . 19
3.4 Discussion . . . 19
3.5 Next Steps . . . 20

4 Clone Tracing at a Single Cell Resolution 27
4.1 Introduction . . . 27
4.1.1 Challenges . . . 27
4.2 Methodology . . . 28
4.3 Results . . . 29
4.4 Discussion . . . 29
4.5 Next Steps . . . 30

5 Conclusion 31

A Differential Expression Results Expanded **32**

References **34**

List of Figures

3.1 Results from quality control analyses. 21
3.2 Distribution of cells displayed with dimensionality reduction. 22
3.3 Top 2 differentially expressed genes in the pre-transplant sample and the post-transplant sample. 23
3.4 Top 2 differentially expressed genes in the pre-transplant sample and the post-transplant sample (continued) 24
3.5 Leiden clustering 25
3.6 Cell type prevalence 26

A.1 Differential expression extended 33

List of Tables

3.1 Sample assignments using hashtag-oligos 18

Chapter 1

Introduction

1.1 Bone Marrow Transplants

Since their inception in the mid-1900s, bone marrow transplants (also known as hematopoietic stem cell transplants) have become increasingly common in modern medicine due to their ability to treat a wide range of diseases, spanning cancerous (leukemia, lymphoma, etc) and non-cancerous conditions (various forms of blood disorders, etc). The goal of a bone marrow transplant is to address dysfunctional bone marrow in a patient by administering healthy bone marrow stem cells to replace or destroy problematic bone marrow cells in the patient [1].

Despite the potential of bone marrow transplants to treat many serious illnesses, there still exist dangers that must be carefully weighed when deciding whether or not to perform a bone marrow transplant. One of the most serious dangers is known as graft failure, which occurs when the administered bone marrow stem cells do not take residence in the host, and thus cannot adequately repopulate the blood system. This presents serious risks to the patient, with increased likelihood of infections and mortality [2].

We wish to study the dynamics of post-transplant bone marrow engraftment in order to better understand what factors contribute to promoting engraftment, as well as to get a sense of how these normal engraftment processes may go awry when problems occur, such as during graft failure. A better understanding of these processes could lead towards potentially better procedures and better patient outcomes in the future. To do this, we propose performing computational analysis on samples taken from bone marrow transplant patients at time points throughout the treatment process. We also propose performing lineage tracing, which involves identifying and tracing cell clone lines across blood samples from different time points from a given patient. This will help us better understand which types of cells tend to persist in a patient's blood system post-transplant.

Chapter 2

Related Work

To our knowledge, there has been no prior work specifically looking at comparing pre-transplant (initial infusion) and post-transplant blood samples in humans to look for differences and potential expression biases in engraftment systematically. Thus, we believe the work outlined in this manuscript will be helpful in guiding future investigation in the area. In this section, we discuss prior work that helped to conceive the experiments discussed in Chapter 4.

One of the most informative works [3] used lentiviral barcoding in mice to track cell clonal lineages across long-term bone marrow reconstitution following a transplant procedure. By probing for the lentiviral barcodes throughout the time course, they were able to identify a single gene, Tcf15 (which codes for a transcription factor), as being "required, and sufficient, to drive HSC quiescence and long-term self-renewal". This demonstrated that novel physiology can be uncovered with the power of clonal lineage tracing.

Since using lentiviral barcodes to perform clone tracing in humans isn't possible for safety and ethical reasons, some works [4] have shown that natural somatic mutations can be used to identify and trace clonal lineages, rather than needing to introduce artificial barcodes. They do this by expanding out clonal populations and performing whole-genome sequencing to identify potential somatic mutations, followed by target sequencing of the variant regions. With this information, they were able to reconstruct a phlegmatic tree of cell clones, estimate the times of embryonic and extra-embryonic tissue divergence during development, and infer the embryonic origin of certain tissues. These results opened the path to clonal lineage tracing in humans.

Other works have shown that it is possible to identify clones at a single-cell resolution while voiding the hassle of expanding colonies, performing whole genome sequencing, and targeted sequencing. Instead, one can simply use a standard single-cell sequencing workflow as shown in this work [5], which uses single-cell ATAC sequencing to identify somatic mutations in mitochondrial DNA. In their procedure, the ATAC sequencing enables high coverage sequencing of the mitochondrial genome, allowing them to perform clone tracing and infer cellular relationships based on mitochondrial DNA variants. Other recent works [6] have built upon this procedure. showing that mitochondrial DNA variants can be identified at a single-cell resolution from even a normal single-cell RNA sequencing cDNA library. They do this by using a specific pool of primers and PCR-amplifying out mitochondrial transcripts from the cDNA library to enrich for mitochondrial RNA. This protocol opens up the

potential to add clonal tracing analysis to any single-cell sequencing experiment.

The final related work we would like to touch upon is the development of the MAS-seq sequencing technology [7], which enables high-throughput long-read sequencing via complementary cDNA concatenation. Since the long-read technology can capture more of the bases of each of the cDNA molecules, we expect that it will provide greater per-base coverage of the exome. In Chapter 4 we will discuss more about how this may be useful.

In Chapter 3, we explore traditional single-cell analysis techniques to explore the blood system following a transplant. In Chapter 4, inspired by the work in clone tracing [3] [4] and armed with the technology of mito-seq [5], MAESTER-seq [6], and MAS-seq [7], we explore utilizing clone tracing to uncover blood system dynamics in the transplant setting.

Chapter 3

Preliminary Analysis of Bone Marrow Transplant Samples

3.1 Introduction

In this chapter, we aim to identify differences between pre-transplant and post-transplant bone marrow samples using standard single-cell sequencing analyses. We hope this will give insight into expression profiles that promote successful engraftment.

3.2 Dataset and Methodology

3.2.1 Dataset Overview

For the transplanting data described in this work, we have pre-transplant and post-transplant bone marrow samples from two different patients. cDNA was isolated from cells in these samples and three distinct types of sequencing were performed on these samples. In addition to normal gene expression sequencing, we performed MAS-seq [8] and MAESTER [6], discussed more in Chapter 4. The regular gene expression library will be the focus of Chapter 3.

It is worth noting that the cDNA from these four samples were stained with hashtag oligos (HTOs) and sequenced together [9]. After sequencing, the cells were demultiplexed based on their HTO counts using the hashsolo [10] function in the Scanpy library [11]. When the demultiplexing was performed, no cells from one of the post-transplant samples were identified. It is unclear if this was because many of the cells from those samples died, or if the HTO used for that sample had low quality. Thus, for the analyses that follow, we do not have data for one of the post-transplant samples.

The first step in the analysis performed in this section was to align the sequencing read data from the regular gene expression library to the human reference genome and perform quality control on the resulting information. To perform the alignment we used the cellranger [12] count function provided by 10x Genomics to generate aligned BAM files from the fastq files. The cellranger count function identified 11117 barcodes belonging to cells in this dataset, and for all downstream analyses we use only these cells (and ignore reads belonging

to other barcodes). Furthermore, there was a mean of 13516 reads per cell and a median of 1779 genes detected per cell.

3.2.2 Dataset Quality Control

After running cellranger counts, we obtain a count matrix with rows representing cells and columns representing genes, where each entry is the number of reads the respective cell has for the respective gene. We next perform quality control on this data, removing potentially bad cells (which could be for a variety of reasons including dead cells or cells with low gene counts). We also view some basic summary statistics to better understand our sequencing data. Once again, we use the Scanpy library to perform these analyses. See Figure 3.1 for some of the results of these preliminary analyses.

In 3.1b we see that the most expressed gene by far is MALAT1, which is unexpected. As mentioned on the 10x genomics website, this generally is inversely associated with cell health, so it's worth keeping in mind that there may be many low-quality/dead cells in this dataset.

The first quality control step we perform is to remove all cells that have less than 200 unique genes and filter out genes that appear in less than 3 unique cells. This allows us to decrease the size of our dataset, since these cells and genes will likely not provide useful information anyway due to their low quality/prevalence. This step filters out 128 cells and 11325 genes. In order to limit the influence of low-quality/dead cells on our analysis, we set a mitochondrial read count cutoff at 10%. As seen in Figure 3.1, the cutoff of 10% is able to keep many of the cells in the lower arm, which are assumed to be high-quality cells. Any cells with a mitochondrial read percentage greater than 10% are assumed to be dead and are not considered in further analyses. Furthermore, we also set a cutoff at 6000 unique genes; any cells with more than 600 unique genes are assumed to be doublets and are excluded from further analyses. These two filters combined removed 952 cells. After all these filtering steps, we had 10035 cells and 25276 genes.

Finally, we demultiplex our four samples using the HTO staining (see table 3.1 for the hashtag assignments). As mentioned above, we don't observe any post-transplant samples from one of the patients. For the preliminary analyses, we only consider cells from the patient with both initial transfusion and post-transplant samples. This leaves us with 6283 cells for the next analysis.

3.2.3 Dimensionality Reduction

The primary type of data that is used in single-cell RNA sequencing data is count data, which represents the number of reads we see of each gene for a given cell. This is typically represented as a count matrix, where each row of the matrix is a vector of counts for a single cell. In order to visualize this data more easily, we can perform *dimensionality reduction* on the data. Typically the total number of dimensions is reduced to two or three dimensions so they can be visualized on a plot. There are a few different methods used for dimensionality reduction, but a few of the common ones are Principle Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP).

Table 3.1: Sample assignments using hashtag-oligos, demultiplexed with the Hashsolo algorithm [10] on Scanpy

Hashtag Assignment	Sample Assignment	Count
Hashtag 1	Patient 1 (initial transfusion)	2424
Hashtag 2	Patient 1 (post-transplant)	3859
Hashtag 9	Patient 2 (initial transfusion)	2172
Hashtag 10	Patient 2 (post-transplant)	0
Doublet	NA	1885
Negative	NA	26

PCA [13] is a linear dimensionality reduction technique that aims to find a (typically lower dimensional) basis such that the reconstruction error upon projection to the basis is minimized. Equivalently, PCA can also be viewed as finding the subspace that maximizes the variance of the data after projection. The basis vectors of this subspace are known as "principle components". Intuitively, this can be thought of as finding the "most informative" linear subspace of the data to then be used in downstream analyses (in this work, for visualization). The PCA basis can be found by finding the SVD (singular value decomposition) of a matrix and taking left singular vectors, or by taking the top eigenvectors of the data covariance matrix. It is worth noting that the principal components are always orthogonal.

UMAP [14] is a nonlinear dimensionality reduction technique that uses manifold approximations to minimize the the cross-entropy between topological representations of the original data and a low-dimensional version of the data. Because, unlike PCA, this is a nonlinear dimensionality reduction technique, UMAP can often find more interesting patterns and differences than linear techniques like PCA. However, this comes at the cost of decreased interpretability. Namely, the resulting lower dimensional UMAP coordinates are not as interpretable as the principal components. This is because the relationship between the original and reduced coordinates complex in a UMAP, compared to a simple matrix multiplication in PCA.

3.2.4 Differential Expression

One useful type of analysis is differential expression, which seeks to determine whether the magnitude of gene expression between different groups of cells is statistically significant (this is done on a per-gene basis). This typically utilizes some type of statistical test like a t-test or a Wilcoxon rank-sum test to determine whether or not the mean expression of a gene is statistically significant between different groups.

This can allow us to determine things like marker genes, or whether certain genes are up-regulated or down-regulated as a result of some type of treatment. In our case, we look to see if any genes are differentially expressed between the pre-transplant and post-transplant samples for a given patient.

3.2.5 Clustering

Clustering is a useful tool to group our cells into meaningful groups. Using the raw count data, or potentially the reduced dimension data from PCA, we can use a clustering algorithm to group the cells into clusters that share similar expression profiles. One common clustering algorithm used in single-cell sequencing analysis is Leiden clustering [15], which makes improvements over previously published clustering algorithms. When using the Leiden clustering algorithm, one doesn't need to specify the number of clusters beforehand. Instead, a resolution parameter is specified which roughly controls how many clusters will be defined. A higher resolution leads to more clusters, while a lower resolution generally yields less clusters.

3.3 Results

Our first step is to use dimensionality reduction to visualize the data. In Figure 3.2 we can see that while the PCA visualization is not able to reveal significant differences between the pre-transplant and post-transplant samples, the UMAP visualization does show some difference between the two. Specifically, in the UMAP visualization we see there are some regions where only either pre-transplant or post-transplant cells are located.

In order to further investigate these potential differences, we perform differential expression (3.2.4) to determine whether particular genes were associated with these differences. To do this, we use the Scanpy rank_genes function with a Wilicoxon-rank sum test (which is non-parametric) rather than the default t-test. The top differentially expressed genes are shown in Figure 3.3a.

To avoid the issues of cell type compositional changes between the pre-transplant and post-transplant samples, we next perform differential expression on a per-cluster basis to control for cell type composition. To generate the clusters, we use Leiden clustering 3.2.5 with a resolution of 0.12, yielding 7 clusters. This resolution was chosen to break up the observed groups of cell types based on cell type marker genes. These marker genes were then used to annotate each of the clusters. The clustering results and annotations are shown in 3.5.

Additionally, we can look at how the relative abundance of cell types changes between the pre-transplant sample and the post-transplant sample. These results are shown in Figure 3.6. The largest changes are an increased proportion of CD8 positive T-cells and a decreased proportion of B-cell in the post-transplant sample compared to the pre-transplant sample.

3.4 Discussion

The results from the all-cell pre-transplant vs post-transplant differential expression are interesting. The topmost differentially expressed genes are STAT1 and FYB1 for the post-transplant sample and BTG1 and CXCR4 for the pre-transplant sample. However, they come with some complications that may inhibit interpretability. For example, FYB1 is typically expressed in T-cells, so the fact that FYB1 is expressed more in the post-transplant samples could be due to the fact that there are more T-cells in the post-transplant sample, rather than

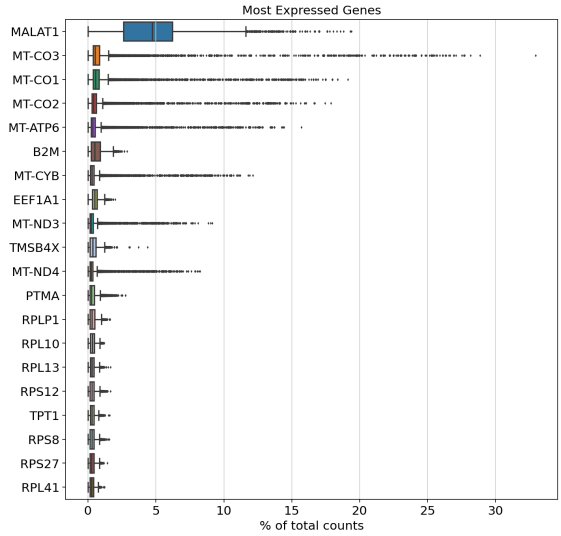
due to some underlying difference in the physiology of the cells. These observations motivated the differential expression analysis by individual Leiden groups. Despite this, the all-cell differential expression analysis could still provide useful information. Genes like STAT1 and CXCR4 are not traditionally associated with particular cell types in hematopoiesis, so they may be worth investigating further.

Looking at the differential expression results at an individual cluster level reveals more potentially interesting genes. It's interesting to note that STAT1 and BTG1 still appear as differentially expressed in the CD8 positive T-cell cluster (3.4b). In fact, multiple cell types besides the CD8 positive T-cells show enriched Jak/Stat associated genes in the post-transplant sample compared to the pre-transplant sample (A.1). Xist is also enriched in the post-transplant sample for multiple groups, which is an unexpected finding as this gene isn't traditionally associated with blood cell functions. Another interesting observation is that both the erythroid cluster and the B-cell progenitor cluster have fewer statistically significant differences. In fact, the erythroid cluster has no statistically significant differences at all (Figure A.1b. However, this could also be due to the fact that there aren't many cells in these clusters.

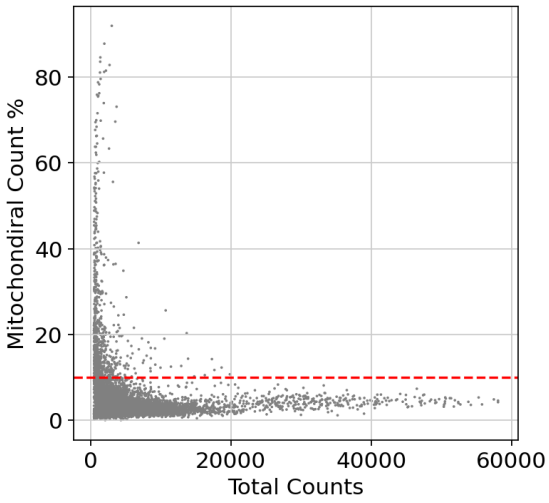
Considering the differences in cell type prevalence between the pre-transplant and post-transplant samples, we see that the largest differences can be found in the T-cell and B-cell compartments. Specifically, we observe that CD8 positive T-cells make up a bigger portion of the post-transplant sample, and that B-cells and CD8 negative T-cells make up a larger proportion of the pre-transplant sample.

3.5 Next Steps

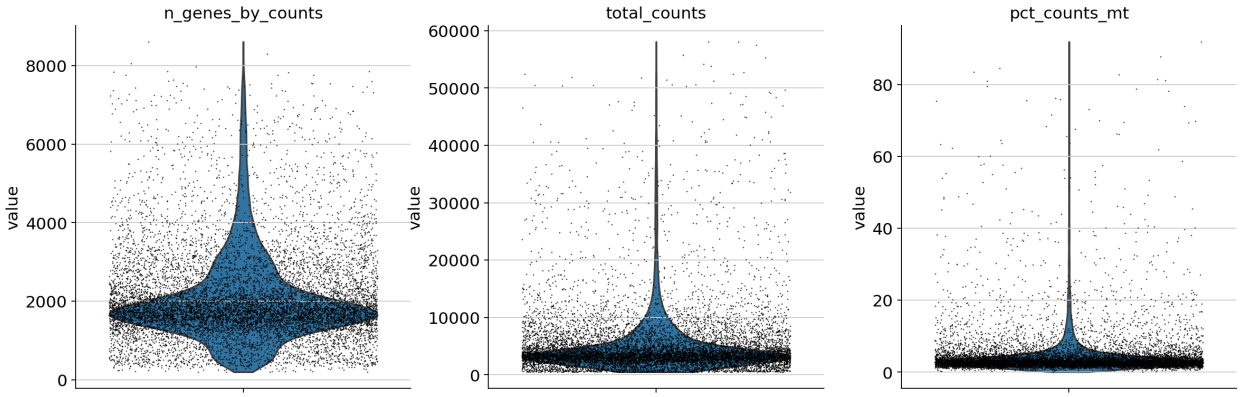
The findings presented lead towards clear follow-up directions. First and foremost, a more systematic search of the differentially expressed genes could reveal patterns that were initially missed. This could be done with tools like gene ontology or gene set enrichment analysis to see if there are genes or pathways that are consistently different between the pre-transplant and post-transplant samples. Similarly, it would be interesting to further investigate the genes we've manually picked out above (including STAT1, JAK2, and XIST). One way this could be done experimentally would be to knock out these genes in a bone marrow transfusion sample and see if they can still engraft following a transplant procedure.



(a) Top genes by read count percentage. MALAT1 is the most expressed gene, which suggests that there may be many low-quality cells in the library. Interestingly lots of the other top genes are mitochondrial genes, which suggests a similar potential for low-quality cells.

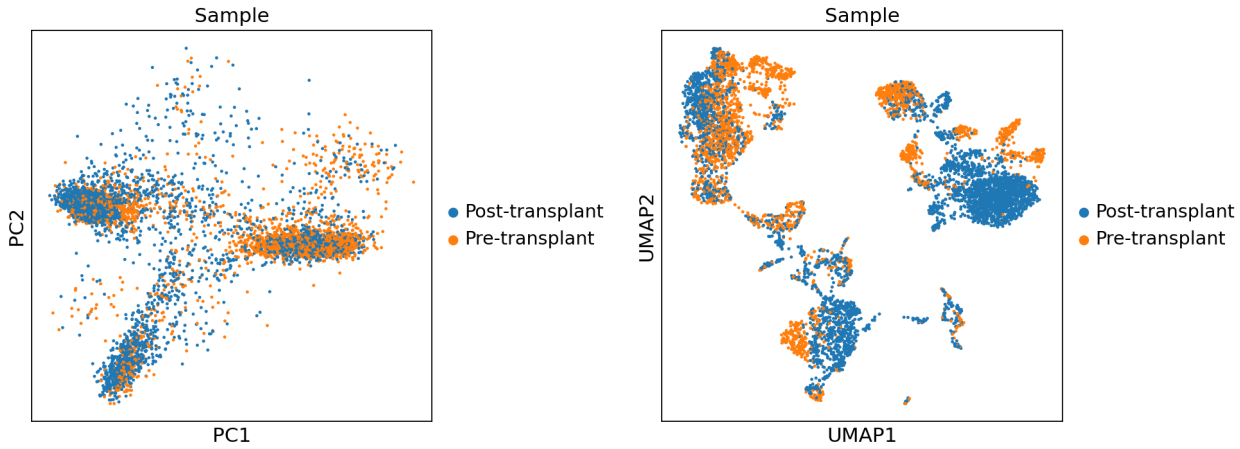


(b) Scatterplot of a cell's total number of counts vs the percentage of those that are mitochondrial reads. The red line shows the 10% cut-off we use to filter cells. All cells with 10% or more mitochondrial reads are removed from further analyses.



(c) High-level statistics

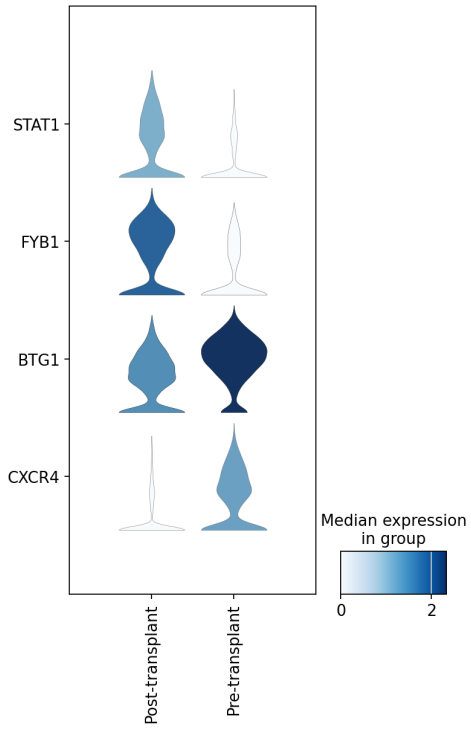
Figure 3.1: Results from quality control analyses. We use these results to inform low-quality cell removal from the dataset.



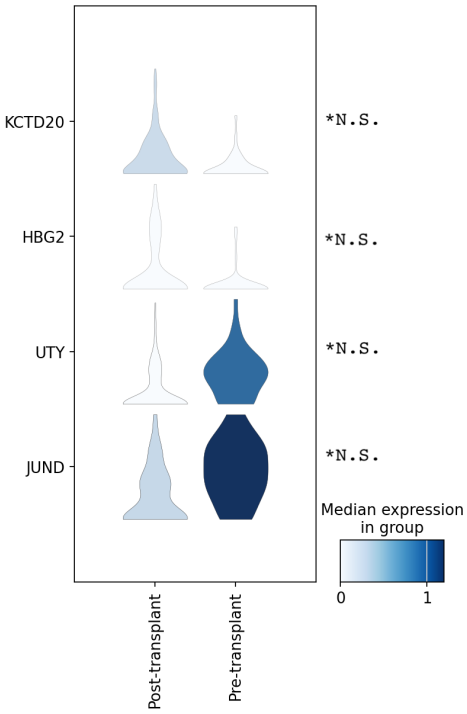
(a) First two PCA coordinates

(b) First two UMAP coordinates

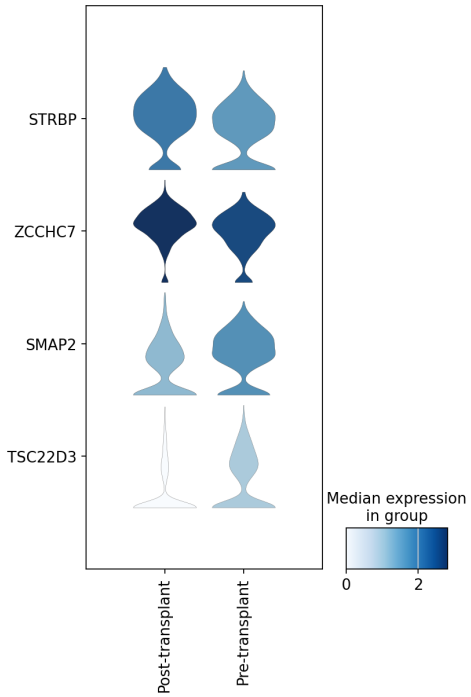
Figure 3.2: Distribution of cells from pre-transplant and post-transplant samples displayed with dimensionality reduction. The two samples seem to distribute similarly when viewed with PCA, but some differences can be observed when using UMAP.



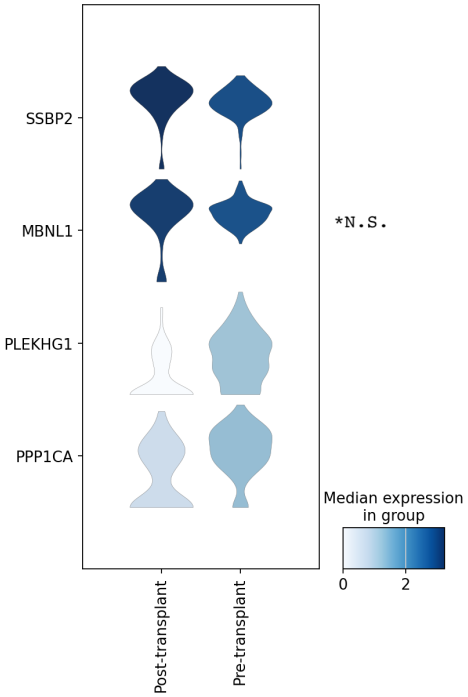
(a) All cells from the patient together



(b) Erythroid cells

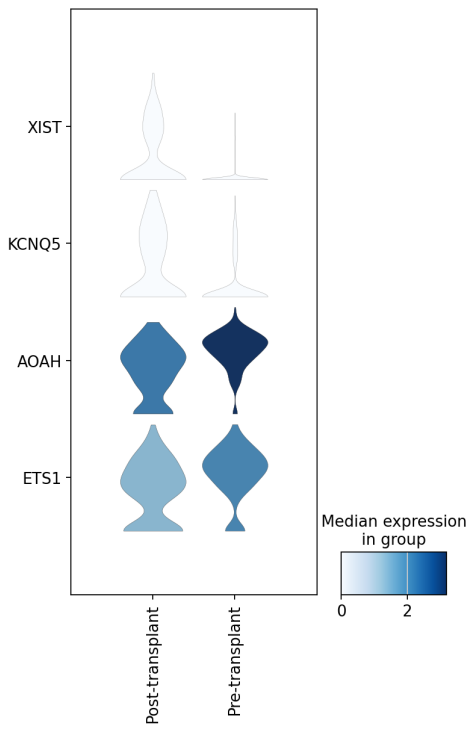


(c) B-cells

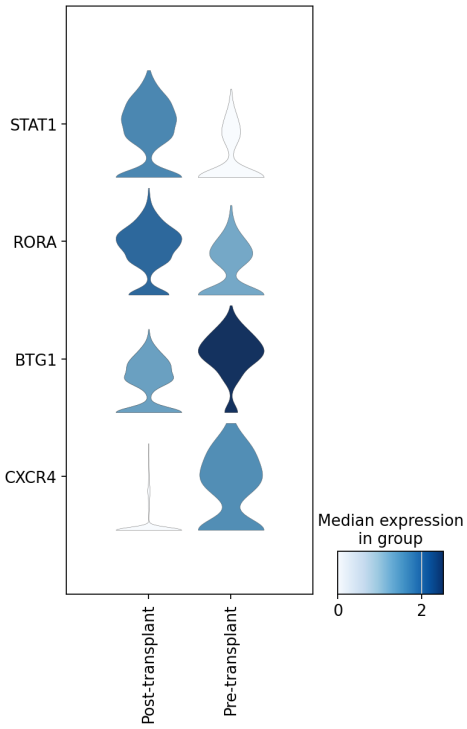


(d) B-cell progenitor

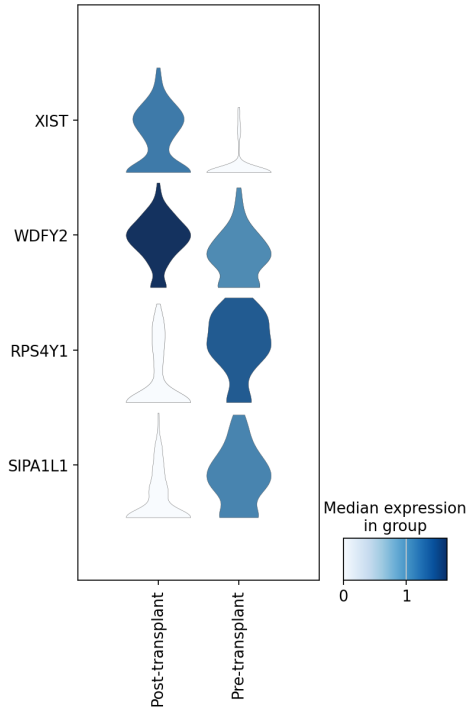
Figure 3.3: Top 2 differentially expressed genes in the pre-transplant sample and the post-transplant sample for all cells (a) and by Leiden cluster (b-h). All of these differences in genes shown here are statistically significant (adjusted p-value below 0.05) except where marked with *N.S., in which case the difference is not statistically significant. For a longer list of differentially expressed genes, see A.1



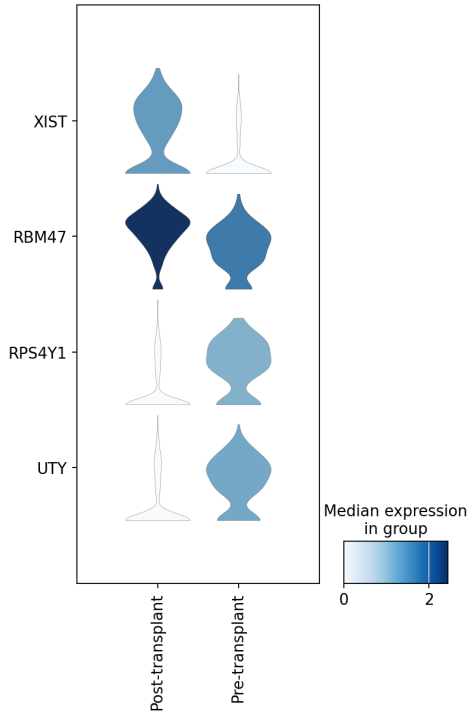
(a) T-cell



(b) CD8 positive T-cell

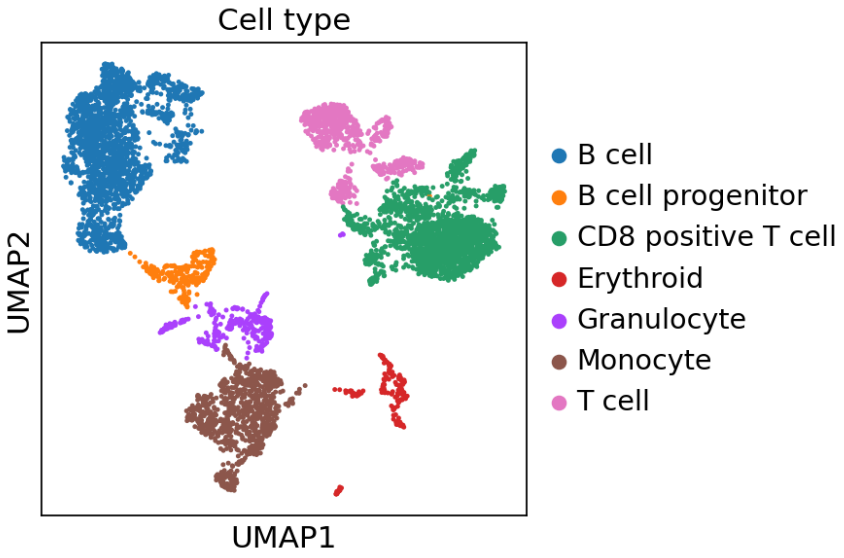


(c) Granulocyte

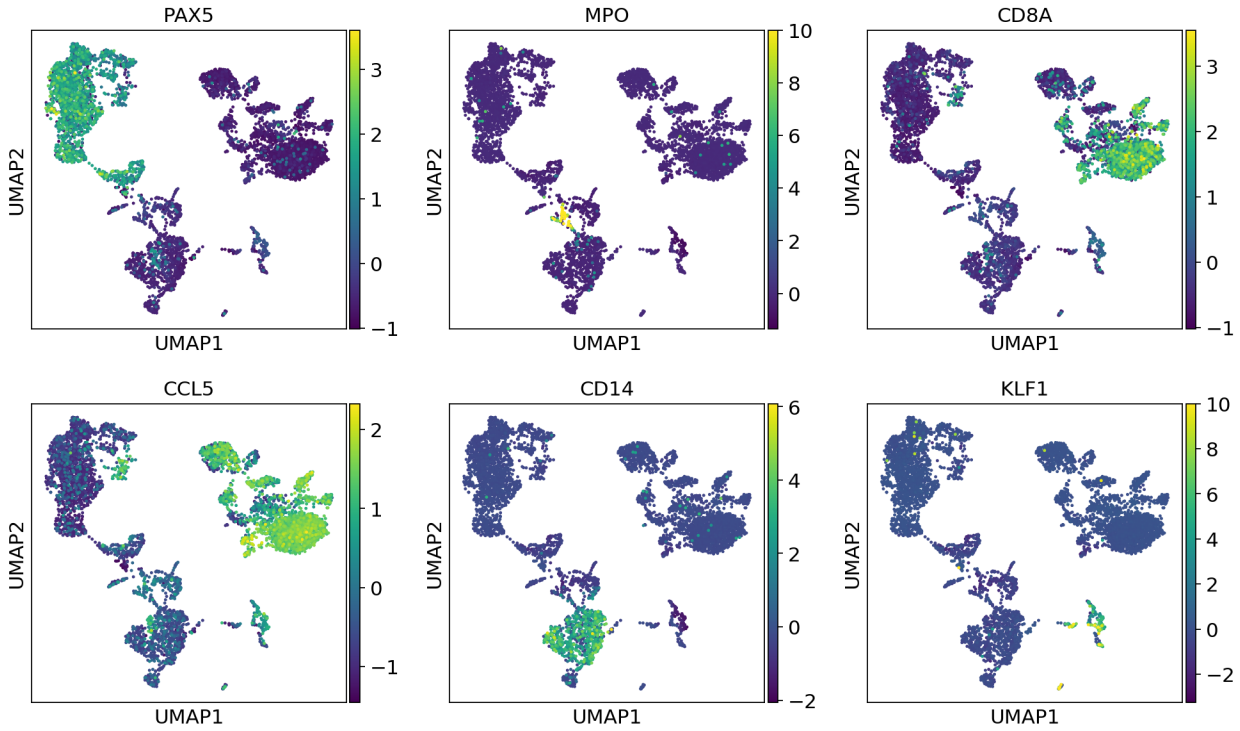


(d) Monocyte

Figure 3.4: Top 2 differentially expressed genes in the pre-transplant sample and the post-transplant sample by Leiden cluster (continued from Figure 3.3. All of these differences in genes shown here are statistically significant (adjusted p-value below 0.05). For a longer list of differentially expressed genes, see A.1



(a) Marker genes were used to identify which cell type of each of the Leiden clusters represented



(b) Some of the marker genes used to identify the clusters. The color scale shows a normalized expression value per cell.

Figure 3.5: Leiden clustering with a resolution of 0.12 yielded the clusters shown here. These clusters were annotated manually using marker genes. There are few granulocytes likely because mature neutrophils lyse during the preparation procedure.

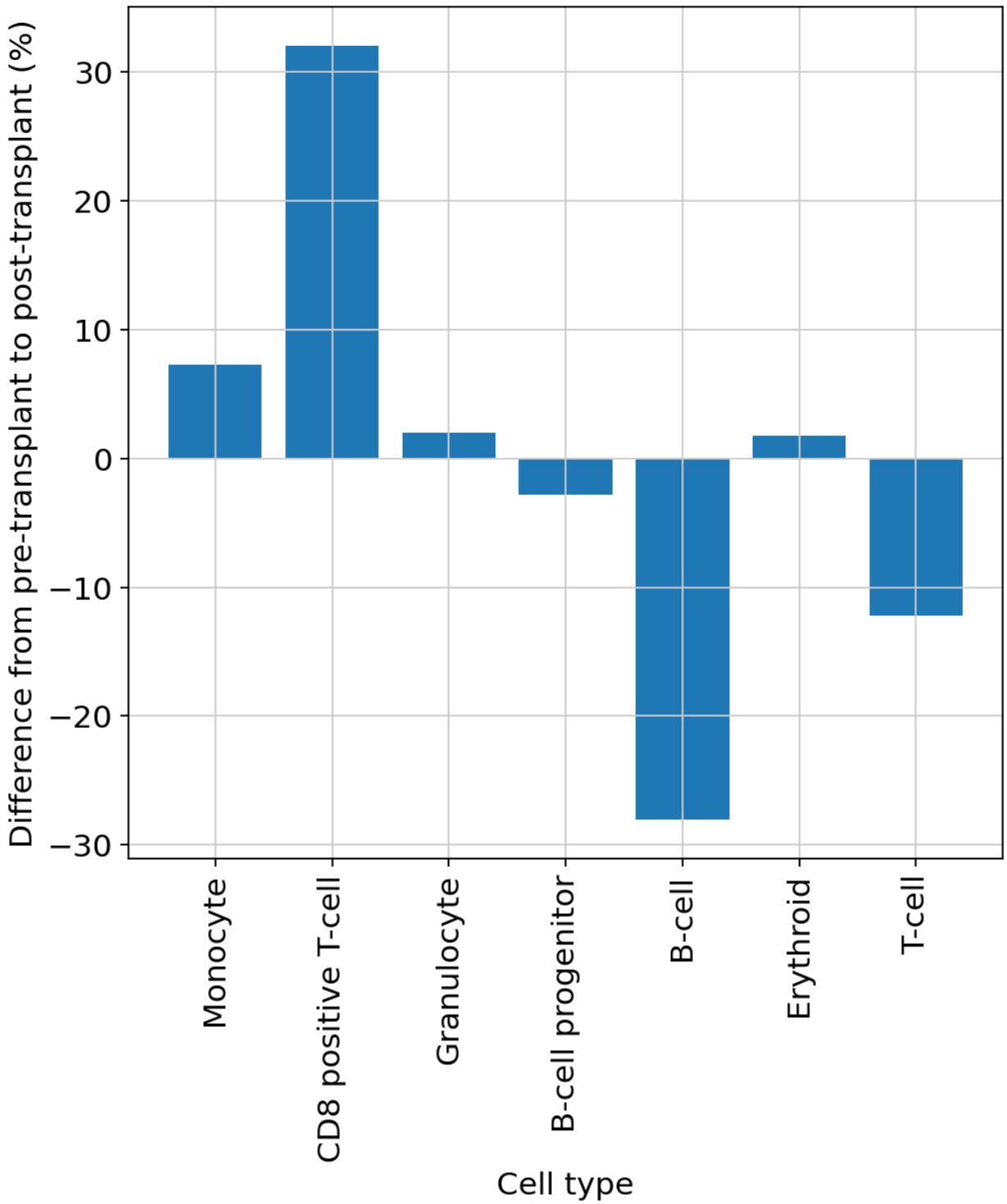


Figure 3.6: Differences in cell type prevalence from the pre-transplant sample to the post-transplant sample. A positive difference means that there is a higher prevalence of the cell type in the post-transplant sample (in terms of percentage).

Chapter 4

Clone Tracing at a Single Cell Resolution

4.1 Introduction

As discussed in the related work (Chapter 2), lineage tracing has been a powerful technique to uncover novel biology [3][4]. Somatic mutations have been used to identify lineages and trace cell clones throughout the human lifetime (2, [4]). In this section, we discuss harnessing the natural variation found in somatic mutations to study the dynamics of the human blood system throughout the transplant process.

Specifically, we want to use lineage tracing to identify clonal lineages that are present in the initial pre-transplant blood sample, as well as in the post-transplant sample. The reason for this is that finding the same clonal linage in both samples suggests that a progenitor of that clone was able to successfully engraft in the patient. With this information, we may be able to learn more about what genes and expression profiles can promote successful engraftment following a bone marrow transplant procedure.

As of the time of publication, the work in this section is still in progress. For now, we describe our high-level motivations and some primary results.

4.1.1 Challenges

It turns out that attempting to identify and assign somatic variants at a single-cell resolution is quite difficult, more so than one may initially expect. In this work we focus on single nucleotide polymorphisms, however, these discussions also extend to other types of variants (copy number variants, etc). Two major obstacles are sequencing errors and sparse/low genome coverage.

The first and perhaps more apparent problem is that sequencing technologies are not 100% accurate. For example, Illumina sequencing (the recommended sequencing platform for 10X Genomics single cell library preps) has an error rate of approximately 0.1% to 0.6% per base depending on the particular instrument used [16]. The existence of sequencing errors makes it difficult to tell whether an observed base difference between two cells is the result of a somatic mutation, or if it is simply a sequencing error. There are ways to reduce the impact of errors; for example, one could omit error-prone sites in the genome, or could potentially require that somatic variants be present in a non-trivial set of cells. However,

it presently seems that it is impossible to fully eliminate the impact of sequencing errors on somatic variant identification and assignment. This is especially true given the fact that sequencing errors can be correlated (the same type of error at the same position may be likely), making them even harder to distinguish from true somatic variants.

The second problem arises from the diploid nature of the human genome. Consider the following hypothetical scenario: we know that there is a somatic variant at some position on some particular chromosome. Some cells are homozygous at the locus (for the sake of argument, say both alleles are adenine), and others contain a somatic variant and are instead heterozygous (say one allele is adenine and the other allele is guanine). Now imagine we are trying to determine whether or not a cell contains the somatic variant by analyzing reads that cover the variant position, and we see that all n of the observed reads are adenine. Should we conclude that the cell contains the somatic variant or not? Both a wild-type and variant cell could potentially give rise to the observed reads.

With some assumptions, we can compute the probability that the cell does indeed contain the variant. Assuming a 50-50 prior and assuming that each allele is equally likely to be sampled, we find via Bayes' Rule that the probability of the cell containing the variant is

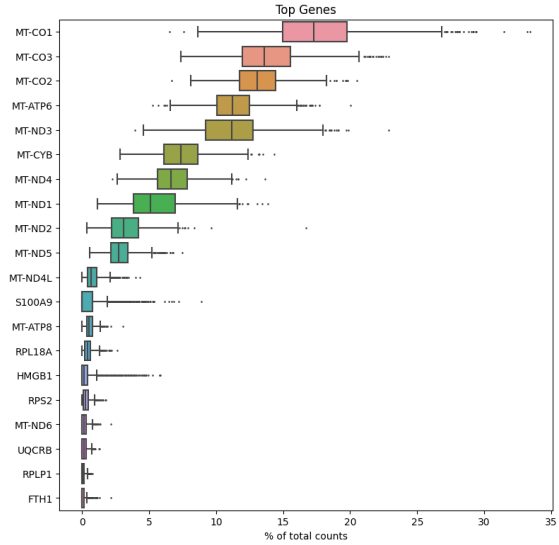
$$\mathbb{P}(variant|reads) = \frac{\mathbb{P}(reads|variant)\mathbb{P}(variant)}{\mathbb{P}(reads)} = \frac{\left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)} = \frac{\left(\frac{1}{2}\right)^n}{\left(\frac{1}{2}\right)^n + 1}$$

We see that if $n = 3$, there is about an 11% chance of the cell containing the somatic mutation. If $n = 5$ the probability of the cell containing the somatic mutation drops to about a 3% chance. While 3% may initially seem sufficiently small, we still have to be wary of our ability to call cells negative for variants. The first reason is that in practice it can be difficult to get three, let alone five, distinct reads covering a specific position for a given cell without targeted sequencing. Another is that when repeating this analysis for many different potential variants across the genome, it becomes increasingly likely that at least one (or more) false negatives will occur throughout the analysis. All this is to say that it is extremely difficult to confidently claim that a cell is negative for a particular variant allele.

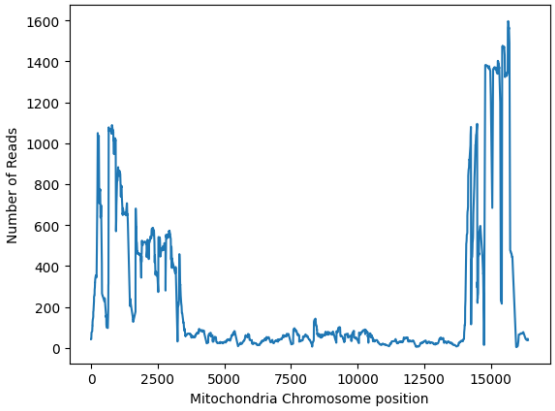
4.2 Methodology

In this section, we utilize the same cDNA library described in Chapter 3.2. From that cDNA library, we perform both MAESTER-seq [6] and MAS-seq [7]. We use the same filters to filter out dead/low-quality cells to exclude from further analysis.

As described in Chapter 2, MAESTER-seq uses specific primers to amplify out mitochondrial DNA transcripts. We plan to identify mitochondrial variants using the high mitochondrial DNA coverage provided by MAESTER-seq. We aim to do this by looking at each of the bases on the mitochondrial chromosome and searching to see if more than one allele is mapped to the base. From these results, we would be able to filter multi-allelic sites for high-confidence variants and use these variants as barcodes to identify lineages. The goal is for the very high coverage to reduce some of the issues discussed above. Similar to the method outlined above, we aim to use MAS-seq to look for multi-allelic sites and



(a) The top genes by read percentage in the library generated with MAESTER-seq. As expected, the top genes are mitochondrial genes.



(b) Number of reads with valid cell barcodes for each position in the mitochondrial genome.

identify high-confidence somatic variants. In particular, we first plan to look at the X and Y chromosomes to avoid the issues of diploidity discussed in section 4.1.1.

4.3 Results

As our first quality check of the MAESTER-seq data, we looked at which genes were the most prevalent by read count percentage (Figure 4.1a. We see that the top 11 genes are mitochondrial genes, and all 13 protein-encoding mitochondrial genes are in the top 20 genes. Next, we computed the coverage of each base (Figure 4.1b) of the mitochondria only considering reads with valid cell barcodes after the filtering described in Chapter 3.

Unfortunately, at the time of publication, the MAS-seq data is not yet available. As such, we do not report results for this data.

4.4 Discussion

The fact that many of the top genes are mitochondrial gives us confidence that the MAESTER-seq prep was able to enrich mitochondrial genes. Despite this, our coverage of the mitochondrial genome is surprisingly lower than expected, with only about 1600 reads for the most covered base position, much lower than reported in the MAESTER-seq paper [6]. Even if all of these reads come from different cells, a large number of the cells in our dataset would still have no read for this position. As is, this low coverage would make lineage tracing very challenging, as we would not be able to confidently assign alleles for many cells.

4.5 Next Steps

The first next step is to sanity check the pipeline used to compute mitochondrial genome coverage, as it seems unlikely the actual coverage is as low as what is reported. Potential issues that need to be looked into are mapping issues, bam (which hold mapped read information) file parsing issues, and read filtering issues (are valid reads accidentally being filtered out?). If one of these issues is causing the seemingly low coverage and can be resolved, clonal lineage tracing with the data will be much easier.

Once the MAS-seq data is available, the next step will be to look at the coverage in a very similar manner to the above. It will be particularly interesting to look at the coverage of the X and Y chromosomes to avoid the issues of diploidity. If the coverage is high enough, we will identify clonal lineages with the data just as we aim to do with the MAESTER-seq data.

Once we identify the clonal lineages, we aim to systematically find differences between the lineages that persist in the post-transplant sample and the ones that vanish. We can also perform more nuanced analyses, such as looking at whether the relative proportion of two lineages changes between the pre-transplant and post-transplant samples.

Once these pipelines are established and verified, the goal is to perform this with more patients and more time points. This will allow us to look at post-transplant dynamics with more resolution, observing how the samples change across many post-transplant time points. Having samples from multiple patients will allow us to determine whether any patterns we identify are consistently displayed.

Chapter 5

Conclusion

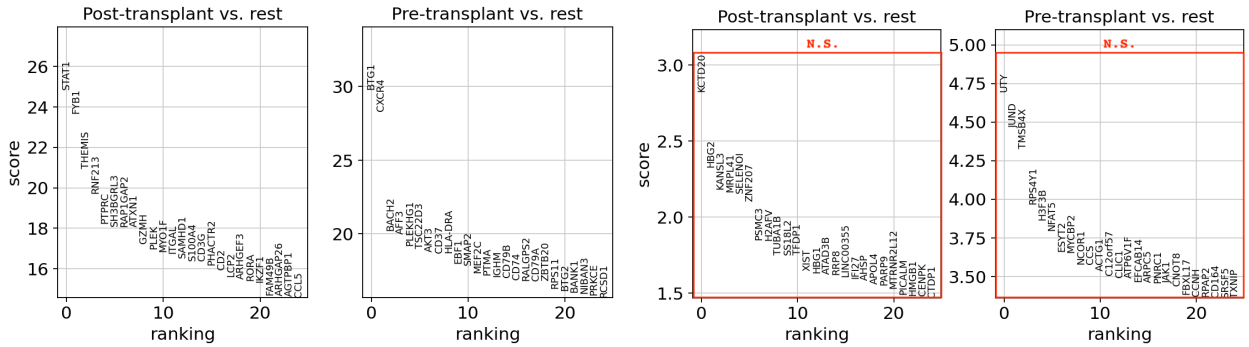
In this work, we present preliminary findings on differences in single-cell sequencing data between matched pre-transplant and post-transplant bone marrow samples. With traditional single-cell analyses, we find differences in both expression profiles and cell type prevalence between the pre-transplant and post-transplant samples. We also report preliminary work on clonal lineage tracing and discuss some next steps for extending these analyses. We are excited to continue this investigation with the proposed follow-up steps and beyond.

We hope that the results and discussions presented here will inspire more work in studying how the blood system evolves following a bone marrow transplant. We believe that a better understanding of these processes could have a direct impact in both basic science as well as in clinical medicine improving patient outcomes.

Appendix A

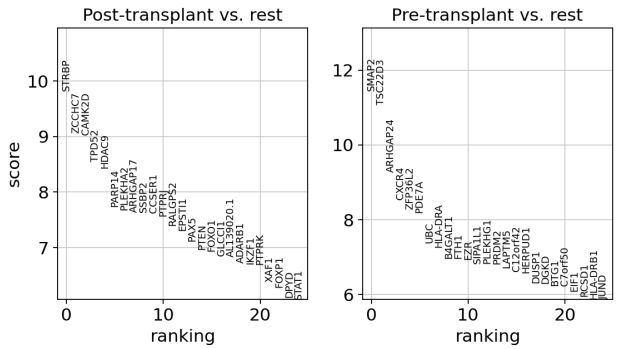
Differential Expression Results Expanded

These graphs showing the top 20 differentially expressed genes (A.1) were generated with the tl.rank_genes_groups function in Scanpy. All differences are significant except for the genes enclosed in red.

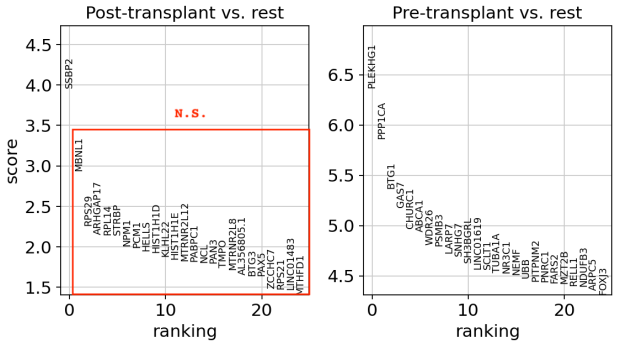


(a) All clusters

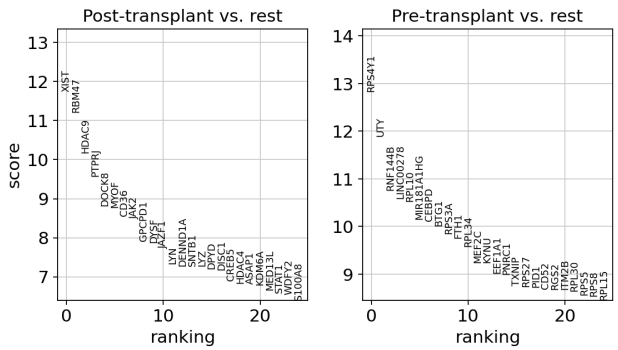
(b) Erythroid cluster



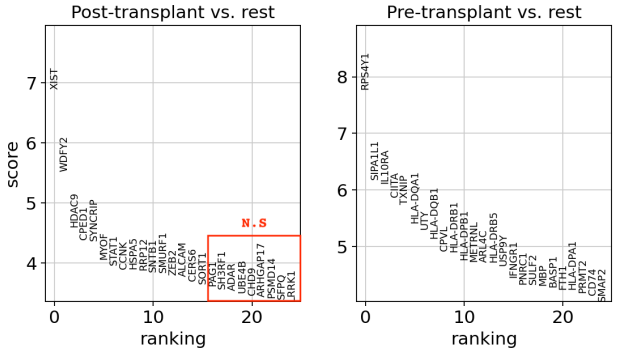
(c) B-cell cluster



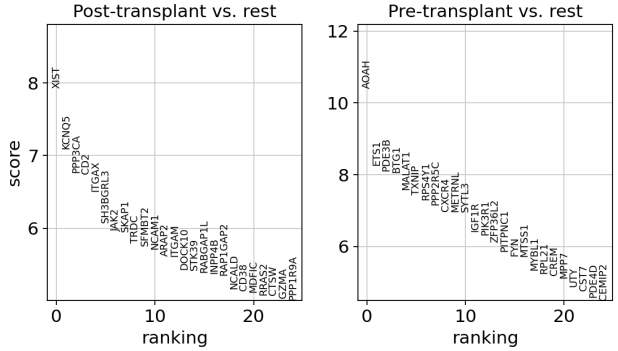
(d) B-cell progenitor cluster



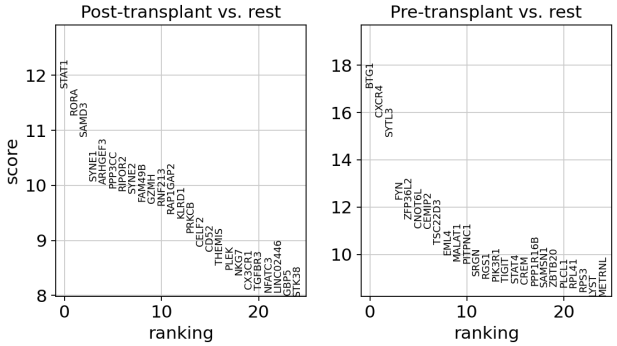
(e) Monocyte cluster



(f) Granulocyte cluster



(g) T-cell cluster



(h) CD8+ T-cell cluster

Figure A.1: Differential expression extended. Differences in genes enclosed in red boxes are not statistically significant.

References

[1] K. Khaddour, C. K. Hana, and P. Mewawalla, "Hematopoietic stem cell transplantation," en, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, Jan. 2023.

[2] Z. Sun, B. Yao, H. Xie, and X. Su, "Clinical progress and preclinical insights into umbilical cord blood transplantation improvement," en, *Stem Cells Transl. Med.*, vol. 11, no. 9, pp. 912–926, Sep. 2022.

[3] A. E. Rodriguez-Fraticelli, C. Weinreb, S.-W. Wang, R. P. Migueles, M. Jankovic, M. Usart, A. M. Klein, S. Lowell, and F. D. Camargo, "Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis," en, *Nature*, vol. 583, no. 7817, pp. 585–589, Jul. 2020.

[4] M. Spencer Chapman, A. M. Ranzoni, B. Myers, *et al.*, "Lineage tracing of human development through somatic mutations," en, *Nature*, vol. 595, no. 7865, pp. 85–90, May 2021.

[5] L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, *et al.*, "Lineage tracing in humans enabled by mitochondrial mutations and Single-Cell genomics," en, *Cell*, vol. 176, no. 6, 1325–1339.e22, Feb. 2019.

[6] T. E. Miller, C. A. Lareau, J. A. Verga, *et al.*, "Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations," *Nature Biotechnology*, vol. 40, no. 7, pp. 1030–1034, Jul. 2022.

[7] A. M. Al'Khafaji, J. T. Smith, K. V. Garimella, *et al.*, "High-throughput RNA isoform sequencing using programmed cDNA concatenation," *Nature Biotechnology*, Jun. 2023.

[8] L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, *et al.*, "Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics," *Cell*, vol. 176, no. 6, 1325–1339.e22, 2019, ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2019.01.022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867419300558.

[9] M. Stoeckius, S. Zheng, B. Houck-Loomis, S. Hao, B. Z. Yeung, W. M. Mauck, P. Smibert, and R. Satija, "Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics," *Genome Biology*, vol. 19, no. 1, p. 224, Dec. 2018.

[10] N. J. Bernstein, N. L. Fong, I. Lam, M. A. Roy, D. G. Hendrickson, and D. R. Kelley, "Solo: Doublet identification in single-cell rna-seq via semi-supervised deep learning," *Cell Systems*, vol. 11, no. 1, 95–101.e5, 2020, ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2020.05.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405471220301952.

[11] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, no. 1, p. 15, Feb. 2018.

[12] G. X. Y. Zheng, J. M. Terry, P. Belgrader, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, no. 1, p. 14 049, Jan. 2017.

[13] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, Dec. 2022.

[14] L. McInnes, J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, 2020. arXiv: 1802.03426 [stat.ML].

[15] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, p. 5233, Mar. 2019.

[16] N. Stoler and A. Nekrutenko, "Sequencing error profiles of Illumina sequencing instruments," *NAR Genomics and Bioinformatics*, vol. 3, no. 1, lqab019, Mar. 2021, ISSN: 2631-9268. DOI: 10.1093/nargab/lqab019. eprint: https://academic.oup.com/nargab/article-pdf/3/1/lqab019/36763060/lqab019.pdf. [Online]. Available: https://doi.org/10.1093/nargab/lqab019.