Single-cell measurement of dynamic cellular behaviors in development, regeneration, and malignancy

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Committee on Higher Degrees in Systems, Synthetic, and Quantitative Biology

have examined a dissertation entitled

Single-cell measurement of dynamic cellular behaviors in development, regeneration, and malignancy

presented by Debra Van Egeren

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Allon Klein

Signature _____

Typed name: Prof. Jason Buenrostro

Signature _____

Typed name: Prof. Douglas A. Lauffenburger

Date: September 27, 2021

Single-cell measurement of dynamic cellular behaviors in development, regeneration, and malignancy

A DISSERTATION PRESENTED
BY
DEBRA VAN EGEREN
TO
THE COMMITTEE ON HIGHER DEGREES IN SYSTEMS BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
SYSTEMS BIOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
SEPTEMBER 2021

Single-cell measurement of dynamic cellular behaviors in development, regeneration, and malignancy

Abstract

While we are now often able to precisely measure the current molecular state of individual cells in mammalian tissues using single cell omics technologies, it can still be difficult to study dynamic behaviors such as differentiation, proliferation, and cell death *in situ* at a similar level of detail. One general strategy for characterizing these behaviors within the native tissue context in multicellular organisms is to measure both current transcriptomic cell state as well as additional information about shared cell ancestry, past molecular states, etc in the same set of individual cells. During my PhD, I worked on multiple projects in which we augmented static single-cell phenotype data with additional information on past cell state in order to study biological systems undergoing rapid cellular changes, such as the bone marrow and the developing mammalian embryo.

Chapter 1 describes our work investigating the effects of the most common driver mutation (*JAK2*-V617F) in myeloproliferative neoplasms on patient hematopoietic stem and progenitor cells. We measured the transcriptome and *JAK2* genotype in single cells from bone marrow from these patients and determined that the mutation had a direct effect on hematopoietic cells, affecting their differentiation behavior and increasing their expression of inflammation-related genes. Additionally, we observed that bone marrow monocytes with the mutation also had higher expression of some pro-inflammatory genes and expressed a cell surface marker associated with the development of fibrosis.

In Chapters 2 and 3, we set out to measure hematopoietic stem and progenitor cell differentiation and division kinetics *in vivo* using two different experimental strategies. In Chapter 2, we developed a new system that temporarily labels cells at the top of the differentiation hierarchy using multiple fluorescent pro-

teins. These proteins are gradually lost due to cell division and degradation after cells differentiate out of the hematopoietic stem cell (HSC) state, allowing us to estimate the amount of time and number of cell divisions that have passed since these cells were HSCs. Unfortunately, we were not able to achieve high enough initial fluorescent protein expression levels for the experimental system to be useful. Instead, we used a complementary approach, described in Chapter 3, to measure HSC differentiation rates using permanent, inducible fluorescent labeling of long-term HSCs. We measured the rate of fluorescent label propagation in downstream hematopoietic progenitor cell populations and used mathematical modeling to conclude that HSCs differentiate slowly into downstream cell types.

Finally, in Chapter 4 we investigated the process of cell fate specification during early mouse embryogenesis by measuring both transcriptional state and lineage history information in individual cells just after gastrulation. To do this, we used a CRISPR barcoding system in which the barcode sequences encoding lineage information are transcribed and therefore able to be read using scRNA-seq. We found that fate restriction occurs gradually during gastrulation and investigated the origins of endothelial cells in different regions of the mouse embryo.

Contents

0 Introduction **1**

1 Single-cell genotyping and transcriptomic profiling of *JAK2*-V617F MPNs **3**
- 1.1 Introduction . . . 5
- 1.2 Results . . . 7
- 1.3 Discussion . . . 21
- 1.4 Methods . . . 23

2 Quantitative measurement of *in vivo* HSPC dynamics using time-resolved lineage tracing **26**
- 2.1 Introduction . . . 27
- 2.2 Results . . . 29
- 2.3 Discussion . . . 34
- 2.4 Methods . . . 35

3 Mathematical modeling to infer hematopoietic differentiation kinetics **39**
- 3.1 Introduction . . . 40
- 3.2 Results . . . 42
- 3.3 Discussion . . . 47
- 3.4 Methods . . . 48

4 Single-cell barcoding to characterize mammalian embryonic development **50**
- 4.1 Introduction . . . 51
- 4.2 Results . . . 52
- 4.3 Discussion . . . 66
- 4.4 Methods . . . 68

5 Conclusion **71**
- 5.1 Common themes and future outlook . . . 72

References **82**

Acknowledgments

First, I'd like to thank my advisor, Franziska Michor. During my time in her lab, I have grown into a much more confident and independent scientist, and I'm grateful for all of her help in that process. I'd also like to thank the current and former members of the Michor Lab for all of the support and discussions about science, life, life as a scientist, and just about everything else. I was particularly fortunate to have entered the lab at around the same time as two other students, Kam and Shayna, to whom I owe much of my sanity during some of the rougher times during my PhD. I'm not sure I would have made it through without them. And I of course have to thank my student Khushi for all of her efforts on our joint projects (even when they didn't exactly work out), for putting up with me for over a year, and for teaching me a lot about management and leadership along the way.

To my (unofficial) co-advisor Fernando Camargo- thank you for taking me (and my "ambitious" new project) under your wing. I have learned so much from you about how to be a rigorous, creative, and persistent scientist, and I'm sure that my time in the lab has shaped the way I will think about science for years to come (in a good way). To the members of the Camargo Lab- thank you for all of the help and support over the years, and for just letting me be part of your community of extremely smart and extremely nice people. I definitely had a lot of questions and issues come up in my experimental work, and you always provided good advice or at least some commiseration. In particular I'd like to thank Sarah and Qi, not only because I've worked with each of them extensively, but also because they are excellent scientists and fantastic people in general to be around.

I'd also like to thank Sahand Hormoz, who again provided quite a bit of unofficial mentorship during much of my PhD. I'm so lucky that a casual meeting about a somewhat unrelated computational analysis

topic blossomed into an interesting and fruitful collaboration spanning a couple of years. That collaboration had an enormous impact on my professional life- it's really what sparked my interest in single-cell work in cancer and directly led me to the research I'm doing in my postdoc and possibly far into the future. I don't think it was just the interesting science that had such an impact on me, but also the time and effort Sahand invested in me.

Finally, I'd like to thank some additional people that have shaped my PhD experience. In particular, I'd like to thank Allon Klein for being a very helpful guide and ally throughout my PhD, especially when I felt like things weren't going so well. I also owe much of my growth and success to the amazing group of PhD students I've met, in particular those in my SysBio cohort. Jacob, David, Allie, June, Ang, Roger, Gemma, Sam, Tessa, Kalki, Anna, and Hailey- getting to know you all was pretty much the best part of graduate school, and I'm so grateful I got to spend my time here with such wonderful scientists and great friends. Daphne (along with Nick) definitely deserves special recognition, since she has managed to live with me and listen to me complain for around four years now (though I guess there are some perks to living with me, like perpetually clean dishes). And of course, this acknowledgements section wouldn't be complete without including my family, who has patiently listened to me talk about work every week on Skype for the past 5 years. Thank you for everything.

0

Introduction

Our ability to quantitatively measure the molecular state of individual cells has dramatically increased over the last ten years. New developments have enabled characterization of the whole genome[1], transcriptome[2,3], and epigenome[4,5] of hundreds to thousands of single cells, leading to revolutionary advances in cancer biology[6,7] and developmental biology[8,9]. While these experimental techniques are quite useful for cataloguing single cell phenotypes and discovering differences in cell state under certain conditions or perturbations, additional tools are often required to answer questions about dynamic changes in cell state. For example, single-cell RNA sequencing or epigenetic characterization alone can determine what cell types are present in the developing embryo, but cannot definitively determine which cell types give rise to other cell types during development. However, in the context of embryonic development, single-cell barcoding has been

used in combination with these techniques to define differentiation hierarchy structures by providing lineage history information[10,11]. More generally, dynamic behavior can often be better assessed *in vivo* using multiple single-cell measurements, such as using single cell lineage history[12,13], divisional history[14], and/or mutational status[15,16] in conjunction with transcriptomic or epigenomic data or combining more than one single-cell sequencing technique[17]. These strategies have the potential to address many longstanding questions, such as how cancer develops from somatic mutations in normal tissues, or what factors regulate tissue regeneration in health and disease.

During my PhD, I helped develop and use some of these techniques to study phenotypic changes that occur in individual cells in cancer, regeneration, and development. Chapter 1 of this thesis describes how we used joint single-cell RNA sequencing and mutation status profiling in bone marrow cells from patients with myeloproliferative neoplasms to determine how the most common driver mutation in this cancer (*JAK2*-V617F) changes the phenotype and differentiation behavior of hematopoietic cells. Chapters 2 and 3 describe two methods of experimentally augmenting static cell state measurements using fluorescent reporters to measure hematopoietic stem and progenitor cell division and differentiation kinetics *in vivo*. Finally, Chapter 4 describes how we used single-cell barcoding and RNA sequencing to study proliferation and fate specification in early mouse embryos.

1

Single-cell genotyping and transcriptomic profiling of *JAK2*-V617F MPNs

Foreword

This project originated in Sahand Hormoz's lab and was conducted primarily under his supervision. My primary role was to analyze the scRNA-seq data, which was collected by Sean Liu and Max Nguyen in the Hormoz Lab. They also helped design the experimental protocol for joint single-cell genotyping and transcriptomic profiling of these samples. Additional follow-up experiments were performed by Baransel Kamaz and supervised by Ann Mullally, both of whom also contributed greatly to the overall study design

and analysis. Additional guidance was provided by Chris Reilly and Franziska Michor. In addition to the scRNA-seq analysis presented here, we performed single-cell whole genome sequencing on hematopoietic stem cells from some of these patients, which was analyzed mainly by Isidro Cortés-Ciriano and Javi Escabi. While I did not describe the results from the whole-genome sequencing here because I wasn't heavily involved in that part of the project, additional details can be found in our published paper[18]. This chapter is adapted from the above-mentioned *Cell Stem Cell* paper and its preprint[19], as well as an additional manuscript in preparation.

Abstract

Myeloproliferative neoplasms (MPNs) are a group of slow-growing blood cancers that are often caused by somatic gain-of-function mutations in *JAK2*. The most common *JAK2* driver mutation, *JAK2*-V617F, causes constitutive JAK-STAT signaling. *In vitro* and *in vivo* mouse experiments have shown that the mutation is sufficient to increase production of red blood cells and/or platelets and recapitulate the phenotype of different MPNs. However, it is still unclear what direct effects the *JAK2*-V617F mutation has on hematopoietic progenitors and differentiated cells in MPN patient bone marrow, and how these molecular changes cause disease in humans. In the past, it has been difficult to identify which individual cells within patient bone marrow carry the mutation, making it difficult to directly compare the phenotypes of mutant and wild-type (WT) cells. We developed a method to simultaneously measure a cell's transcriptome and determine whether it has a mutation in *JAK2* by specifically amplifying the mutation locus during scRNA-seq sample processing. Using this approach on bone marrow samples from 7 MPN patients, we compared the gene expression profiles of *JAK2*-WT and *JAK2*-mutant HSPCs and found that *JAK2*-V617F cells have a megakaryocyte-erythroid fate bias and have increased expression of proinflammatory and antigen presentation genes. We also found that bone marrow monocytes have a high *JAK2* variant allele fraction, and that *JAK2*-mutant monocytes have increased expression of intermediate monocyte genes and of the cell surface marker *SLAMF7*. Our

results suggest that *JAK2*-mutant monocytes may have a role in MPN pathogenesis that should be further investigated.

1.1 Introduction

Myeloproliferative neoplasms (MPNs) are a group of chronic blood malignancies that are diagnosed in approximately 2.7 out of 100,000 people in the US annually[20]. These malignancies are often diagnosed later in adulthood and are generally incurable, though patients often live for decades after diagnosis without major complications[21]. As their name suggests, MPNs are defined by aberrant myeloid cell proliferation, and there are three different subtypes of MPN that affect different myeloid cell lineages. Essential thrombocythemia (ET) is defined by increased platelet counts, polycythemia vera (PV) is defined by increased red blood cell counts, and primary myelofibrosis (PMF) has variable effects on peripheral blood counts and is defined by bone marrow fibrosis and extramedullary hematopoiesis. All of these subtypes often have an indolent disease course; however, PMF is the most aggressive of the three and is most likely to lead to bone marrow failure or progression to acute leukemia[21]. Treatment for all of these conditions is often supportive and aimed at normalizing blood counts, although some newer therapies (e.g., ruxolitinib) target hematopoietic stem cells (HSCs) that harbor the mutation(s) causing the disease.

Most MPNs are caused by a mutation in one of three genes: *JAK2*, *CALR*, and *MPL*[21]. Gain-of-function mutations in *JAK2* are the most common driver mutation, especially in PV patients[22]. *JAK2*-V617F is the most common gain-of-function mutation in *JAK2*, followed by mutations in exon 12[22,23]. These mutations lead to constitutive kinase activity and downstream signaling via the JAK-STAT pathway. In mouse models of *Jak2*-V617F driven disease, pathogenesis was found to be particularly dependent on the activity of Stat1, Stat3, and/or Stat5[24]. In MPN patients, the *JAK2*-V617F mutation increases the expression of STAT1 and STAT5 targets, including components of the interferon response pathway[25]. These changes also lead to activation of the erythropoietin and thrombopoietin response pathways, resulting in cytokine-independent

erythroid and megakaryocyte progenitor expansion[26]. Blocking the effect of the *JAK2*-V617F mutation can reverse this phenotype in some cases. The JAK1/2 inhibitor ruxolitinib is used in PV and PMF patients with persistent symptoms or hematopoietic abnormalities to normalize blood counts and reduce the size of the *JAK2*-mutant clone in peripheral blood[27,28].

While *Jak2* mutations have been experimentally characterized in the context of malignant blood disorders, *in vivo* models of *JAK2*-V617F MPNs have some limitations. For example, in both transgenic and HSC transplant mouse models, the disease phenotype (as determined by red blood cell and platelet counts and the presence of bone marrow fibrosis) can be hard to predict and control[29]. Variation in the level of *Jak2*-V617F expression, background genotype, and mouse species all contribute to differences in disease phenotype[29]. Collecting and analyzing samples directly from MPN patients avoids the problem of setting up an experimental system that faithfully reproduces the disease, but introduces other challenges. In particular, it is difficult to study the direct effect of somatic *JAK2* mutations in MPN patients, since few techniques can identify JAK2-mutant and *JAK2*-WT HSPCs in MPN patient samples and characterize their phenotype on a single-cell level. To overcome this challenge, we established a single-cell RNA-sequencing (scRNA-seq) protocol that can measure the whole transcriptome and the genotype at a specific locus in individual cells. A similar method has been used to characterize the effect of *CALR* mutations in MPN patients, and showed that the unfolded protein response and NF-κB pathway activity was upregulated in *CALR*-mutant HSPCs[15]. Here, we studied the effect of the *JAK2*-V617F mutation in ET and PV patients, and found significant transcriptional changes in HSPCs and bone marrow monocytes with the mutation. More generally, our work shows that joint genotyping and phenotype measurement at a single cell level can reveal information about the molecular mechanisms underlying the development of cancer from oncogenic somatic mutations.

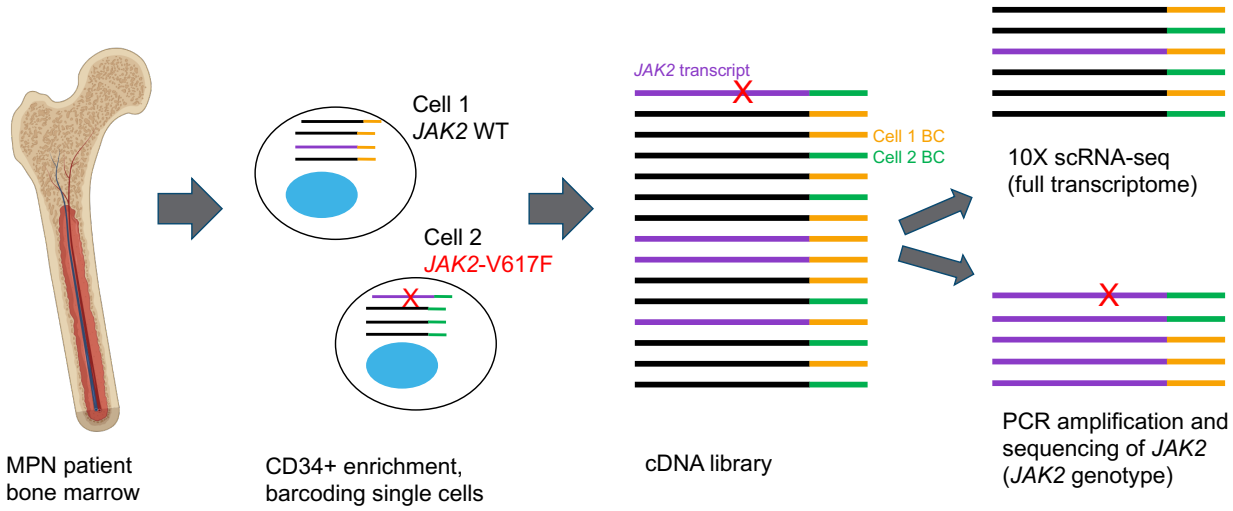

Figure 1.1: Joint single-cell *JAK2* genotyping and scRNA-seq was performed on patient bone marrow samples. Schematic describing experimental procedures. To increase our ability to sequence the site of the *JAK2*-V617F mutation, *JAK2* transcripts (purple) were amplified using locus specific primers after cell barcodes are added (yellow and green sequences) during 10x library preparation.

1.2 Results

To study the effect of the *JAK2*-V617F mutation on individual cells in patients with MPNs, we developed a method that combines single-cell *JAK2* genotyping and scRNA-seq (Fig. 1.1, Methods) and used it to compare the transcriptomes of *JAK2*-mutant and *JAK2*-WT within the same patient. This experimental method is a modified version of the standard 10x scRNA-seq protocol, but after cell barcoding, the cDNA library is split into two parts. One part of the library is prepared with the standard 10x protocol for sequencing of the whole transcriptome, while the other part undergoes specific PCR amplification of the *JAK2* mutation locus. While the two resulting libraries are sequenced separately, both share the same set of cell barcodes, allowing later computational processing steps to match up reads from the two libraries that come from the same cell. Therefore, this method was able to both identify *JAK2* mutations and measure the transcriptome in individual cells.

We applied this method to CD34+ cells from bone marrow aspirates from 7 newly-diagnosed MPN patients-4 with ET and 3 with PV (Table 1.1). All of these patients had *JAK2* mutations detected in peripheral blood;

	ET 1	**ET 2**	**ET 3**	**PV 1**	**PV 2**	**PV 3**	**ET V617L**
Age/sex	34 M	63 F	41 M	56 M	56 M	69 M	49 F
PB *JAK2* VAF	16.6% (V617F)	25.3% (V617F)	10.6% (V617F)	62.7% (V617F)	68.0% (V617F)	40.9% (V617F)	45.0% (V617L)
Additional PB mutations	-	-	-	*EZH2* (4.2%)	*TET2* (32.0%)	*TET2* (31.1%)	-
# cells	7868	7349	7091	6732	7472	7911	7704
# *JAK2* mutant	174	217	68	308	469	541	107
# *JAK2* WT	646	525	567	216	329	687	285

Table 1.1: MPN patient characteristics. Patient PV 1 had an *EZH2* E745* mutation, PV 2 had a *TET2* S268* mutation, and patient PV 3 had a frameshift mutation in *TET2*. Note that PV 3 also had a smaller VAF *TET2* missense mutation (C1289S) detected in peripheral blood. Abbreviations used: PB, peripheral blood; VAF, variant allele fraction.

6 had the *JAK2*-V617F mutation, while 1 ET patient had a variant previously unreported in humans (*JAK2*-V617L). Experiments in Ba/F3 cells have previously shown that the *JAK2*-V617L mutation can induce cytokine independence and constitutive JAK-STAT signaling *in vitro*[30]. Two PV patients also had truncating mutations in *TET*2 and one patient with PV had a mutation in *EZH*2; a clinical NGS assay detected no additional MPN-associated mutations in peripheral blood of these patients[31]. Two bone marrow samples from healthy donors (22 and 29 year old females) were also collected and sequenced, though without *JAK2* mutation amplification.

We isolated and sequenced 6000-8000 CD34+ enriched bone marrow cells from each patient and were able to amplify and detect the mutation locus in at least one *JAK2* transcript from 5-15% (mean 9.5%) of cells. As expected, cells with higher *JAK2* expression (i.e., more *JAK2* transcripts detected in the whole transcriptome sequencing) were more likely to have the mutation locus detected during amplicon sequencing (Fig. 1.2A). In particular, cells with a mutant *JAK2* transcript call had higher average numbers of *JAK2* transcripts detected, since most *JAK2*-mutant cells are erythroid and megakaryocyte progenitors (see below, Fig. 1.4), which have particularly high total transcript counts and hight *JAK2* expression (shown later in Fig. 1.9C). In most cells with a *JAK2* mutation locus call, only one *JAK2* mutation locus UMI was detected (Fig. 1.2B). This low coverage of the mutation locus made it challenging to definitively determine whether a cell was *JAK2*-WT

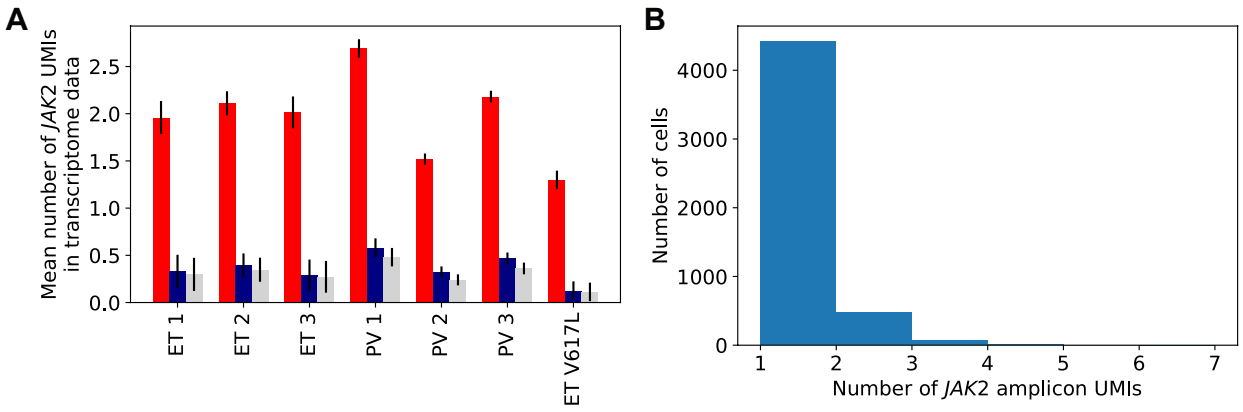

Figure 1.2: Detection of the *JAK2* mutation locus in scRNA-seq libraries depends on expression of *JAK2*. A. Average number of *JAK2* UMIs detected in the whole transcriptome library in cells with a mutant *JAK2* amplicon call (red), a WT *JAK2* amplicon call (blue), or no *JAK2* amplicon call (grey). Error bars are +/- SEM. B. Number of *JAK2* mutation locus UMIs detected per cell by amplicon sequencing. The data include cells from all patients with at least one *JAK2* amplicon call.

or had a heterozygous *JAK2* mutation. In a cell heterozygous for the mutation, it would be possible that only WT transcripts were detected, particularly if very few *JAK2* UMIs were sequenced. This source of error would cause us to underestimate the fraction of cells in our single-cell data that have heterozygous *JAK2* mutations. To correct for this, we assumed most cells had only a single *JAK2* mutation call and that patients with peripheral blood *JAK2* variant allele fraction (VAF) $< 50\%$ had primarily heterozygous *JAK2* mutations. Under these assumptions, *JAK2*-mutant cells have a 50% chance of having only a WT transcript, so we doubled the fraction of cells in a population that had a *JAK2*-mutant called to estimate the frequency of mutant cells. Importantly, we only applied this correction when computing population-level statistics, not when examining the transcriptomes of individual cells (as in our differential expression analysis).

1.2.1 *JAK2*-V617F induces an erythroid/megakaryocyte fate bias in MPN patient HSPCs.

To determine how the *JAK2*-V617F mutation affects HSPC fate commitment, we first identified HSPC cell types in the scRNA-seq whole transcriptome data (Fig. 1.3). We first merged the data from all 7 patients and the healthy controls and performed batch correction to better align cell types between the samples (Fig. 1.3B). We then clustered the data and manually assigned clusters to HSPC types based on expression of marker genes

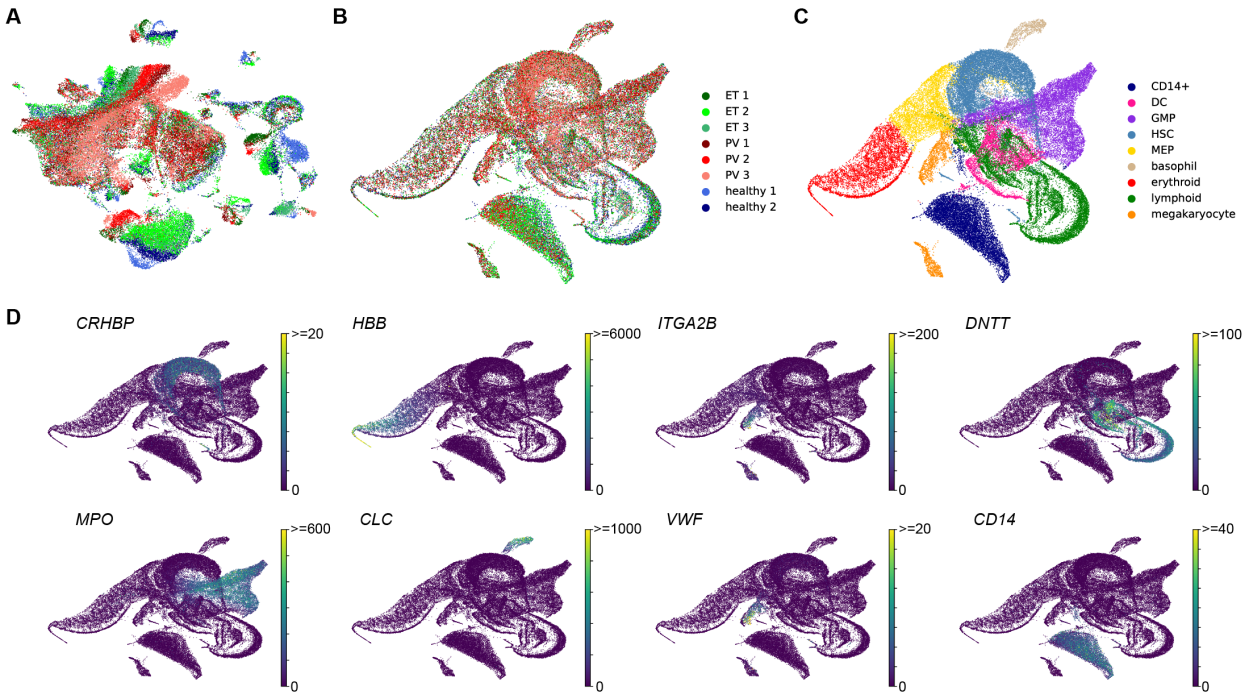

Figure 1.3: Several HSPC cell types were identified in all scRNA-seq bone marrow samples. A. UMAP of all patients and healthy controls combined, before batch correction. Colors denote sample ID, as shown in B. B. UMAP of all patients and healthy controls combined, after batch correction with Seurat v3, colored by sample ID. C. Cell type classifications of all cells sequenced. D. Expression of selected marker genes used to identify cell types in the combined, batch corrected UMAP.

(Fig. 1.3C-D). MPN patients had an HSPC differentiation hierarchy that was similar in overall structure to that of healthy controls, with the same cell types represented in similar proportions. Of note, a CD14+CD34- population was captured in our data, which was likely a contaminating bone marrow monocyte population, as they express typical monocyte marker genes (see Section 1.2.4 of these results for more details).

We used the *JAK2* amplicon calls to determine whether cells with the *JAK2*-V617F mutation produced different hematopoietic progenitor types than WT cells. In all patients, we found that *JAK2*-mutant cells and *JAK2*-WT cells were found in the same regions of the UMAP plots, suggesting that the transcriptional changes caused by *JAK2* mutations are subtle and do not result in the creation of new, distinct cell types outside of the normal hematopoietic hierarchy (Fig. 1.4). However, in patients with the *JAK2*-V617F mutation, cells with a mutant *JAK2* transcript call were more likely to be megakaryocyte, erythroid, or megakaryocyte-erythroid progenitors (MEPs) (Fig. 1.4B-C). The frequency of cells with a *JAK2*-V617F transcript out of all cells with

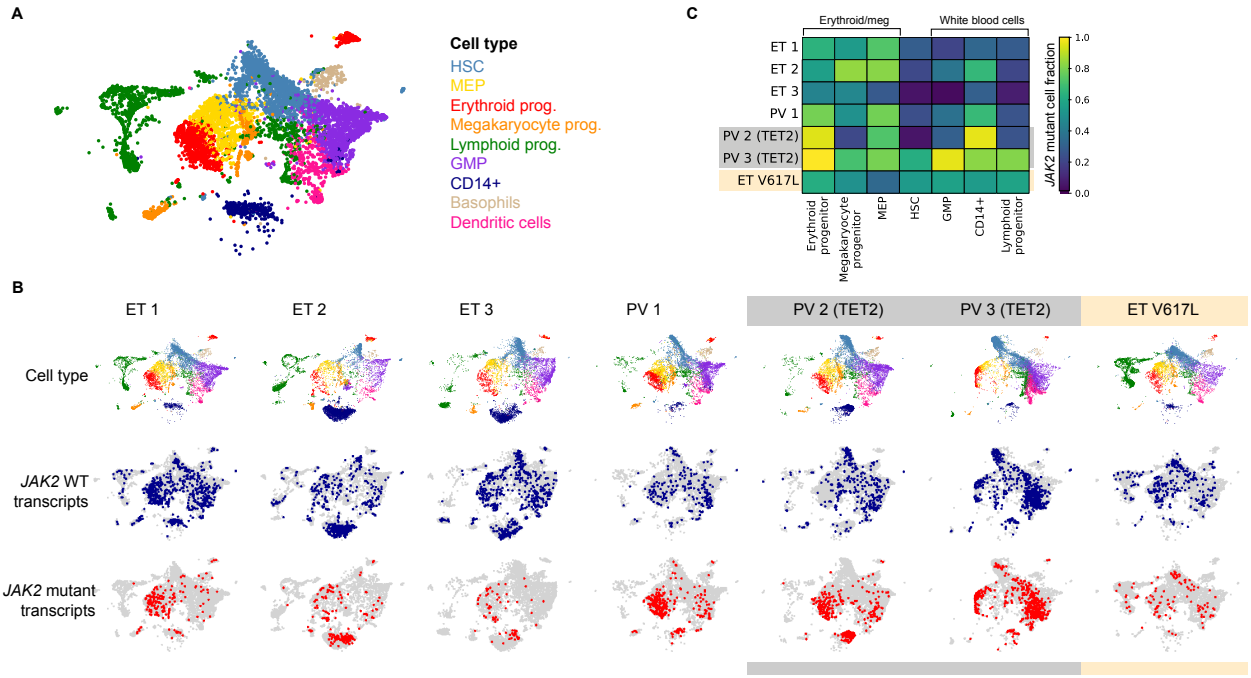

Figure 1.4: *JAK2*-V617F HSPCs were more likely to give rise to megakaryocytes and erythroid progenitors than WT cells. A. Example UMAP of CD34+ bone marrow cells from patient ET 1 alone, colored by cell type. B. CD34+ cell UMAPs for all patients individually. The top row is colored by cell type (same colors as in A). The bottom two rows show cells with at least one *JAK2*-WT or textitJAK2-mutant transcript detected. C. *JAK2*-mutant cell frequency of each cell type in MPN patients. Note that for patients presumed to have heterozygous mutations (ET 1, ET 2, ET 3, PV 3, and ET V617L), the fraction of cells with at least one mutant transcript out of all cells with a *JAK2* amplicon call was doubled to estimate the mutant cell fraction.

at least one mutation locus call increased as cells became more committed to an erythroid or megakaryocyte fate, suggesting that cells with the mutation differentiate faster towards an erythroid/megakaryocyte fate and/or proliferate faster than WT cells. This observation is consistent with the clinical presentation of MPNs (increased red blood cell or platelet count), although we did not see substantial differences between patients with different MPN subtypes (ET and PV). We also found a small number of HSCs with the mutation from each patient, indicating that the disease-causing mutation originated in an HSC.

Interestingly, *JAK2*-mutant HSPCs in the ET patient with the noncanonical *JAK2*-V617L mutation did not have a fate bias that was discernibly different from WT cells (Fig. 1.4B-C). While the V617L mutation does not seem to be directly affecting the differentiation behavior of CD34+ HSPCs, the fact that this mutation has a high peripheral blood allele fraction (45.0%) and was the only MPN-associated mutation found in the blood of this patient suggests that it may still be causing disease, though perhaps in a different way than the

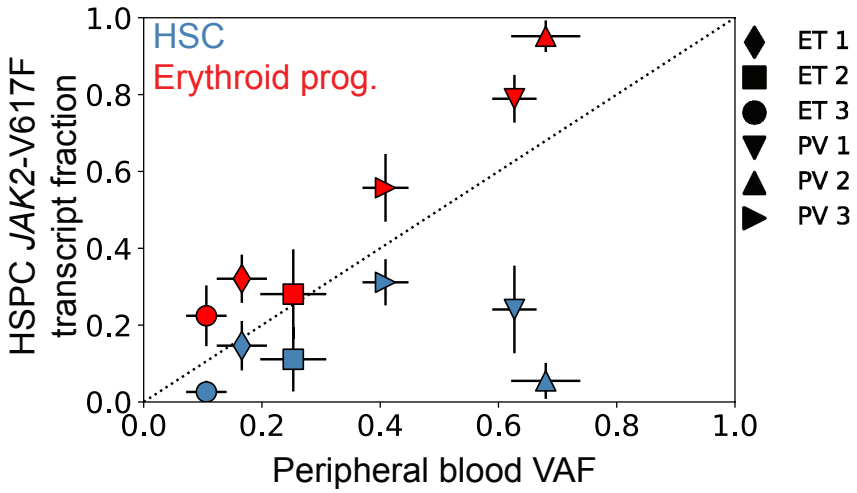

Figure 1.5: HSCs have a lower *JAK2*-V617F VAF than measured by peripheral blood sequencing. *JAK2*-V617F variant allele frequency in two HSPC cell types (HSCs: blue, erythroid progenitors: red) as compared to peripheral blood, measured by clinical NGS assay. Error bars indicate 95% binomial confidence intervals.

JAK2-V617F mutation would.

The *JAK2* mutation VAF in peripheral blood is often used in clinical practice as a diagnostic marker and as a metric to track disease burden. This test largely samples mature neutrophils, which may have a different variant allele burden from bone marrow cell populations, particularly those that do not directly produce granulocytes. In fact, we found that the peripheral blood VAF underestimated the variant transcript fraction in erythroid progenitors (Fig. 1.5), suggesting that peripheral blood measurements may not adequately reflect the contribution of the mutant clone to erythropoiesis. Additionally, the *JAK2* mutation burden in HSCs was significantly lower than that measured by peripheral blood sequencing in all patients, indicating that peripheral blood VAF also does not accurately reflect disease burden at the very top of the HSPC differentiation hierarchy.

Our analysis suggests that the *JAK2*-V617F mutation likely causes disease not by dramatically altering the structure of the hematopoietic differentiation hierarchy, but rather by changing differentiation, proliferation, and survival dynamics along the existing differentiation hierarchy. To investigate these phenotypic changes more closely, we compared the transcriptomes of *JAK2*-V617F and *JAK2*-WT cells. To maximize the power of this comparison, the differential expression analysis was performed with an expanded set of mutation calls that leveraged single-cell whole-genome sequencing data to identify additional mutations associated with the

JAK2-V617F HSPC population.

1.2.2 Single-cell whole genome sequencing data can be used to confidently detect somatic mutations in scRNA-seq data.

In a separate part of this study, we sequenced the whole genomes of individual single-HSC-derived colonies from patients ET 1 and ET 2 to infer the history of expansion of the *JAK2*-mutant HSCs. While that work will not be discussed here in detail (see our paper[18] for more information), we used the somatic mutations identified by the whole genome sequencing to improve our ability to identify cells that likely belong to the *JAK2*-mutant population. Mutations that were found in *JAK2*-V617F HSC colonies sequenced and none of the *JAK2*-WT colonies most likely represent mutations that are shared by some or all *JAK2*-mutant cells.

We used these *JAK2*-associated mutations to improve our identification of *JAK2*-mutant cells in our scRNA-seq data from ET 1 and ET 2 in two ways. First, we identified a set of expressed somatic mutations found exclusively in all *JAK2*-V617F HSCs and PCR amplified these mutations in the scRNA-seq libraries using a similar protocol as used to amplify the *JAK2* mutation locus. This procedure was performed successfully for 3 mutations in patient ET 1 (in *ASHL1*, *HSPA9*, and *UPF1*) and 1 mutation in ET 2 (in *NRROS*). Using this extra amplicon data, we were able to identify an additional 2107 cells with a *JAK2*-associated mutation call in the scRNA-seq data from ET 1 and 1057 more cells with a call in ET 2. Example calls for highly-expressed mutations (*UPF1* for ET 1 and *NRROS* for ET 2) are shown in Fig. 1.6A. *JAK2*-mutant cells identified using these additional clade-specific mutations also showed a megakaryocyte-erythroid bias similar to that observed in the *JAK2*-mutant cells identified through *JAK2* amplicon sequencing (Fig. 1.6A). Also, in cells which also had a *JAK2* amplicon call, around 50% of cells with a mutant non-*JAK2* amplicon call had a *JAK2*-V617F mutation detected (Fig. 1.6B), which would be expected if all cells that had a *JAK2*-associated mutation also were heterozygous for *JAK2*-V617F.

Second, we looked for these mutations in the raw 10x scRNA-seq data from these patients (Methods). It

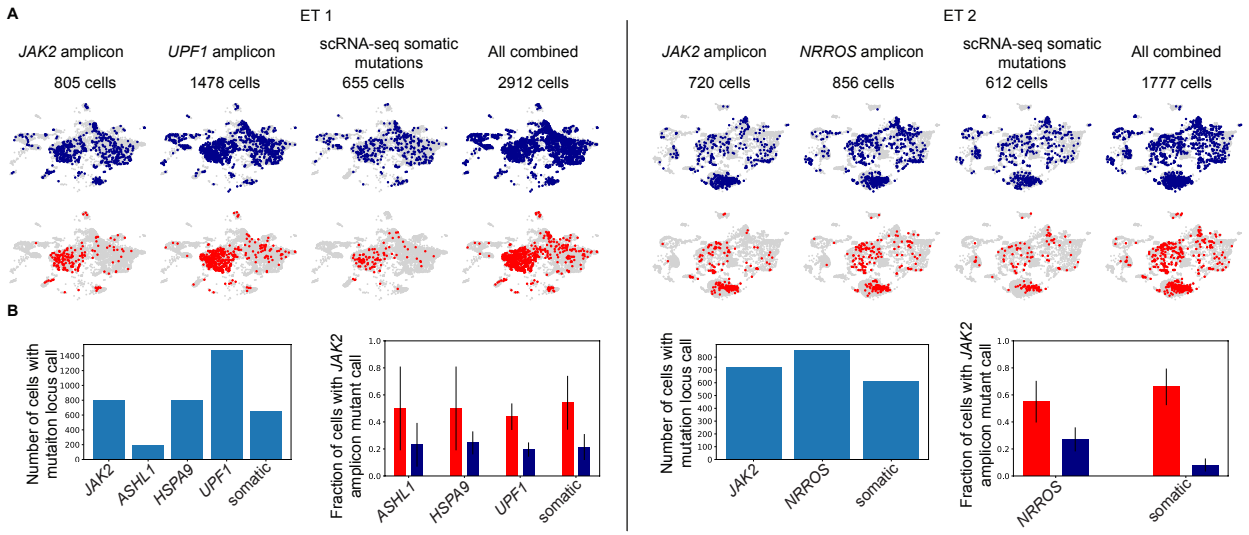

Figure 1.6: Additional *JAK2*-V617F HSPCs were identified in the scRNA-seq data by amplifying other somatic mutations exclusively found in all *JAK2*-mutant HSCs. A. UMAPs showing cells with WT (blue, top row) or mutant (red, bottom row) transcript calls from *JAK2* amplicon sequencing, textitJAK2-associated amplicon sequencing (*UPF1* and *NRROS*), and raw 10x transcriptome reads ("scRNA-seq somatic mutations"). B. Left panels: number of cells with detectable WT or mutant transcript calls from amplicon sequencing and direct detection from scRNA-seq reads ("somatic"). Right: Of cells with both a *JAK2* amplicon call and a *JAK2*-associated mutation locus call, the fraction of cells with a *JAK2*-V617F call. For cells with a *JAK2*-associated mutation (red bars), approximately 50% of cells should have a *JAK2*-V617F amplicon call, since the *JAK2*-V617F mutation is heterozygous in these patients and most cells have only a single *JAK2* transcript sequenced. Blue bars indicate the fraction of *JAK2*-V617F amplicon calls in cells with WT calls at *JAK2*-associated mutation loci. Error bars are 95% binomial confidence intervals.

is ordinarily quite challenging to accurately call somatic mutations directly from scRNA-seq data for many reasons, some of which are mitigated by our specific objectives and the availability of the single-clone whole genome sequencing data. One issue is that the low sequencing coverage when using 3'-biased single-cell transcriptome data makes it difficult to call mutations in most of the genome. However, we are searching for a specific set of *JAK2*-V617F-associated mutations in each patient (220 are shared by all *JAK2*-mutant cells in ET 1, 398 mutations in ET 2), the detection of any of which marks a cell as part of the *JAK2*-V617F clade. By limiting our search to only these known somatic mutations, the overall false discovery rate is reduced. Another issue is that many single nucleotide variants that are identified in a subset of cells in an scRNA-seq dataset are caused by PCR or sequencing errors and are not true somatic mutations. This is especially true for very low-frequency variants, making it especially difficult to confidently detect low-frequency somatic mutations that are present only in the descendants of a particular HSC, for example. Again, by leveraging the

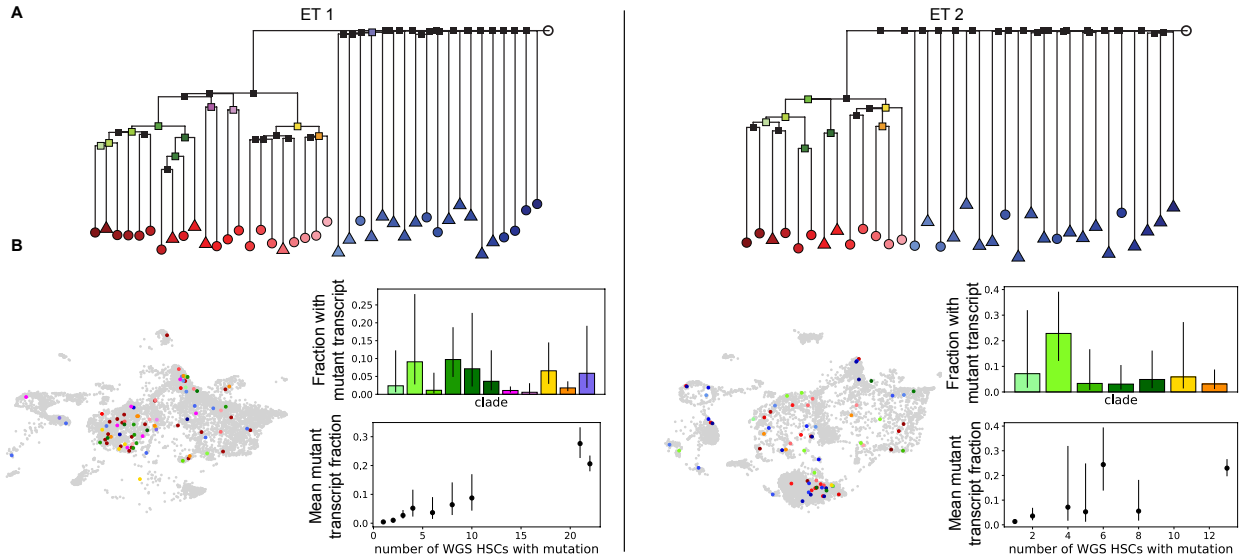

Figure 1.7: We identified cells in our scRNA-seq data that are likely descendants of specific individual HSCs or groups of related HSCs. A. Single-cell phylogenies inferred from single colony whole-genome sequencing. Sequenced colonies derived from HSCs are denoted as circles, and those derived from MPPs are triangles. The procedures used to generate and validate these phylogenies is described elsewhere [18]. Individual HSC/MPPs (leaves) and inferred common ancestors of HSC subsets (internal nodes) with unique somatic mutations that could be potentially detected in the scRNA-seq are colored (red: *JAK2*-mutant colonies, blue: *JAK2*-WT colonies, others: inferred ancestors). B. Left: UMAP showing cells with a mutant transcript corresponding to descendants of a single HSC or an HSC clade. Colors correspond to colors of leaves or nodes in A. Right: fraction of cells with a mutant transcript corresponding to an HSC clade out of all cells in which either a WT or mutant version of that clade-specific transcript was detected (top). The fraction of mutant transcripts averaged over all mutations shared by HSC clades with the same number of HSCs (bottom). The mutant fraction increases with clade size. All error bars are 95% Bayesian credible intervals.

whole genome sequencing data we can restrict our search space for somatic mutations to those which have previously been confidently identified to be present in one or more HSCs from that patient's bone marrow sample, reducing the potential for false positives.

These additional somatic mutation calls from the transcriptome reads were used to identify additional *JAK2*-mutant cells (Fig. 1.6). Around 5-10% of cells in the scRNA-seq data had a detectable somatic mutation locus in the transcriptome (Fig. 1.6B). Cells with a mutation in one of these *JAK2*-associated mutation loci have a megakaryocyte-erythroid lineage bias (Fig. 1.6A) and have approximately a 50% *JAK2*-mutant VAF (Fig. 1.6B), as would be expected for patients a heterozygous mutation.

In addition to identifying more *JAK2*-mutant cells for our differential expression analysis, we also used this method of calling somatic mutations from transcriptome sequencing reads to identify HSPCs that are

clonal descendants of specific HSCs or groups of HSCs. We detected somatic mutations unique to individual HSCs or small clades of HSCs in our scRNA-seq data (Fig. 1.7). While we were not able to identify enough cells per HSC clone to do a full analysis comparing clonal output, we did find that mutations found in more HSCs in the whole genome sequencing dataset were also found in more cells in the scRNA-seq data, which we would expect if our method were accurately calling descendants of individual cells (Fig. 1.7B). Improving detection of somatic mutations specific to individual HSCs by amplifying those mutation loci, for example, could increase the number of cells we can confidently identify as descendants of a specific stem cell. However, given that descendants of each individual HSC make up a low fraction of HSPCs overall ($<< 1\%$ expected for each HSC), even with perfect detection of all HSC-specific somatic mutations we would likely have to sample tens to hundreds of thousands of cells to confidently compare the number and phenotype of differentiated descendants of specific HSCs, which was beyond the scope of this work.

1.2.3 *JAK2*-mutant HSPCs have increased expression of antigen-presentation genes.

Using the additional *JAK2*-V617F-associated mutation calls, we compared the transcriptomes of *JAK2*-mutant and *JAK2*-WT HSPCs within individual patients to investigate the cell-intrinsic effect of *JAK2* mutations on HSPC function. Some significantly differentially expressed genes were specific to individual patients, but many were found to be shared across all ET or all PV patients by combining p-values from multiple patients (Fig. 1.8A-B). Genes previously shown to be upregulated in MPN patient *JAK2*-V617F hematopoietic progenitors[25] were found to be significantly upregulated in our *JAK2*-V617F MEPs, although we didn't detect a similar significant change in HSCs, likely due to the smaller number of sequenced HSCs (Fig. 1.8C-D). We found that the *JAK2*-V617F mutation increased the expression of antigen presentation genes in MEPs and erythroid progenitors, particularly in ET patients (Fig. 1.8A-B, E). The expression of antigen presentation genes is upregulated by JAK-STAT signaling (particularly through STAT1)[32,33], so there is likely a direct link between constitutive JAK2 activity in these mutant HSPCs and increased MHC expression.

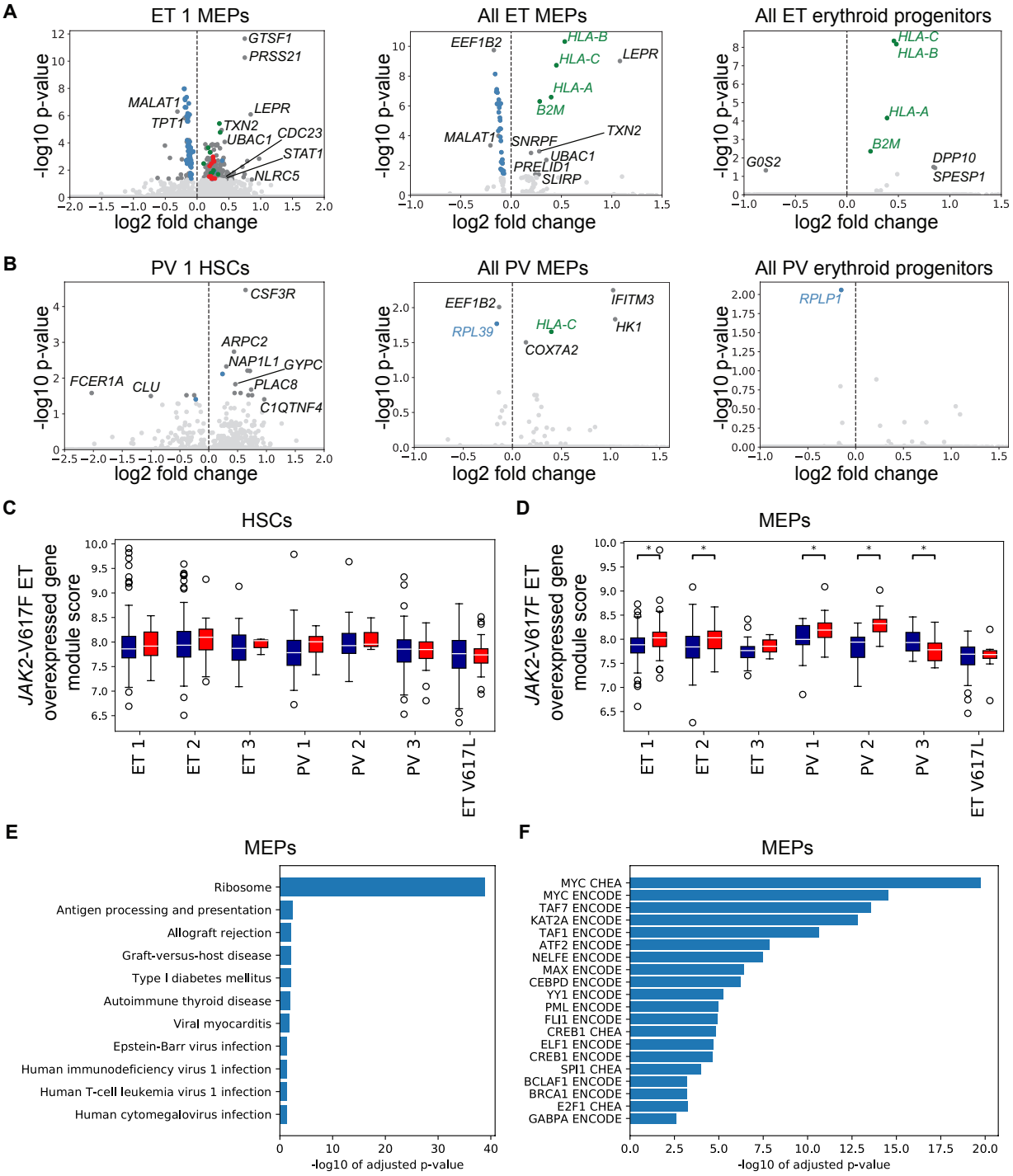

Figure 1.8: *JAK2*-mutant and *JAK2*-WT HSPCs have different gene expression profiles in MPN patient bone marrow. A-B. Volcano plots showing differential expression results comparing *JAK2*-mutant and *JAK2*-WT HSPCs. Log fold changes are positive if expression is higher in mutant cells, negative if expression is lower in mutant cells. Ribosomal genes are marked in blue, antigen presentation genes are marked in green, and proteasomal genes are marked in red. Selected significant genes are also labeled. C-D. Expression scores of a gene set previously found to be upregulated in *JAK2*-mutant HSPCs[25], in *JAK2*-V617F (red) and *JAK2*-WT (blue) HSCs (C) and MEPs (D). White center line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile range, points are outliers. * indicates $p < 0.05$, Wilcoxon rank sum test. E-F. Gene set enrichment analysis for differentially expressed genes between *JAK2*-mutant and WT MEPs, using the KEGG 2019 Human Biological Processes gene sets (E) and ChEA/ENCODE transcription factor target sets (F).

We also found that the *JAK2*-V617F mutation increased expression of the leptin receptor (*LEPR*) in ET patient MEPs (Fig. 1.8A). In normal hematopoiesis, the protein hormone leptin is secreted by bone marrow adipocytes and supports HSPC proliferation[34,35,36]. High expression of the leptin receptor in HSCs has been previously shown to be associated with increased proliferation, self-renewal capacity, and transplant engraftment potential, as well as a proinflammatory transcriptomic profile[37]. All of these characteristics are consistent with the observed phenotype of the malignant *JAK2*-V617F HSPCs in MPN patients.

Finally, we found an enrichment of MYC/MAX-signaling associated genes in the differential expression results for MEPs (Fig. 1.8F), suggesting that *JAK2* V617F may be associated with dysregulation of MYC/MAX signaling. MYC and MAX play a key role in controlling cell cycle and specifying cell fate in MEPs[38], so changes in MYC/MAX activity could contribute to the increased megakaryocyte or erythroid progenitor production found in MPN patients.

While ribosome content and translational activity is crucial to erythroid progenitor differentiation and expansion[39,40], the relative expression of ribosomal genes has been shown to decrease during erythroid differentiation[41,42]. In our data, we found that *JAK2*-mutant cells had lower normalized ribosomal gene expression than WT cells (Fig. 1.8A-B, E), which is consistent with these cells having a more differentiated phenotype. While we observed a decrease in the normalized fraction of ribosomal transcripts per cell in *JAK2*-mutant HSPCs, we also observed an increase in the total number of transcripts detected per cell in *JAK2*-mutant HSPCs (Fig. 1.9A-B). Therefore, it is possible that the absolute ribosomal gene expression is not lower in mutant cells, since these cells have more mRNA transcripts overall. This change in total number of transcripts is also consistent with a more differentiated phenotype for *JAK2*-V617F HSPCs, since more differentiated cell types had greater total transcript counts overall (Fig. 1.9C).

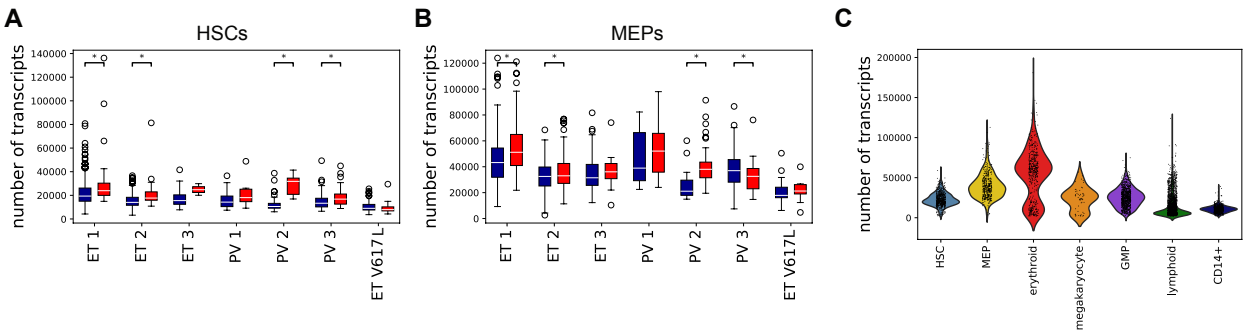

Figure 1.9: Total transcript count per cell is increased in more differentiated HSPCs and in *JAK2*-mutant cells. A. Total number of transcripts per cell detected by scRNA-seq for *JAK2* WT (blue) and *JAK2* mutant (red) HSCs and MEPs from each MPN patient. White center line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile range, points are outliers. * indicates $p < 0.05$, Wilcoxon rank sum test. B. Total number of transcripts per cell for each HSPC cell type from healthy donor 1.

1.2.4 *JAK2*-mutant bone marrow monocytes have an intermediate monocyte phenotype and may differentiate into pathogenic fibrocytes.

We further characterized the CD34- bone marrow monocyte population, adding scRNA-seq data from two additional *JAK2*-V617F MPN patients (one ET and one PV) to the 6 patients with *JAK2*-V617F mutations analyzed above. As expected, the monocyte cluster was made up of mostly CD14+ CD16- classical monocytes, with smaller populations of CD16+ CD14low nonclassical monocytes and CD14int CD16int intermediate monocytes (Fig. 1.10A). In constrast to *JAK2*-WT monocytes, *JAK2*-mutant monocytes were mostly intermediate monocytes (Fig. 1.10B-D). Intermediate monocytes are marked by high MHC class II expression[43] and may be a transitional state between classical and nonclassical monocytes. While they are dysregulated in some autoimmune and infectious conditions, their biological role in many contexts remains unclear[44,45]. *JAK2*-mutant PMF patients have higher levels of circulating intermediate monocytes, which have some abnormal cytokine secretion patterns, suggesting these cells may play a direct role in the pathogenesis of MPN[46].

We found that *JAK2*-mutant monocytes had higher expression of proinflammatory/interferon response genes than WT cells, including MHC genes, STATs, and TNF family genes (Fig. 1.10E-F). Gene set enrichment analysis showed that these significantly differentially expressed genes are associated with infec-

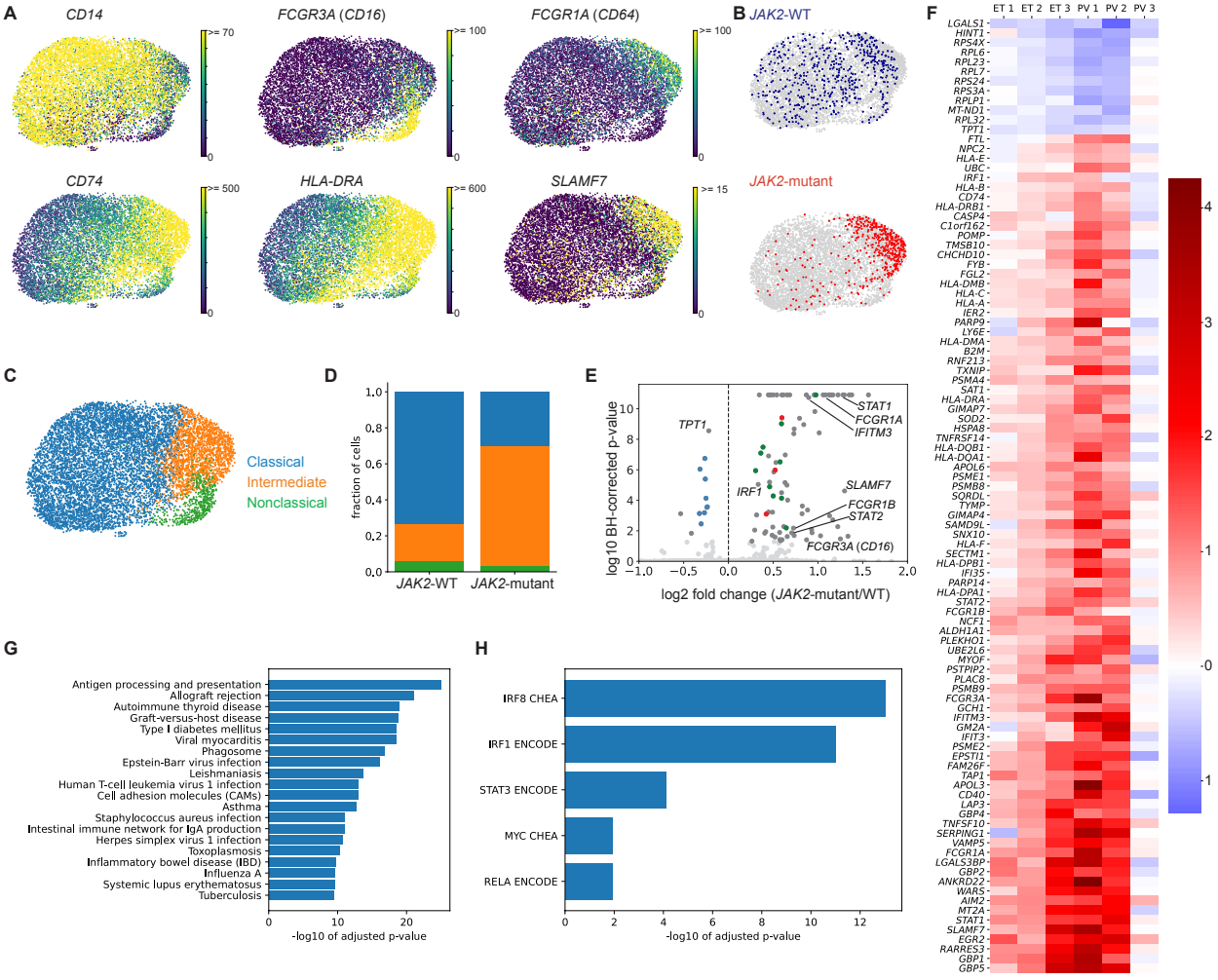

Figure 1.10: The *JAK2*-V617F mutation increased expression of inflammation-related genes and *SLAMF7* in MPN bone marrow monocytes. A. Batch corrected combined UMAPs showing expression of marker genes in monocytes from 8 MPN patients and 2 healthy donors. B. UMAPs showing expression of *JAK2*-V617F and WT *JAK2* transcripts in monocytes. C. Batch corrected UMAP showing monocyte subset classifications. D. Fraction of *JAK2*-mutant and *JAK2*-WT monocytes from MPN patients of each monocyte type. Colors are the same as in C. E. Volcano plot showing differential expression results comparing *JAK2*-mutant and *JAK2*-WT monocytes. Log fold changes are positive if expression is higher in mutant cells, negative if expression is lower in mutant cells. Ribosomal genes are marked in blue, antigen presentation genes are marked in green, and proteasomal genes are marked in red. Selected significant genes are also labeled. F. Heatmap of log2 mean expression fold changes in *JAK2*-mutant vs. *JAK2*-WT monocytes from patients ET 1-3 and PV 1-3. Only genes that are significantly differentially expressed in the combined dataset between the six patients are shown. Colors denote log2 fold changes, where positive values mean that the gene is more highly expressed in *JAK2*-mutant cells. G-H. Gene set enrichment analysis for differentially expressed genes between *JAK2*-mutant and WT monocytes, using the KEGG 2019 Human Biological Processes gene sets (G) and ChEA/ENCODE transcription factor target sets (H).

tious/inflammatory biological processes (Fig. 1.10G), and activation of STAT3 and interferon response factors (Fig. 1.10H). These gene expression changes were largely consistent between all PV and ET patients except for patient PV 3 (Fig. 1.10F), who also had a smaller population of sequenced monocytes. Interestingly, we also found increased expression of *SLAMF7*, a cell surface protein associated with differentiation of monocytes into fibrocytes in *JAK2*-V617F PMF[47]. This finding suggests that *JAK2*-mutant monocytes in MPN patients without overt bone marrow fibrosis may have the potential to differentiate into fibrocytes, leading to disease progression.

In summary, we found that the *JAK2*-V617F mutation increased expression of antigen presentation and proinflammatory genes in HSPCs and bone marrow monocytes. In HSPCs, these changes are quite subtle, suggesting that the main effect of the mutation in MPN patients is to promote differentiation and/or expansion of HSCs and megakaryocyte/erythroid progenitors following largely the same molecular trajectories as normal HSPCs. On the other hand, our results suggest that *JAK2* V617F may lead to a pathogenic monocyte phenotype, possibly through abnormal cytokine secretion or promoting bone marrow fibrosis.

1.3 Discussion

We developed a high-throughput method to jointly perform scRNA-seq and genotype a specific locus in individual cells and used this method to investigate the direct effect of *JAK2* mutations on the phenotype of HSPCs in MPN patients. We found that the *JAK2*-V617F mutation leads to increased production of committed megakaryocyte and/or erythroid progenitors, although this fate bias (megakaryocyte vs. erythroid) did not necessarily correspond with MPN subtype. Differential expression analysis revealed that *JAK2*-V617F MEPs express higher levels of antigen presentation and other inflammation-associated genes, as well as the leptin receptor. Additional experimental work is being conducted to further investigate the role of the leptin receptor in *JAK2*-mutant MPNs. Follow-up studies are also being performed to better understand the role of *JAK2*-mutant monocytes in the pathogenesis of PV and ET. The expression of SLAMF7 we observed in the

JAK2-V617F monocytes is particularly intriguing for two reasons. First, it represents a possible functional association of these cells with bone marrow fibrosis[47]. Second, if the SLAMF7+ monocytes prove to cause disease progression, treatment with the SLAMF7-targeting monoclonal antibody elotuzumab (currently approved to treat multiple myeloma) could improve MPN patient outcomes.

Our work has several limitations that affected our ability to measure the effects of the *JAK2*-V617F mutation in MPN patients. One important limitation was our ability to capture and sequence *JAK2* transcripts in the sampled cells. We were not able to read the *JAK2* locus genotype in most cells (85-95%) which we encapsulated for scRNA-seq. This was partly due to lack of expression of *JAK2* in some cell types. To improve our genotyping efficiency, we used additional whole genome sequencing information on other mutations shared by all *JAK2*-mutant HSCs to identify additional *JAK2*-mutant or *JAK2*-WT cells. However, improving the genotyping efficiency of our experimental method further would increase our ability to detect differences between mutant and WT cells. Additionally, since many of our patients had heterozygous mutations, some cells with *JAK2*-WT transcripts detected likely have heterozygous *JAK2*-V617F mutations and were misclassified as *JAK2*-WT in our differential expression analysis. Increasing the number of *JAK2* mutation locus UMIs detected per cell would improve our ability to distinguish between true *JAK2*-WT cells and *JAK2*-mutant heterozygotes. Finally, given that we only collected a single sample of bone marrow from each patient, the static nature of the scRNA-seq measurement made it difficult to learn what particular changes in the dynamic behaviors (differentiation, proliferation, and cell death were caused by the *JAK2*-V617F mutation.

To further understand the dynamic processes that lead to the development of *JAK2*-V617F MPN, we also performed single-colony whole-genome sequencing and phylodynamic analysis on two of the patients (ET 1 and ET 2), which is described fully in our published article[18]. Briefly, we built single-cell phylogenies from both *JAK2*-mutant and *JAK2*-WT MPPs and HSCs from these patients and used the structure of the *JAK2*-mutant clade phylogenies to reconstruct the history of *JAK2*-V617F HSC expansion from the time at which the mutation occurred to the time of diagnosis. We estimated the time at which these patients first acquired

the *JAK2* mutation, and found that it occurred decades before MPN diagnosis (approximately 25 years before diagnosis in ET 1 and 44 years before diagnosis in ET 2). The structures of the inferred phylogenies of the *JAK2*-mutant populations were quite different from the phylogenies of WT cells and of HSCs from healthy individuals[8]. Using these phylogeny structures, we also found that the *JAK2* mutation conferred a fitness advantage onto HSCs, with an estimated 63% advantage over WT in ET 1 and 44% in ET 2. This part of the study showed that single-cell lineage information can be used to infer cell population dynamics that occurred in the past (in this case, during the MPN premalignant phase), which could be a powerful tool when applied to other cancer types.

More generally, our method combining scRNA-seq and *JAK2* genotyping could also be applied to other cancer types and oncogenic mutations to measure the effects of genomic changes on the phenotype of individual cells. This approach has been used by multiple groups to study MPNs[15,48] and relapsed CLL[49] and could be extended to solid tumors as well. Ultimately, the results from these studies can be used to better understand the molecular mechanisms driving clonal expansion and other pathogenic changes that occur during the development and progression of cancer.

1.4 Methods

Detailed descriptions of all of the methods can be found either in our published article[18] or our preprint[19]. Some key methods are summarized here for convenience. Code used for the scRNA-seq analyses can be found in our GitLab repository.

1.4.1 Experimental procedures

Bone marrow aspirates were collected from newly-diagnosed MPN patients with a *JAK2* mutation detectable by clinical NGS assay in their peripheral blood, in accordance with DFCI IRB protocol 01-206 and MGH IRB protocol 13-583. CD34+ mononuclear cells were isolated from the bone marrow samples using the EasySep

Human CD34 Positive Selection Kit II (StemCell Technologies). The CD34+ cell suspensions were used to generate cDNA libraries using the 10x Chromium controller and the 10x Chromium Single Cell 3' Library & Gel Bead Kit v3. Part of this library was used for whole transcriptomic sequencing according to the standard 10x protocol. Another part was used for targeted amplification and sequencing of the *JAK2* mutation locus using a triple-nested PCR protocol with primers that flank the mutation locus.

1.4.2 Identification of *JAK2*-mutant amplicon sequencing transcripts

The raw sequencing files were initially processed using CellRanger 4.0.0 to generate fastq files for the *JAK2* amplicon library. Transcripts with quality < 30, those with low read support, and those that did not match a whitelisted 10x cell barcode within a 2 bp difference were discarded. Similar cell barcodes (within a 2 bp difference) were error-corrected and merged. Reads with either the expected WT nucleotide or mutant nucleotide were used to determine whether each transcript had the mutation. Transcripts with $>$50% mutant reads were designated as mutant transcripts, and $<$50% were designated as WT transcripts.

1.4.3 scRNA-seq data preprocessing

The raw sequencing files were initially processed using CellRanger 4.0.0 to generate fastq files and count matrices for the whole transcriptome libraries. Count matrices were loaded and analyzed in scanpy. Genes expressed in $<$ 3 cells and cells with $<$ 2000 UMIs or $>$ 20% mitochondrial transcripts were filtered out, and total count normalization was performed. Log-transformed expression values with mitochondrial transcript percentage and total counts regressed out were used to generate UMAP visualizations and perform clustering. Clustering was performed on the combined dataset including all MPN patients and healthy controls. This dataset was batch-corrected using the Seurat v3 integration function[50] and was clustered using the Louvain algorithm. Each cluster was manually assigned to a cell type using its expression level of marker genes.

1.4.4 Differential gene expression analysis

JAK2-mutant and *JAK2*-WT cell total-count normalized gene expression profiles were compared using the Wilcoxon rank sum test (scanpy) for each cell type in each patient separately. The resulting p-values were corrected for multiple hypotheses within each patient using the Benjamini-Hochberg procedure. These adjusted p-values for each cell type were combined for groups of patients with the same disease subtype (PV or ET) using Fisher's method. The ET patient with the V617L mutation was not included in this combined analysis. Gene set enrichment analysis on the list of significantly differentially expressed genes for the combined ET or PV patients was performed using GSEApy, using the 2019 KEGG biological processes and the ChEA/ENCODE transcription factor target gene sets.

1.4.5 Calling additional somatic mutations in scRNA-seq data

Cells with somatic mutations associated with *JAK2* V617F were identified using the scRNA-seq transcriptome data from patients ET 1 and ET 2. All 10x reads that mapped to genomic locations found to be somatically mutated in *JAK2*-mutant HSCs from the patient in the whole-genome sequencing data with unambiguous cell barcodes were selected. Reads with base quality greater than 30 were used to determine whether each transcript had the mutant or WT sequence. To minimize the number of false positive mutant calls, we eliminated mutation loci which had errant mutant reads identified in a bank of 36 bone marrow or peripheral blood 10x scRNA-seq datasets from healthy individuals.

2

Quantitative measurement of *in vivo* HSPC dynamics using time-resolved lineage tracing

Foreword

I designed and implemented all parts of this project, with supervision from Fernando Camargo and Franziska Michor. Experimental assistance from Alejo Rodriguez-Fraticelli and other members of the Camargo Lab was much appreciated.

Abstract

Quantitative measurement of *in vivo* differentiation and cell division dynamics in individual hematopoietic stem and progenitor cells (HSPCs) is challenging, since the cellular processes occurring over time cannot be easily directly observed *in situ*. To estimate the rate at which these processes occur in unperturbed hematopoiesis, we developed an experimental system that uses fluorescent protein degradation to estimate the amount of time that has elapsed since a differentiation event has occurred in an HSPC and fluorescent protein dilution to estimate the number of cell divisions that have occurred. We designed and characterized a fluorescent timer reporter with three fluorescent proteins with different degradation kinetics. While we initially found that fluorescent protein levels were correlated with time and cell divisions in cell lines *in vitro*, we were not able to achieve high enough fluorescent protein expression levels in mouse HSPCs to accurately measure differentiation kinetics.

2.1 Introduction

Hematopoietic stem cells (HSCs) are the source of all new blood cells after birth. In order to produce enough blood cells to maintain homeostasis over the animal's lifetime and respond to injury and infection, HSCs and the their hematopoietic progenitor cell (HPC) progeny must divide and differentiate at appropriate rates. Dysregulation of this process can lead to cytopenias or malignancy. Understanding the kinetics of hematopoietic stem and progenitor cell (HSPC) differentiation and the mechanisms that regulate these rates may allow us to better treat these conditions.

Measuring the rates of dynamic behaviors such as differentiation, cell division, and cell death *in vivo* is challenging. Most experimental techniques cannot assess the phenotype of individual cells without removing them from their native context and/or destroying them. Therefore, methods to measure differentiation and division in hematopoietic cells *in situ* often augment static cell state measurements with additional dy-

namic information. For example, groups have used fluorescent markers to track the progeny of particular cell populations over time, which can be combined with mathematical modeling to estimate population-level differentiation and division rates[51,52]. To track cell divisions during tissue regeneration, previous studies have used dilution of fluorescently-tagged histones (e.g., H2B-GFP)[53,54]. These fluorescently-tagged histones have a long half-life once they are incorporated into chromatin[55] (although recent work has challenged this notion in HSPCs[56]). After production of the H2B-conjugated fluorescent protein stops, each division cuts the amount of fluorescence in each cell by half. Therefore, by measuring the fluorescence in each cell, the number of divisions that each individual cell has undergone since expression of the protein stopped can be estimated. Stable cytoplasmic fluorescent dyes such as CFSE can also be used to track divisions, but require *ex vivo* cell manipulation and are therefore more useful in the transplant setting[57].

While fluorescent proteins that degrade very slowly can be used to track cell divisions, proteins with shorter half lives can be used to measure the amount of time that has passed since each individual cell has undergone a specific phenotype change. These shorter-lived proteins degrade in the cell over time according to predictable first-order kinetics, so by measuring the level of fluorescence in each cell after transcription of the fluorescent proteins have stopped, we can estimate the amount of time since the cell stopped expressing the proteins. Simultaneously using two or more proteins with different degradation kinetics improves the accuracy of this estimate and provides a way to correct for fluorescent protein loss through dilution during cell division. Such a strategy with two fluorescent reporters with different half lives (specifically engineered using a protein degron sequence) has been used to measure the kinetics of differentiation in intestinal organoids[58]. In this intestinal organoid system, expression of the fluorescent reporters was restricted to a specific set of multipotent enteroendocrine progenitors expressing *Neurog3*, so the fluorescence measurements were used to estimate the time since the cell lost *Neurog3* expression and differentiated out of the multipotent enteroendocrine state. Since most fluorescent proteins have half lives of a few hours to days, this general strategy can be adapted to assess differentiation kinetics in regenerative tissues in which important differentiation events

happen over the course of a few days.

Here, we designed an experimental system to measure HSPC kinetics with three fluorescent reporters- one conjugated to histone H2B with a long degradation half life to measure cell divisions, and two cytoplasmic proteins used to estimate the amount of time that has passed since expression of the reporters ended. We intended to use this system to measure division and differentiation kinetics in individual HSCs as they differentiate into multipotent progenitors (MPPs) and other fate-restricted hematopoietic cell types. However, although the system seemed to work in the 293T cell line *in vitro*, expression of the timing construct was too low in HSPCs to make the system useful *in vivo*.

2.2 Results

2.2.1 We designed a reporter construct with multiple fluorescent proteins to estimate time and number of divisions after differentiation *in vivo*.

To estimate the amount of time that has elapsed since a cell has turned on or off expression of a specific gene, we designed a fluorescent reporter containing three different fluorescent proteins with different degradation rates. While the degradation rates of fluorescent proteins in mammalian cells is not well studied, the half lives of most GFP and RFP derivatives is likely 24 hours or longer[59,60,58], suggesting that it would be feasible to use these proteins to study HSPC differentiation events, many of which likely happen over a few days[51]. Cells would constitutively express high levels of these proteins until a differentiation event stops transcription of the fluorescent reporter. We designed two strategies that control expression of the reporter in response to different biological events. The first links expression of the reporter to the expression of another gene by knocking the rtTA tetracycline-inducible transcription factor into the endogenous locus for that gene (Fig. 2.1A). By choosing a gene expressed exclusively in stem cells, under continuous exposure to doxycycline (dox), stem cells will build up the fluorescent reporter protein levels. After the cells differentiate and lose expression of

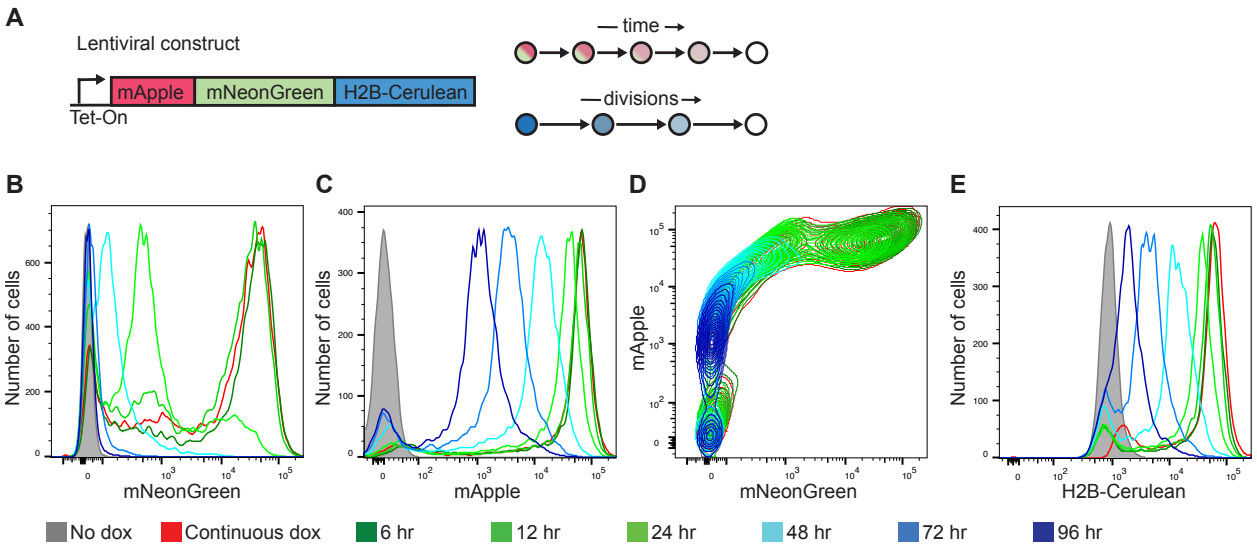

Figure 2.1: Levels of both cytoplasmic and histone-tagged fluorescent proteins decrease hours after stopping fluorescent protein expression. A. Genetic architecture of the Tet-On lentiviral fluorescent timer. B-E. Fluorescent protein levels in 293T cells at various timepoints (color legend for panels B-E at bottom) after dox withdrawal following continuous dox exposure.

the gene, transcription of the fluorescent proteins will stop and the timer will start, as the existing fluorescent proteins in each cell degrade and dilute via cell division. The other approach we used was to flank the fluorescent protein sequences with loxP sites that cause the intervening sequence to invert irreversibly after Cre recombination. If Cre expression occurs when the cell starts expressing a protein specific to differentiated cells, transcription of the fluorescent proteins will turn off upon expression of that gene. We tested both of these general approaches in 293T cells and mouse HSPCs.

2.2.2 The level of fluorescent timer protein expression decreases over time in a predictable way in 293T cells *in vitro*.

We started by implemented a drug-inducible timer system and used it to test the kinetics of fluorescent protein loss in 293T cells *in vitro* (Fig. 2.1A). When introduced into a cell line with constutitive rtTA transcription factor expression, all three fluorescent proteins will be expressed as long as the cells are exposed to dox. We created a lentivirus containing the reporter construct and used it to infect 293T cells stably expressing rtTA. We found that these cells expressed high levels of mNeonGreen, mApple, and H2B-Cerulean, and that levels

of all three fluorescent proteins decreased over time after dox withdrawal (Fig. 2.1B-E). mNeonGreen fluorescence dropped off sharply at around 12 hours after dox withdrawal, suggesting a half life of a few hours. In contrast, mApple fluorescence decreased more slowly, suggesting it is more stable in the cytoplasm of these cells. mNeonGreen and mApple fluorescence levels were correlated in these cells, further supporting our hypothesis that degradation of these proteins over time is causing fluorescence loss after dox withdrawal (Fig. 2.1D). We did not see discrete peaks in the H2B-Cerulean fluorescence corresponding to cell populations that divided different numbers of times after dox withdrawal (Fig. 2.1E), in contrast to what some studies with long-lived fluorescent proteins have shown[60,53]. This might have been due to the low overall level of H2B-Cerulean fluorescence detected, even when the protein was constitutively expressed. This low expression may have been due to the fact that the Cerulean fluorescent protein is less bright than most red or green fluorescent proteins.

2.2.3 Hematopoietic cells did not express the fluorescent timer proteins highly enough to allow for accurate estimation of time or number of divisions.

Encouraged by the cell line results, we modified the design of the fluorescent timer construct and created a transgenic mouse that constitutively expressed the reporter in all HSPCs. Rather than use the rtTA transcription factor to control expression of the reporter, we flanked the polycistronic fluorescent protein sequence with loxP sites, so that the coding sequence is inverted irreversibly upon Cre recombination (Fig. 2.2A). To increase the brightness of the histone-conjugated protein, we switched the Cerulean protein for TagBFP2 and added an inverted iRFP 670 sequence that is expressed in cells only after Cre recombination. If this construct were stably inserted into cells that express active Cre recombinase upon expression of a cell-type specific gene, transcription of the fluorescent protein would stop only after differentiation into that cell type. We used homologous recombination to insert this construct into mouse embryonic stem (ES) cells in the TIGRE constitutive expression locus[61]. These gene-targeted ES cells expressed all three fluorescent proteins,

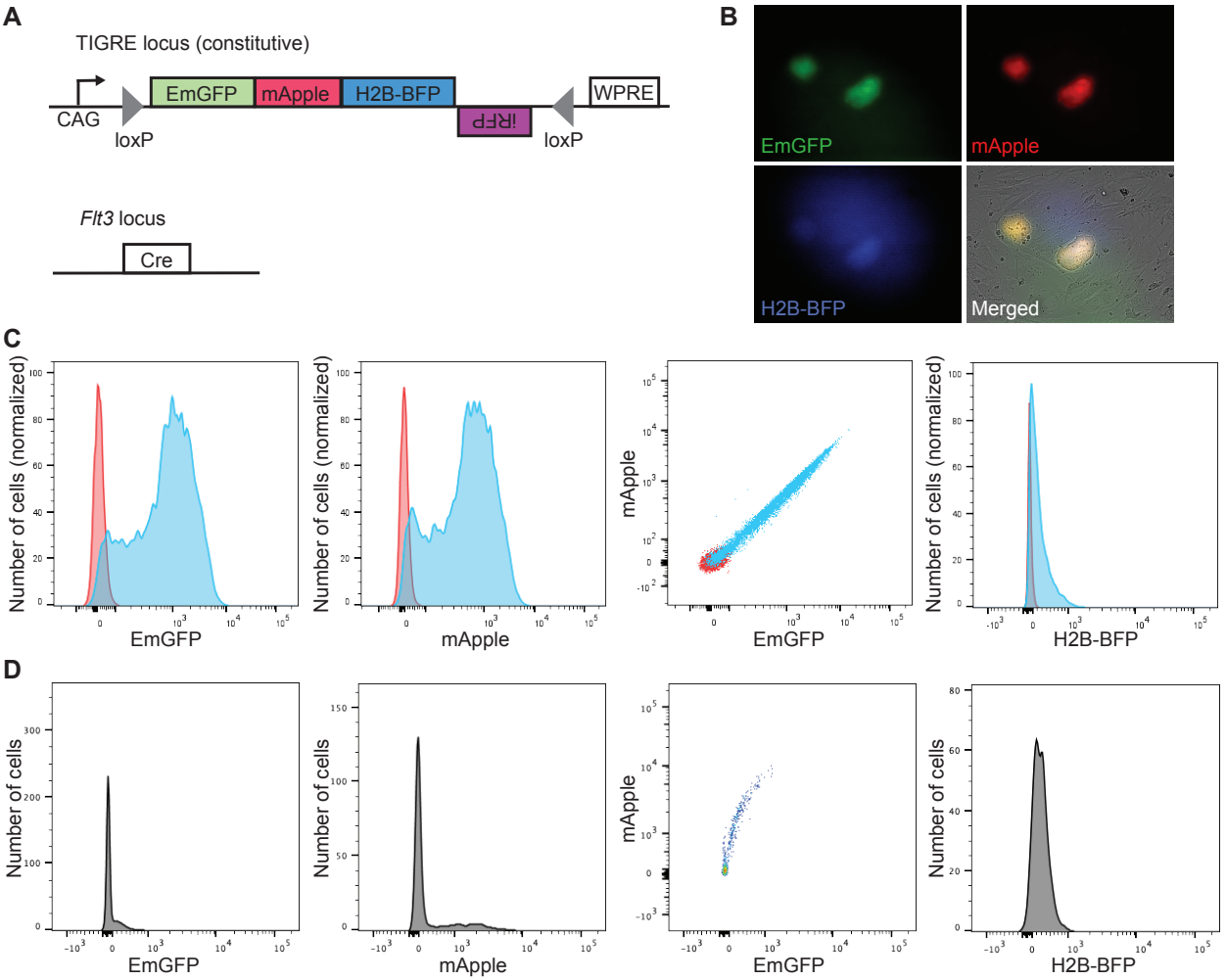

Figure 2.2: The fluorescent reporter proteins were expressed at high levels in mouse ES cells but not HSPCs in the transgenic mouse. A. Genetic architecture of the transgenic mouse line with the fluorescent timing construct inserted via homologous recombination into the TIGRE expression locus. The loxP sites shown cause irreversible inversion of the sequence between them when recombined with Cre. Experiments were done on chimeras that did not have the Flt3-Cre element, which would be necessary to start the timer. B. Fluorescence microscopy of the ES cell clone used to create the chimeras. C. Flow cytometry measurements of fluorescent protein levels in the ES cell clone used to create the chimeras. Blue: gene targeted ES clone; red: wild-type negative control. D. Flow cytometry measurements of fluorescent protein levels in Lin- cells from the bone marrow of a chimera. Abbreviations used: EmGFP, mEmerald fluorescent protein; H2B-BFP, H2B-TagBFP; CAG, CMV enhancer chicken β-actin promoter; WPRE, WHV Posttranscriptional Regulatory Element.

though the H2B-BFP expression was quite low (Fig. 2.2B-C). We injected these ES cells into wild-type C57BL/6J mouse blastocysts with the intention of creating a stable transgenic mouse line. This line would then be crossed with a Flt3-Cre mouse line, creating a mouse in which the fluorescent timer would start recording after expression of Flt3 starts in MPPs[62]. However, we found that the chimeras that resulted from the blastocyst injections had low expression overall in bone marrow HSPCs (Fig. 2.2D). We therefore decided

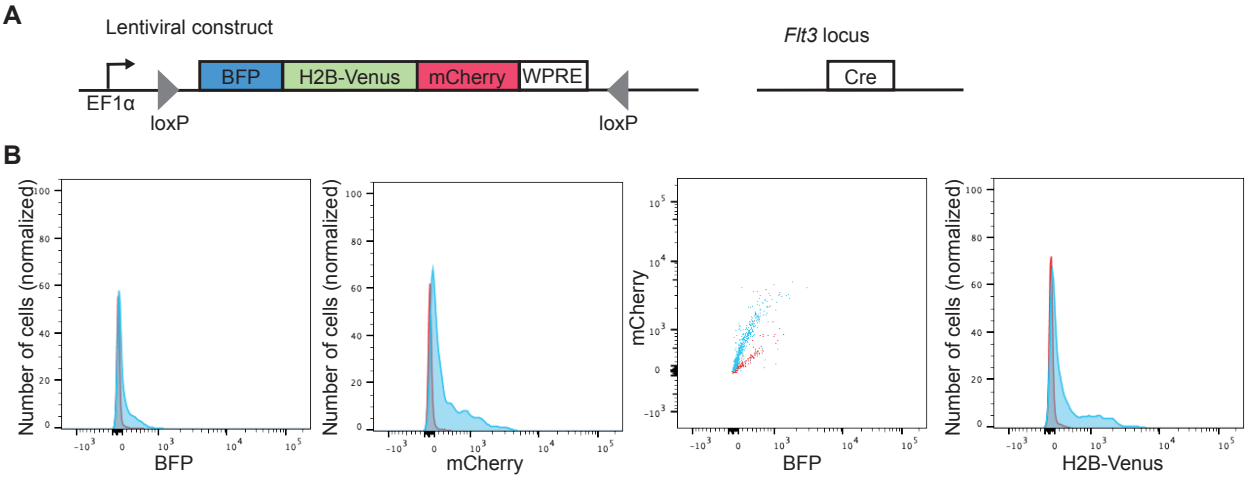

Figure 2.3: Lentiviral infection did not produce HSPCs that express the fluorescent reporter construct highly. A. Genetic architecture of lentivirus used to infect mouse HSPCs with the fluorescent timing construct constitutively expressed. The loxP sites shown cause irreversible inversion of the sequence between them when recombined with Cre. For the timer to start, the HSPCs used for the lentiviral infection must already have the germline Flt3-Cre element. However, we initially infected wild-type HSPCs (which would constitutively express the construct) to assess the maximum expression level of the proteins. B. Fluorescent protein levels in lentivirus-infected (blue) and uninfected control (red) Lin- mouse HSPCs, as measured by flow cytometry. Abbreviations used: BFP, TagBFP; EF1α, EF1α promoter; WPRE, WHV Posttranscriptional Regulatory Element.

to abandon this attempt to create a line with this construct inserted into the germline, as the low constitutive expression of all three fluorescent proteins in HSPCs would make it difficult to accurately estimate the number of divisions and the amount of time that elapsed after differentiation using fluorescence measurements alone.

We made one final attempt to stably express the multicolor fluorescent reporter in HSPCs. Since lentiviral infection produced high expression of a three-protein reporter in 293T cells (Fig. 2.1), we decided to infect HSPCs *ex vivo* with lentivirus rather than create a transgenic mouse line. Since we saw relatively low expression of the H2B-Cerulean division counter in the previous iteration of the lentiviral approach, we created a new plasmid with H2B conjugated to the brighter Venus fluorescent protein (Fig. 2.3A). Unfortunately, we found that none of the fluorescent proteins were highly expressed when the lentivirus was used to infect mouse HSPCs, despite a reasonably high infection rate in these cells (Fig. 2.3B). Because of this low expression, we did not pursue this strategy further.

2.3 Discussion

We designed and tested an experimental system that uses the degradation and dilution of fluorescent proteins to estimate the kinetics of HSPC differentiation and division in mice. While the levels of the fluorescent proteins did decrease over time after inserting the fluorescent reporter construct into 293T cells *in vitro*, we could not implement a strategy that initially expressed high enough levels of these proteins in HSPCs to make this a viable strategy to quantitatively measure differentiation and division rates. If we had successfully created such an experimental system, we would have used it to study the initial events after HSC differentiation *in vivo* by starting the timer and division counter just after differentiation into Flt3+ MPPs by using cells with both the timer and Flt3-Cre. Cells could then be sorted using their fluorescence measurements into populations that had undergone different numbers of divisions since the cell first differentiated from an HSC to an MPP (using the H2B-conjugated fluorescent protein) or into populations that had differentiated from HSCs at different times in the past (using the cytoplasmic timer fluorescent proteins). The single-cell phenotypes of these sorted populations would then be measured using flow cytometry with surface marker staining and/or scRNA-seq. We would then use the timing/division estimates and the cell state measurements together to determine how long it takes for newly-created MPPs to differentiate into downstream HSPC types. These data could be used to answer questions such as whether cell division is required for specification of particular hematopoietic lineages[57,63] or how division and differentiation kinetics change due to different perturbations.

Several factors could be responsible for the low fluorescent protein expression we observed in mouse HSPCs using both germline and lentiviral reporter constructs that prevented further development of this measurement system. Expression of multiple fluorescent proteins at high levels may be detrimental to HSPCs, leading to silencing of fluorescent protein expression. Cleavage of the P2A and T2A sites linking the three fluorescent protein sequences together in one continuous mRNA sequence may also be inefficient, leading to lower protein levels. Changing the fluorescent proteins, the order of the protein coding sequences within

the construct, and/or the location in the genome in which the constructs inserts might improve expression. Sufficient cytoplasmic fluorescent protein expression can be achieved with a very similar dual-color reporter in intestinal organoids *in vitro*[58], which, along with our 293T data, suggests that cell-type specific effects may limit the expression level of fluorescent reporters. Indeed, recent studies have suggested that, in contrast to other cell types, HSCs constitutively expressing H2B-conjugated fluorescent proteins have somewhat low fluorescence levels with considerable intercellular variation[64,56].

In general, augmenting single cell phenotypic measurements (scRNA-seq, flow cytometry, etc) with additional information on past dynamics using fluorescent proteins is a useful strategy to investigate differentiation and cell division kinetics[53,58,54,14]. While we had difficulty achieving the high, uniform expression of these proteins in HSPCs that our experimental system required, other mammalian cell types or reporter construct designs might prove more amenable to our overall approach. Such strategies allowing for precise quantification of single-cell kinetics is essential for investigating the molecular factors driving differentiation in regenerative tissues, cell-cell variation in fate commitment, and coupling of cell division and differentiation.

2.4 Methods

2.4.1 Fluorescent timing construct design and synthesis

All fluorescent constructs were designed with two cytoplasmic fluorescent proteins and one human-H2B-conjugated fluorescent protein all expressed on the same transcript. P2A/T2A linker sequences were used between the three coding sequences, which are cleaved during translation to produce three separate polypeptides. This cleavage also produced polypeptides without methionine at their N-termini, which likely increases the degradation rate of the protein in the cytoplasm[65]. Expressing all three proteins from the same transcript reduces the amount of cell-cell variation when measuring the relative amounts of the two cytoplasmic fluores-

cent proteins, improving the accuracy of the timing estimates. The designs were commercially synthesized in a lentiviral backbone by VectorBuilder.

2.4.2 Lentiviral production and concentration

To produce lentivirus that can insert the fluorescent reporter into cells of interest, lentiviral plasmids containing the reporter sequence were transfected into Lenti-X 293T cells (Takara Bio). In each 15 cm plate of Lenti-X cells, we transfected 27 μg of the fluorescent protein lentiviral plasmid along with the packaging plasmids psPAX2 (Addgene #12260; 10.8 μg per plate) and pMD2.G (Addgene #12259; 5.4 μg per plate) using the TransIT-293 Transfection Reagent (Mirus Bio). For each 15 cm plate, 130 μL of TransIT reagent was mixed with 1 mL Opti-MEM (Gibco) and left at room temperature for 10 min. Then, the TransIT mixture was mixed with the plasmids dissolved in another 1 mL Opti-MEM, and left at room temperature for 15 min. The media in the culture plates was changed to 10 mL DMEM complete (supplemented with 10% FBS, pen-strep, L-glutamine) and the TransIT-plasmid solution was added to the culture plate dropwise. Cells were incubated for 48 hours, after which 5 harvests were performed, each 12 hours apart. For each harvest, the culture supernatant was collected and kept at 4°C and an additional 12 mL DMEM complete was added to the culture.

To concentrate the lentivirus, the viral supernatant was filtered through a 0.45 μm filter. Filtered supernatant was ultracentrifuged at 16,000 rpm for 90 min at 4°C and the supernatant was discarded. The pellets were resuspended with the remaining residual liquid and stored at -80°C.

2.4.3 *In vitro* timecourse to test kinetics of the lentiviral Tet-On system in 293T cells

Unconcentrated lentivirus was used to produce monoclonal 293T lines stably transfected with the Tet-On reporter construct (Fig. 2.1A). 293T cells stably expressing rtTA and a puromycin resistance gene were maintained in puromycin and infected with viral supernatant at a 1:2 dilution with 5 μg/mL polybrene for 48 hours.

Doxycycline (1 μg/mL) was then added to induce expression of the fluorescent proteins. Single cells that highly expressed all three proteins were sorted into individual wells of a 96-well plate and monoclonal lines were chosen from the surviving single-cell cultures for further experiments.

To assess the kinetics of fluorescent protein loss, clones expressing the reporter and rtTA were kept in media with 1 μg/mL dox for $>$ 1 week to reach steady-state fluorescent protein concentrations within the cells. Dox was then withdrawn from the cells and flow cytometry was performed at 6, 12, 24, 48, 72, or 96 hours after withdrawal to measure fluorescent protein levels.

2.4.4 ES cell gene targeting to insert the timing construct into the TIGRE locus

Gene targeting via homologous recombination was used to insert the Cre-driven reporter construct (Fig. 2.2A) into the TIGRE expression locus. The targeting vector was created by cloning the reporter sequence into a TIGRE targeting vector with homology arm sequences (Addgene #92142). 25 μg each endotoxin-free linearized TIGRE targeting plasmid and a CRISPR spCas9 plasmid based on the PX459 backbone (Addgene #62988) with TIGRE locus sgRNA sequences were transfected into 5 million C57BL/6J agouti ES cells using the Mouse Embryonic Stem Cell Nucleofector Kit (Lonza Bioscience). One ES cell clone with PCR-confirmed targeting that was observed to express all three fluorescent markers was chosen for injection into C57BL/6J blastocysts for chimera production.

2.4.5 Lentiviral spin infection and flow cytometry of HSPCs

Bone marrow from a wild-type C57BL/6J was isolated and lineage depleted with MACS Cell Separation (Miltenyi BIotec), using CD8, CD4, Ter119, B220, and Gr1 biotin-conjugated antibodies and Anti-Biotin MicroBeads (Miltenyi Biotec). For spin infection, 20,000 Lin- cells were plated in 100 μL F12 media (Gibco) with 5 μg/mL polybrene. 10 μL of concentrated virus was added and cells were centrifuged at 800xg for 90 min at room temperature. 100 μL F12 media was added to each well and TPO (100 ng/mL; BioLegend),

SCF (10 ng/mL; Miltenyi), and ITS-X (1X; Gibco) were added and the cells were cultured at 37°C. 24 hours later, an additional 10 uL of concentrated virus was added. Flow cytometry was performed on the culture following fixation with 4% paraformaldehyde 4 days after the initial spin infection.

3

Mathematical modeling to infer hematopoietic differentiation kinetics

Foreword

Fernando Camargo conceived of and supervised this project. My role was to construct the mathematical model of hematopoiesis and use it to estimate differentiation and proliferation rates from the experimental lineage tracing data. The analysis was performed with input and additional supervision from Thomas Höfer and Franziska Michor. Qi Yu performed all experiments, with initial assistance creating the transgenic mouse line from Basanta Gurung and Constantina Christodoulou. Qi also performed the initial gating and analysis

of the flow cytometry data. We will likely use some of these results in a paper with additional data from Hans-Reimer Rodewald's lab investigating how hematopoietic stem cell differentiation rates change in response to various stimuli.

Abstract

While hematopoietic stem cells (HSCs) are the only cell type that can fully reconstitute the entire blood system, much of the cell proliferation required to maintain sufficient differentiated output is hypothesized to occur in other downstream hematopoietic cell types. To quantitatively estimate how quickly the least-differentiated population of HSCs contributes to blood production, we developed a transgenic mouse model that specifically labels HSCs with a heritable fluorescent marker after administration of tamoxifen. We measured this fluorescent label frequency in different bone marrow cell types and used a mathematical model of stem cell differentiation and proliferation to estimate the rate at which HSCs differentiate in their native context in mouse bone marrow. We found that HSCs differentiate slowly, and that this differentiation rate is lower in older mice, which is consistent with some previous reports using a similar overall HSC labeling approach. This new mouse model will be used to assess the impact of different perturbations (e.g., inflammation, irradiation) on HSC output.

3.1 Introduction

Hematopoietic stem cells are the ultimate source of all blood cells in the adult mammal. However, experimental data suggest that more downstream hematopoietic progenitors- namely, multipotent progenitors (MPPs)- are the cells that directly contribute to production of most mature blood cell types over short and intermediate timescales[66,51,67]. According to this hypothesis, these downstream transit-amplifying populations proliferate rapidly to provide sufficient numbers of differentiated cells to maintain homeostasis, while true HSCs are more quiescent and represent a reserve population of long-lasting stem cells that can be activated in response

to illness or injury to provide additional hematopoietic support if necessary. In contrast, other studies suggest that HSCs are more active contributors to steady-state hematopoiesis[68,69,70]. Accurately determining the contribution of HSCs to steady-state hematopoiesis is important for understanding how factors such as aging and infection affect hematopoietic output and risk of malignancy.

Many studies that quantitatively measure the contribution of HSCs to unperturbed *in vivo* hematopoiesis employ similar lineage tracing strategies, though small differences between each particular mouse model used may lead to different conclusions. The general approach ideally involves specifically labeling the most upstream HSCs, which give rise to all other hematopoietic stem and progenitor cell (HSPC) types, with a heritable marker (often expression of a fluorescent protein). The speed at which downstream populations acquire labeled descendants of HSCs can be used to estimate the contribution of HSCs to downstream populations, and mathematical modeling can be used to provide quantitative estimates of HSC differentiation rate using these data. However, this approach has some limitations that could contribute to the controversy surrounding the contribution of HSCs to hematopoiesis. First, some labeling strategies do not specifically label HSCs, and/or do not label a high enough fraction of HSCs to reliably assess their contribution to downstream populations. For example, while the *Pdzk1ip1*-CreER mouse model initially labels approximately 30% of the most undifferentiated long-term HSCs (LT-HSCs) upon induction with tamoxifen, approximately 10% of short-term HSCs (ST-HSCs) and 3% of MPPs are also labeled[68]. This initial population of labeled cells in other compartments can expand and contribute to downstream blood production, interfering with the estimate of the LT-HSC contribution to hematopoiesis. Second, most approaches to inferring HSC differentiation rates assume that HSPC differentiation rates and the number of HSCs and other stem and progenitor types are constant over the lifetime of the animal[51,68]. However, previous studies have shown that older mice have more HSCs than younger mice[52] and that aging affects the rates at which different blood lineage (e.g., myeloid and lymphoid) are produced[71,72], suggesting that an approach that can account for changes in differentiation rates over time may produce more accurate estimates[52].

To address these issues, we developed a new transgenic mouse model that labels LT-HSCs with a heritable fluorescent marker more specifically than previous models. We used the observed label frequencies in different HSPC compartments to estimate the rate at which LT-HSCs differentiate into ST-HSCs and thereby contribute to blood production. The mathematical model we used to infer this rate allows for changes in the LT-HSC differentiation rate over time. We found that LT-HSCs contribute slowly to blood production, and that their per capita differentiation decreases over time. We are currently using this lineage tracing mouse model to determine whether perturbation of the hematopoietic system by infection, inflammatory stimuli, or irradiation changes LT-HSC differentiation kinetics.

3.2 Results

3.2.1 We developed an inducible labeling system that specifically marks the least-differentiated Mds1+ HSCs.

In order to measure the *in vivo* rate of HSC differentiation in steady state and perturbed conditions, we developed a transgenic mouse which specifically marks LT-HSCs and their progeny after induction with tamoxifen. In this mouse model, cells expressing Mds1, which is a part of a transcription factor complex (Mds1-Evi1, part of the *Mecom* locus) expressed specifically in HSCs[73], express the fluorescent protein mEOS2 (Fig. 3.1A). Expression of mEOS2 is restricted to HSPCs that have not yet expressed Flt3, a marker of differentiation into multipotent progenitors (MPPs) that is not expressed in mouse HSCs[62]. Cells expresssing mEOS2 from the *Mds1* locus also express FlpOERT2, which causes these Mds+Flt3- HSCs to irreversibly turn on expression of the tdTomato fluorescent marker upon tamoxifen exposure. A subset of Mds+Flt3- HSCs is therefore permanently labeled with tdTomato, and this label is eventually propagated to downstream cell types when Tom+ HSCs divide and differentiate (Fig. 3.1B). We found that, one week after tamoxifen induction, expression of tdTomato was highly restricted to the population of LT-HSCs (CD150+CD48- Lin-

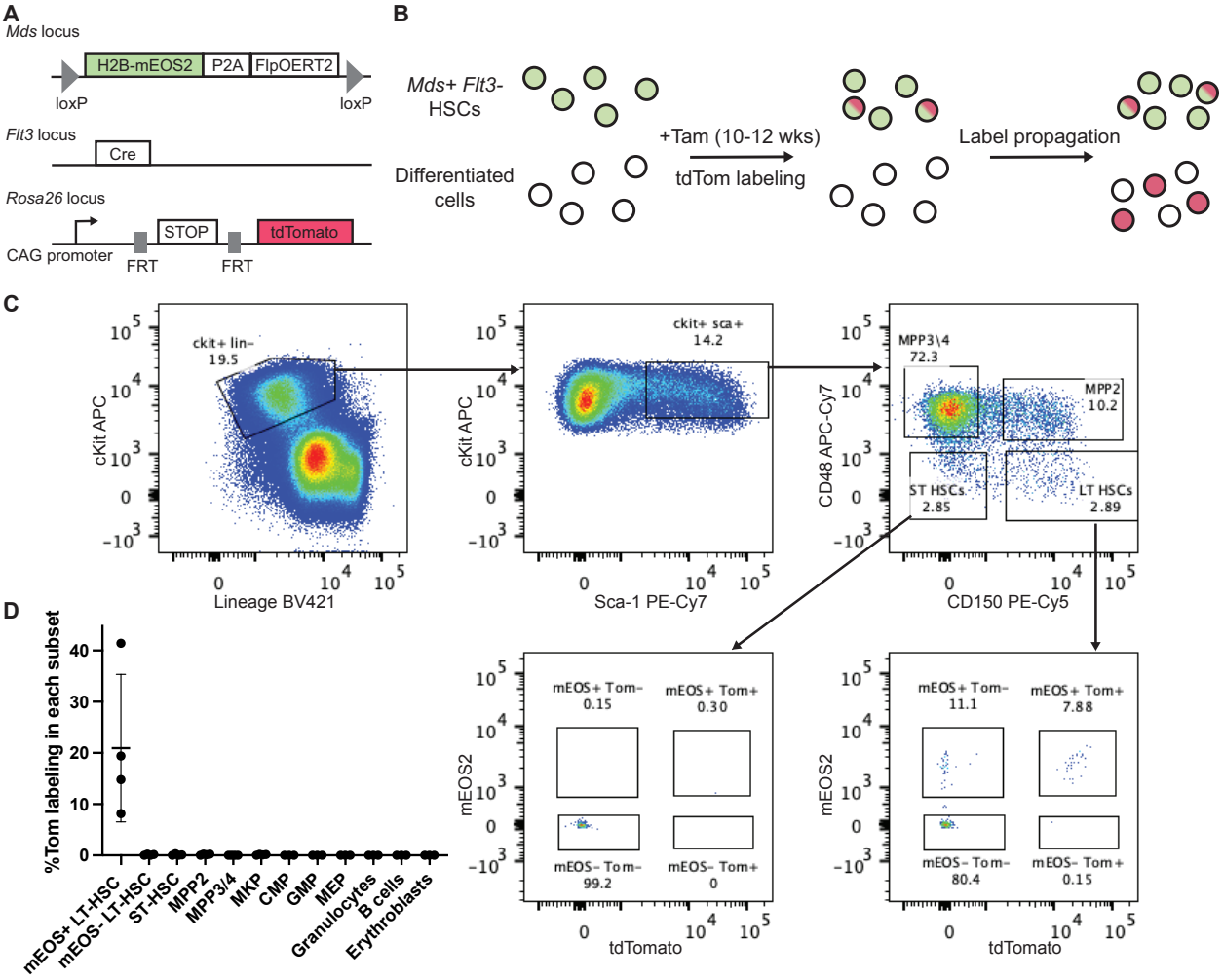

Figure 3.1: Our novel inducible fluorescent lineage tracing system specifically labels HSCs *in vivo*. A. Genetic architecture of the Mds dual color labeling transgenic mouse model. B. Diagram describing fluorescent label propagation after induction with tamoxifen. Red denotes cells that express tdTomato, while green denote cells that express mEOS2. C. Gating strategy to identify Tom+ HSCs and MPPs from an initial population of live single HSPCs, D. Percent of cells of each HSPC type expressing tdTomato 1 week after label induction with 2 mg tamoxifen.

Sca1+cKit+ (LSK) cells) that also expressed mEOS2 (Fig. 3.1C-D). Our model has less nonspecific labeling of downstream cell compartments, including ST-HSCs (CD150-CD48- LSKs) and MPPs (CD150+/-CD48+ LSKs), than a previously-published mouse model that labeled Tie2+ HSPCs[51].

The rate at which the tdTomato label frequency increases in more-differentiated HSPC types is related to the rate at which HSCs contribute to hematopoiesis. Faster HSC differentiation would lead to faster replacement of Tom- cells in downstream compartments with Tom+ HSC progeny. Unfortunately, we were not able to repeatedly sample the bone marrow in individual mice multiple times, so we could not determine how the

Tom+ frequency in bone marrow cell types changed over time within the same mouse. Therefore, to track the rate of label propagation after tamoxifen induction, we induced Tom labeling in a large cohort of adult mice (n=49) and sacrificed these mice at different times 1 - 52 weeks after labeling. We then used flow cytometry to measure how the label frequency in bone marrow HSPC types changed after HSC labeling.

3.2.2 A mathematical model of HSPC dynamics can be used to estimate time-dependent differentiation and net growth rates of HSPC types.

We used a deterministic model based on a previously-published method[51] to estimate the rates of steady-state HSPC differentiation and division in mice using our tdTomato labeling data. This model is a system of differential equations that describe how the frequency of tdTomato labeled cells change over time after tamoxifen administration. We assumed that each HSPC type could undergo two main types of dynamic behaviors (Fig. 3.2A):

1. Differentiation, in which a cell from an upstream compartment becomes a different downstream cell type. Cells spend an exponentially-distributed amount of time in each compartment and differentiation from compartment i into compartment j at a rate $\alpha_{i\to j}$, which has the same value for all cells in the compartment.
2. Net growth of cells in the compartment. Cells can undergo symmetric self-renewal, in which a cell divides into two new cells of the same type, or can be eliminated from the compartment through cell death. However, we only explicitly model the net growth of cells in the compartment (self-renewal - cell death). Expansion of the compartment occurs at the net growth rate β_i, which has the same value for all cells in a particular compartment i.

Assuming that Tom+ and Tom- cells have the same differentiation and division rates, the number n_i of labeled or unlabeled cells in a specific compartment is

$$\frac{dn_i}{dt} = \alpha_{u\to i} n_u + (\beta_i - \alpha_{i\to d}) n_i \tag{3.1}$$

if compartment i has exactly one contributing upstream compartment u and one downstream compartment d. The fraction of cells $f_i = \frac{n_{i,\text{lab}}}{n_{i,\text{total}}}$ in each compartment with the tdTomato label is governed by the differential

equation

$$\frac{df_i}{dt} = \alpha_{u\to i} f_u \left(\frac{n_{u,\text{total}}}{n_{i,\text{total}}}\right) + (\beta_i - \alpha_{i\to d}) f_i - \frac{f_i}{n_{i,\text{total}}} \frac{dn_{i,\text{total}}}{dt} \tag{3.2}$$

If we assume that the total (Tom+ and Tom-) number of cells in each compartment is approximately constant ($\frac{dn_{i,\text{total}}}{dt} \approx 0$), then[51]

$$\frac{df_i}{dt} = \frac{\alpha_{u\to i} n_{u,\text{total}}}{n_{i,\text{total}}} (f_u - f_i). \tag{3.3}$$

This equation can be used to estimate a constant differentiation rate $\alpha_{u\to i}$ into compartment i by using the experimentally-measured label frequencies of the u and i compartments and the total population sizes of the two compartments. However, if the upstream compartment size $n_{u,\text{total}}$ is not constant, Equation 3.3 can also be rearranged to provide a time-dependent estimate of the differentiation rate[52]

$$\hat{\alpha}_{u\to i} = \frac{\hat{df}_i}{dt} \frac{n_{i,\text{total}}}{n_{u,\text{total}} (f_u - f_i)} \tag{3.4}$$

using the numerical derivative of label frequency $\frac{\hat{df}_i}{dt}$.

3.2.3 HSCs differentiate slowly into downstream cell types.

We measured the frequency of Tom+ cells over time in different HSC types using flow cytometry (Fig. 3.2B). The least-differentiated HSC population (mEOS+ Mds1+Flt3- LT-HSCs) was the only population with significant tdTomato labeling initially, and this label frequency varied between individual mice. The Tom+ cell frequency in the mEOS+ LT-HSC population did not decrease over time and did gradually increase in all other HSC compartments, indicating that the mEOS+ LT-HSCs were the ultimate source of all HSC types. Since label frequency data from all mice were combined to estimate the HSC differentiation rates, we normalized the label frequency in all compartments to the mEOS+ LT-HSC Tom+ frequency to reduce the effect of interindividual variation in initial label induction efficiency (Fig. 3.2C). While we found that the LT-HSC

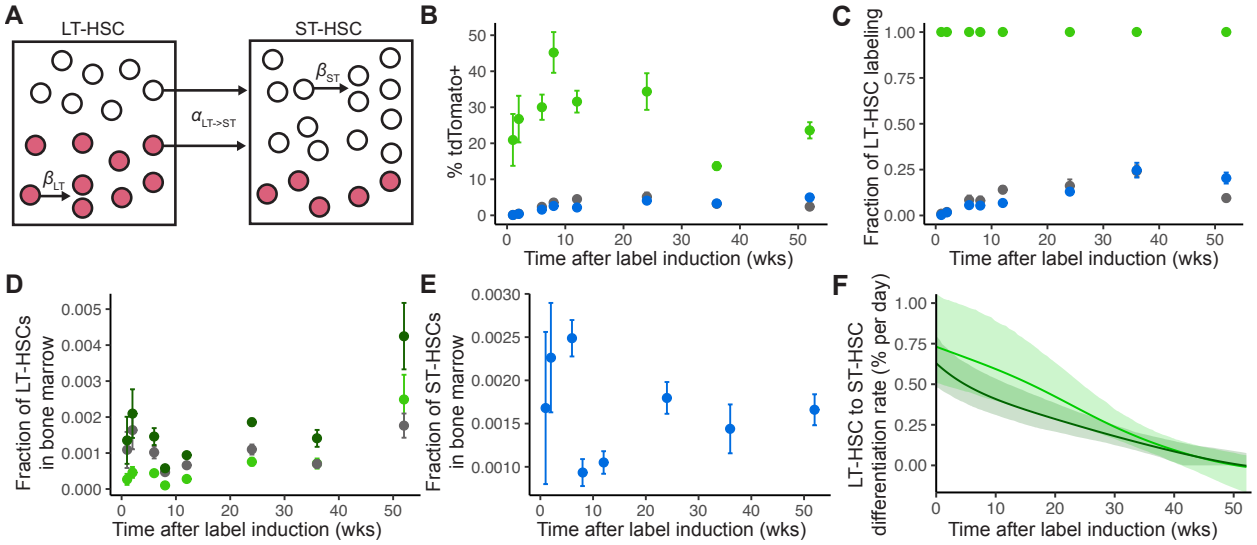

Figure 3.2: Slow propagation of the tdTomato label suggests that LT-HSCs do not contribute much to steady-state hematopoiesis. A. Differentiation and self-renewal model diagram showing two example compartments (LT-HSCs and ST-HSCs). Each compartment has a net proliferation rate (β) and a differentiation rate (α) that are assumed to be the same for all cells (Tom+ and Tom-) in that population. B. Frequency of tdTomato-labeled cells in mEOS+ LT-HSCs (green), mEOS- LT-HSCs (grey), and ST-HSCs (blue). Each mouse yielded a single measurement per bone marrow cell type (49 total mice). Error bars are +/- SEM. C. Normalized frequency of tdTomato-labeled cells in mEOS- LT-HSCs (grey), ST-HSCs (blue), and MPP3/4s (orange), relative to the labeling frequency in mEOS+ LT-HSCs (green). Error bars are +/- SEM. D-E. Fraction of mEOS+ LT-HSCs (D, light green), mEOS- LT-HSCs (D, grey), all LT-HSCs combined (D, dark green), or ST-HSCs (E, blue) in bone marrow. Error bars are +/- SEM. F. Estimated differentiation rate for all LT-HSCs (dark green) or mEOS+ LT-HSCs only (light green) into ST-HSCs. The shaded regions are bootstrapped 95% confidence intervals (n=1000 samples for each).

population size (in particular, the mEOS+ LT-HSCs) increased over time (Fig. 3.2D), as has been noted previously [52], the ST-HSCs did not seem to monotonically increase in number over time (Fig. 3.2E).

We used Equation 3.4 to estimate the time-dependent rate at which LT-HSCs differentiated into ST-HSCs from the measured tdTomato label frequencies (Fig. 3.2F). If we assumed that all LT-HSCs directly contributed equally to ST-HSC production, we found that the maximum LT-HSC differentiation rate was relatively low ($< 0.7\%$ of the LT-HSC population differentiated into ST-HSCs per day). This differentiation rate decreased over time. Since the total number of LT-HSCs increased over time (Fig. 3.2D), our results suggest that the observed HSC expansion that occurs during aging is balanced by a decrease in the per capita HSC differentiation rate. When we assumed that only mEOS+ LT-HSCs contribute to downstream blood production, we estimated an HSC differentiation rate similar to the one estimated assuming all LT-HSCs contribute equally (Fig. 3.2F). While we cannot formally distinguish between these two possible configurations for

the hematopoietic differentiation hierarchy from the labeling data alone, the inferred differentiation rates from both hierarchy structures we tried suggest that LT-HSCs do not directly contribute much to steady-state hematopoiesis over short timescales.

3.3 Discussion

Using our new lineage tracing mouse model that very specifically labels LT-HSCs, we found that LT-HSCs differentiate slowly ($< 1\%$ per day) to create other hematopoietic progenitor types. This rate decreased over time, though the total differentiated output of the entire LT-HSC compartment remained relatively constant during aging since the number of LT-HSCs increased over time. Our inferred differentiation rates are similar to those estimated recently using a different HSC lineage tracing mouse model[52,51], and suggest that expansion in more differentiated compartments (ST-HSCs, MPPs, etc.) drives hematopoietic production at steady state.

Other direct measurements of HSPC dynamic behaviors could be used in conjunction with our label propagation data to improve our understanding of HSPC division and differentiation kinetics. In particular, orthogonal experimental measurements of cell division rates in HSPCs (i.e., EdU labeling) have been recently used with Tie2 HSC labeling data to improve the accuracy of differentiation rate estimates[52]. Importantly, this strategy can also be used to estimate rates of asymmetric division and cell death in HSPCs[52], which cannot be specifically estimated using fluorescent label propagation data alone. Our model could be extended to incorporate similar measurements of cell division rate.

Our approach and other similar approaches combining fluorescent label propagation and mathematical modeling share a few important assumptions that limit their ability to accurately describe hematopoietic differentiation and division kinetics. First, these approaches require separating HSPCs into discrete cell types and modeling transitions between these separate populations. While earlier work defined discrete hematopoietic progenitor cell types according to surface marker expression and differentiation potential[74], more recent

single-cell sequencing results suggest that HSPC differentiation is better described by a continuum[63]. Second, this type of compartment model also assumes that all cells within a compartment are equivalent and all have the same division and differentiation probabilities at all times. Finally, these approaches often require some prior knowledge of the structure of the hematopoietic differentiation hierarchy. While many parts of the hierarchy have been well-established experimentally, recent studies have raised questions about the existence or origin of some HSPC types (e.g., megakaryocyte progenitors[67,75], common lymphoid progenitors[76]). Strategies that measure differentiation kinetics without resorting to prior knowledge of specific HSPC types or differentiation hierarchies, such as joint single cell barcoding and scRNA-seq[77].

We are currently using this transgenic mouse model to determine how stimuli such as irradiation, inflammation, and infection change the rate of HSPC differentiation and/or cell division. Sublethal irradiation eliminates many HSPCs, requiring expansion of the surviving cells to compensate for the loss, while infection and inflammation increase the rate of HSC proliferation to increase the production of immune cells[78,79]. Determining which cell types directly respond to these stimuli and what specific kinetic effects they cause (increased differentiation, proliferation, or cell survival) could help us understand how the HSC compartment is maintained under stress and how to better treat stress-related cytopenias.

3.4 Methods

3.4.1 Experimental procedures

Adult mice (10-12 weeks old) with the germline transgenes shown in (Fig. 3.1A) were induced with 2 mg tamoxifen via intraperitoneal injection. After 1 - 52 weeks (>= 4 mice per timepoint), these mice were sacrificed and bone marrow mononuclear cells from the pelvis, spine, and both femurs and tibiae were isolated. Flow cytometry was performed to measure the frequency of tdTomato labeling in each HSC type, following surface staining with the following antibodies:

1. cKit APC

2. Lineage (Ter119, CD4, CD8, B220, Gr1) BV421
3. Sca-1 PE-Cy7
4. CD150 PE-Cy5
5. CD48 APC-Cy7

HSC/MPP cell types were identified using the gating strategy in Fig. 3.1C, after manual compensation.

3.4.2 Estimation of HSPC differentiation and proliferation rates

Our modeling approach is adapted from previous work estimating constant[51] and time-dependent[52] HSPC differentiation rates from other HSC lineage tracing mouse models. We used Equation 3.4 to numerically estimate the time-dependent LT-HSC to ST-HSC differentiation rate $\hat{\alpha}_{u\to i}$. We first normalized the Tom+ label frequency with the label frequency of the mEOS+ LT-HSCs so that the normalized mEOS+ LT-HSC label frequency was 1 for each mouse. Then, we smoothed the mEOS+ and mEOS- LT-HSC and ST-HSC Tom+ label frequencies over time using local polynomial smoothing (degree = 1). We also smoothed the measured frequency of these cell types in the bone marrow samples over time using the same local polynomial smoothing procedure. The time derivative of the ST-HSC Tom+ frequency $\frac{\hat{df}_i}{dt}$ was estimated numerically using the smoothed frequency data using the symmetric difference quotient. The LT-HSC differentiation rate was estimated up to one year after label induction using these smoothed curves.

Confidence intervals on the LT-HSC differentiation rate estimates were calculated by bootstrapping. Label frequency data and total compartment size were resampled 1000 times, preserving the number of samples from each timepoint during each resampling. Differentiation rates were estimated using the same procedure outlined above and the empirical 95% confidence intervals were computed from this sampled data.

4

Single-cell barcoding to characterize mammalian embryonic development

Foreword

This project was conceived by Fernando Camargo and Sarah Bowling. Sarah performed all of the experiments under Fernando's supervision. I performed the bulk of the data analysis with input from Sarah and Fernando. Additional assistance with cell type identification, data preprocessing, and access to mouse embryo atlas data, as well as general advice, was provided by Mai-Linh Ton, Ivan Imaz-Rosshandler, and Bertie Göttgens. Help with the computational analysis was also provided by Duluxan Sritharan and Sahand Hormoz, who

developed the original analysis pipeline for the CARLIN barcoding system[80]. Experiments, data analysis, and interpretation of results are still ongoing.

Abstract

Many important cell fate specification events in mammalian embryogenesis occur during gastrulation. While lineage tracing experiments and single-cell technologies have revealed much about how embryonic progenitor cells expand and differentiate during this process, questions about the origins and differentiation trajectories of specific cell types (e.g., hematopoietic cells, endothelium, the heart) remain. To assess the fate potential and proliferative capacity of all embryonic cell types during gastrulation, we used a CRISPR barcoding system (CARLIN) to identify the progeny of individual cells that existed at specific timepoints in the early mouse embryo. The inducible CARLIN barcodes are transcribed and can be captured and amplified during scRNA-seq library preparation, enabling the joint measurement of the transcriptome and lineage information in the same set of single cells. We used the data from this system to investigate the kinetics of fate restriction and the origins of endothelial cells in the early mouse embryo.

4.1 Introduction

In mouse embryos, specification of most tissue types occurs during gastrulation at approximately embryonic day (E) 6.5-8.5. During this process, pluripotent cells of the epiblast are thought to first differentiate into the three germ layers of the embryo proper (endoderm, mesoderm, and ectoderm), which go on to eventually give rise to all of differentiated tissues of the adult animal[81,82]. Recently, this highly dynamic developmental event has been investigated by several groups using single cell omics technologies to determine what cell states are present in the embryo at different times during development. These studies have revealed new information about the origins of the germ layers[83], blood progenitors[84], gut and other endodermal tissues[85,84], and the somites[86] in the mouse. One important limitation of most of these single-cell studies that focus on the

current molecular state of each cell is that they often cannot definitively identify specific cell state changes that occur during embryogenesis. Instead, often the best they can do is to infer differentiation dynamics using static phenotype data under some set of restrictive assumptions, usually that cell state changes are continuous and occur gradually. While these approaches have yielded important new biological insights, some of which have been formalized into quantitative computational models of development[87,88], a more direct single-cell experimental measurement of cell state changes during early mouse embryogenesis would allow us to rigorously test these models and findings and may reveal unexpected differentiation behaviors.

To this end, we used a CRISPR-based single-cell inducible barcoding system[80] to simultaneously measure both cell lineage histories and transcriptomic profile in the mouse embryo after gastrulation has completed. This system is capable of identifying and characterizing the phenotype of the descendants of individual progenitor cells that existed before or during gastrulation, and has more specific temporal control of barcode editing than a previous CRISPR-barcoding used in mouse embryos[89]. Using these data, we explored questions related to the fate potential and expansion kinetics of individual embryonic progenitors, and specifically investigated the origins of endothelial cells in different regions of the mouse embryo. In the future, we intend to use these data to investigate lineage convergence and divergence in other tissues, as well as to validate and/or refine other quantitative models of mouse embryonic development that were generated without lineage data.

4.2 Results

4.2.1 Expressed single-cell barcoding allows for joint transcriptomic and lineage information in mouse embryos.

To investigate the dynamics of proliferation and cell fate specification in the early mouse embryo, we improved the CARLIN CRISPR-based single-cell barcoding system[80] previously developed in the Camargo

Lab and used it to uniquely mark clones derived from individual progenitor cells during embryogenesis. To improve our ability to read out barcodes in each cell in our in scRNA-seq data, an additional CRISPR target array was added to the TIGRE expression locus[61] with the same 10 target sequences as in the original CARLIN target array (Fig. 4.1A). This TIGRE CARLIN target array provided a second opportunity to capture lineage information in each cell when performing single-cell transcriptomic profiling, in addition to the original CARLIN barcodes still present. A single set of 10 matching constitutively expressed guide RNAs permits random editing of both the original and TIGRE CARLIN arrays when Cas9 expression is induced by doxycycline (dox) exposure. Therefore, a single pulse of dox creates random barcodes in both target arrays which uniquely mark the descendants of individual progenitor cells at the time of CARLIN array editing.

By administering dox to early (pre-gastrulation) mouse embryos and using scRNA-seq to simultaneously read transcriptional phenotype and CARLIN lineage information in individual cells, we determined how many descendants each individually-barcoded early-stage progenitor had, and measured the phenotype of these descendants in the embryo after gastrulation and tissue specification (Fig. 4.1B). We induced CARLIN barcoding in four mouse embryos at three different timepoints in embryonic development (one embryo at E5.5, SB800; two embryos at E6.5, SB490 and SB361; and one embryo at E7.5, SB984) and performed scRNA-seq on all embryos at E9.25 (after gastrulation) to determine how the fate potential of individual progenitors changes during development. It is important to note that this CRISPR editing process occurred gradually over a period of 36 hours after dox administration in mouse embryos (Fig. 4.1C), suggesting that the progenitors are likely being marked with CARLIN barcodes several hours after dox administration and that it is difficult to precisely estimate the time at which each barcode was created. Therefore, the cells in which the barcodes were originally generated represent a slightly more variable and more mature progenitor pool than if CARLIN barcoding occurred instantly at the time of dox administration.

By specifically amplifying the CARLIN transcripts in the scRNA-seq libraries (as was performed previously[80]), we were able to detect at least one original CARLIN or TIGRE CARLIN array transcript in the

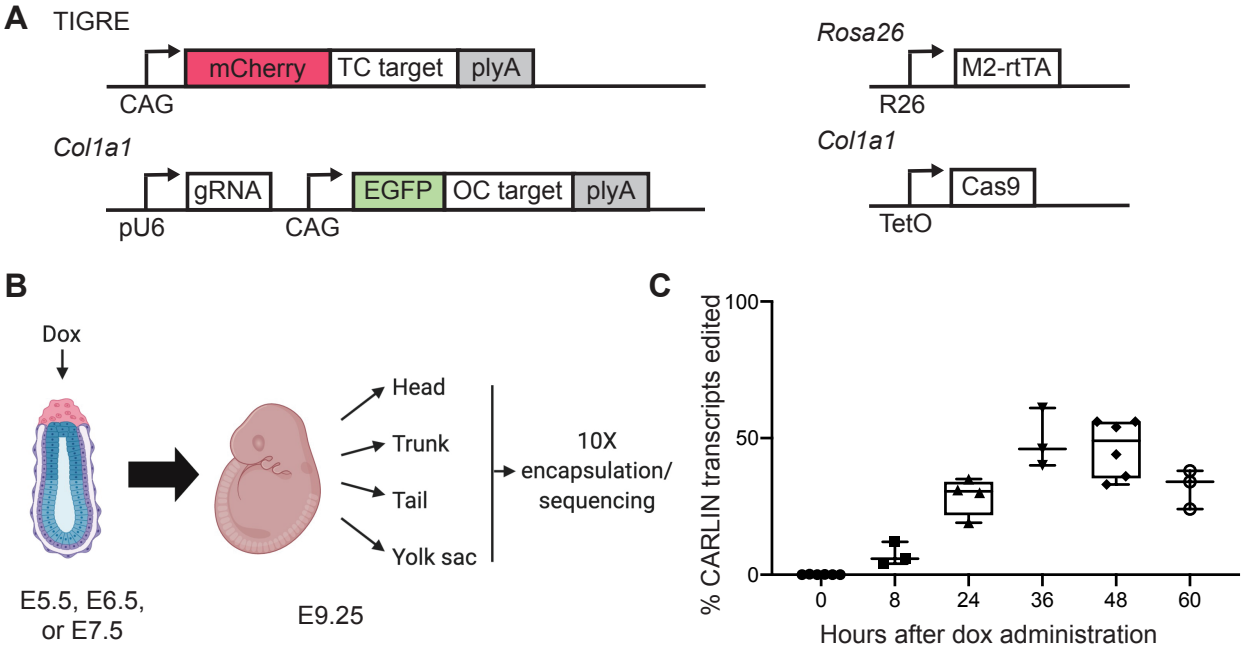

Figure 4.1: The CARLIN CRISPR-Cas9 barcoding system was used to investigate the dynamics of early mouse embryo fate specification. A. Genomic architecture of the transgenic mice with two separate CARLIN target arrays (original CARLIN, OC; and TIGRE CARLIN, TC). Editing of these arrays is induced by dox administration, which leads to Cas9 expression and random CRISPR editing of both target arrays. B. Experimental schematic, showing early *in utero* induction of CARLIN barcodes with one dox injection at E5.5, E6.5, or E7.5 and subsequent measurement of single-cell transcriptional phenotype and CARLIN lineage information using scRNA-seq at E9.25. C. CARLIN barcode editing kinetics in mouse embryos, as measured by bulk sequencing. Panels B and C were used with permission from Sarah Bowling.

majority of cells sampled (Fig. 4.2A). The TIGRE CARLIN capture efficiency was slightly higher than that of the original system, likely because it was expressed at higher levels than the original array. However, only a minority of CARLIN arrays were edited (Fig. 4.2B), reducing the number of cells with available CARLIN lineage information. While editing and capture efficiency was variable between embryos, there was not an obvious systematic effect of barcode induction time on either metric.

Since some CRISPR edit patterns are far more likely to be generated than others (especially deletions of entire target sequences within an array)[80], not all edited barcode sequences were likely to represent unique editing events that happened in a single progenitor in the embryo at the time of labeling, and therefore would not mark clonal cell populations. To filter out these high-frequency edits, we estimated the probability of generating each edit pattern by sequencing a large pool of granulocytes in which CARLIN editing was induced, and removed CARLIN barcodes that occurred very frequently in this allele bank (Methods)[80]. Therefore,

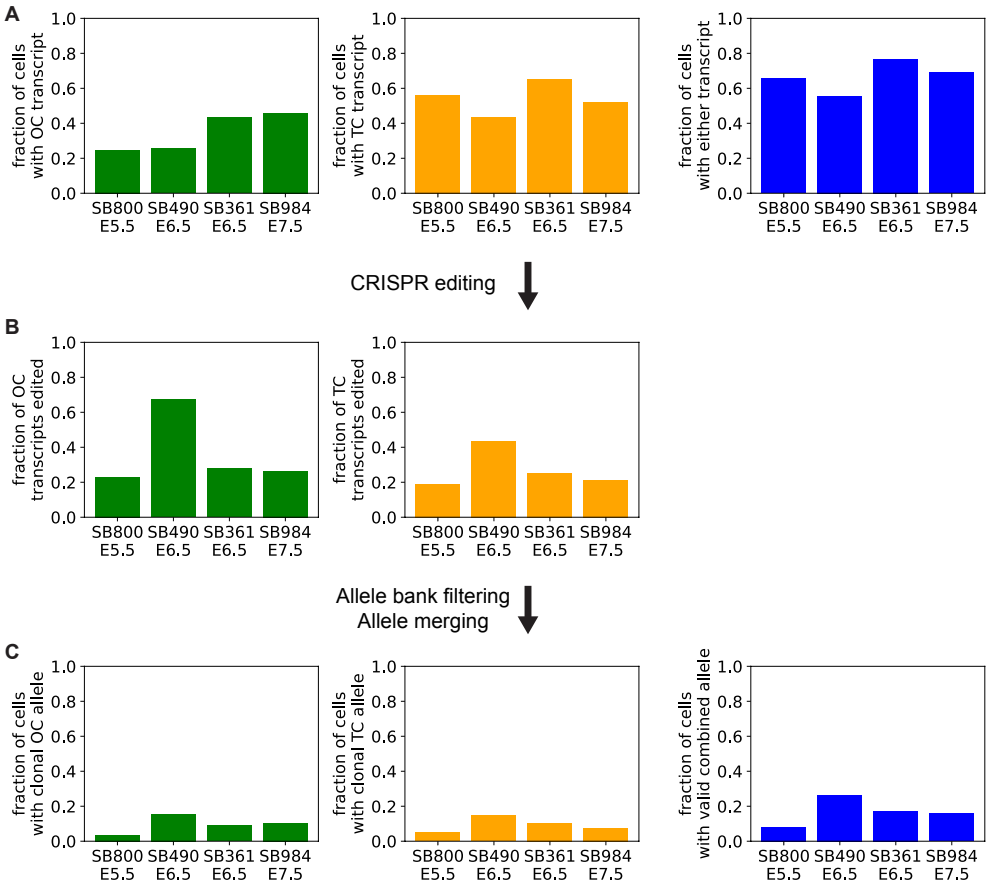

Figure 4.2: Approximately 10-20% of cells have a valid CARLIN barcode call. A. Capture efficiencies for original CARLIN (green), TIGRE CARLIN (orange), or either target array (blue) in scRNA-seq data from each embryo. B. Editing efficiencies (fraction of CARLIN transcripts that were edited) for original CARLIN (green) and TIGRE CARLIN (orange). C. Overall fraction of cells in each embryo with an original CARLIN (green), TIGRE CARLIN (orange), or combined OC/TC allele (blue) detected that was determined to be a clonal barcode that marked the descendants of a single progenitor at the time of barcode induction.

only a small fraction of cells overall (approximately 10-20%) expressed at least one edited CARLIN transcript that could be used to identify clonal cell populations (Fig. 4.2C).

To maximize the number of cells with available clonal lineage information, we took into account the sequences of both the original (OC) and TIGRE CARLIN (TC) arrays for each cell to create a unique merged barcode for each OC/TC allele pair (Methods). For each embryo sequenced, we defined the set of all unique (OC, TC) allele pairs found in at least one sampled cell as the set of all possible merged CARLIN barcodes. Cells with both OC and TC transcript calls were assigned to the corresponding merged barcode category, while cells with only one CARLIN transcript (either OC or TC) detected was assigned a merged barcode

category at random from the available merged barcodes that share the edited CARLIN transcript that was detected in the cell. Cells without a valid edited TC or OC call were not assigned a merged barcode and were therefore discarded from further analyses on CARLIN barcoded clones. For some embryos we sequenced, this merging procedure nearly doubled the number of cells with a valid barcode detected over using either OC or TC information alone (Fig. 4.2C).

4.2.2 We identified cell types in our mouse embryo scRNA-seq data with the help of an annotated atlas.

We identified embryonic progenitor cell types in our scRNA-seq data from the four embryos sequenced at E9.25 by mapping our data onto an unpublished E9.25 scRNA-seq atlas from Ivan Imaz-Rosshandler and the Göttgens Lab. This atlas extends a previous atlas from the same group[84] and was generated using similar experimental and computational methods. We used Harmony[90] to integrate our data with the E9.25 atlas data and assigned each cell in our data the cell type label of its nearest neighbor in the reference atlas (Methods). After merging some of similar lower-frequency cell types into larger categories, we found that the inferred cell types represented transcriptionally-similar cells in our dataset (Fig. 4.3A) and expressed appropriate marker genes (Fig. 4.3B). We also used these cell type labels to identify the germ layer or other general early progenitor tissue type associated with each cell (Fig. 4.3C).

Cells from different gross anatomical regions of each embryo (head, tail, trunk and yolk sac) were physically separated and sequenced separately. While some cell types (e.g., erythroid progenitors) were found predominantly in a specific region of the embryo, most cell types were found in multiple anatomical regions in the embryo proper (Fig. 4.3D). After batch correction between individual embryos in our dataset (Methods), there weren't striking differences in the transcriptional profiles between the four embryos sequenced (Fig. 4.3E).

Most cell types had similar CARLIN transcript capture efficiencies and editing efficiencies (Fig. 4.4).

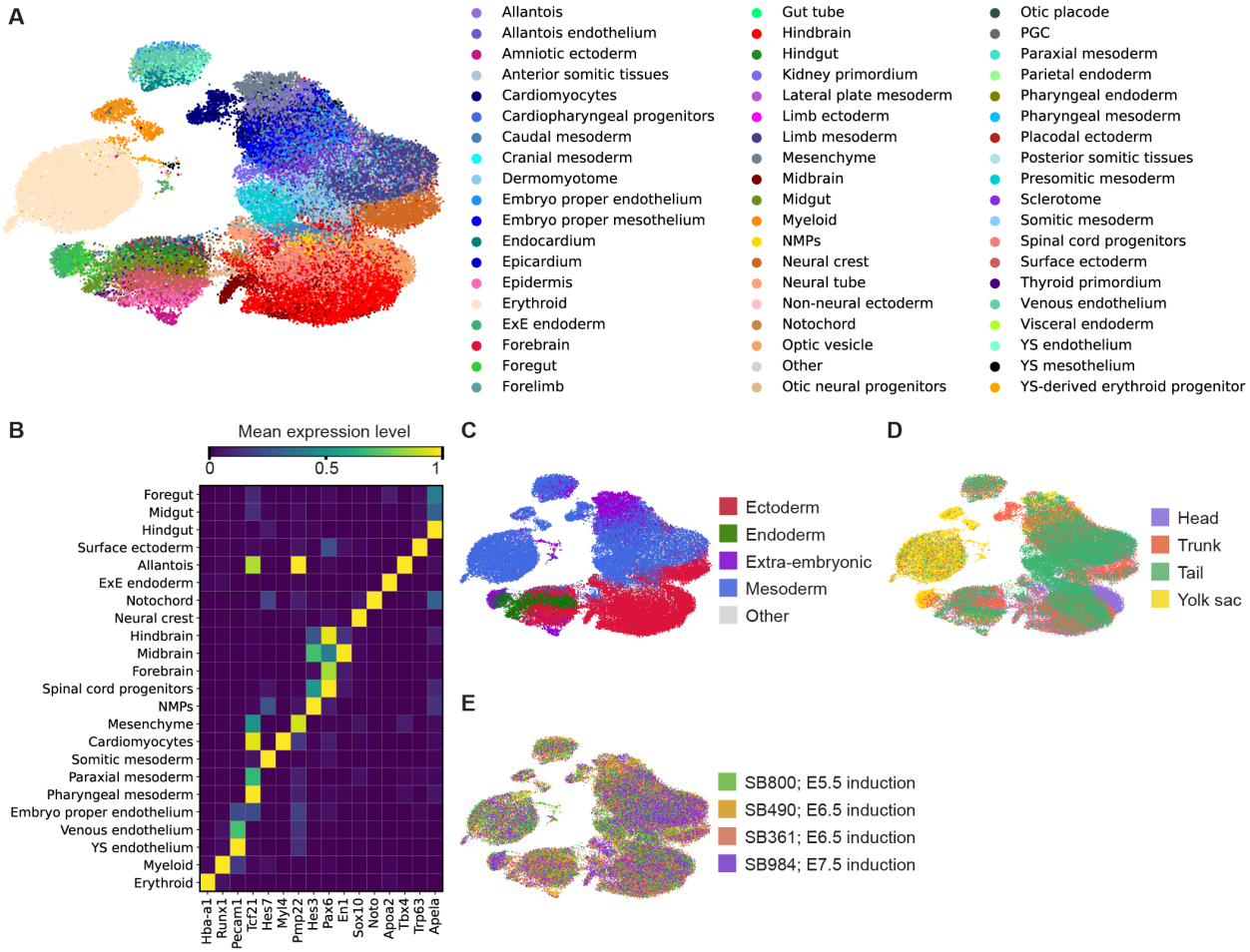

Figure 4.3: We used a mouse embryo atlas to identify cell types in our merged scRNA-seq data. A. UMAP of batch-corrected scRNA-seq data from four embryos sequenced at E9.25, colored by cell type. B. Mean expression of marker genes in select cell types. Marker genes were selected from the published scRNA-seq embryo atlas sampled at earlier developmental times (E6.5-E8.5)[84]. Expression levels are scaled so that values for each gene range between 0 and 1 across the cell types selected. C. UMAP of merged embryo scRNA-seq data, colored by germ layer/tissue type identity. D. UMAP of merged embryo scRNA-seq data, colored by dissected anatomical region of origin. E. UMAP of merged embryo scRNA-seq data, colored by mouse embryo.

However, there were a few notable exceptions- erythroid cells had lower CARLIN capture efficiencies and the extra-embryonic endoderm had unusually low CARLIN editing efficiencies.

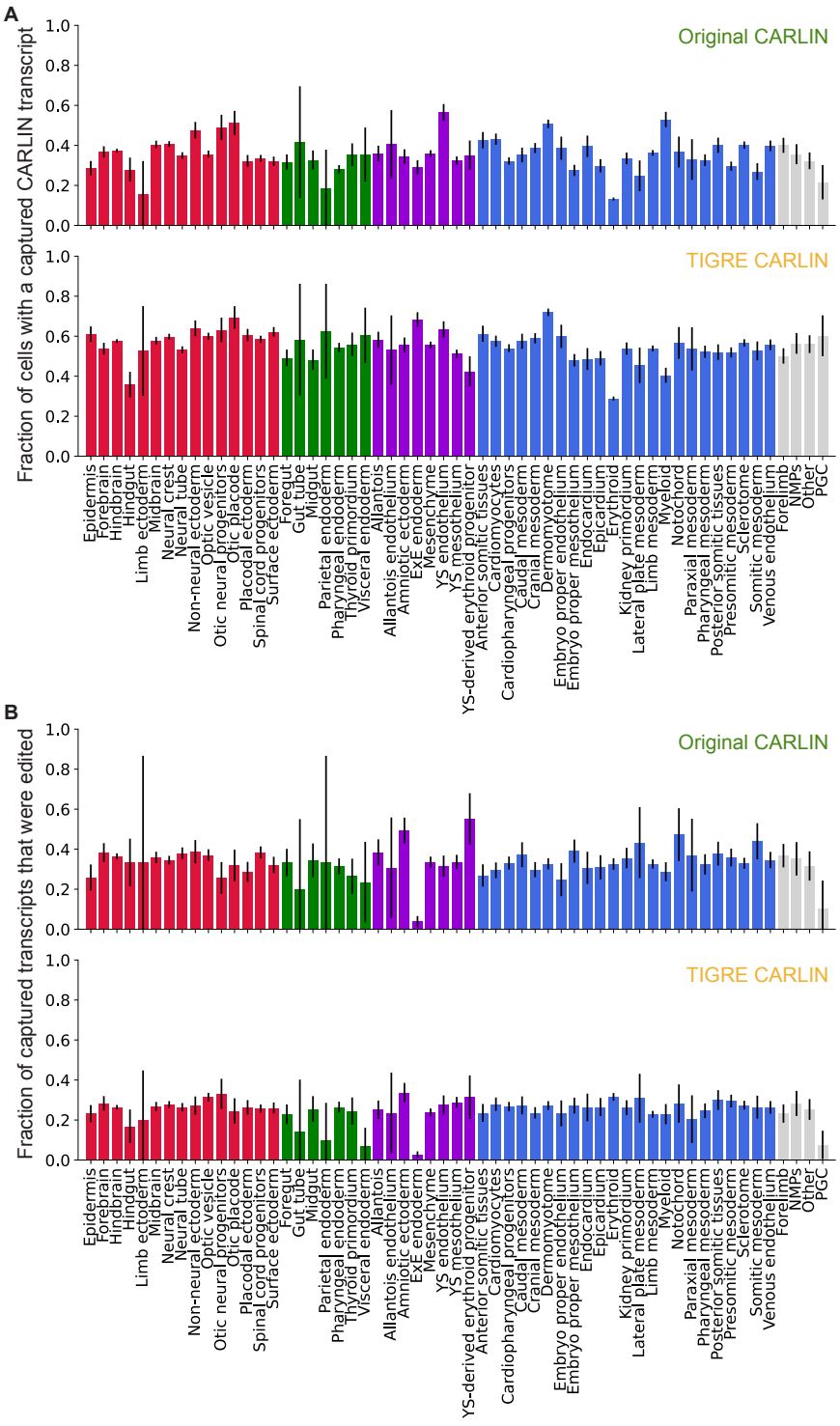

Figure 4.4: CARLIN barcode capture and editing efficiencies are similar across all cell types sampled. A. Capture efficiencies in all four embryos combined, for each cell type identified. B. Editing efficiencies in all four embryos combined, for each cell type identified. For all plots, bars are colored by the germ layer of the cell type, as given in Fig. 4.3C (red, ectoderm; green, endoderm; purple, extra-embryonic; blue, mesoderm; grey, other). Error bars are 95% binomial confidence intervals.

4.2.3 Progenitors that arise earlier in development each produce more descendants and have more diverse cell fates than progenitors that arise later.

The frequency of each merged CARLIN barcode is proportional to the number of surviving descendants produced by the embryonic progenitor cell in which the barcode originated. As expected, embryos in which barcoding occurred earlier had larger clone sizes and fewer unique clones (Fig. 4.5A), since there is a longer time window in which these progenitors can expand. In all embryos, there were CARLIN barcodes that were found in unusually large numbers of cells, which could represent either particularly proliferative progenitors or CARLIN editing events that occurred very early in the 36-hour CRISPR editing window observed in this system.

Larger clonal populations were also more likely to contribute to multiple germ layers. Particularly in the embryo induced at E5.5, barcoded clones that were found in multiple germ layers and/or extra-embryonic tissue were much larger than unipotent clones (Fig. 4.5B). This could be because progenitors that expanded more after barcoding arose earlier in development and were therefore less fate restricted, because progenitors which are more proliferative are also more multipotent, or simply because larger clones are more likely to be found multiple tissue types by chance. This question could be further investigated by resampling the CARLIN barcode data to eliminate the statistical impact of larger clone sizes on observed clone potency.

Additionally, embryos that were induced later had more barcoded clonal populations that only contributed to a single germ layer than embryos that were induced earlier and a smaller fraction of multipotent clones (Fig. 4.5C). Conversely, the embryo induced at E7.5 (SB984) had a much higher fraction of unipotent clones than the embryos induced at E6.5 and E5.5. This effect could again be partially due to the smaller clone sizes in the embryos in which CARLIN editing was induced later. However, this observation is also consistent with the hypothesis that cells become more fate restricted as development progresses, reducing the diversity of cell types each progenitor is capable of creating.

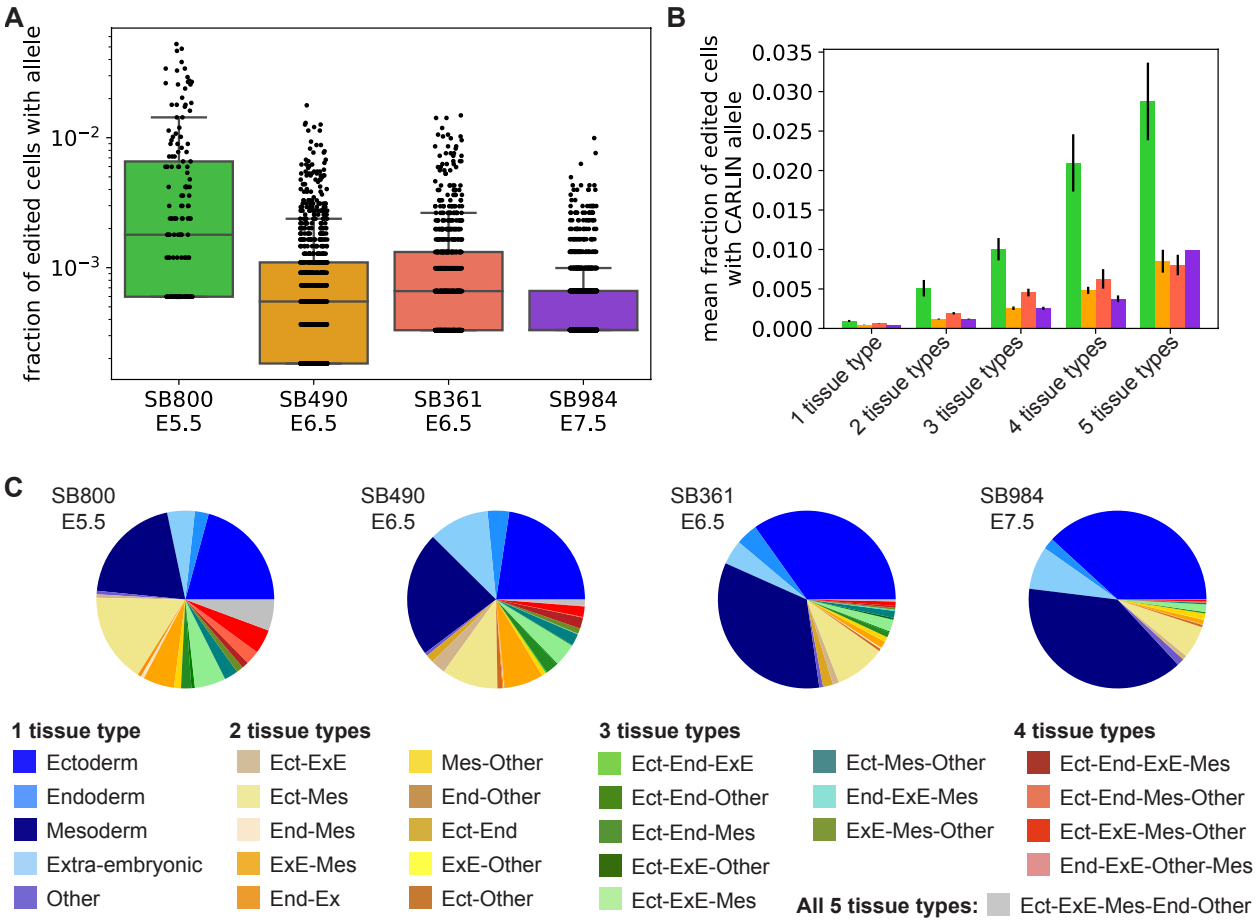

Figure 4.5: Embryos that were induced with dox earlier have larger barcoded clone sizes with more diverse cell fates in each clone. A. Frequency of each CARLIN barcode clone in individual embryos. Boxes show median and interquartile ranges, whiskers are at 1.5x interquartile range. Black dots show all data points. B. Frequency of CARLIN barcodes, according to their presence in cells from different germ layers or tissue types (x-axis) and mouse embryo (colors: green, SB800; orange, SB490; red, SB361; purple, SB984; as in A). CARLIN barcode alleles that appeared in more tissue types represent progenitors with less restricted fate potentials at the time of barcode induction. C. Fraction of CARLIN barcodes from each embryo that contribute to different numbers of germ layers/tissue types. Abbreviations used: Ect, ectoderm; End, endoderm; Mes, mesoderm; ExE, extra-embryonic.

Finally, we found that most cell types sampled had similar numbers of cells per unique CARLIN-barcoded progenitor (Fig. 4.6), suggesting that proliferation rates of individual progenitors during the period from approximately E6.0-E8.0 are relatively similar between progenitors with different cell fates. This observation is consistent with previous results indicating that cells are proliferating exponentially at a uniform rate during early mouse development[91]. If proliferation rates are similar across all progenitor cells in the embryo, then the particularly large clones we observed (Fig. 4.5A) likely represent barcodes that arose earlier during development, rather than individual progenitors that were more proliferative. Notably, each individual progenitor

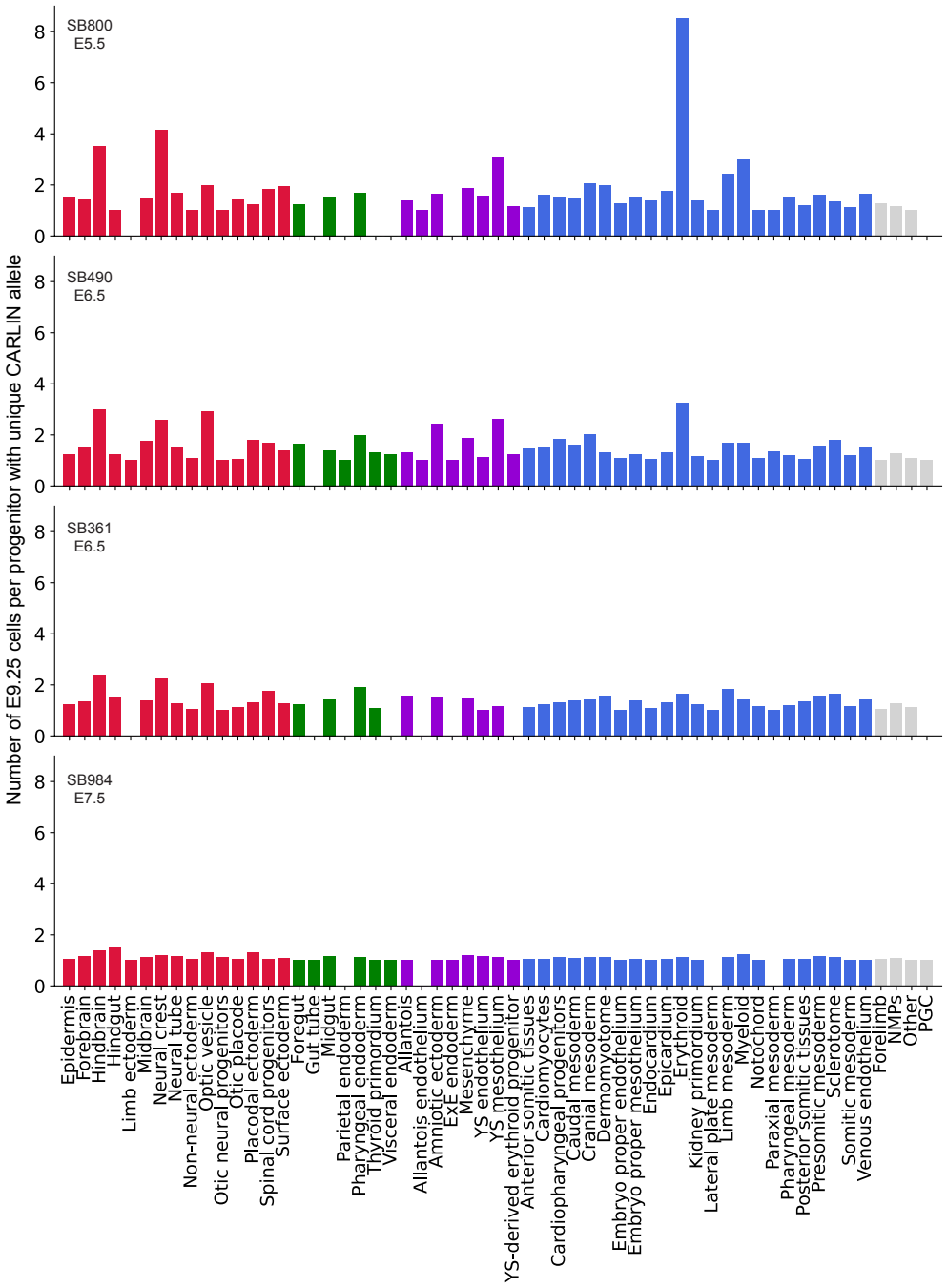

Figure 4.6: Different cell types at E9.25 originate from similar numbers of progenitor cells per differentiated cell. Mean number of sampled cells of each cell type with an edited CARLIN barcode read that arise from each sampled progenitor. Each plot represents data from a single mouse embryo. For all plots, bars are colored by the germ layer of the cell type, as given in Fig. 4.3C (red, ectoderm; green, endoderm; purple, extra-embryonic; blue, mesoderm; grey, other). Error bars are +/- SEM.

did give rise to more erythroid cells than to other cell types, particularly when barcodes were induced earlier in development. This suggests that progenitors (at this stage, mostly yolk sac derived cells) that give rise to erythroid cells may divide more rapidly than other cells in the embryo.

4.2.4 Lineage history is associated transcriptional similarity in embryonic cells sampled after gastrulation.

Cells with similar transcriptional phenotypes in the developing embryo are thought to have similar lineage histories and possibly share a recent common ancestor. To investigate the relationship between cell lineage histories and phenotypic similarity, we computed the correlations between the CARLIN barcode frequency distributions of different cell types (Fig. 4.7). If the barcode frequencies were highly correlated between two different cell types, the two cell types likely shared a similar set of progenitors at the time of CARLIN barcode induction and therefore were closely related in terms of their lineage histories. We found that ectodermal cell types had higher lineage correlations with other ectodermal cell types in embryos induced at E6.5 (Fig. 4.7B-C). Mesodermal cell types may have had shared lineage histories in these embryos as well. However, similar patterns are harder to discern in the embryo induced at E5.5 (Fig. 4.7A), since there is a high degree of barcode sharing in general, likely due to the larger average clone size. Similarly, the overall degree of barcode sharing between cell types was too low in the embryo induced at the latest timepoint to observe any obvious patterns (Fig. 4.7D). Taken together, these data suggest that progenitors that existed in the embryo at approximately E6.5-E7.5 exhibited some degree of fate restriction, at least at the level of germ layers, and that the descendants of each progenitor had similar transcriptional phenotypes.

The lineage correlation analysis was designed to determine whether cells that were transcriptionally similar also had similar barcode sharing patterns. A complementary approach to investigating the relationship between transcriptional similarity and shared lineage history is to determine whether groups of barcodes that had similar tissue contribution patterns were found in cells that had similar transcriptional phenotypes.

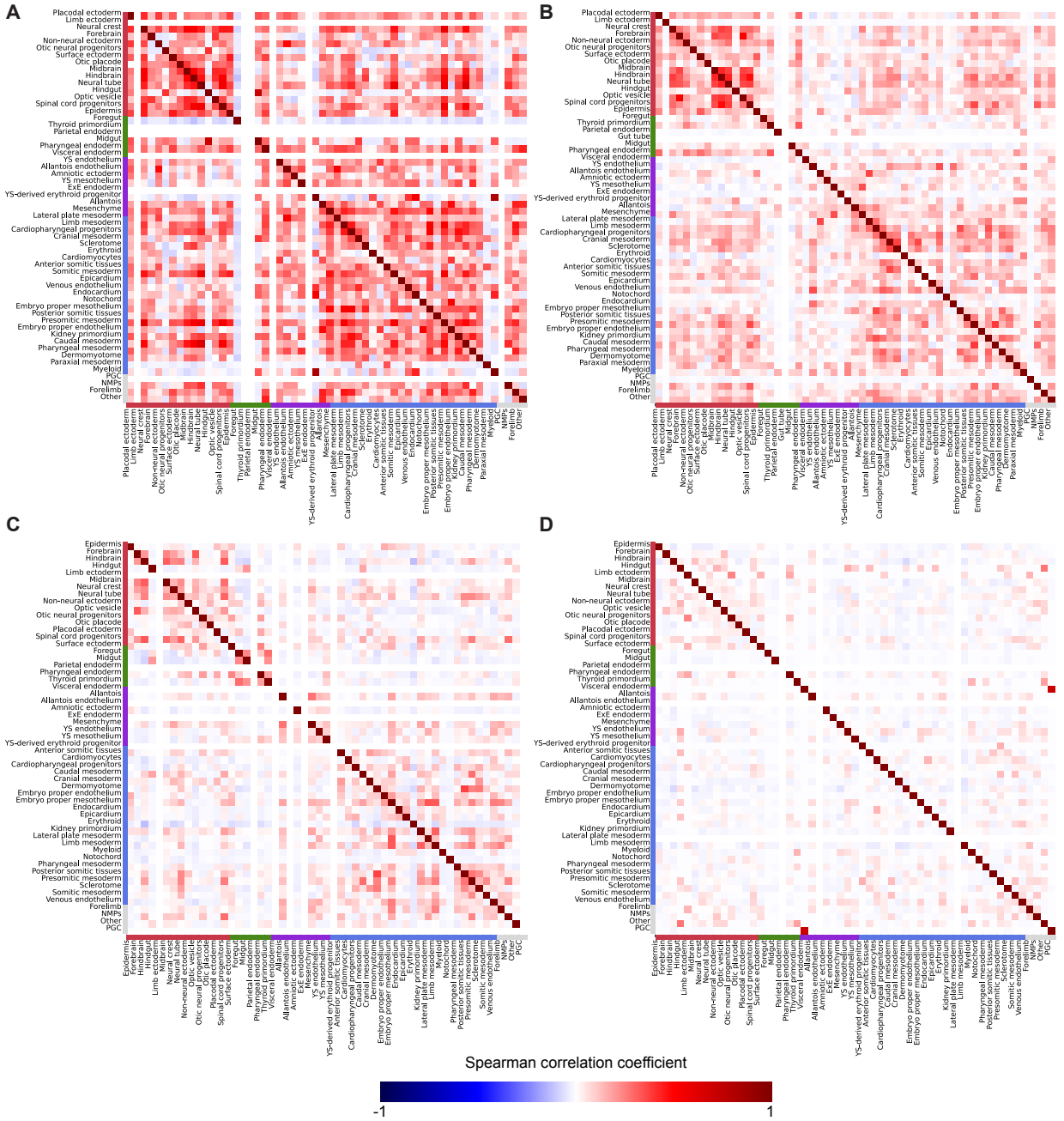

Figure 4.7: Cell types that originate from the same germ layer have similar CARLIN barcode patterns. A-D. Spearman correlation coefficients of the CARLIN barcode frequencies for each cell type for SB800 (A), SB490 (B), SB361 (C), and SB984 (D). Axes are ordered by germ layer, which is indicated by the colored bars at the left and bottom edges of each heatmap (red, ectoderm; green, endoderm; purple, extra-embryonic; blue, mesoderm; grey, other).

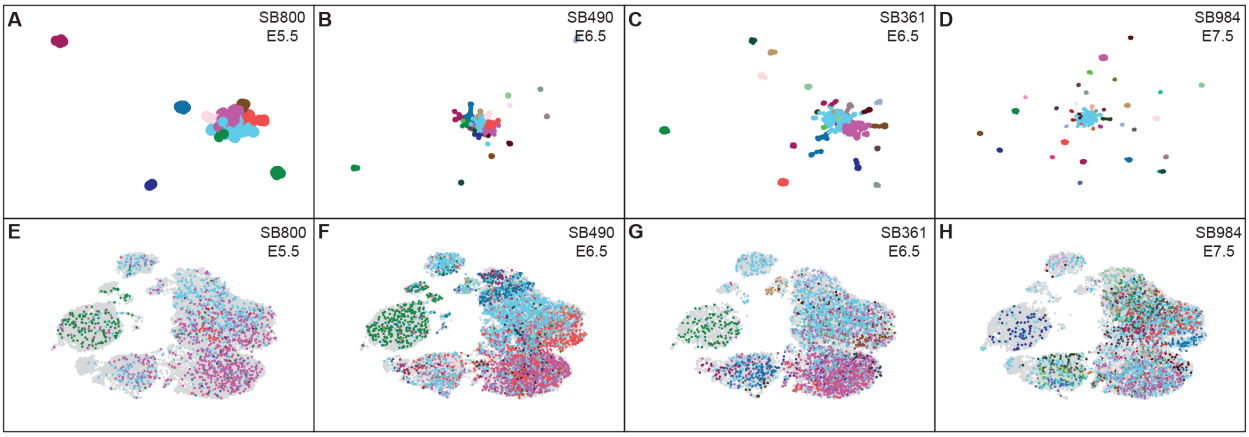

Figure 4.8: CARLIN barcodes separate into clusters based on their contribution to different cell types. A-D. UMAPs of CARLIN allele cell type frequency data for SB800 (A), SB490 (B), SB361 (C), and SB984 (D). Each point is a CARLIN allele and is colored by CARLIN allele cluster membership. Clusters represent CARLIN alleles that have similar contribution patterns to E9.25 cell types. E-H. UMAPs of transcriptome data for SB800 (E), SB490 (F), SB361 (G), and SB984 (H). Each point is a single cell, colored by to which CARLIN barcode cluster the CARLIN transcript sequenced for that cell belongs. Colors are the same as the corresponding plots in A-D. Cells without a CARLIN barcode are shown in grey.

To perform this analysis, we constructed a cell type contribution distribution for each barcoded clone and used this vector to cluster the barcodes according to their cell type contributions in each embryo. We visualized these relationships between CARLIN barcodes using UMAP (Methods). In this analysis, CARLIN barcodes that were within the same cluster and located closer together on the UMAP plots had more similar cell type contribution profiles. We found that these barcode clusters were smaller and more distinct in embryos induced later in development, which is consistent with the observations that clones derived from later progenitors become smaller and more fate restricted (Fig. 4.8A-D).

Importantly, we also observed that CARLIN barcodes found in the same cluster were generally found in specific regions of the UMAPs using the scRNA-seq data (Fig. 4.8E-H), suggesting that cells that have similar transcriptomic profiles also are part of CARLIN barcoded clones that have similar contribution patterns in the embryo. However, there are some cell types that appear quite transcriptionally homogeneous which include cells from multiple different barcode clusters. For example, the endoderm/surface ectoderm cluster at the lower left of each UMAP in Fig. 4.8E-H contains cells belonging to multiple barcode clusters, some of which also contribute to the mesoderm and/or neural tissue, suggesting there might be distinct populations

of multipotent progenitors that give rise to these cells. Furthermore, in embryo SB490, there are multiple barcode clusters contributing to the endothelial cell population in the upper left of the UMAP (Fig. 4.8F), suggesting that there might be multiple embryonic origins for endothelial cells.

4.2.5 Single-cell lineage information reveals that endothelial cells have multiple origins.

To further investigate the origins of endothelial cells, we noted that while all endothelial cells had similar transcriptional phenotypes and formed a single cluster in the transcriptomic data, there were some important subgroups of endothelial cells we identified. First, our mapping of the scRNA-seq data to the reference atlas separated our endothelial cells into four transcriptionally-defined subtypes (Fig. 4.9A). Interestingly, the transcriptionally-defined yolk sac endothelium corresponded with the endothelial cells that were found in the yolk sac during dissection of these embryos for sequencing (Fig. 4.9B). Endothelial cells that originated from the three other anatomical regions isolated from the main part of the embryo (head, trunk, and tail) had similar transcriptional profiles.

However, endothelial cells isolated from different parts of the embryo had different progenitor cells of origin at E6.5-E7.5. In particular, endothelial cells isolated from the head were more likely to share barcodes with the cranial mesoderm than endothelial cells from other regions (Fig. 4.9C-D). Likewise, the endothelial cells from the yolk sac share more CARLIN barcodes with other yolk sac tissues than endothelial cells found in other parts of the embryo (Fig. 4.9C-D). These results suggest that endothelial cells arise *in situ* from separate local progenitor pools that exist at approximately E6.5-E7.5, which subsequently converge to create transcriptionally similar endothelial cells by the time of sampling at E9.25.

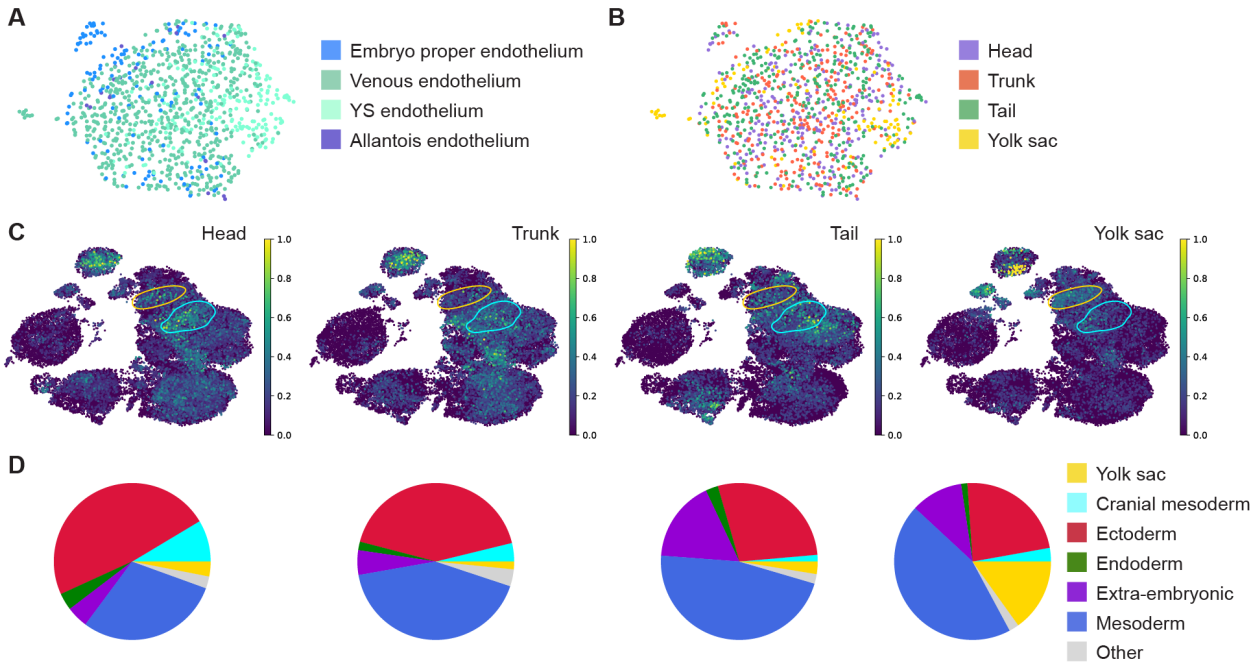

Figure 4.9: Endothelial cells in different regions of the embryo have different ancestors. A-B. UMAP of endothelial cell types in merged data from E6.5-induced embryos (SB490 and SB800), colored by endothelial cell type (A) or anatomical region from where the cell originated (B). C. UMAP of merged E6.5 embryo data showing cells which share CARLIN barcodes with endothelial cells found in the head, tail, trunk, or yolk sac. Colors indicate the normalized density of cells which share a barcode with the queried endothelial cell type, averaged over the 15 nearest neighbors of each cell. Contours show regions of the UMAP where cranial mesoderm (cyan) or yolk sac cells (yellow) are located (defined as regions within which 60% of cells of that type are located). D. Fraction of CARLIN barcodes found in endothelial cells in each anatomical region (head, trunk, tail, yolk sac from left to right) that are shared with specific cell types or germ layers. The mesoderm and extra-embryonic categories do not include cranial mesoderm or yolk-sac-derived cells, respectively.

4.3 Discussion

Here, we investigated the dynamics of mouse embryonic development using an expressed CRISPR-Cas9 cell barcoding system and scRNA-seq. We found that cell fate restriction occurs gradually during gastrulation (approximately E6.0-E7.5), which is reflected in a correspondence between transcriptional similarity and barcode sharing between individual cells in the sampled E9.25 embryo. Furthermore, our results indicated that there are region-specific lineage histories for endothelial cells, suggesting that there are multiple local multipotent progenitor populations that create a population of endothelial cells that is relatively transcriptionally homogeneous. These findings are consistent with previous studies showing that endothelial cells can arise *in situ* from mesodermal progenitors in the allantois and yolk sac[84,92], and point to the existence of similar

local progenitor pools in different regions of the embryo proper.

Our CARLIN barcode data are a direct measurement of lineage relationships and clonal histories in the developing mouse embryo. Therefore, these data could be compared to existing qualitative and quantitative models of mouse embryonic development to identify inconsistencies which could indicate weaknesses in our current understanding of mouse development. To do this, we plan to compare our combined phenotypic and lineage data to predictions from quantitative computational models of embryonic development created from scRNA-seq data[87]. Our collaborators in the Göttgens Lab have built an optimal-transport model[93] using their embryo atlas data that computes predicted developmental trajectories of individual embryonic progenitors over time. These developmental trajectories can be directly compared to our CARLIN data to assess the accuracy of the optimal transport model, adjust the existing optimal transport model to create a more accurate description of mouse embryonic development, and to infer previous molecular states of clonal populations marked by individual CARLIN barcodes. In general, the results from these two approaches (scRNA-seq developmental time series modeling and ground truth lineage data from barcoding systems) complement each other well, since the scRNA-seq time series results provide a draft model to which the sparse barcoding data can be fitted, and the barcoding data provides orthogonal information that can be used to test the assumptions of the computational model. Consequently, new computational tools that leverage both lineage tracing and cell state data from the same system can offer additional insights into developmental dynamics that are not accessible from either data source alone[94].

Currently, there are a few features of the CARLIN barcoding system that limit its ability to derive accurate cell lineage histories. First, while we can detect CARLIN transcripts in the majority of cells using scRNA-seq, most of these transcripts are not edited and therefore many cells do not contain a barcode that can be used for identification of clonal populations. Changing the dose or route of doxycycline administration may improve the editing efficiency, but may also change the editing kinetics unfavorably. Second, the length of the CRISPR editing period *in vivo* makes it difficult to precisely estimate when a given progenitor cell

was barcoded, and means that within a single experiment the barcoded clonal populations all originated at different times within an approximately 36 hour long window. Unfortunately, the lower bound of the length of this editing period is set by pharmacokinetics of doxycycline within the mouse, so this period probably cannot be shortened further. However, lengthening the period of time in which editing occurs by using different dox dosing strategies (e.g., multiple doses, exposure through drinking water) could enable us to track developmental dynamics over extended periods of time. These data would allow us to reconstruct single cell barcode phylogenies and use these trees to estimate rates of dynamic behavior[89,95].

In summary, our work using transcribed inducible CRISPR barcodes and scRNA-seq revealed information about mouse embryogenesis previously unattainable through investigation of static transcriptional profiling alone. While similar barcoding systems cannot be used in humans to collect single-cell lineage information during human development, methods to extract lineage information using somatic mutations either in the nuclear genome or the mitochondrial genome[13] may provide sufficient information to reconstruct major cell fate decisions. Indeed, this approach has already been used to trace developmental trajectories in the blood system[96] and the brain[97] in humans, suggesting the somatic mutation rate is high enough to reliably reconstruct lineage histories during early development. Combining this lineage information with precise single-cell phenotype information, as we did in the mouse embryo, could result in a similarly high resolution map of human developmental dynamics.

4.4 Methods

4.4.1 Experimental procedures

Timed pregnancies resulting in transgenic mouse embryos with the genetic architecture delineated in Fig. 4.1A were set up. At E5.5, E6.5, or E7.5, pregnant females were injected retroorbitally with 25 μg/g of a 10 mg/mL solution of doxycycline to induce barcode editing. At E9.25, pregnant females were sacrificed and embryos

were dissected into head, trunk, tail, and yolk sac regions and dissociated. 10X scRNA-seq encapsulation, library preparation, and sequencing was performed on the single-cell suspensions as published[80], with part of the cDNA library used for whole transcriptome analysis and another part reserved for targeted CARLIN transcript amplification and sequencing using a nested PCR approach[80]. 10,000 cells from each embryo region were sequenced for each embryo.

4.4.2 CARLIN transcript identification and alignment

Identification and alignment of CARLIN transcript reads were performed as published[80], using both original CARLIN and TIGRE CARLIN primer sequences to identify CARLIN reads.

4.4.3 CARLIN barcode filtering and merging

To remove highly frequent barcodes that were likely to be independently generated multiple times within a single embryo, we created and analyzed original CARLIN and TIGRE CARLIN granulocyte allele banks as previously described[80]. CARLIN alleles that occurred very frequently in these banks and had a greater than 5% chance of being independently generated more than once in each embryo at the time of barcode induction were discarded from our scRNA-seq dataset.

To merge the OC and TC alleles into a single, clonal identifier for each cell, we first identified all cells without a clonal OC or clonal TC allele call as having no valid barcode information. Cells with both an OC and a TC allele call, with at least one of the two being a valid clonal OC or TC allele call after allele bank filtering, were assigned the barcode identifier ([OC ALLELE], [TC ALLELE]). Cells with only an OC or only a TC transcript read were assigned an existing paired barcode identifier from the set of valid barcodes that included the corresponding sequenced allele call, randomly with probability proportional to the frequency of each of these barcodes. If no other valid barcodes with the OC or TC allele existed, cells were assigned the barcode ([OC ALLELE], none) or (none, [TC ALLELE]), respectively.

4.4.4 scRNA-seq data preprocessing and visualization

Transcriptome reads were aligned to mm10 using CellRanger v6.1.0 and processed in scanpy. Scrublet[98] was used to identify cell doublets, and cells with doublet score > 0.15 were discarded. Cells with fewer than 200 genes or more than 1000 genes expressed were discarded, along with cells with more than 5000 UMIs. Cells with mitochondrial read fraction greater than 5% or less than 0.05% were also discarded. After total count normalization, data from all four embryos were merged and batch corrected using Harmony[90]. These batch-corrected coordinates were used to generate all UMAP plots.

4.4.5 Embryo cell type identification

Cell types were identified in our scRNA-seq data by mapping the data onto an annotated E9.25 reference atlas (unpublished, from Ivan Imaz-Rosshandler and Bertie Göttgens). The atlas had been previously batch corrected using fastMNN[99]. We centered and scaled the batch-corrected atlas data and our total-count normalized transcriptome data so that each gene had mean 0 and variance 1, and then batch corrected the combined CARLIN embryo data with the atlas using Harmony. Cell type labels were directly transferred from the nearest neighbor in the atlas to each cell in the CARLIN dataset.

4.4.6 Clustering of CARLIN barcode patterns

For each OC/TC merged CARLIN barcode in each embryo, we computed the fraction of cells with that barcode that was in each of the 57 cell types. These cell type frequency vectors for each CARLIN barcode were used as a data matrix for clustering and UMAP visualization in scanpy. Using scanpy, the neighborhood graph and UMAP coordinates were computed for the barcodes in each embryo. Louvain clustering was used to identify groups of barcodes with similar cell type contribution profiles.

5

Conclusion

During my PhD, I worked on multiple projects with the goal of characterizing differentiation and/or cell division within intact multicellular tissues. The capacity to quantitatively study these behaviors is especially important for understanding biological systems in which individual cell phenotypes change rapidly, such as in regenerative tissues (e.g., hematopoiesis, see Chapters 1-3) or in development (e.g., mammalian embryogenesis, see Chapter 4). Not only are these measurements important for describing unperturbed biological systems, but they are also valuable to determine how specific conditions affect these dynamic behaviors (e.g., malignancy, mutations, exposure to drugs or environmental stimuli). There are some general strategies and themes that emerge from the work I did during my PhD regarding quantitative measurement of cell dynamics.

5.1 Common themes and future outlook

Since intact mammalian tissues often are not amenable to direct longitudinal observation of individual cell-level behaviors, direct measurement of changes in individual cell phenotype (e.g., differentiation) often rely on simultaneously measuring current cell state and estimating previous cell state in the same sample. This can be done either by collecting additional naturally-occurring phenotypic information from each cell to learn about the past, or by designing an experimental system to specifically record useful information about cell histories.

5.1.1 Inferring molecular or cell division histories in unaltered tissues

Some biological processes naturally occur at longer timescales or cause permanent molecular changes and therefore contain information about the previous state of a cell. By measuring both current cell state and this naturally-occurring historical record in the same population of cells, we can directly relate prior molecular events to current phenotype. For example, the pattern of somatic mutations in the human genome provide a (sparse) record of past transcriptional state over the lifetime of the organism, and can be used to infer the cell-of-origin of tumors years after they were initiated [100,101]. This general approach is particularly important when studying human tissues, since the capacity for experimental manipulations using human samples is often limited. For instance, in Chapter 1, we developed and implemented a strategy to measure both genotype and transcriptional phenotype information at a single cell level. While we used this method in our work mostly to discover transcriptional differences between cells with and without a particular driver mutation in myeloproliferative neoplasm patients, the general approaches we developed to detect naturally-occurring somatic mutations in scRNA-seq data could be used to identify lineage relationships in developing human tissues or human cancer samples.

5.1.2 Experimental systems for recording past events

New experimental systems designed to record prior molecular information can enable more efficient and accurate observation of cell histories than would be possible from unmodified tissues. For example, multiple groups have developed single-cell barcoding systems to track cell lineage histories [95,89,102] similar to the one we developed to study mammalian embryogenesis in Chapter 4. These barcoding systems have predictable and tunable barcode generation kinetics that often result in higher-resolution lineage maps than could be reconstructed from somatic mutation data alone. Other systems can record the time at which specific events occurred in the past, such as differentiation of a stem cell (e.g., HSC differentiation, as studied in Chapter 2), exposure to a particular stimulus [103], or even prior interactions with other cells [104]. However, these approaches have their own challenges. First, a bespoke system must often be designed for each new event to be recorded, each biological context, etc, and tracking multiple events in the same cell can often be difficult. Furthermore, estimating the precise time at which the molecular and cellular event in question occurred in the past can be difficult in these systems, since in many cases these historical events are recorded with a single permanent change (e.g., CRISPR editing, Cre recombination). Therefore, developing better techniques to measure time in cells and record information in a time-resolved manner, such as systems based on degradation of fluorescent proteins [58] (also see Chapter 2) could substantially improve our ability to measure differentiation and cell division rates.

In summary, our work developing and implementing quantitative measurement techniques to characterize dynamic cell behaviors *in situ* could be used in the future to study regeneration, development, and malignancy in humans and/or model organisms. These particular biological processes are fundamentally defined by changes in cell state (e.g., differentiation, cell division, and cell death), so directly investigating dynamic behavior on a single-cell level is required to fully understand them.

References

[1] Minussi, D. C. *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**, 302–308 (2021).

[2] Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

[3] Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015). URL https://www.sciencedirect.com/science/article/pii/S0092867415005498.

[4] Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **23**, 2126–2135 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3847781/.

[5] Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

[6] Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4123637/.

[7] Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835–849.e21 (2019).

[8] Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018). URL https://www.nature.com/articles/s41586-018-0497-0.

[9] Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).

[10] Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

[11] Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* **36**, 442–450 (2018).

[12] Gaiti, F. *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).

[13] Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325–1339.e22 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419300558.

[14] Oren, Y. *et al.* Cycling cancer persister cells arise from lineages with distinct programs. *Nature* **596**, 576–582 (2021). URL https://www.nature.com/articles/s41586-021-03796-6.

[15] Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019). URL https://www.nature.com/articles/s41586-019-1367-0.

[16] Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).

[17] Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020). URL https://www.sciencedirect.com/science/article/pii/S0092867420312538.

[18] Van Egeren, D. *et al.* Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms. *Cell Stem Cell* **28**, 514–523.e9 (2021). URL https://www.cell.com/cell-stem-cell/abstract/S1934-5909(21)00051-5.

[19] Van Egeren, D. *et al.* Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in JAK2-mutant myeloproliferative neoplasms (2020). URL https://www.biorxiv.org/content/10.1101/2020.08.24.265058v1. BioRxiv preprint.

[20] Shallis, R. M. *et al.* Epidemiology of the classical myeloproliferative neoplasms: The four corners of an expansive and complex map. *Blood Reviews* **42**, 100706 (2020). URL https://www.sciencedirect.com/science/article/pii/S0268960X20300564.

[21] Spivak, J. L. Myeloproliferative Neoplasms. *N Engl J Med* **376**, 2168–2181 (2017).

[22] Lundberg, P. *et al.* Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood* **123**, 2220–2228 (2014).

[23] Scott, L. M. *et al.* JAK2 Exon 12 Mutations in Polycythemia Vera and Idiopathic Erythrocytosis. *New England Journal of Medicine* **356**, 459–468 (2007). URL https://doi.org/10.1056/NEJMoa065202.

[24] Skoda, R. C., Duek, A. & Grisouard, J. Pathogenesis of myeloproliferative neoplasms. *Experimental Hematology* **43**, 599–608 (2015). URL https://www.sciencedirect.com/science/article/pii/S0301472X15002155.

[25] Chen, E. *et al.* Distinct Clinical Phenotypes Associated with JAK2V617F Reflect Differential STAT1 Signaling. *Cancer Cell* **18**, 524–535 (2010). URL https://linkinghub.elsevier.com/retrieve/pii/S1535610810004198.

[26] Lu, X., Huang, L. J.-S. & Lodish, H. F. Dimerization by a Cytokine Receptor Is Necessary for Constitutive Activation of JAK2V617F. *Journal of Biological Chemistry* **283**, 5258–5266 (2008). URL https://www.jbc.org/article/S0021-9258(20)57245-1/abstract.

[27] Harrison, C. *et al.* JAK Inhibition with Ruxolitinib versus Best Available Therapy for Myelofibrosis. *New England Journal of Medicine* **366**, 787–798 (2012). URL https://doi.org/10.1056/NEJMoa1110556.

[28] Vannucchi, A. M. *et al.* Ruxolitinib versus standard therapy for the treatment of polycythemia vera. *N Engl J Med* **372**, 426–435 (2015).

[29] Li, J., Kent, D. G., Chen, E. & Green, A. R. Mouse models of myeloproliferative neoplasms: JAK of all grades. *Dis Model Mech* **4**, 311–317 (2011). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3097453/.

[30] Dusa, A. *et al.* Substitution of Pseudokinase Domain Residue Val-617 by Large Non-polar Amino Acids Causes Activation of JAK2. *J. Biol. Chem.* **283**, 12941–12948 (2008). URL http://www.jbc.org/content/283/19/12941.

[31] Kluk, M. J. *et al.* Validation and Implementation of a Custom Next-Generation Sequencing Clinical Assay for Hematologic Malignancies. *The Journal of Molecular Diagnostics* **18**, 507–515 (2016). URL https://www.jmdjournal.org/article/S1525-1578(16)30024-1/abstract.

[32] Christova, R. *et al.* P-STAT1 mediates higher-order chromatin remodelling of the human MHC in response to IFNγ. *Journal of Cell Science* **120**, 3262–3270 (2007). URL https://doi.org/10.1242/jcs.012328.

[33] Brutkiewicz, R. R. Cell Signaling Pathways That Regulate Antigen Presentation. *The Journal of Immunology* **197**, 2971–2979 (2016). URL https://www.jimmunol.org/content/197/8/2971.

[34] Wang, H., Leng, Y. & Gong, Y. Bone Marrow Fat and Hematopoiesis. *Frontiers in Endocrinology* **9**, 694 (2018). URL https://www.frontiersin.org/article/10.3389/fendo.2018.00694.

[35] Umemoto, Y. *et al.* Leptin Stimulates the Proliferation of Murine Myelocytic and Primitive Hematopoietic Progenitor Cells. *Blood* **90**, 3438–3443 (1997). URL https://doi.org/10.1182/blood.V90.9.3438.

[36] Bennett, B. D. *et al.* A role for leptin and its cognate receptor in hematopoiesis. *Current Biology* **6**, 1170–1180 (1996). URL https://www.sciencedirect.com/science/article/pii/S0960982202706842.

[37] Trinh, T. *et al.* Leptin receptor, a surface marker for a subset of highly engrafting long-term functional hematopoietic stem cells. *Leukemia* **35**, 2064–2075 (2021). URL https://www.nature.com/articles/s41375-020-01079-z.

[38] Lu, Y.-C. *et al.* The Molecular Signature of Megakaryocyte-Erythroid Progenitors Reveals a Role for the Cell Cycle in Fate Specification. *Cell Rep* **25**, 2083–2093.e4 (2018). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6336197/.

[39] Le Goff, S. *et al.* p53 activation during ribosome biogenesis regulates normal erythroid differentiation. *Blood* **137**, 89–102 (2021). URL https://doi.org/10.1182/blood.2019003439.

[40] Khajuria, R. K. *et al.* Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis. *Cell* **173**, 90–103.e19 (2018). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5866246/.

[41] Athanasiadis, E. I. *et al.* Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nature Communications* **8**, 2045 (2017). URL https://www.nature.com/articles/s41467-017-02305-6.

[42] Caron, M. *et al.* Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity. *Scientific Reports* **10**, 8079 (2020). URL https://www.nature.com/articles/s41598-020-64929-x.

[43] Wong, K. L. *et al.* Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets. *Blood* **118**, e16–e31 (2011). URL https://doi.org/10.1182/blood-2010-12-326355.

[44] Ziegler-Heitbrock, L. & Hofer, T. P. J. Toward a Refined Definition of Monocyte Subsets. *Front Immunol* **4**, 23 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3562996/.

[45] Kapellos, T. S. *et al.* Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases. *Front. Immunol.* **10** (2019). URL https://www.frontiersin.org/articles/10.3389/fimmu.2019.02035/full.

[46] Barone, M. *et al.* The role of circulating monocytes and JAK inhibition in the infectious-driven inflammatory response of myelofibrosis. *Oncoimmunology* **9**, 1782575 (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7458658/.

[47] Maekawa, T. *et al.* Increased SLAMF7high monocytes in myelofibrosis patients harboring JAK2V617F provide a therapeutic target of elotuzumab. *Blood* **134**, 814–825 (2019). URL https://ashpublications.org/blood/article/134/10/814/260635/Increased-SLAMF7high-monocytes-in-myelofibrosis.

[48] Tong, J. *et al.* Hematopoietic Stem Cell Heterogeneity Is Linked to the Initiation and Therapeutic Response of Myeloproliferative Neoplasms. *Cell Stem Cell* **28**, 502–513.e6 (2021). URL https://www.cell.com/cell-stem-cell/abstract/S1934-5909(21)00018-7.

[49] *Single-Cell Multi-Omics Reveals Distinct Paths to Survival of Admixed BTK C481 Mutant Vs. Wild-Type Cells in Clinically Progressing Chronic Lymphocytic Leukemia* (American Society of Hematology, 2020). URL https://ash.confex.com/ash/2020/webprogram/Paper138374.html.

[50] Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019). URL https://www.sciencedirect.com/science/article/pii/S0092867419305598.

[51] Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542–546 (2015). URL https://www.nature.com/articles/nature14242.

[52] Barile, M. *et al.* Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation (2020). URL https://www.biorxiv.org/content/10.1101/2020.08.06.239186v1. BioRxiv preprint.

[53] Foudi, A. *et al.* Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nature Biotechnology* **27**, 84–90 (2009). URL http://www.nature.com/articles/nbt.1517.

[54] Buczacki, S. J. A. *et al.* Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature* **495**, 65–69 (2013). URL https://www.nature.com/articles/nature11965.

[55] Toyama, B. H. *et al.* Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell* **154**, 971–982 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3788602/.

[56] Morcos, M. N. *et al.* Continuous mitotic activity of primitive hematopoietic stem cells in adult mice. *Journal of Experimental Medicine* **217** (2020). URL https://doi.org/10.1084/jem.20191284.

[57] Grinenko, T. *et al.* Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice. *Nat Commun* **9**, 1898 (2018). URL https://www.nature.com/articles/s41467-018-04188-7.

[58] Gehart, H. *et al.* Identification of Enteroendocrine Regulators by Real-Time Single-Cell Differentiation Mapping. *Cell* **176**, 1158–1173.e16 (2019).

[59] Corish, P. & Tyler-Smith, C. Attenuation of green fluorescent protein half-life in mammalian cells. *Protein Eng Des Sel* **12**, 1035–1040 (1999). URL https://academic.oup.com/peds/article/12/12/1035/1567742.

[60] Upadhaya, S. *et al.* Kinetics of adult hematopoietic stem cell differentiation in vivo. *Journal of Experimental Medicine* jem.20180136 (2018). URL http://jem.rupress.org/content/early/2018/10/04/jem.20180136.

[61] Madisen, L. *et al.* Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. *Neuron* **85**, 942–958 (2015). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4365051/.

[62] Boyer, S. W., Schroeder, A. V., Smith-Berdan, S. & Forsberg, E. C. All hematopoietic cells develop from hematopoietic stem cells through Flk2/Flt3-positive progenitor cells. *Cell Stem Cell* **9**, 64–73 (2011). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103692/.

[63] Tusi, B. K. *et al.* Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).

[64] Bernitz, J. M., Kim, H. S., MacArthur, B., Sieburg, H. & Moore, K. Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. *Cell* **167**, 1296–1309.e10 (2016). URL https://www.cell.com/cell/abstract/S0092-8674(16)31405-2.

[65] Gonda, D. K. *et al.* Universality and structure of the N-end rule. *J. Biol. Chem.* **264**, 16700–16712 (1989).

[66] Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).

[67] Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).

[68] Sawai, C. M. *et al.* Hematopoietic Stem Cells Are the Major Source of Multilineage Hematopoiesis in Adult Animals. *Immunity* **45**, 597–609 (2016). URL https://www.cell.com/immunity/abstract/S1074-7613(16)30308-9.

[69] Säwen, P. *et al.* Murine HSCs contribute actively to native hematopoiesis but with reduced differentiation capacity upon aging. *eLife* **7**, e41258 (2018). URL https://doi.org/10.7554/eLife.41258.

[70] Chapple, R. H. *et al.* Lineage tracing of murine adult hematopoietic stem cells reveals active contribution to steady-state hematopoiesis. *Blood Adv* **2**, 1220–1228 (2018).

[71] Beerman, I. *et al.* Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A* **107**, 5465–5470 (2010). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2851806/.

[72] Grover, A. *et al.* Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat Commun* **7**, 11075 (2016). URL https://www.nature.com/articles/ncomms11075.

[73] Zhang, Y. *et al.* PR-domain–containing Mds1-Evi1 is critical for long-term hematopoietic stem cell function. *Blood* **118**, 3853–3861 (2011). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3193263/.

[74] Zhang, Y., Gao, S., Xia, J. & Liu, F. Hematopoietic Hierarchy – An Updated Roadmap. *Trends in Cell Biology* **28**, 976–986 (2018). URL https://www.cell.com/trends/cell-biology/abstract/S0962-8924(18)30102-8.

[75] Morcos, M. N. F. *et al.* Hematopoietic lineages diverge within the stem cell compartment (2020). URL https://www.biorxiv.org/content/10.1101/2020.08.21.261552v1. BioRxiv preprint.

[76] Jensen, C. T., Strid, T. & Sigvardsson, M. Exploring the multifaceted nature of the common lymphoid progenitor compartment. *Curr Opin Immunol* **39**, 121–126 (2016).

[77] Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367** (2020).

[78] Baldridge, M. T., King, K. Y., Boles, N. C., Weksberg, D. C. & Goodell, M. A. Quiescent haematopoietic stem cells are activated by IFN-γ in response to chronic infection. *Nature* **465**, 793–797 (2010). URL https://www.nature.com/articles/nature09135.

[79] Kaufmann, E. *et al.* BCG Educates Hematopoietic Stem Cells to Generate Protective Innate Immunity against Tuberculosis. *Cell* **172**, 176–190.e19 (2018).

[80] Bowling, S. *et al.* An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* **181**, 1410–1422.e27 (2020).

[81] Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113**, 891–911 (1991).

[82] Tam, P. P. L. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. *Mechanisms of Development* **68**, 3–25 (1997). URL https://www.sciencedirect.com/science/article/pii/S0925477397001238.

[83] Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

[84] Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019). URL https://www.nature.com/articles/s41586-019-0933-9.

[85] Lohoff, T. *et al.* Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat Biotechnol* (2021).

[86] Guibentif, C. *et al.* Diverse Routes toward Early Somites in the Mouse Embryo. *Dev Cell* **56**, 141–153.e6 (2021).

[87] Mittnenzweig, M. *et al.* A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell* (2021). URL https://www.sciencedirect.com/science/article/pii/S0092867421004396.

[88] Qiu, C. *et al.* Systematic reconstruction of the cellular trajectories of mammalian embryogenesis. *bioRxiv* 2021.06.08.447626 (2021). URL https://www.biorxiv.org/content/10.1101/2021.06.08.447626v1.

[89] Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019). URL https://www.nature.com/articles/s41586-019-1184-5.

[90] Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019). URL https://www.nature.com/articles/s41592-019-0619-0.

[91] Tzouanacou, E., Wegener, A., Wymeersch, F. J., Wilson, V. & Nicolas, J.-F. Redefining the Progression of Lineage Segregations during Mammalian Embryogenesis by Clonal Analysis. *Developmental Cell* **17**, 365–376 (2009). URL https://www.cell.com/developmental-cell/abstract/S1534-5807(09)00339-6.

[92] Downs, K. M., Gifford, S., Blahnik, M. & Gardner, R. L. Vascularization in the murine allantois occurs by vasculogenesis without accompanying erythropoiesis. *Development* **125**, 4507–4520 (1998).

[93] Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928–943.e22 (2019). URL https://www.sciencedirect.com/science/article/pii/S009286741930039X.

[94] Forrow, A. & Schiebinger, G. LineageOT is a unified framework for lineage tracing and trajectory inference. *Nat Commun* **12**, 4940 (2021).

[95] McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016). URL https://www.science.org/doi/10.1126/science.aaf7907.

[96] Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).

[97] Bizzotto, S. *et al.* Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249–1253 (2021).

[98] Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**, 281–291.e9 (2019). URL https://www.cell.com/cell-systems/abstract/S2405-4712(18)30474-5.

[99] Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421–427 (2018). URL https://www.nature.com/articles/nbt.4091.

[100] Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015). URL https://www.nature.com/articles/nature14221.

[101] Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state. *bioRxiv* (2019). URL https://www.biorxiv.org/content/10.1101/517565v1.

[102] Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, eabc1944 (2021).

[103] Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018). URL https://www.science.org/doi/10.1126/science.aap8992.

[104] Ombrato, L. *et al.* Metastatic-niche labelling reveals parenchymal cells with stem features. *Nature* **572**, 603–608 (2019). URL https://www.nature.com/articles/s41586-019-1487-6.