Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Alessandro Greco

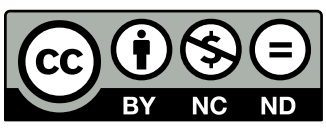
born in: Taranto

Oral examination: 27-06-2024

Lineage dynamics of haematopoiesis: insights from single-cell multi-omics and lineage tracing

Referees: Prof. Dr. Thomas Höfer

Dr. Angela Teresa Filimon Goncalves



Abstract

In haematopoiesis research, the description of the molecular states involved in differentiation from tip stem cells to unipotent progenitors is an ongoing effort that incorporates newly available technologies.

In particular, the application of single-cell sequencing technologies enables the dissection of transcriptional, epigenetic and immunophenotypic heterogeneity of stem and progenitor cells. As importantly, in-vivo fate mapping allows to trace the progeny of stem cells and their progeny and infer dynamical properties of the haematopoietic system.

However, a unified view of haematopoietic states that incorporates molecular heterogeneity in the context of differentiation relationships between subsets is still lacking.

In this thesis, I integrate information from multiple assays to obtain a high-resolution description of haematopoietic heterogeneity and link it to differentiation insights from lineage tracing evidence using statistical testing. The results presented here support a hierarchical model of the haematopoietic system, in which tip stem cells undergo an ordered sequence of regulatory events involving exit from a primitive state, proliferative activation, followed by intermediate lineage restriction steps that result in uni-potent progenitors.

In Chapter 2, fate mapping and index-sorted scRNA-seq data are combined to investigate how transcriptional and immuno-phenotypical heterogeneity affect the direct differentiation pathway between LT-HSCs and MkPs.

In the following chapters, a multi-omic dataset featuring paired scRNA-seq and scATAC-seq layers is analysed to investigate the mappability between the two and unveil modality-specific states related to proliferation, extrinsic signalling, and three-dimensional chromatin re-modelling.

Next, I utilise probabilistic trajectory inference and validate it using transcriptional mapping to lineage tracing datasets. The output of this procedure is used to inform a lineage potential model that enables statistical testing of fate association and thus the degree of hierarchy in lineage choices. Unbiased grouping of mature cell fates results in a branching model that features early lineage split within the stem cell compartment into intermediate oligo-potent progenitors before further specification into uni-potent progenitors.

To gain a molecular understanding of the detected branching events, I make systematic

use of differential expression and differential accessibility analysis to investigate the overall mechanisms that govern commitment at both transcriptional and chromatin level, thus use mathematical modelling to detect the hierarchical order of regulatory events. Finally, I utilise a recently developed gene regulatory network inference algorithm to reveal a highly dynamical regulatory landscape that links haematopoietic specification to key transcription factors.

In summary, the work presented in this thesis makes use of bioinformatic analysis, mathematical modelling and lineage tracing to obtain a robust description of the cellular states that comprise the haematopoietic system and their relation to differentiation potential and proposes a computational framework that can be used to enable the quantitative description of differentiation systems starting from single-cell sequencing data.

Zusammenfassung

Das hämatopoetische System erneuert kontinuierlich eine große Vielfalt von Zelltypen, die für den Sauerstofftransport, die Immunabwehr und die Aufrechterhaltung der Integrität der Blutgefäße zuständig sind. Das Verständnis der molekularen Zustände und Übergänge, während der Differenzierung von hämatopoetischen Stammzellen zu unipotenten Progenitoren, ist eine fortlaufende Anstrengung, die erheblich von neuen Technologien profitiert. Insbesondere ermöglicht die Anwendung von Einzelzell-Sequenzierungstechnologien die Analyse der transkriptionellen, epigenetischen und immunphänotypischen Variationen innerhalb von Stamm- und Vorläuferzellen. Im Gegensatz dazu ermöglicht es das in vivo "fate mapping", die Nachkommenschaft von Stammzellen zu verfolgen und die dynamischen Eigenschaften des hämatopoetischen Systems abzuleiten. Jedoch fehlt noch eine einheitliche Sicht auf die hämatopoetischen Zustände, welche die vielfältigen molekularen Zustände mit den Differenzierungsbeziehungen zwischen den Stamm- und Vorläuferzellsubsets integriert.

In dieser Dissertation integriere ich Informationen aus mehreren Assays Experimenten, um eine hochauflösende Beschreibung der zellulären Heterogenität im hämatopoetischen System zu erhalten. Die Ergebnisse diskutiere ich im Kontext von Erkenntnissen aus lineage tracing Studien. Diese Analyse unterstützt ein mehrschichtiges hierarchisches Modell des hämatopoetischen Systems: Spitzenstammzellen durchlaufen eine geordnete Abfolge von regulatorischen Prozessen, die den Austritt aus einem primitiven Zustand und die proliferative Aktivierung beinhalten. Es folgen schrittweise Einschränkungen der möglichen Zelltypspezifikationen, die schließlich zur vollständigen Festlegung des Zelltyps führen.

Zunächst werden Schicksalsmapping und scRNA-Seq-Daten von indexsortierten Stamm- und Vorläuferzellen kombiniert, um zu untersuchen, wie transkriptionelle und immunphänotypische Heterogenität den direkten Differenzierungsweg zu Megakaryozytenvorläufern beeinflussen. Diese Analyse deckt zwei unabhängige Wege der Thrombopoese auf und kartiert diese auf molekularer Ebene.

In den folgenden Kapiteln wird ein Datensatz analysiert, der für jede Zelle, sowohl deren mRNA-Profil (scRNA-seq), als auch Informationen zur Chromatinzugänglichkeit (scATAC-seq) enthält. Mit diesem Multi-Omics-Datensatz wird die Abbildbarkeit zwischen den bei-

den Ebenen untersucht und modalitätsspezifische Zustände im Zusammenhang mit Proliferation, extrinsischer Signalgebung und dreidimensionaler Chromatin-Umgestaltung werden aufgedeckt.

Anschließend, nutze ich eine Methode zur probabilistischen Inferenz von Trajektorien und validiere diese mit Transkriptom-Daten aus Tracing-Experimenten. Das Ergebnis dieses Verfahrens dient zur Konstruktion eines Modells, das es ermöglicht, die Assoziation verschiedener Differenzierungswege statistisch zu quantifizieren.

Eine unvoreingenommene Gruppierung von unipotenten Entwicklungsstadien führt zu einem nicht baumartig verzweigten Modell, das eine frühe Aufspaltung innerhalb des Stammzellkompartiments in intermediäre oligopotente Progenitoren zeigt, bevor eine weitere Spezifikation zu unipotenten Progenitoren erfolgt.

Um ein molekulares Verständnis der Verzweigungsvorgänge zu gewinnen, verwende ich systematisch Differential Expression (DE) und Differential Accessibility (DA) Analysen, um die Mechanismen zu identifizieren, die die Genregulation während der Spezifizierung beeinflussen. Zuletzt verwende ich einen neu entwickelten Algorithmus zur Analyse von Genregulationsnetzwerken. Dieser zeigt eine sehr dynamische regulatorische Landschaft auf, in der die anfängliche Zusammenarbeit spezifischer Transkriptionsfaktoren zur Differenzierung übergeht und durch gegenseitige Hemmung Unipotenz erreicht wird.

Zusammenfassend nutzt die in dieser Arbeit vorgestellte Forschung bioinformatische Analysen, mathematische Modellierung und Lineage-Tracing, um eine robuste Beschreibung der Zellzustände im hämatopoetischen System und ihrer Beziehung zum Differenzierungspotenzial zu erlangen. Zudem wird ein Rechenmodell vorgeschlagen, das zur quantitativen Beschreibung von Differenzierungssystemen ab Einzelzell-Sequenzierungsdaten verwendet werden kann.

Contents

Chapter 1

Introduction

Haematopoiesis is one of the most important processes that sustains life in complex organisms. It allows the continuous replenishment of a remarkable variety of cell types that perform crucial functions in the body, ranging from oxygen transport to wound healing and defence against pathogens. An impressively large number of cells are involved in the execution of these functions, estimated to be around 10^{13} for humans and 10^{10} for mice [1]. These disparate cell types, which include erythrocytes, granulocytes, platelets, and lymphocytes, are not static but require constant replenishment. Humans produce 10^{14} cells per year, whereas mice produce 10^{11}. This massive flow of newly specialised cells is ultimately maintained by a comparatively rare population of stem and progenitor haematopoietic cells through intermediate steps of amplification and differentiation.

The discoveries that led to the current understanding of haematopoietic stem cells (HSCs) retain a strong footprint on the current research questions in the field. The notion of a unique stem cell type that could differentiate into any mature blood lineage [2] started as a controversial hypothesis. Long after the initial formulation, foundational work from Till and McCulloch [3] demonstrated that bone marrow contains haematopoietic clones that could divide and give rise to mixed myeloerythroid progeny. This remarkable finding was achieved by administering lethal irradiation to mice and then transplanting cells marked by a distinct chromosomal aberration induced using radiation. In the following years, clinical bone marrow transplantation became widespread and, although HSCs were thought to be present in bone marrow, they could not yet be isolated.

A fundamental change ensued when it became possible to isolate cells based on the expression of surface markers [4]. After a series of experiments proposing several combinations of surface markers for different progenitors, a hierarchy of different progenitors emerged [5]. LT-HSCs represent the apex population, capable of self-renewing and reconstituting the entire haematopoietic system in lethally irradiated hosts. Downstream of LT-HSCs ($Lin^-Kit^+Sca1^+CD48^-CD150^+$), equally multipotent progenitors with lower self-renewal and

reconstitution capabilities (ST-HSCs and MPPs) generate lineage-restricted subsets (CMPs and CLPs) with myeloerythroid and lymphoid potential. From these oligopotent progenitors, another step of lineage restriction yields unipotent specialised progenitors. Successive refinement of the gating strategy allowed the isolation of LMPPs, a subset of MPPs whose output is mostly lymphoid [6], and more primitive LT-HSCs that express the EPCR marker [7]. As the definition of populations became more refined and experimental techniques evolved, it became possible to transplant single HSCs and confirm their ability to replenish the entire haematopoietic system [8].

1.1 Lineage Tracing

While these studies allowed the discovery of the fundamental properties of haematopoiesis, they are mainly based on in vitro assays and transplantation in highly disrupted animals. As such, they are more likely to indicate what HSCs and downstream progenitors can do, rather than the function they actually perform under physiological conditions. For this reason, a more recent wave of research has focused on investigating how HSCs and progenitor cells behave in vivo. A paramount development that enabled lineage tracing was the development of Cre-inducible systems, in which Cre recombinase expression is driven by a specific cell-type marker, causing the excision of a stop cassette that permanently turns on the expression of a reporter gene [9]. This technique allows to measure the temporal interval between label induction in HSCs and the detection of labelled progeny in downstream compartments. This line of research revealed that, contrary to transplantation settings, tip HSCs rarely differentiate into ST-HSCs, and that the bulk of haematopoiesis is dependent on less self-renewing MPPs [10], as exemplified by the absence of phenotype in mice whose HSCs are selectively ablated [11]. Moreover, different mature lineages receive label at different times: platelets receive labels first, followed by granulocytes and erythrocytes, while labelled lymphocytes are not detected until much later [12]. Population-level lineage tracing was shortly followed by a more refined class of lineage tracing systems, in which random unique barcodes were induced in single cells, thus allowing investigation of the clonal aspects of haematopoiesis, highlighting how multiple clones contribute to haematopoiesis at different times [13]. In a transposon-based system, barcodes whose detection was restricted to the Megakaryocyte lineage introduced the hypothesis of a platelet-restricted subset of HSCs [14].

1.2 Single-cell RNA-seq

While informative on the lineage output of haematopoietic stem and progenitor cells (HSPCs), lineage tracing alone cannot offer molecular insight into the different clonal behaviours observed solely based on the handful of surface markers used to sort cells. An in-depth snapshot of molecular heterogeneity in haematopoiesis can be accessed through RNA-sequencing technologies, that allow to collect mRNA from a selected population and, more recently, from single cells (scRNA-seq). A typical scRNA-seq is performed on hundreds to millions of cells, for which the transcript count for the 48000 genes comprising the mouse genome is available. Such large amount of data requires careful computation to extract meaningful biological insights [15]. A fundamental step in a scRNA-seq sequencing pipeline is dimensionality reduction: how can thousands of cells and genes be summarised and visualised? Typically, the first linear step consists of detecting the main axis of variation using Principal Component Analysis (PCA), and 20 to 50 components are used to build a k-neighbors graph, in which the k closest neighbours are linked in a network structure. Subsequently, based on the objective of the analysis, cells can be divided into similarity clusters, further reduced or visualised in 2 dimensions using non-linear visualisation techniques, such as t-SNE, UMAP, or more specialised methods such as PHATE [16].

However, drawing conclusions about the dataset solely based on 2-dimensional visualisations leads to erroneous interpretations [17], and quantification of the metric of interest in higher dimensions is recommended. In this regard, it is usually helpful to resort to non-linear dimensionality reduction methods that better capture nuanced features of the transcriptional landscape. Diffusion maps [18] model the transcriptional landscape as a stochastic diffusion process and are particularly indicated in datasets comprised of a continuum of states such as haematopoiesis. Particularly relevant in haematopoiesis research are methods that attempt to leverage proximity relations in transcriptional landscapes to infer differentiation trajectories. The field of trajectory inference bloomed in the last decade and is built on the assumption that, as cells differentiate, they occupy all intermediate states between the progenitor and final state in monotonic order. Therefore, cells can be ordered based on the transcriptional distance from an initial cell; this process is called pseudotemporal ordering and the inferred quantity pseudotime [19]. Another necessary assumption of pseudotemporal ordering is that the variation in the data can be imputed to the differentiation process for which pseudotime is estimated. This assumption is clearly problematic in scRNA-seq, whose variation is affected by other processes, such as cell cycle, signalling state, and cell death. However, haematopoietic stem cells do not differentiate into one lineage only, but branch into several mature cell types. Detection of branching points (and, in general, differentiation topology) is another primary task in trajectory inference that inspired several

classes of methods based on different theoretical assumptions, starting from minimum spanning trees [20], fitting of principal curves to low-dimensional embeddings [21], and linking clusters based on modularity properties of the neighbourhood graph [22]. Once again, methods that make use of stochastic process theory [18, 23, 24] are of particular relevance in haematopoiesis, as they are particularly indicated to describe datasets that cannot be easily divided into clusters, but rather form a relatively continuous landscape of intermediate states. Using these methods, a large number of studies (reviewed in [25]) have detected and described heterogeneity in the stem cell compartment with regard to proliferation, lineage bias, and metabolic state. When considered on its own, scRNA-seq evidence seems to challenge the classic hierarchical model of haematopoiesis, in favour of a more fuzzy view in which stem cells gradually acquire bias towards the major lineages. If two-dimensional embeddings of the transcriptional landscape are interpreted at face value, they seem to indicate that HSCs and progeny do not undergo discrete choices that restrict their potential in a stepwise manner, but rather drift towards one branch or the other and end up committing to the closest one [26]. However, it can be argued that this conclusion relies heavily on one of the previously cited limitations on the interpretation of scRNA-seq data, in particular the identification of molecular heterogeneity with developmental ordering of cells. In other words, observing a transcriptional state does not necessarily equate to a distinct identity of such a state in terms of differentiation output. In summary, while lineage tracing can track progeny in hematopotic clones but cannot resolve their heterogeneity aside from surface markers, single cell sequencing data describe heterogeneity down to the molecular level, but with no direct link to functional heterogeneity in terms of output.

1.3 Integration of multiple data modalities

A few studies have successfully merged lineage tracking and scRNA sequencing. In [27], uniquely barcoded cells are cultured and their progeny is analysed in ex vivo culture and in transplantation settings, thus trying to link the transcriptional state of a cell with its future lineage output. Surprisingly, the transcriptional state is a poor predictor of a cell output, even in ex-vivo settings in which the fate of cells is only determined by culture conditions and their internal state. In a successive study [28], unique barcodes were induced in-vivo and lineage output was evaluated after 7-20 weeks, and unique barcodes were classified as differentiation-inactive, myeloid-restricted, and multipotent. Similarly to the previous example, the authors tested whether a transcriptional signature obtained by comparing cycling and non-cycling HSCs [29] could discriminate the subsets of differentiation defined by barcode, with a negative result. However, the paucity of barcodes in each category lacked the statistical power to obtain robust signatures for each of these subsets, except for a few genes. If transcriptional data alone cannot describe lineage choices in haematopoiesis, where is this information available? One possibility is that such information is simply not contained in the cells that undergo the decisions. In an instructive model of fate choice, external cues such as cytokines enforce lineage decisions in HSPCs, which trigger specific programmes depending on the external cues. Although evidence for the effect of cytokines on commitment has long been available [30], the intrinsic contribution to this mechanism is still debated. Rather than directly instructing decisions, cytokines might act on a heterogeneous population of stem cells and promote survival of the ones that had already made the intended choice [31]. This defines a permissive model of commitment at the opposite end of the instructive model. Initially, evidence leaned in favour of the instructive model [32, 33], but the mass adoption of single-cell sequencing technologies has returned the spotlight to molecular heterogeneity. Moreover, direct evidence for the relevance of the intrinsic state of stem cells for their differentiation output has been shown using a combination of lineage tracing and transplantation [34].

If not in the transcriptional layer, researchers have often hypothesised that the epigenetic state of a stem cell carries more reliable information about its lineage fate. In a successful biological metaphor to describe differentiation [35], cell-intrinsic dynamics that lead to differentiation are described as a downhill path in an epigenetic landscape, that defines what states are available to a cell as it rolls down and carries out decisions. Epigenetic modifications, such as DNA methylation, covalent histone modifications, nucleosome remodelling, have shown primary relevance in the regulation of stem cell function [36]. Among the several technologies developed to probe the epigenetic state of a biological sample, the most relevant to the scope of this thesis is the assay for transposase-accessible chromatin with

sequencing (ATAC-seq) [37], which captures open chromatin sites, thus allowing to measure the net impact of epigenetic modifications on the folding structure of DNA [38]. A few years after its bulk implementation, single-cell ATAC-seq was developed [39] and deployed on human haematopoietic populations to dissect the regulatory heterogeneity underlying differentiation from HSCs to mature populations [40]. If not in the transcriptional layer, researchers have often hypothesised that the epigenetic state of a stem cell carries more reliable information about its lineage fate. In a successful biological metaphor to describe differentiation [35], cell-intrinsic dynamics that lead to differentiation are described as a downhill path in an epigenetic landscape, that defines what states are available to a cell as it rolls down and carries out decisions. Epigenetic modifications, such as DNA methylation, covalent histone modifications, and nucleosome remodelling, have shown primary relevance in the regulation of stem cell function [36]. Among the several technologies developed to probe the epigenetic state of a biological sample, the most relevant to the scope of this thesis is the assay for transposase-accessible chromatin with sequencing (ATAC-seq) [37], which captures open chromatin sites, thus allowing to measure the net impact of epigenetic modifications on the folding structure of DNA [38]. A few years after its bulk implementation, single-cell ATAC-seq was developed [39] and deployed on human haematopoietic populations to dissect the regulatory heterogeneity underlying differentiation from HSCs to mature populations [40].

The transcriptional and chromatin states of a cell are closely related to each other, as a locally open chromatin configuration is required for polymerases and transcription factors to bind DNA and exercise their function. At the same time, chromatin remodelling depends on the transcription and translation of multiple chromatin factors [41]. Although causality between the two layers flows in both directions, chromatin states are thought to precede transcriptional and phenotypical changes in differentiation and development systems. This phenomenon is known as lineage priming (a more in-depth introduction is given in Chapter 5) and is associated with the activation of enhancers that lead to massive changes in transcription [42]. In haematopoiesis, chromatin priming has so far been supported by bulk population analysis using Chip-seq in combination with ATAC-seq and RNA-seq [43, 44]. Despite its advantages in describing regulatory states over transcriptomics, there are technical and conceptual drawbacks in only using ATAC-seq to describe cellular states: as mentioned previously, chromatin and transcriptional states interact causally with each other without a predefinite hierarchy. A typical example is the cell cycle: although it has little to no impact on the chromatin landscape at the single-cell level, proliferative activation of HSPCs has been repeatedly associated with commitment to one or more mature lineages [29]. Ideally, a complete description of events that determine lineage commitment includes both transcriptional (e.g., cell cycle, signalling cues, metabolic activation) and chromatin states. Technologies

that include more than one -omic layer in a biological sample are termed multi-omics, and respond to the need to integrate information across assays [45]. While computational tools to align samples across modalities have developed, they rely on the assumption that different assays are underlied by the same factors of variation, which arguably defeats the purpose of a multi-omic approach if the research questions concern priming or other assay-specific states. More recently, it became possible to profile multiple layers from single-cells simultaneously [46–48], unlocking a new set of possibilities and challenges to interpret such rich datasets [49]. One of the most common research questions addressed using these data is whether chromatin priming can be detected by ordering cells along a pseudotemporal trajectory. In particular, systematic analysis of RNA features paired with correlated chromatin peaks in [48] enabled the detection of chromatin unfolding before the onset of RNA expression during the development of the hair follicle in mice. Another crucial insight that can be discerned using paired multiomic data involves the regulatory interactions that originate the heterogeneity in the data. In particular, transcription factors (TFs) link the transcriptional and chromatin configurations of a cell: they either require open chromatin sequencing or cause chromatin opening (pioneer TFs) and bind target DNA sequences in enhancers, thus causing three-dimensional alterations and affecting the transcription rates of target genes. The net impact of regulatory interactions between TFs and target genes can be summarised in gene regulatory networks (GRNs), in which nodes correspond to genes and links are weighted based on the intensity and direction of regulation. Although dynamical analysis of simple regulatory circuits can be promptly analysed using dynamical systems theory, the inverse task of inferring the correct gene regulatory network underlying observed transcriptional states has not yet been achieved [50]. Multi-omic data have the potential to dramatically improve GRN inference algorithms by leveraging chromatin level information, such as TF motif enrichment in chromatin peaks. Several methods for GRN inference from multiomic data have been published and deployed to dissect regulatory dynamics in multiomic datasets [51–53].

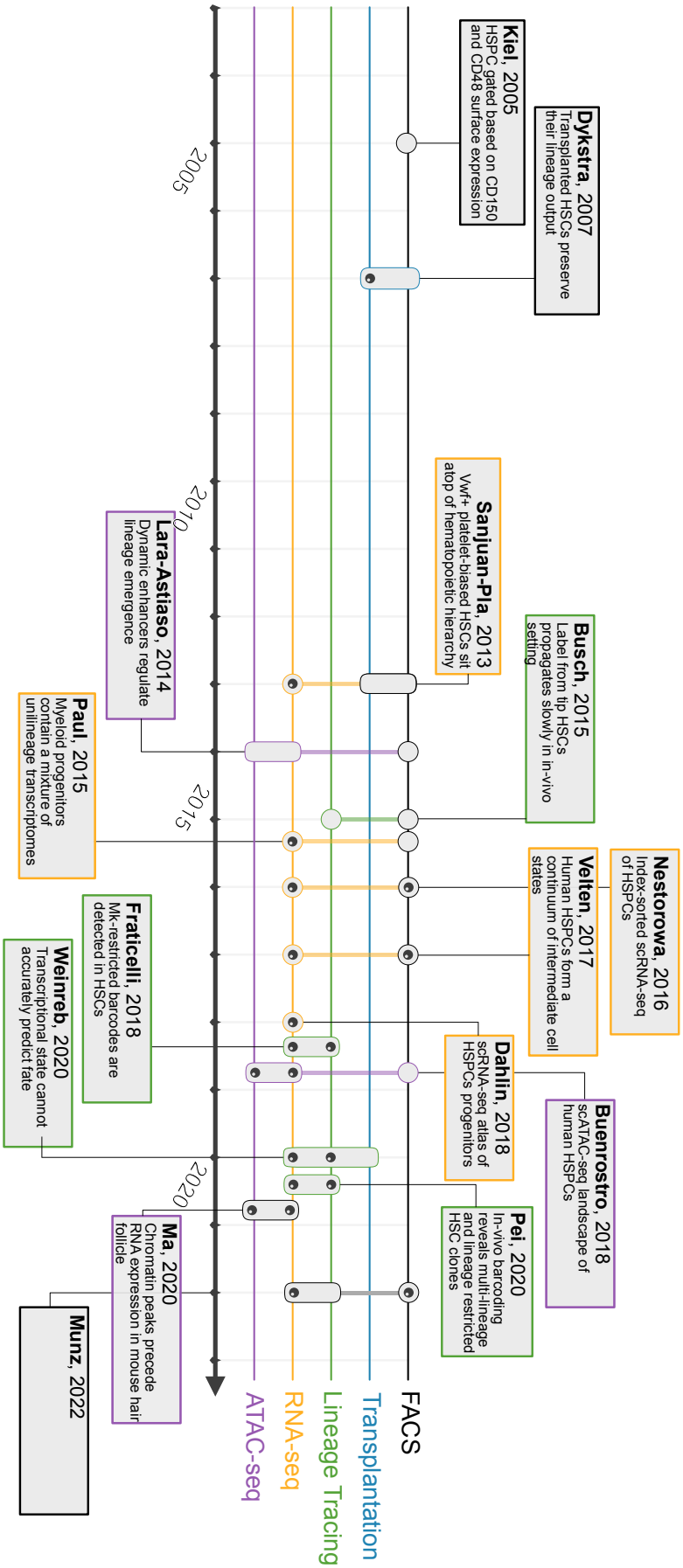


Figure 1.1. Selected haematopoiesis publications annotated based on the technologies used to obtain key results. Single-cell assays are indicated by a dark dot.

1.4 Overview of this thesis

Throughout its historical development, haematopoietic research has moved from considering LT-HSCs as a homogeneous cell type to observing many aspects of their molecular heterogeneity using single-cell sequencing technologies. However, the amount of diversity that can be ascribed to a system could merely be a consequence of the resolution used to look at the data. Lineage tracing evidence seems to indicate that functional heterogeneity in terms of lineage output is a true feature of haematopoiesis. However, a satisfactory mapping between lineage output heterogeneity and molecular features available using modern single-cell sequencing technology is lacking. The aim of this thesis is to investigate how single-cell sequencing experiments can be paired with lineage tracing assays to detect lineage bias in sequencing data.

In the first chapter of this thesis, I describe how the combination of bulk-level lineage propagation, mitotic tracking, index sorting, and scRNA-seq data generated by Clara Munz and N. Morcos at the TU of Dresden has helped me inform a mathematical model that links the megakaryocyte bias detected within the LT-HSC compartment using lineage tracing and scRNA-seq. In this work, I leveraged index-sorting to identify the LT-HSC subset that originates the shortcut toward the Megakaryocytic lineage and contextualise it with respect to stem cell heterogeneity. Using trajectory inference on index-sorted defined subsets of HSPCs, I generated a candidate topology for haematopoietic differentiation that was further refined by Congxin Li by integrating label propagation and mitotic tracking data. The results of this collaboration were published in [54].

The remainder of the thesis is dedicated to the explorative analysis of a multiomic dataset obtained thanks to work from Larissa Frank (Division of Cellular Immunology, DKFZ), Nina Claudino and Jonas Metz (Division of Theoretical Systems Biology, DKFZ) using the commercially available 10x Single Cell Multiome ATAC + Gene Expression kit. In the second chapter, I introduce the general features of the data and leverage the matched transcriptional and chromatin assays to quantify the mappability between the two and detect and obtain a molecular description of modality-specific states.

Chapter three focuses on trajectory inference. First, I devise a sampling strategy to overcome previous issues in haematopoietic datasets and detect an early stem cell population that contains non-activated stem cells. Next, I use pseudotemporal ordering and mathematical modelling to estimate the sequence of events that lead to lineage commitment. Next, I use probabilistic trajectory inference on chromatin data and validate its results against lineage tracing and transplantation datasets using transcriptional annotations. Next, I propose a lineage potential model to obtain a discrete partition of data based on differentiation probabilities and perform statistical tests on its outcome to achieve a quantitative assessment

of lineage coupling and hierarchy within the haematopoietic system. Lastly, I use statistical associations between single lineages to group trajectories into branches and compare them against immunophenotypic gates, thus retrieving a phenotypical description of lineage emergence.

In Chapter 4, I make systematic use of differential expression and differential accessibility analysis to investigate the overall mechanisms that govern commitment at both the transcriptional and chromatin level, thus I use mathematical modelling to detect the hierarchical order of regulatory events in haematopoiesis. Next, I test two methods to quantify chromatin priming, use a recently developed GRN modelling tool, and reveal a highly dynamical regulatory landscape that enables haematopoietic specification.

Chapter 2

Integration of scRNA-seq and lineage tracing evidence to reveal distinct pathway of thrombopoiesis

2.1 Introduction

As introduced in the previous Chapter, the populations that constitute the haematopoietic system and their differentiation relationship are a matter of active research since the discovery of the existence of multipotent haematopoietic stem cells [3]. Advancements in genetic barcoding allowed to progress from transplantation-based assessment of lineage potential to investigation of in vivo realisations of differentiation trees. However, many aspects of haematopoietic differentiation remain debated.

The earliest models of differentiation, based on transplantation and in vitro experiments, represented the haematopoietic hierarchy as a tree in which HSCs occupy the top, followed by intermediate progenitors lacking self-renewal and with restricted lineage potential [55]. According to this model, to differentiate into MkPs, HSCs must go through a phenotypic sequence that includes ST-HSCs, MPPs, CMPs, and MEPs. However, the hypothesis of a direct differentiation path from HSCs to MkPs started to emerge based on transplantation experiments [56], response to challenges [57], and finally lineage tracing [14]. Furthermore, successive attempts to chart the landscape of haematopoiesis by single cell RNA sequencing revealed a high similarity between stem cells and MkPs [56, 58]. However, several aspects of this connection remain unresolved. Does a specific subset of HSCs exclusively generate Mks, or does direct differentiation occur from genuinely multipotent stem cells? Are platelets produced solely in this manner, or do both the classical and direct pathways coexist? Does inflammation influence differentiation through this path?

In a joint effort with Clara Munz (who performed mouse experiments) and Congxin Li (who performed mathematical modelling on lineage tracing data), I integrated evidence from fate mapping data and mitotic tracking with scRNA-seq analysis of index-sorted haematopoietic populations to investigate lineage relations between HSCs and progenitor populations. I investigated transcriptional proximity and used partition-based trajectory inference to propose a differentiation topology to be compared against fate mapping and mathematical modelling.

In addition to SLAM-defined haematopoietic populations [59], surface expression of Sca-1 and CD201 (EPCR) was used to further resolve HSPC populations. Mathematical modelling indicates that tip LT-HSCs (ES HSC) are $EPCR^{+}$, multipotent, and self-renewing. In addition, Sca-1 expression divides HSCs, MPPs, and HPC-1 into subpopulations that differ in their lineage bias: $Sca1^{hi}$ cells retain lymphoid potential, while loss of Sca-1 ($Sca\text{-}1^{-/lo}$) shifts HSPCs toward myelo-erythroid and megakaryocyte lineages. Transcriptional and fate mapping analysis reveals that, in addition to the canonical route of thrombopoiesis through the CMP, MEP and MkP gates, a direct differentiation path between $Sca\text{-}1^{lo}$ HSCs and $CD48^{-/lo}$ MkPs is responsible for the production of roughly half the platelets, and its usage

increased in emergency setting to replenish the platelet pool more rapidly. Pseudotemporal analysis of megakaryocyte marker expression further corroborates independent maturation along the two pathways. Furthermore, I analysed bulk RNA-seq samples of the two newly defined MkP subsets to investigate their molecular differences and concluded that while platelet-related genes are substantially equal, myeloid markers are more expressed in $CD48^{hi}$ MkPs, possibly due to rare myeloid contamination or transient myeloid potential during the intermediate stages.

2.2 Fate mapping data implicate loss of Sca-1$^{-/lo}$ HSCs, fast labelling of platelets

Haematopoietic populations were isolated from bone marrow, as described in [59]. The addition of EPCR to the marker panel allowed further subdivision of the LT-HSC gate into CD201^{+} LT-HSC (ES HSC), while all populations were further subdivided based on the surface expression of Sca-1 (Fig. 2.1 A). Fate mapping evidence was obtained by generating a mouse model Fgd5ZsGreen: CreERT2 / R26LSL-tdRFP (Fig. 2.1 B), which introduces a fluorescent RFP label in LT-HSCs with high specificity (Fig. 2.1C). The system was complemented with H2B-GFP label dilution measurements, allowing the estimation of proliferation and differentiation rates across populations. A large cohort of mice (n=82) was induced and data was collected at different timepoints (Fig. 2.1D). By observing label propagation trends in HSC subsets, one can observe how the percentage of labelled cells among ES HSCs does not grow over time, indicating that this population self-renews completely and represents the ultimate source of HSCs (Fig. 2.1 E).

Fate mapping and mitotic tracking evidence were modelled using ordinary different equations by Congxin Li. Each of the previously defined populations is characterised by rates of proliferation, death, and differentiation toward downstream progenitors. Of all possible models that can be fitted to the data, those that perform best using Akaike's information criterion are marked by three elements (Fig. 2.1F) :

- Commitment to the myeloid branch (CMP) is accompanied by loss of Sca-1 surface expression;
- Commitment to the lymphoid branch requires retention of Sca-1 surface expression;
- A large number of Sca-1$^{-/lo}$ HSCs exit the system, by either death or differentiation into an external compartment.

To further elucidate the nature of the loss term originating from Sca-1$^{-/lo}$ HSCs obtained by mathematical modelling, it is worth considering fate mapping evidence downstream of the populations considered so far, specifically in megakaryocyte progenitors (MkP) and platelets. Since labelled cells descend from ES HSCs, it is expected that each haematopoietic population has a lower portion of labelled cells with respect to its direct progenitor. Surprisingly, a clear violation of this constraint is observed between platelets and MPPs and CMPs (Fig. 2.1G), suggesting that an HSC subpopulation feeds into the megakaryocytic differentiation pathway, bypassing some of the phenotypical steps usually ascribed to thrombopoiesis. Furthermore, successive inclusion of MkPs in the dataset revealed that their label accumulation is also slower than platelets (Fig. 2.1H). Given that platelets only

descend from MkPs, the only explanation for this observation is that the latter are composed of subsets of different origins labelled at different speeds from the ES HSC source, requiring an in-depth analysis of molecular heterogeneity and similarities across HSPC populations. For this reason, I analysed a single-cell transcriptomic dataset obtained using the Smart-seq2 protocol on index-sorted HSCs, MPPs, HPCs-1, and LS^{-}K progenitors (including MkPs) performed at TU Dresden by the group of Alexander Gerbaulet.

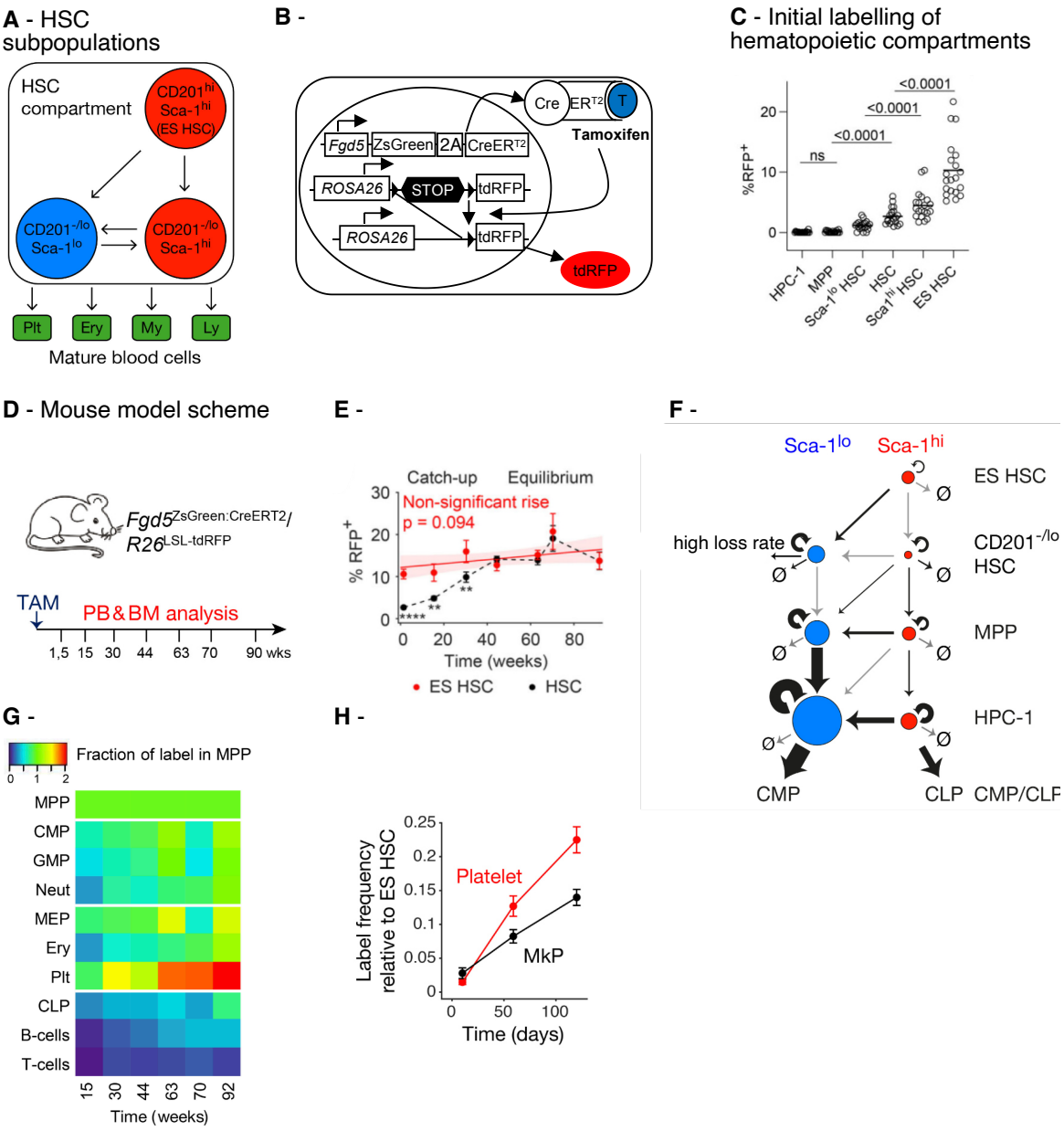


Figure 2.1. **A**, Stratification of LT-HSC populations based on surface expression of CD201 and Sca-1. **B**, Mouse model schematics. *Fgd5* expression drives fluorescent marker ZsGreen and Cre-estrogen receptor fusion protein Cre-ERT2. Tamoxifen administration triggers the excision of the stop cassette from the R26$^{LSL\text{-}tdRFP}$ allele, causing *Fgd5*-expressing cells and their progeny to be permanently labelled by RFP expression. **C**, Percentage of RFP^{+} cells after TAM administration. **D**, Timepoints of analysis. **E**, Labelling frequency over time in HSCs and ES HSCs. The non-increasing frequency in the latter indicates that the labelled cells are the ultimate source of adult haematopoiesis. **F**, Results of mathematical models performed by Congxin Li. Arrow thickness is indicative of the flux between populations. **G**, Labelling frequency over time relative to MPP. Platelets equilibrate the label faster than MPPs, suggesting an alternative route of differentiation. **H**, Labelling dynamics in platelets and MkPs. Faster platelet labelling points to heterogeneity in the MkP compartment. **All panels in this figure are adapted from [54]**, data produced by Mina N.F. Morcos, Clara M. Munz, Congxin Li.

2.3 scRNA-seq embedding unveils heterogeneity in LT-HSCs, MkPs

The data were preprocessed using a standard Bioconductor preprocessing pipeline (see Methods), while populations were annotated using the gates introduced previously. To gain a first understanding of similarities across populations, I used PHATE [16], a visualisation method based on information-geometric principles that is particularly indicated to embed continuous differentiation processes in low dimensions. PHATE arranges haematopoietic populations consistently with results from mathematical modelling: transcriptional states progress from LT-HSCs, through MPPs and HPCs, to myeloid and erythroid progenitors. Megakaryocyte progenitors are placed close to both LT-HSCs and erythroid progenitors (Fig. 2.2A). Accordingly, scoring of transcriptional lineage sets obtained from [60] shows higher expression of differentiation markers downstream of LT-HSCs (Fig. 2.2B).

As reported previously [56, 61], *Vwf* and *Itga2b* are expressed by LT-HSCs close to the megakaryocytic region of the landscape (Fig. 2.2 B). Surface marker expression, measured through index sorting, demonstrates the link between transcriptional and immunophenotypical heterogeneity in LT-HSCs: as inferred using fate mapping evidence, EPCR (CD201) and Sca1 expression marks tip stem cells. CD41 and CD150 are associated with megakarayocyte signal, while MkPs population show heterogeneous CD48 surface expression (Fig. 2.2C).

In summary, the initial exploratory analysis of the single-cell RNAseq dataset suggests transcriptional continuity between HSC subpopulations and the MkP and MPP subsets.

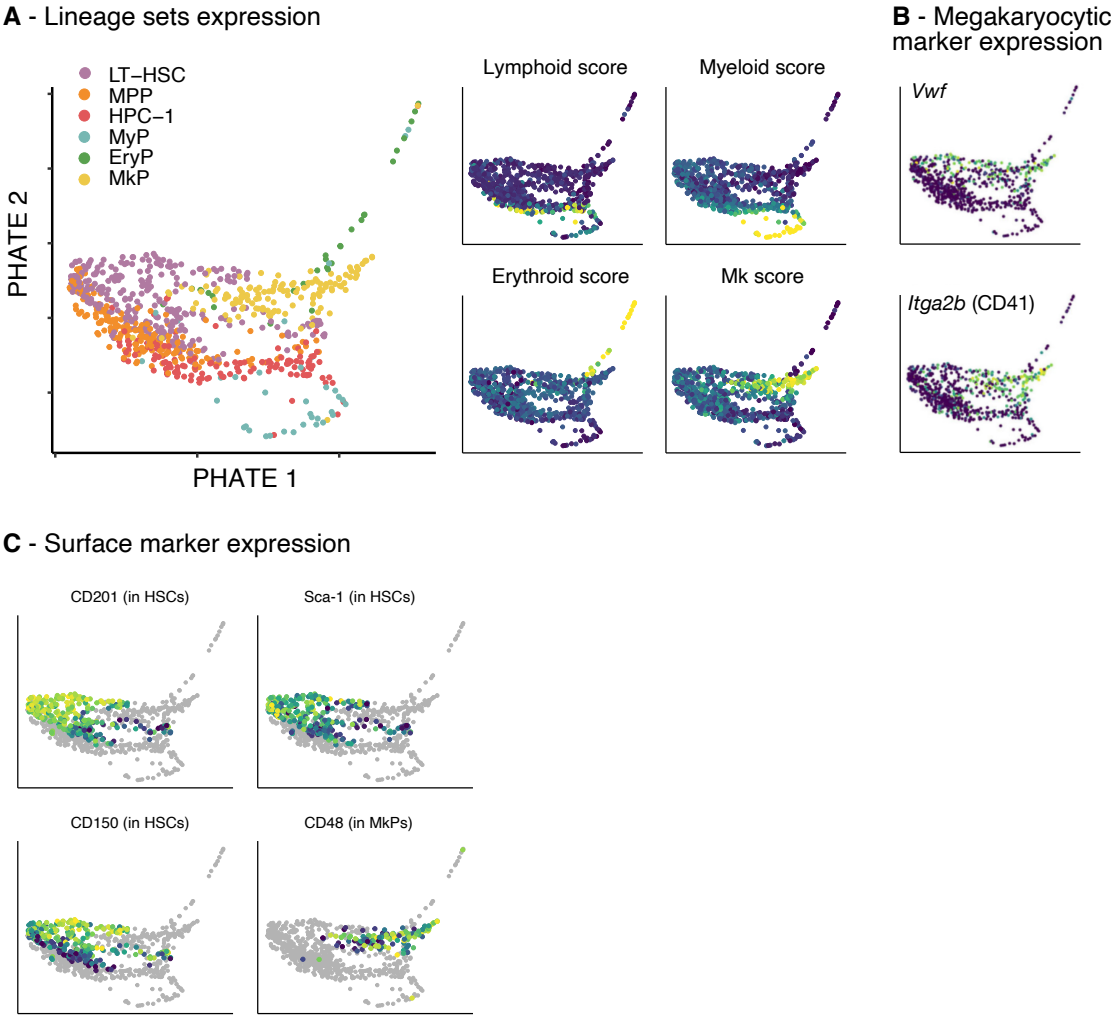


Figure 2.2. **A**, Sorted populations and expression of lineage markers obtained from [60]. **B**, Transcriptional expression of *Vwf* and *Itga2b*(Cd41). **C**, Surface marker expression of SLAM markers obtained through index sorting. **All panels in this figure are adapted from [54]**

2.4 Transcriptional connectivity graph generates candidate links for mathematical modelling

To make the results of the scRNAseq analysis comparable to the modelling framework, I divided the populations based on Sca-1 and CD201 surface expression and annotated a subpopulation of HPC-1 cells as lymphoid based on the transcriptional score of relevant markers. Additionally, I split MkP into $CD48^{+}$ and $CD48^{-/lo}$, as fate mapping strongly suggested heterogeneity within this population (Fig. 2.3A). Next, I sought to use transcriptional proximity relations between populations to inform mathematical modelling. As a first step, I computed diffusion pseudotime [18] and ordered cell states based on transcriptional similarity to LT-HSCs. By comparing pseudotime and H2B-GFP expression, I observed how the pseudotemporal succession of populations mirrors mitotic history: cells with higher pseudotime have undergone more divisions, while early populations exhibit higher levels of H2B-GFP protein. Notably, MkP subsets display similar pseudotemporal scores, but the $CD48^{-/lo}$ group has divided significantly less. (Fig. 2.3B). Next, I sought to focus on HSCs and MkP subsets to generate candidate topological configurations to compare against lineage tracing data. By measuring the Euclidean distance on the first ten principal components, I observed that:

- $CD48^{-/lo}$ MkPs are transcriptionally closer to LT-HSCs than their $CD48^{hi}$ counterpart (Fig. 2.3C);

- In turn, $CD48^{-/lo}$ MkPs are nearest to the $Sca\text{-}1^{-/lo}$ LT-HSC subset (Fig. 2.3D).

These two observations on the transcriptional proximity of MkP and LT-HSC subsets provide further evidence of a direct differentiation route between $Sca\text{-}1^{-/lo}$ LT-HSCs and $CD48^{-/lo}$ MkPs. To develop a more comprehensive and systematic approach to elucidate similarities in this dataset, I used PAGA [22]. This method draws links between discrete partitions based on a statistical comparison between the number of links between cells that belong to distinct partitions and a null model connectivity. The resulting graph shows connections among sorted populations (Fig. 2.3E) with high confidence (Fig. 2.3F). Embedding using Fruchterman–Reingold layout summarises and expands the previous proximity analysis by assigning higher similarity between $Sca\text{-}1^{-/lo}$ subsets and myelo-erythroid-megakaryocytic subsets. Moreover, $CD48^{hi}$ MkPs are connected to HPC-1 via $LS^{-}K$ progenitors.

2.5 Pseudotemporal analysis corroborates independent MkP maturation in direct pathway

At first, this evidence seems to suggest that MkPs progress from a $CD48^{-/lo}$ state to a $CD48^{hi}$ one as they differentiate into Mk. To dig deeper into the potential relationship between the MkP subsets, I computed the expression of megakaryocytic markers as a function of pseudotime progression. The correlation analysis shows that there is no significant difference in the accumulation of Mk transcripts in relation to pseudotime between the two subpopulations (Fig.2.3G).

The contradiction between PAGA connectivity and independent maturation of the two pathways is only apparent, since links inferred using snapshot transcriptomic data are indicative of proximity and do not necessarily signify a direct descent relationship between the linked populations. To corroborate the PAGA topology with fate mapping data, Congxin Li expanded the previous mathematical modelling using PAGA links as candidate lineage connections to fit fate mapping and mitotic tracking evidence. Results were successively validated using culture and transplantation assays performed by Clara Munz. This inference reveals a scheme remarkably similar to that obtained with PAGA, identifying a direct thrombopoiesis pathway that progresses from $CD201^{-/lo}$ $Sca1^{-/lo}$ LT-HSCs to $CD48^{-/lo}$ MkPs in addition to the typical path through MPPs and MEPs (Fig.2.3H). Of note, further mathematical modelling work on thrombopoietin-stimulated HSPCs revealed how the $CD48^{-/lo}$ subset replenishes platelets faster during challenges.

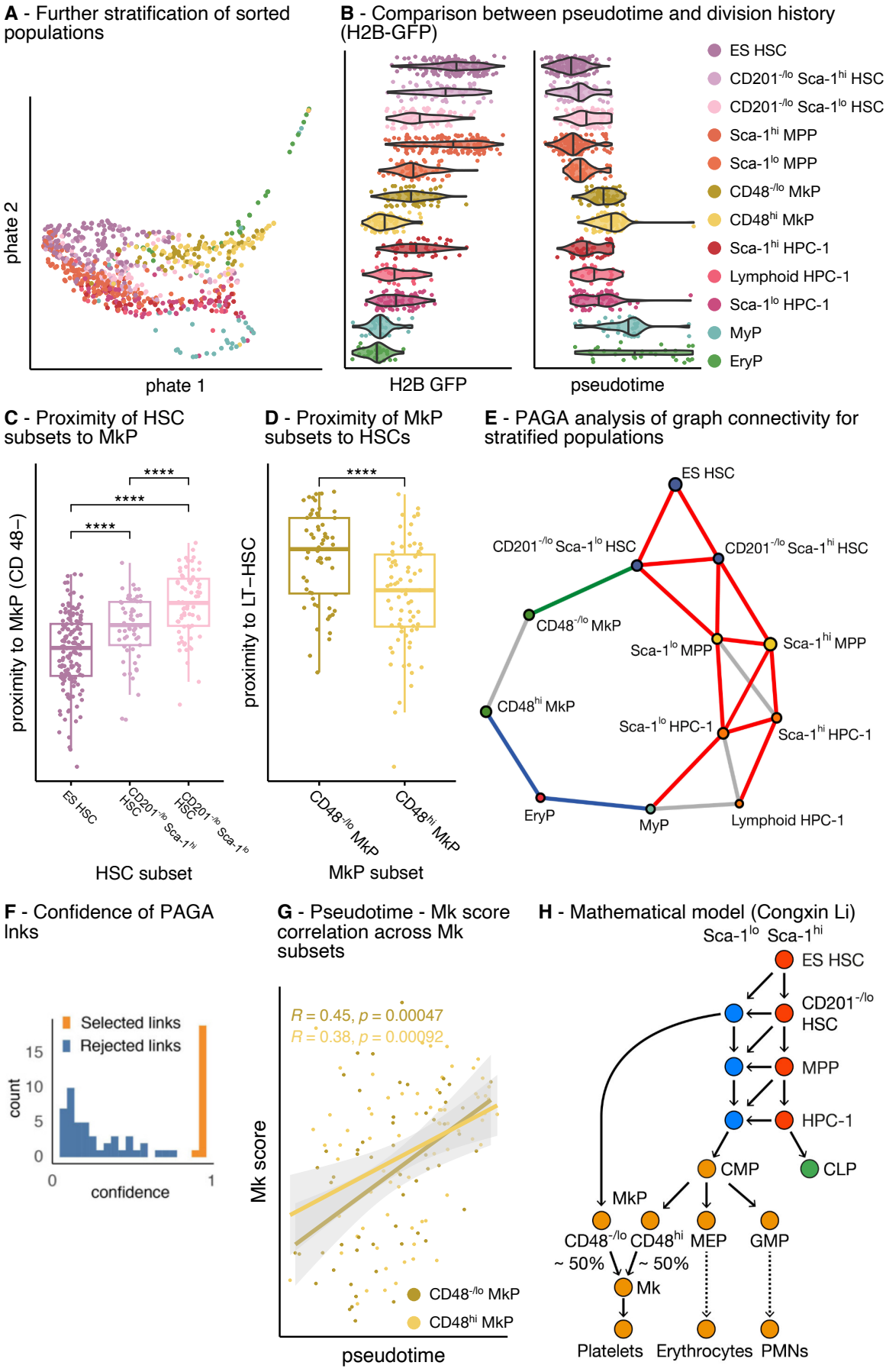

A - Further stratification of sorted populations
phate 2
phate 1
B - Comparison between pseudotime and division history (H2B-GFP)
H2B GFP
pseudotime
ES HSC
CD201-/lo Sca-1hi HSC
CD201-/lo Sca-1lo HSC
Sca-1hi MPP
Sca-1lo MPP
CD48-/lo MkP
CD48hi MkP
Sca-1hi HPC-1
Lymphoid HPC-1
Sca-1lo HPC-1
MyP
EryP
C - Proximity of HSC subsets to MkP
proximity to MkP (CD 48−)
****
ES HSC
CD201-/lo Sca-1hi HSC
CD201-/lo Sca-1lo HSC
HSC subset
D - Proximity of MkP subsets to HSCs
proximity to LT−HSC
CD48-/lo MkP
CD48hi MkP
MkP subset
E - PAGA analysis of graph connectivity for stratified populations
ES HSC
CD201-/lo Sca-1lo HSC
CD201-/lo Sca-1hi HSC
CD48-/lo MkP
Sca-1lo MPP
Sca-1hi MPP
CD48hi MkP
Sca-1lo HPC-1
Sca-1hi HPC-1
EryP
MyP
Lymphoid HPC-1
F - Confidence of PAGA lnks
Selected links
Rejected links
count
confidence
G - Pseudotime - Mk score correlation across Mk subsets
R = 0.45, p = 0.00047
R = 0.38, p = 0.00092
Mk score
CD48-/lo MkP
CD48hi MkP
pseudotime
H - Mathematical model (Congxin Li)
Sca-1lo Sca-1hi
ES HSC
CD201-/lo HSC
MPP
HPC-1
CMP
CLP
MkP
CD48-/lo
CD48hi
MEP
GMP
~ 50%
~ 50%
Mk
Platelets
Erythrocytes
PMNs

Figure 2.3

Figure 2.3 ***(previous page).*** **A**, Further stratification of sorted populations based on surface expression of Sca-1, CD201 (EPCR) and CD48. **B**, Comparison of the dilution marker H2B-GFP and pseudotime in stratified populations. **C**, Proximity of the HSC subsets to MkP population in transcriptional space, computed using Euclidean distance on the first 10 principal components. **D**, Proximity of MkP subsets to LT-HSC population in transcriptional space computed on first 10 principal components. **E**, PAGA connectivity graph of sorted populations. Link colours indicate whether the connections were confirmed (red) or disproved (grey). Blue links could not be tested; while the direct connection from HSCs to MkPs was highlighted in green. **F**, PAGA link confidence. **G**, Pseudotemporal trend for the expression of megakaryocyte markers in the MkP subsets. **H**, Mathematical modelling performed by Congxin Li integrating CD48-based MkP subsets reveals that each CD48 MkP subset produces roughly half of the Mks. **All panels in this figure are adapted from [54]**, data produced by Mina N.F. Morcos, Clara M. Munz, Congxin Li

2.6 MkP subsets express platelet markers, but different levels of myeloid markers

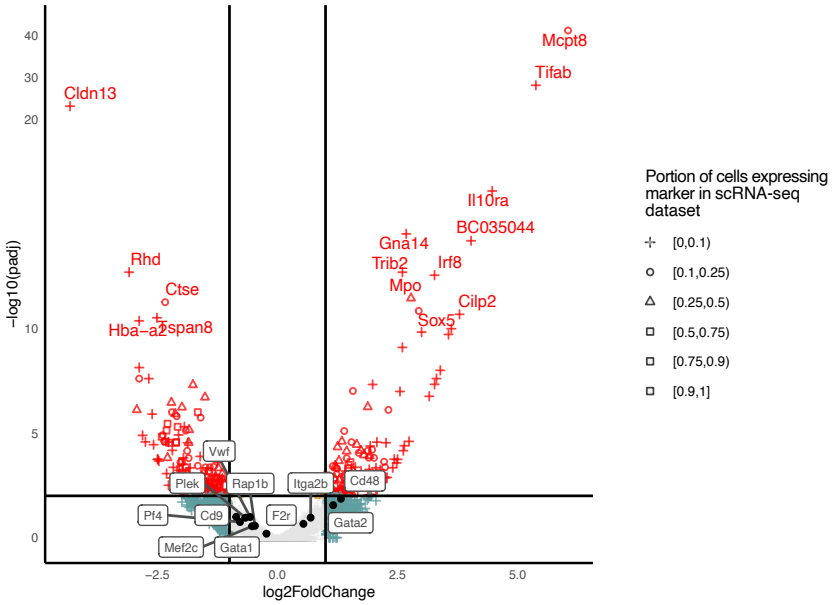
Once these MkP subsets were established, my focus shifted to identifying molecular features that distinguish them. To achieve this, I analysed bulk RNA-seq data of $CD48^{hi}$ and $CD48^{-/lo}$ MkPs generated and normalised using DeSeq2 [62] at the TU of Dresden and performed pathway enrichment analysis on the data. Differential expression analysis reveals that multiple myeloid markers, including *Irf8* and *Mpo*, were upregulated in $CD48^{hi}$ MkP; however, comparison with scRNA-seq samples indicated that these were only expressed in a small portion of cells. At the same time, the core markers of megakaryocytic differentiation, such as *Pf4*, *Cd9*, and *Vwf*, were expressed similarly across subsets, suggesting a comparable ability of producing platelets (Fig.2.4A). I used GSEA [63] to compare transcriptional profiles with Gene Ontology gene sets [64], and found that $CD48^{hi}$ MkP are enriched in genes associated with cell cycle (in agreement with the reported differences in H2B-GFP concentrations) and with epigenetic regulation (Fig.2.4B).

Next, the results of the differential expression analysis were put in the context of other haematopoietic cell types via comparison of expression profiles of the two subsets with relevant cell types from Msigdb. The results indicate a higher enrichment of $CD48^{-/lo}$ MkPs with respect to erythroid cell types, while $CD48^{hi}$ MkPs were enriched in myeloid signatures (Fig.2.4C). Comparison with curated gene sets from the Msigdb database revealed a strong enrichment of *Cbfa2t3* targets in both subsets, albeit in opposite direction (Fig.2.5A). This gene is a transcriptional co-repressor that recruits histone-modifying enzymes [65] and interacts with several master regulators of haematopoiesis such as *Gata1*, *Gata2* and *Tal1* during lineage specification of erythroid, megakaryocytic and myeloid lineages [66, 67].

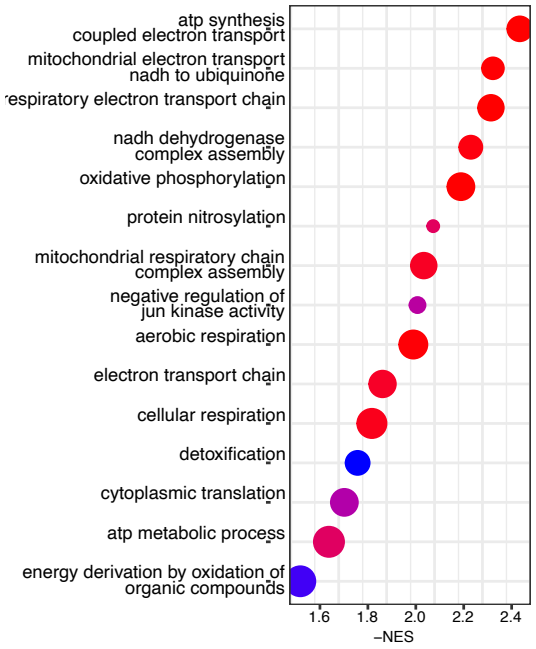
Lastly, I included a comparison with a recently published dataset that defines distinct Megakaryocyte subtypes, including a newly discovered subpopulation with immunologic

and phagocytic capabilities [68]: in agreement with the previous myeloid characterisation of $CD48^{hi}$ MkP, this subset shares a high similarity with the immunologic Mk subset (Fig.2.5B).

A - Differential expression analysis of CD48stratified MkP samples



B - GSEA enrichment of GO terms in CD48$^{-/lo}$ MkP



C - GSEA enrichment of GO terms in CD48^{hi} MkP

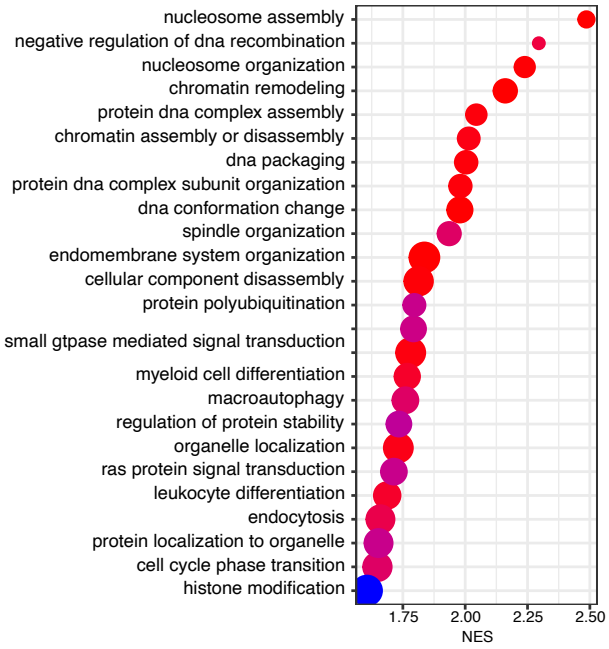


Figure 2.4. **A**, Differentially expressed genes between CD48$^{-/lo}$ MkP (left) and CD48^{hi} MkP subsets. The comparison was performed on bulk samples, while the shapes were assigned based on the percentage of cells expressing the markers in the scRNA-seq dataset. **B, C**, GO terms enrichment in MkP subsets. **All panels in this figure are adapted from [54]**.

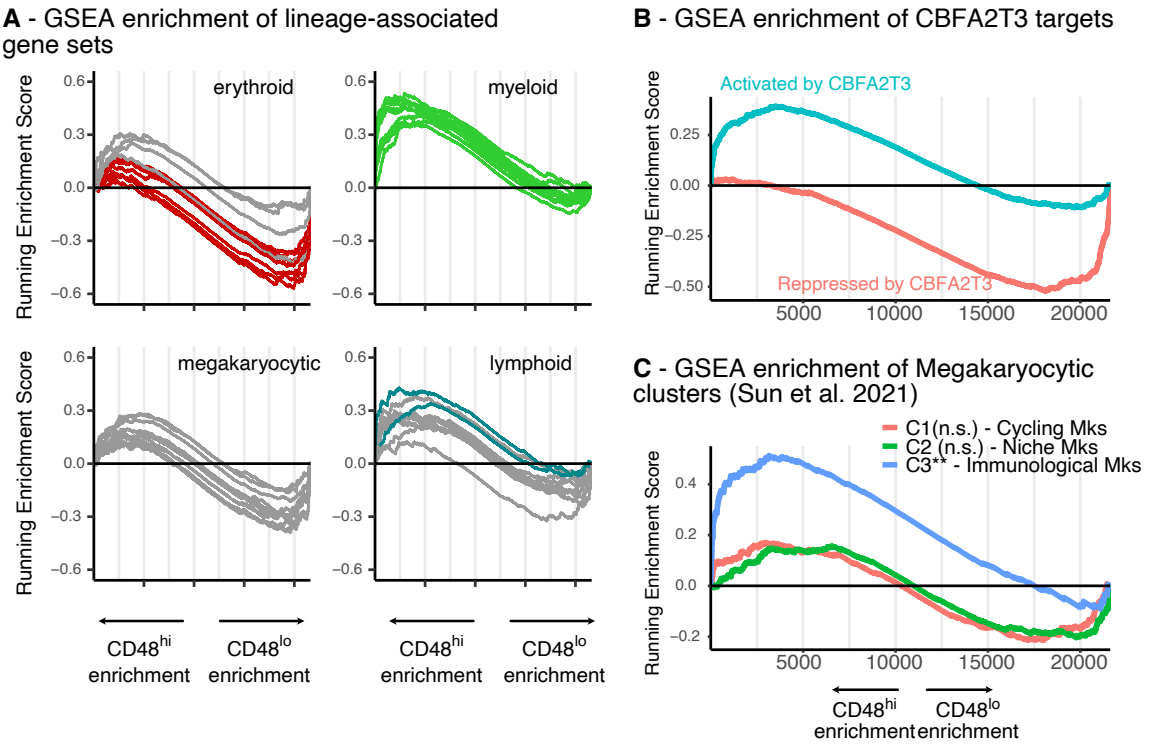


Figure 2.5. **A**, GSEA enrichment plots for terms related to haematopoietic lineages. Coloured lines are significant terms, while grey ones are not significant. **B-C**, GSEA enrichment plots of CBFA2T3 targets (**A**) and for Mk subsets detected in [68] (**B**). **Panel A in this figure is adapted from [54]**.

2.7 Discussion

The debate about the position of megakaryocytes within the haematopoietic system has gained traction in recent years: their similarity to HSCs in transcriptional [58], immunophenotypic [69], and labelling dynamics [14] has confounded the characterisation of their differentiation pathway. To address this, I integrated mathematical modelling by Congxin Li with transcriptomic analysis to uncover two native pathways to thrombopoiesis. Specifically, I employed the PAGA method on sorted cell populations rather than on transcriptionally defined clusters, which has proven successful in representing necessary conditions for a differentiation link between populations in this and subsequent works.

Transcriptional analysis of CD48-based MkP subsets has revealed that the two subsets share a substantial equivalence in platelet-related genes, but mostly differ in the expression of myeloid-related markers. There could be different explanations for this observation, listed below in increasing order of speculation: $CD48^{hi}$ MkPs are phenotypically close to some myeloid cell types, with which they share CD41 expression. It is possible that the gate is contaminated by these progenitors, resulting in differential expression of myeloid markers. This contamination would also explain the high significance enrichment of myeloid markers in $CD48^{hi}$ paired with a low number of expressing cells in the scRNA-seq dataset. Given that $CD48^{hi}$ MkPs share a developmental path with more myeloid cell types, expression of myeloid genes could be a residual trace of the transient myeloid potential of these progenitors. An increasing number of authors report the existence of novel Mk subsets that warrant a re-evaluation of the functions performed by this cell type. After initial reporting in [68], several works analysing Mk subsets have been published [70, 71], including a paper identifying CD48 itself as a marker of immune Mks that assume immunomodulatory functions during infection [72]. This finding resonates with results from this Chapter, which assigns to $CD48^{-/lo}$ MkPs the complementary role of platelet generation upon challenges. However, evidence for Mk heterogeneity is still accumulating, and further validation of these findings is needed before considering its implications.

In general, the integration of index-sorted transcriptomic datasets with fate mapping evidence presents a powerful approach to enhance our understanding of haematopoietic populations and their differentiation pathways.

Chapter 3

Single-cell multiomic analysis of haematopoiesis differentiation: comparison between chromatin and transcriptional landscapes

3.1 Introduction

In the previous chapter, the integration of transcriptional data with fate mapping experiments enabled new insights into haematopoietic differentiation.

Among the numerous mechanisms that contribute to the molecular decisions that determine fate commitment, epigenetic regulation of gene expression is considered one of the most important [36]. As described in the Introduction, the scATAC-seq can be used to read out the net contribution of epigenetic modifications to the chromatin configuration of single cells [39].

Until now, due to technological limitations, most single-cell sequencing studies have focused on transcriptional profiling. However, the transcriptional and chromatin states are strongly interdependent: to enable transcription, chromatin needs to be open; in turn, transcription factors and other chromatin factors can cause DNA folding or unfolding around nucleosomes in relevant regions [41]. While gene expression profiles provide a detailed snapshot of several molecular processes at work in a cell, chromatin configurations capture the large-scale effects of cis-regulatory elements affecting the genome, rendering it an ideal technology to investigate differentiating systems, in which enhancer dynamics play a prominent role. Indeed, chromatin profiles obtained using ATAC-seq have often been reported to be more accurate than RNA-seq in describing differentiation trajectories [43, 73].

Multi-omic datasets, in which multiple omic layers are sequenced from the same biolog-

ical sample, offer a window into the study of how gene expression and epigenetic landscapes are related to each other.

Applied to single-cell paired datasets (in which sequencing of both RNA and ATAC is performed on the same cells), multi-omic analysis has fuelled research into the field of trajectory inference, lineage priming and gene regulatory network inference [74, 75].

While each of these topics will be addressed in depth in the following chapters, this chapter focuses on the exploratory analysis of an HSPC multi-omic dataset obtained using the commercially available 10x multi-omics kit. Mouse experiments were kindly performed by Larissa Frank (Division of Cellular Immunology, DKFZ), while Jonas Metz and Nina Claudino (Division of Theoretical Systems Biology, DKFZ) handled sample processing and sequencing using the 10x multiomic protocol. I contributed by designing the experiment and performing all the computational analyses shown in this Chapter.

In particular, I focused on the detection and quantification of the differences between the RNA and ATAC landscapes. A typical example is proliferation, a process reported to affect gene expression more prominently than chromatin configuration [74]. This observation is confirmed in this dataset, whose RNA layer clearly clusters based on the proliferative state of cells, while the chromatin landscape remains substantially unaffected. I leveraged this difference to select the most appropriate cell cycle removal method for the transcriptome.

Aside from the cell cycle, gene expression and chromatin landscapes are overall very similar, displaying the same branching behaviour in correspondence of lineage decisions.

Interestingly, I observed some other differential states in both layers: a signalling state that can only be observed in transcriptional features, and a chromatin cluster dominated by high CTCF motif enrichment that is undetectable in gene expression space.

Despite the mostly technical nature of the work presented here, the focus is on the biological annotation of the data; thus bioinformatic aspects are discussed in detail in the Methods section.

3.2 Experimental design and pre-processing

In previous experiments, the primary objective of immunophenotypic gating was to obtain highly pure cell populations with large cell numbers. This approach was adopted to ensure that the experimental results were not confounded by the presence of other cell types. Consequently, the gates were intentionally designed to be non-contiguous, to minimise the risk of contamination by unwanted cell types.

However, this gating strategy may result in inadvertent loss of information from cells that reside in the transitional zones between two gates. To address these issues, the experiment was designed to achieve two primary objectives: first, enrich rare cell types, such as long-term haematopoietic stem cells (LT-HSC), and second, sample all intermediate cell states spanning the range of phenotypes, i.e., not leaving any gaps between gates (Fig.3.1A, Fig.3.1B). This gating strategy preserves the definition of LT-HSCs, ST-HSCs, and LS-K as in [5], while extending the MPP phenotype to $CD150^{+}CD48^{+}$ LSK cells, defined as MPP2 in [76]. Importantly, sampling cells from all the intermediate $LS^{-}K$ regions improves the performance of trajectory inference algorithms, as described in more detail in the next chapters.

Sequencing of the samples was performed simultaneously to minimise batch effects. Alignment using CellRanger yielded similar values in terms of sequencing saturation, library size, and the number of features detected in all samples (Fig.3.1B, Fig.3.1C). Once aligned reads are available, obtaining feature matrices (i.e., a matrix containing quantification of each feature in single cells) for RNA-seq and ATAC-seq requires separate pipelines (sketched in Fig.3.1D, Fig.3.1E more in-depth in Methods). From a computational point of view, the scRNA-seq literature has developed accurate statistical models to select features and normalise counts [77, 78], while scATAC-seq features are much more sparse, and preprocessing is based on techniques adapted from text mining [79]. At the same time, some core biological differences set the two layers apart: while it is relatively straightforward to obtain a gene expression matrix in RNA-seq, ATAC-seq reads can generate multiple feature matrices with information regarding transcription factor motif enrichment, genome tiles or peaks accessibility, or gene body accessibility (reviewed in [80]). In my work, I used either of these matrices based on the biological question at hand.

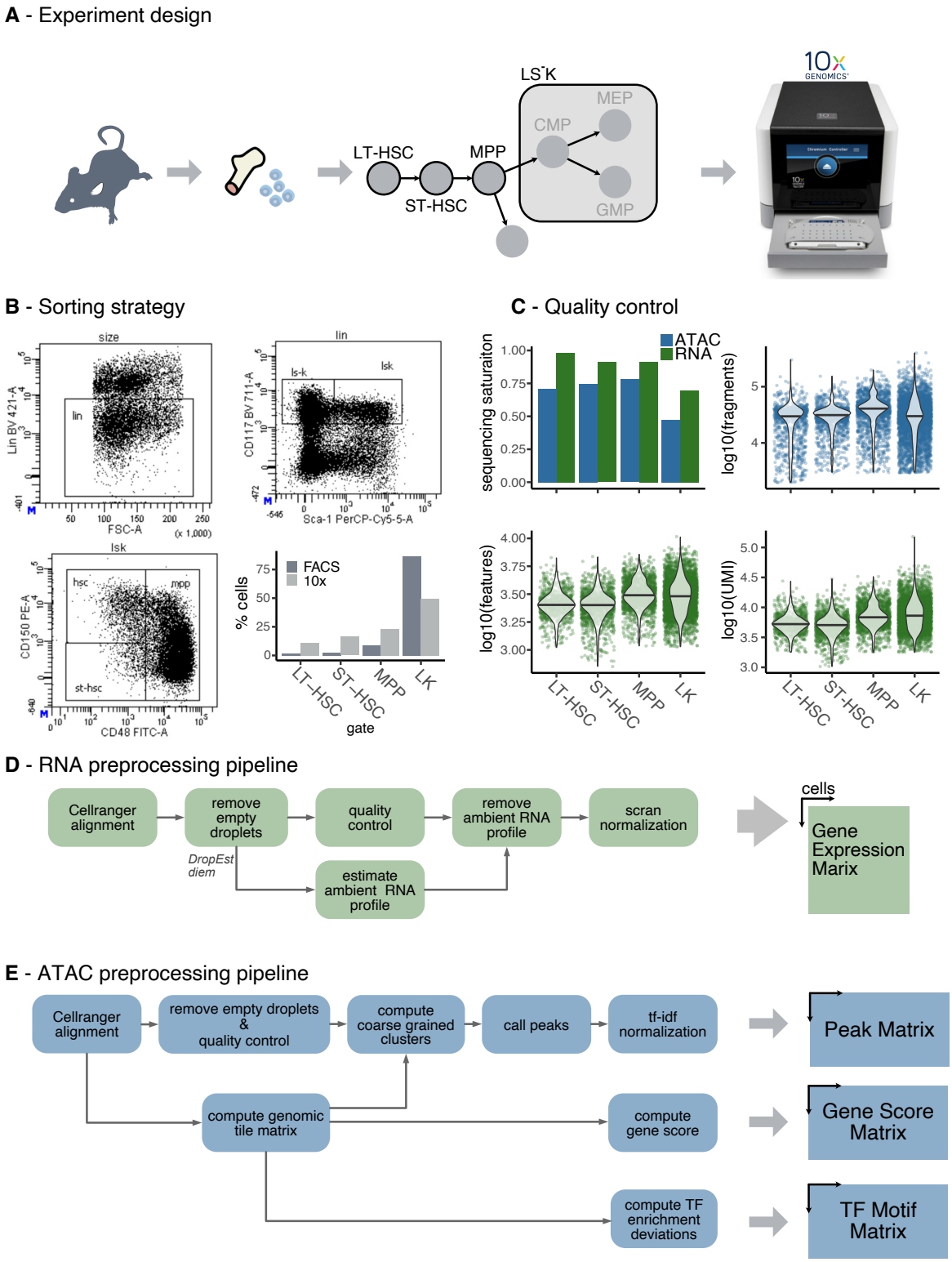

A - Experiment design
LT-HSC
ST-HSC
MPP
LSˉK
MEP
CMP
GMP
10x GENOMICS
B - Sorting strategy
size
lin
FSC-A
Lin BV 421-A
lsk
ls-k
CD117 BV 711-A
Sca-1 PerCP-Cy5-5-A
hsc
mpp
st-hsc
CD150 PE-A
CD48 FITC-A
FACS
10x
% cells
gate
C - Quality control
ATAC
RNA
sequencing saturaiton
log10(fragments)
log10(features)
log10(UMI)
D - RNA preprocessing pipeline
Cellranger alignment
remove empty droplets
DropEst diem
quality control
estimate ambient RNA profile
remove ambient RNA profile
scran normalization
cells
Gene Expression Marix
E - ATAC preprocessing pipeline
Cellranger alignment
remove empty droplets & quality control
compute coarse grained clusters
call peaks
tf-idf normalization
Peak Matrix
compute genomic tile matrix
compute gene score
Gene Score Matrix
compute TF enrichment deviations
TF Motif Matrix

Figure 3.1

Figure 3.1 ***(previous page).*** **A**, Experiment design. Bone marrow cells are harvested from a male adult mouse, sorted into LT-HSCs, ST-HSCs, MPPs and LS$^-$K and sequenced using the 10x multiomics kit. The experiment was carried out by Larissa Frank, Nina Claudino, and Jonas Metz. **B**, Cells were sorted in broad gates, so there was no phenotypical gap between populations. *Bottom right*: proportion of cells in each gate in the FACS experiment, compared to the number of cells in the 10x experiment. The frequency of LSK populations is enriched to compensate for their paucity in the bone marrow. **C**, Quality control plots for RNA (green) and ATAC (blue) libraries across samples. ATAC fragments, RNA library size and number of RNA features are comparable across the samples, while the saturation is slightly lower in the LS$^-$K gate. **D,E**, Diagram of bioinformatic processing of RNA and ATAC data to obtain feature matrices. While both pipelines involve steps of quality control, normalisation, and feature selection, multiple feature matrices can be extracted from ATAC reads.

3.3 Cell cycle signature distinguishes chromatin and transcriptional landscapes

The first analysis I performed on the data was a comparison between the manifold structure of the RNA and the ATAC landscape. Initial visualisation using UMAP embeddings obtained independently from each layer shows how the structure of the data is overall similar: a progression from LT-HSC to MPP culminates in the highly heterogeneous progenitor gate, in which branching toward mature cell types appears prominent (Fig. 3.2A). However, some differences between UMAP embeddings are also noticeable. In particular, the RNA embedding features a group of cells from the LSK gates that occupy an intermediate region between the LSK and LS$^-$K progenitors.

The colouration of cells based on transcriptional inference of cell cycle activity highlights the impact of proliferative status on transcriptional, but not chromatin state, as previously reported in the multiomics literature [74]. To compare the two layers while accounting for this effect, it is necessary to remove the cell cycle signature.

A common approach to achieve this is through linear regression models. However, this method may inadvertently remove correlates of cell cycle status that hold significant biological insights, particularly in developmental and differentiation systems. When applied to this dataset, linear correction for cell cycle appears to successfully remove the cycling cluster, although not entirely, as cycling cells are still more clustered than in the ATAC embedding (Fig. 3.2C).

Alternatively, Independent Component Analysis (ICA) can be used to decompose data into independent sources of variation. This technique has proven to be effective in isolating cell cycle signatures in transcriptomics data [81]. In the present dataset, while the association of principal components with the proliferative signature is moderately present in more components that are also associated with other biological signals, two ICs can be directly

associated with S and G2M signatures (Fig. 3.2 D, Fig. 3.2 E), allowing cell cycle removal simply by excluding these components in subsequent analyses.

To measure the extent of the correction more quantitatively, I measured the assortativity (i.e., the tendency of vertices in a network to be connected to similar vertices, defined in [82]) of the cell cycle phase in the neighbourhood graph, confirming that ICA-based cell cycle correction outperforms linear methods to regress out cell cycle (Fig. 3.2F).

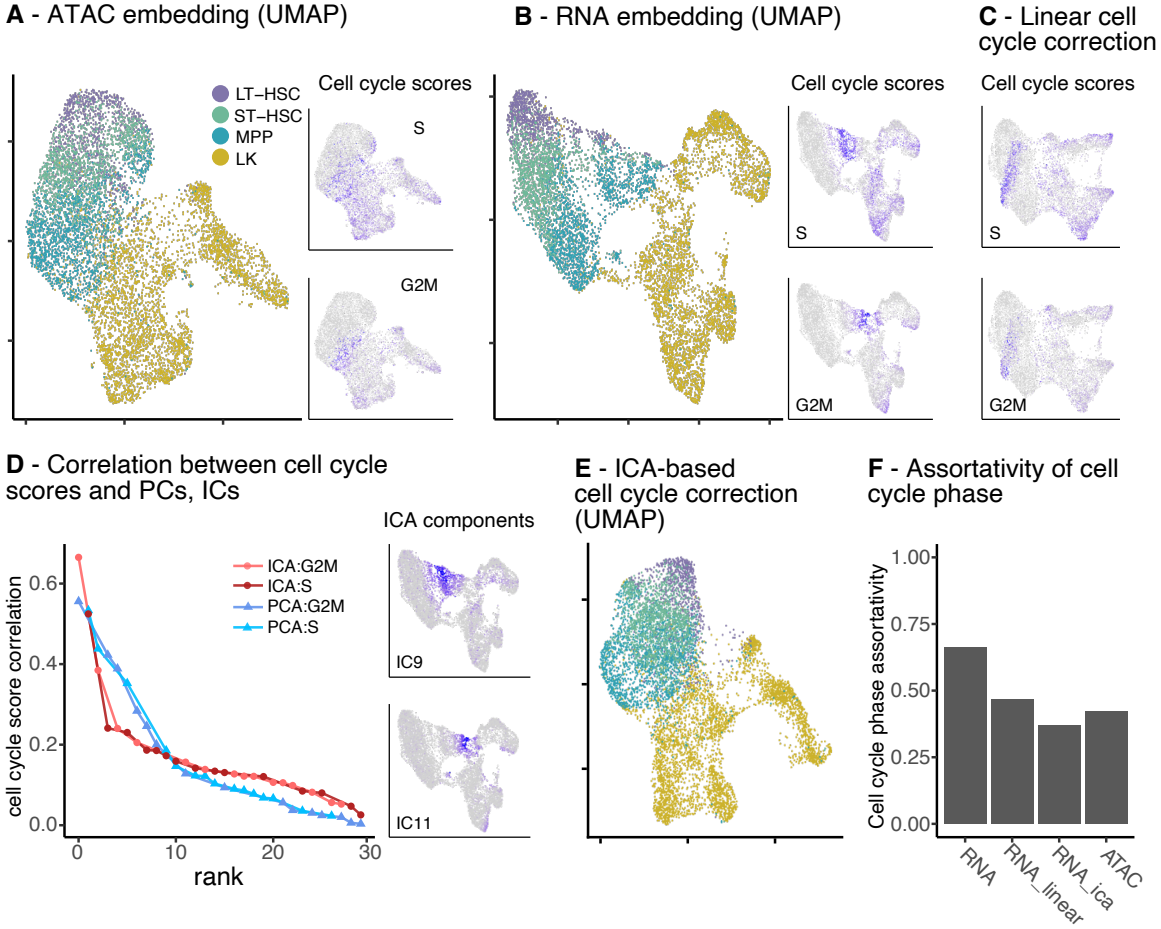


Figure 3.2. **A**, *Left*: UMAP embedding of Peak Matrix, coloured by sorted population. *Right*: Transcriptional estimate of cell cycle activity. **B**, UMAP embedding of gene expression matrix, coloured by sorted population. *Right*: Transcriptional estimate of cell cycle activity. **C**, UMAP embedding of GEX matrix after removal of cell cycle signature using a linear method. **D**, *Left:* Correlation between linear dimensionality reduction coordinates and transcriptional estimate of cell cycle activity: while multiple principal components are correlated with proliferative score, each phase-specific score can be assigned to one specific independent component. *Right*: RNA embedding coloured by projection of ICs maximally correlated with cell cycle. The patterns obtained closely resemble the cell cycle scores shown in panel **B**. **E**, UMAP embedding of GEX after exclusion of cycle-related independent components. **F**, Assortativity of cell cycle phase in neighbourhood graph across different cell cycle removal strategies, compared with ATAC baseline, in which cell cycle is not expected to influence the embedding.

3.4 Quantitative comparison yields modality-specific states

Once cell cycle has been successfully removed, I investigated the similarities between the RNA and ATAC landscape in higher dimensions. To achieve this, I used Diffusion Maps, a non-linear dimensionality reduction technique that models the dataset as a diffusion process and is particularly suitable for capturing differentiation-like dynamics [18]. Of note, differently from UMAP, t-SNE, and PHATE, the purpose of this method is to summarise the data in a lower dimension space, rather than 2D visualisation. While similar to PCA and ICA in this purpose, its non-linear nature allows to de-noise the data and obtain a low-dimensional representation that reconstructs the geometrical manifold underlying the data.

1. I operated on the premise that if RNA and ATAC show similarities, then it is feasible to predict the diffusion map coordinates of one based on the other. To put this into practice, I used a random forest algorithm to predict RNA coordinates from ATAC data, and vice versa, and then I quantified the accuracy of these predictions using the r-squared statistic (see Methods for a detailed description). As shown in Fig.3.3A, it is in general possible to reconstruct the diffusion coordinates of a cell in one assay using the coordinates of the other one, although less precisely for higher components. A noticeable exception is represented by RNA:DC5, which seems to have no relation to ATAC diffusion components. To investigate this, I visualised the RNA:DC5 coordinate as a colour scale on RNA UMAP coordinates, revealing that this component captures the variation associated with a small subset of LSK cells in RNA coordinates.

The existence of a mapping between RNA and ATAC for each diffusion component is not sufficient to conclude that the two landscapes are equivalent. It can be the case, indeed, that regions of high cell density in one landscape might not correspond to regions of high density in the complementary one. When this condition occurs, clusters that are detected using one assay will not be detected using the complementary one. Therefore, I obtained clusters in both RNA and ATAC, annotated them, and thus compared their homogeneity in the complementary assay (see Methods).

Since the two datasets are preprocessed using different methods (ICA for RNA, LSI for ATAC), I opted for clustering on diffusion map coordinates to ensure that neighbourhood graphs across assays are similarly denoised. Clusters obtained in early haematopoiesis populations are generally poorly separated and not suited for the description of the slow kinetics of lineage emergence. For this reason, I did not put emphasis on a precise annotation of these clusters but rather used them to compare high-density regions of the transcriptional and chromatin landscape. An in-depth analysis of lineage emergence in the dataset is presented in Chapter 4.

Clusters obtained using the Leiden graph clustering algorithm on RNA diffusion components are distinct but overlap in RNA coordinates (Fig.3.3B). Marker detection yields a set of genes that enable annotation of states such as early stem (rna5, rna3, rna12), lympho-myeloid (rna1), lymphoid (rna10), myelo-erythroid (rna2), myeloid (rna4, rna6, rna9), megakaryocytic (rna11), erythroid (rna7, rna8). Similarly, I computed clusters using ATAC coordinates and detected TF motifs associated with each cluster, generating the annotation of the stem (atac7, atac9, atac1), lympho-myeloid (atac3), myeloid (atac4), and lineage-specific progenitor clusters (Fig.3.3C).

Next, I sought to compare the homogeneity of these clusters between RNA and ATAC by measuring cluster connectivity in the nearest-neighbourhood graph. To do so, I used modularity, a graph metric that quantifies the significance of links across clusters compared to a null connectivity model, returning a normalised connectivity measure for each pair of clusters [22]. In graph-based approaches such as the one I utilised, cluster solutions are determined by maximising the number of connections within a cluster while minimising the number of intra-cluster connections. Consequently, the clusters computed on RNA can be regarded as optimally separated on the RNA neighbourhood graph; however, these clusters are not optimally separated on the ATAC graph. As a result, RNA clusters are anticipated to exhibit greater connectivity in the ATAC graph. Accordingly, in Fig.3.3D, chromatin connectivity across clusters is consistently higher than in the transcriptional counterpart and vice versa (Fig.3.3E).

Aside from this global trend, I observed a remarkable difference in connectivity in two occurrences:

- rna12 links to rna1 and rna3 are sensibly higher in ATAC than in RNA, indicating that this transcriptional state is not discernible in ATAC. This observation is in line with the results from the previous section, which assigned low mappability between RNA:DC5 and ATAC components;

- atac7 displays a large number of RNA links to cells from the atac1 cluster, indicating that these two clusters are virtually indistinguishable in RNA coordinates. The lack of a distinguishable transcriptional state associated with atac7 contributes to the high ATAC connectivity observed in clusters rna1, rna3, and rna5, as numerous cells within these RNA clusters belong to the atac7 cluster (Fig.3.3F).

To interpret the features that distinguish these cells from their immediate neighbours, I first compared the clusters with the most connections to rna12 and atac7 in the complementary modality (Fig.3.3G, Fig.3.3I, see Methods), thus performed marker detection using the neighbouring clusters as contrast. The vast majority of genes associated with rna12 are related to interferon response (*Iigp1*, *Stat1*, *Ifi44*, *Ifi203*), indicating that these cells are re-

sponding to interferon-mediated inflammatory signals (Fig.3.3H). A cluster with a very similar transcriptional signature has been described in response to poly(I:C) in [83] and steady state [84].

On the other hand, atac7 is clearly defined by a strong CTCF motif enrichment (Fig.3.3J). This DNA-binding protein is involved in the formation of long-range chromatin interactions and is often associated with the formation of promoter-enhancer complexes. In the context of haematopoiesis, it has been associated with several key transcription factors, such as *Runx1* and [85], *Tal1* [86], linked to erythroid development [87] and both activation [88] and quiescence [89] in HSPCs.

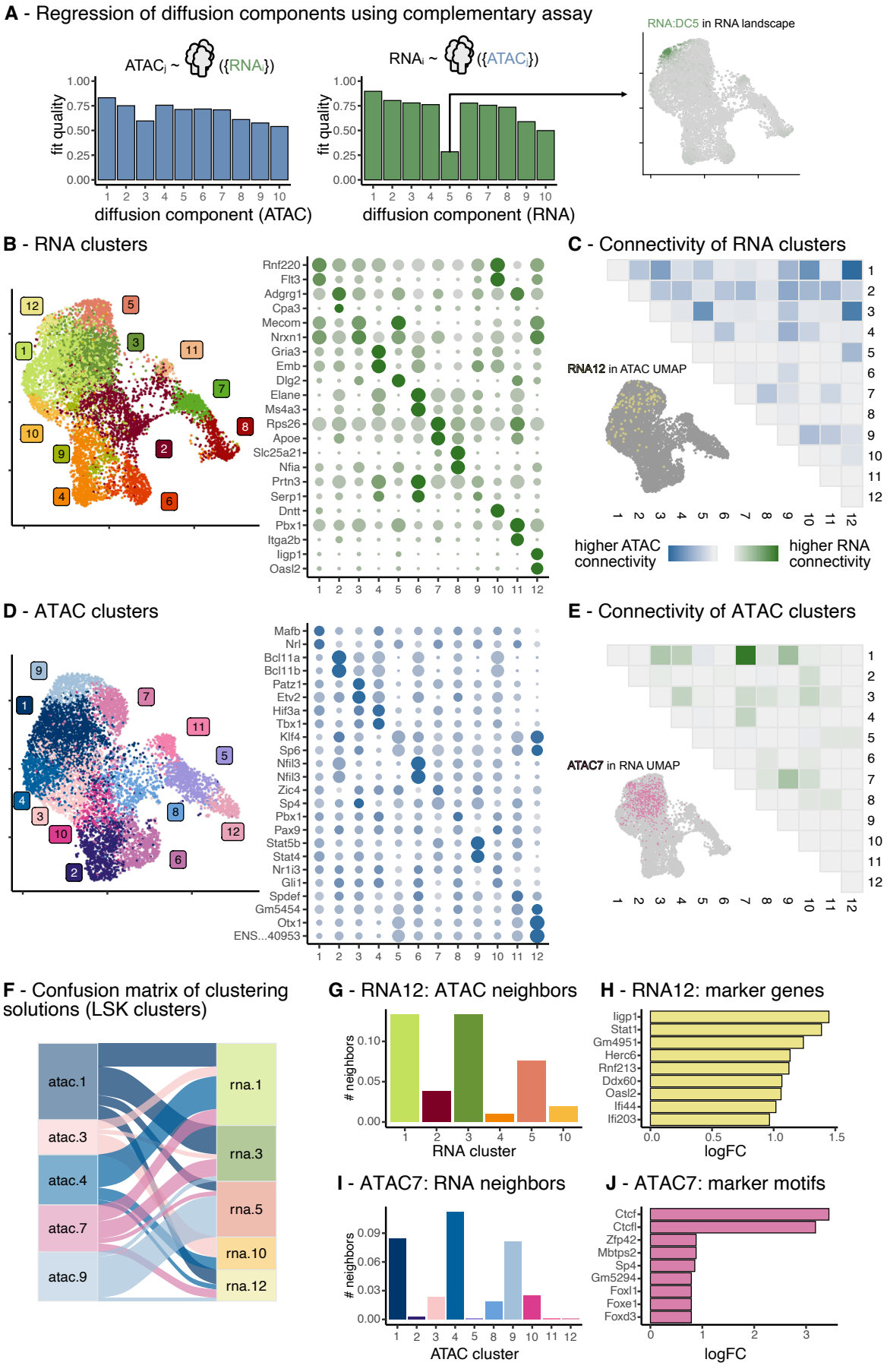

A - Regression of diffusion components using complementary assay
ATAC_j ~ ({RNA_i})
RNA_i ~ ({ATAC_j})
fit quality
diffusion component (ATAC)
diffusion component (RNA)
RNA:DC5 in RNA landscape
B - RNA clusters
C - Connectivity of RNA clusters
RNA12 in ATAC UMAP
higher ATAC connectivity
higher RNA connectivity
D - ATAC clusters
E - Connectivity of ATAC clusters
ATAC7 in RNA UMAP
F - Confusion matrix of clustering solutions (LSK clusters)
atac.1
atac.3
atac.4
atac.7
atac.9
rna.1
rna.3
rna.5
rna.10
rna.12
G - RNA12: ATAC neighbors
neighbors
RNA cluster
H - RNA12: marker genes
logFC
I - ATAC7: RNA neighbors
ATAC cluster
J - ATAC7: marker motifs

Figure 3.3

Figure 3.3 ***(previous page).*** **A**, *Left*: Each ATAC diffusion component is regressed using a random forest algorithm trained on RNA components. The quality of each of each regressor is reported using the mean of the R-squared metric between the target values (ATAC diffusion components) and their prediction based on RNA diffusion components. *Centre*: Same as previous, swapping RNA and ATAC. Diffusion component 5 cannot be regressed using ATAC information. *Right*: Visualisation of DC5 from RNA assay on RNA landscape. **B**, Clusters detected in the GEX assay and their relative markers. Dot size indicates detection percentage, while dot colour refers to the average expression value. **C**, Difference in the connectivity of RNA clusters between the RNA and ATAC neighbourhood graphs. Given that RNA clusters minimise intra-cluster connectivity in the RNA layer, more connections between clusters are detected in the ATAC neighbourhood graph. In particular, cluster rna12 has numerous connections with cluster 1 and cluster 3 in the ATAC layer, indicating poor detectability of this cluster in the ATAC layer. **D** Clusters detected in the ATAC assay and their relative marker motifs. **E**, Difference in the connectivity of the ATAC clusters between RNA and ATAC neighbourhood graphs. The atac7 cluster shows a high number of connections to cluster atac1 in the RNA layer. **F**, Comparison of LSK clusters across assays. Each line represents a single-cell and connects its own classification in ATAC and RNA. The high mixing between clusters indicates that there is no clear correspondence between RNA and ATAC clusters in this region of the landscape. **G**, For cells in poorly rna12, the RNA cluster of its ATAC neighbours is shown. **H**, logFC of top significant marker genes for rna12. **I**, Same procedure as in **G** for atac7. **J**, logFC of enriched atac7 motifs.

3.5 Discussion

In this chapter, I introduced a coarse grain description of the multi-omic data generated using 10x, which will serve as a basis for the inference of trajectories and gene regulatory networks, and quantification of lineage priming. By pre-processing the available modalities separately, I was able to compare RNA and ATAC landscapes and annotate them.

First, proliferative status has been shown to impact the transcriptional state, but not its chromatin counterpart, as previously reported in [74]. I leveraged this observation to use the ATAC-seq state as a benchmark for evaluating various cell-cycle correction methods. The chosen ICA-based correction that achieves the removal of proliferation-associated clusters while retaining secondary groups of cycling cells was also observed in ATAC-seq embeddings.

Mitigation of the proliferative signature is of primary importance, as it allows for the evaluation of the differential impact of other processes on each modality. Of note, the availability of paired data is only a recent development in multiomics. Due to the predominance of unpaired datasets, multi-omic methods have so far chiefly focused on aligning different modalities based on similarity, rather than detecting differences across layers [90]. In this work, the availability of paired data allowed me to investigate subtle differences between the transcriptional and chromatin layer. To do so, the random-forest regressor approach shown here is well-suited for evaluating how much information is shared across the two modalities, and can be extended to compare any kind of cell-level annotation across modalities in any

multi-omic dataset.

Mapping single diffusion components is necessary, but not sufficient, to establish the equivalence of the transcriptional and chromatin landscape, since co-localisation across multiple components can still differ across layers. This prompted me to compute clusters and compare their connectivities using modularity-based metrics. Both clustering solutions offer a description of the main lineages included in the dataset: myeloid, megakaryocytic, erythroid, and lymphoid. Furthermore, the comparison of cluster connectivities allowed the annotation of modality-specific states.

The RNA cluster associated with interferon response has recently been described in [84]. In addition to its well-documented role during inflammation [91], interferon signaling plays a role in tuning differentiation and proliferation during homeostasis [92]. The low coherence of this cluster in ATAC coordinates suggests that the transcriptionally detected signaling state has not affected cells at the level of chromatin organisation, possibly due to the short timescale of such a mechanism. However, one of the clusters detected on the chromatin coordinates has no transcriptional equivalent. Interestingly, this cluster has a strong association with CTCF, a transcription factor that plays a key role in the formation and stabilisation of loops that define interacting DNA domains and has been associated with some key transcription factors in haematopoiesis [85, 86]. It is possible that the chromatin rearrangement mediated by CTCF factor has not propagated to the transcriptional level, thus resulting in poor detection of this cluster in RNA coordinates. Further analysis of this cluster will be performed in the context of trajectory inference and gene regulatory network inference in the following chapters.

Chapter 4

Trajectory inference quantifies hierarchy and branch emergence in haematopoiesis

4.1 Introduction

The current notion of haematopoiesis as a deeply hierarchical system in which multipotent stem cells are capable of differentiating into all types of mature blood cells emerged gradually in the second half of the XXth century [55]. These findings were obtained chiefly through transplantation experiments, followed by iterative refinement of the definition of the tip stem cell population by stratifying progenitors based on the expression of surface markers. More recently, technological advances introduced the possibility of assessing the fate of the line without the need for transplantation by using in-vivo lineage tracing [93]. On the other hand, the immunophenotypic description of progenitor populations has been combined with single-cell sequencing technology to explore the potential heterogeneity hidden in FACS-defined populations [94]. In particular, the tree-like topology, in which each population descends from one compartment only, has been questioned by lineage tracing evidence that showed how megakaryocytes [14,95], basophils [96], and monocytes [27] are likely to have more than one source.

On the other hand, obtaining putative lineage relationships from single-cell sequencing data is not as straightforward: the task of inferring and parametrising the most likely topology to generate a dataset has caused the flourishing of a class of algorithms known as trajectory inference, which aim to order single cells based on a putative developmental time (termed pseudotime) and infer a graph-like structure that recapitulates the developmental relationship between populations [50]. Many such methods rely on clustering to reduce the complexity of the dataset and define the cell types that constitute the progression from stem to terminally differentiated cells [21,22]. However, the discrete labels produced by clusters are resolution parameter-dependent and suboptimal for describing the various fate combina-

tions that can occur within the haematopoietic system. Alternatively, cluster-free methods model differentiation as a stochastic random walk on the neighbourhood graph and allocate a differentiation probability into terminal lineages for each cell [23, 24, 97]. Although these methods accurately depict continuous states within the haematopoietic landscape, they lack a straightforward way of grouping cells into branches. Moreover, all of these methods are based on the underlying assumption that the measured heterogeneity is the result of differentiation. In the previous chapter, I showed how the interpretation of transcriptome data is affected by processes that are not directly related to differentiation, such as cell cycle progression and interferon signalling. By contrast, ATAC-seq is sensitive to processes that are likely involved in lineage commitment, such as enhancer dynamics and CTCF-mediated chromatin remodelling reported in the previous chapter. Another issue that affects trajectory inference in haematopoiesis is the dependence of the observed landscape on the relative abundance of the populations that make up the data. In the haematopoietic field, this results in a trade-off between an unbiased sample of the progenitor gate, resulting in a data set dominated by committed progenitors [58, 95, 98] or data in which the upper compartments are enriched [99–101], resulting in unnaturally dense neighbourhood graphs that confound probabilistic trajectory inference.

In this Chapter, I used a novel subsampling strategy to overcome these limitations and obtain a reliable probabilistic description of lineage commitment on the chromatin data, resulting in the detection of an early metastable stem cell state. I validate the outcome of the trajectory inference pipeline using transcriptional mapping to fate mapping and transplantation datasets. Next, I compare RNA-derived cell cycle estimates with fate probabilities to uncover the impact of proliferation on lineage choice, resulting in the detection of the first molecular bifurcation between lymphoid fate and proliferative activation. Additionally, I combined trajectory inference with a newly designed procedure to iteratively assign lineage potentials to single cells and quantify the association between mature lineages along the haematopoietic system. I used statistical testing to infer a quantitative description of haematopoietic topology that is close to a fully hierarchical lineage restriction model, with the notable exception of the basophil and monocyte lineage, which are statistically associated with multiple lineages.

As in the previous chapter, I selected the content for this chapter to emphasise the biological relevance of the reported results. A more in-depth discussion of the technical and computational aspects of the analysis is contained in the Methods appendix.

4.2 Detecting of an early metastable state by subsampling of HSPC populations

As a preprocessing step for trajectory inference, I considered the impact of the gating strategy on the estimation of the transition matrix. The FACS experiment was designed to acquire a substantial number of LSK cells, notwithstanding their rarity in the bone marrow, to increase the statistical power of the analysis. However, this approach results in the overrepresentation of these cell populations within the dataset (4.1A). This disproportion can negatively impact the performance of probabilistic trajectory inference methods, as they may incorrectly assign elevated transition probabilities towards these states owing to their artificially increased density in the landscape. To overcome this limitation and preserve the statistical power offered by a high number of LSK cells, I first subsample the dataset based on FACS proportions between cells to compute pseudotime and fate probabilities. Subsequently, the results were projected back on the complete data set using the shared nearest neighbourhood (SNN) graph (see **??**, Fig. 4.1B). Importantly, feature selection is not repeated on the subsampled dataset, as the paucity of stem cells would penalise features that distinguish them. Next, I used the diffusion pseudotime method to estimate the succession of states that follow an initial one, detected here using the diffusion component maximally correlated with stem markers (see 6.3). Two features of the pseudotemporal distribution of cells stand out (Fig.4.1C):

- As cells differentiate, the pseudotemporal gap between gated populations becomes narrower. This aligns with endogenous fate mapping kinetics, which suggests that it takes much longer for label equilibration to occur from LT-HSCs to ST-HSCs than it takes from ST-HSCs and MPPs [102];

- An initial peak in cell density, comprising a subset of LT-HSCs (48% of the total), is followed by a monotonic increase in cell numbers downstream. In unbiased sampling conditions, cell densities reflect the speed at which cells transition towards higher pseudotime [18]. This means that the early high-density region represents a metastable state associated with stem cells.

In summary, the sampling strategy described in this paragraph enables the detection of an early metastable stem cell populations that reconciles single-cell sequencing description with evidence from fate mapping.

4.3 Trajectory inference unveils lineage emergence within a short pseudo-temporal window

Next, I investigated the proliferative and differentiative behaviour of cells along pseudotime. To do so, I compared pseudotime against transcriptional proliferation estimates and differentiation probability. Of note, this analysis is viable only on chromatin data, as proliferative status impacts pseudotemporal estimates on scRNA-seq data, confounding proliferation and lineage emergence signals.

The proliferative trend in pseudotime is characterised by a gradual increment in the portion of cycling cells in proximity of the ST-HSC/MPP transition. I fitted this trend using monotonic splines (see Methods) and reported the half-saturation constant in Fig.4.1D.

Next, I performed trajectory inference using cellRank, which enables the detection of terminal states and computation of fate probabilities by integrating nearest-neighbourhood graph information with cell-ordering information such as pseudotime (in this case diffusion pseudotime [18]) to infer a directed cell-to-cell transition matrix (see Methods). I chose to compute these results on the ATAC layer, as it is the most sensitive to differentiation-relevant processes such as enhancer dynamics and chromatin remodelling [103, 104].

When plotted against pseudotemporal ordering, probability trends for single lineages exhibit a consistent pattern: early cells display a flat 'baseline' probability, followed by increasingly variable values that generate a bifurcation between cells with a near-certain differentiation probability and cells that commit to other lineages, for which the chances of differentiating into the considered lineage reach zero (Fig. 4.1E). I parametrised this bifurcating pattern using a monotonic spline fitting on binned pseudotemporal trends (see 6.3). Thus, I repeated the estimation of half-saturation pseudotime for the detected lineages (Fig.4.1E, Fig.4.1F). In Figure 4.1F, G, I juxtaposed the pseudotemporal values of the events described so far, resulting in a sequence of events that characterise the early haematopoietic landscape: a nonproliferative subset of LT-HSCs gives rise to increasingly proliferative states that include ST-HSCs and MPPs, followed shortly by a series of branching events that establish the major blood lineages (Fig.4.1F).

To ensure that this result is not a consequence of how I measured the half-saturation constant, I repeated the comparison using an alternative method, i.e. computing the pseudotemporal coordinate at which the pseudo-derivative of the reported quantities exceeds one, indicating the point at which the growth is higher than the identity function (Fig.4.1G). This measure confirms the ordering obtained using half-saturation constants, while the onset timing is shifted to an earlier pseudotime. The difference between the onset values obtained using the two different measures is unimportant, as pseudotime measures an ordering rather than physical time (this topic is discussed further in the Discussion section). Finally,

I jointly visualised the differentiation probabilities using a method developed in [95] to embed the six-dimensional matrix containing lineage fate probabilities inferred using cellRank. The resulting reduction provides an intuitive understanding of the pairing between emerging lineages radiating from the central region. Monocytic and neutrophil, and erythroid and megakaryocytic lineages share common progenitors downstream of stem cells, while lymphoid and basophil fates do not belong to a larger group. Lineage association will be discussed more in-depth in section 4.5. To summarise, in this section, I used mathematical modelling on proliferation estimates and trajectory inference, resulting in a pseudotemporally resolved picture of haematopoietic differentiation and proliferation, in which exit from a primitive cell state is followed by proliferative activation and lineage commitment.

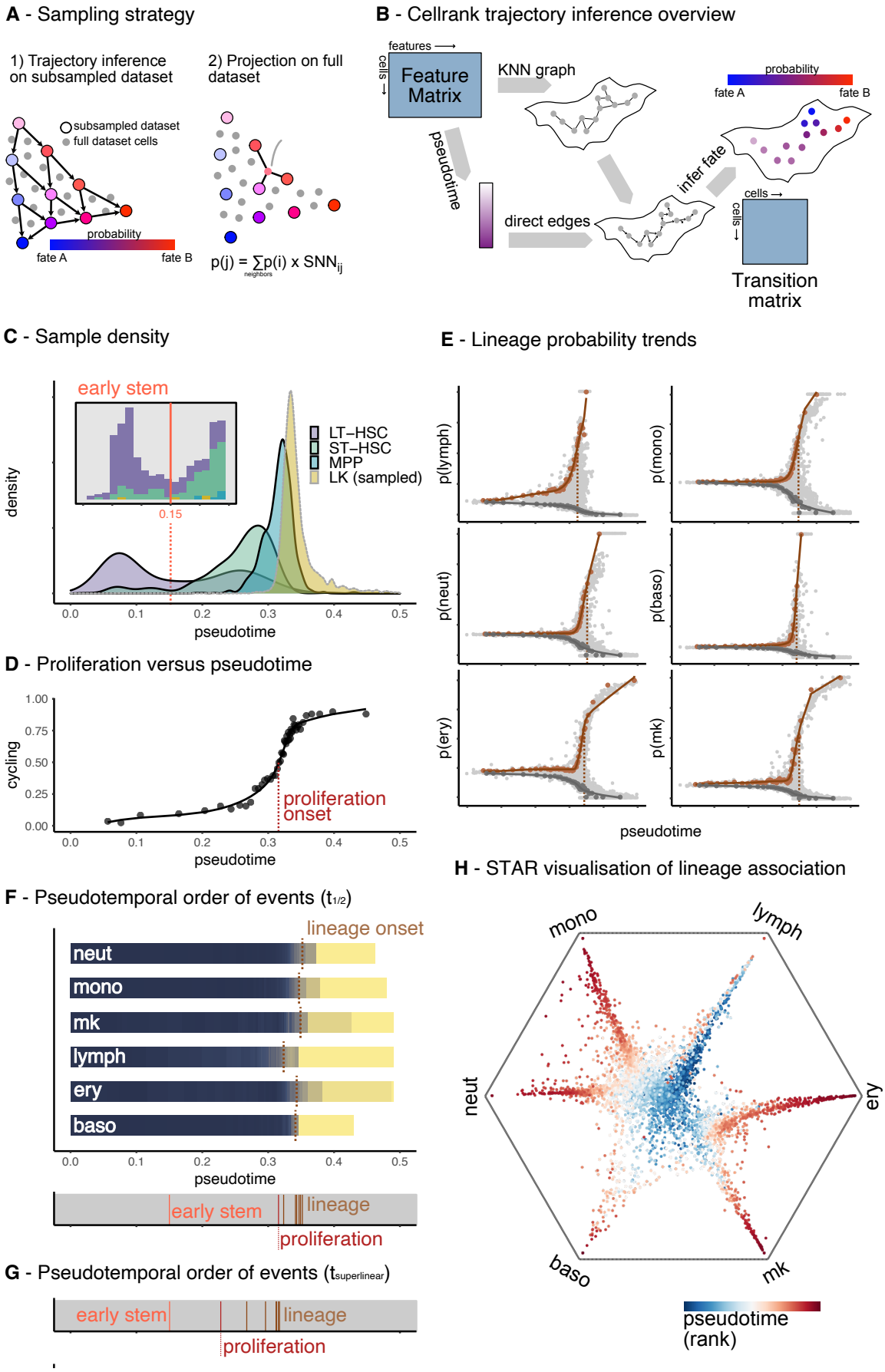

A - Sampling strategy
1) Trajectory inference on subsampled dataset
2) Projection on full dataset
subsampled dataset
full dataset cells
probability
fate A
fate B
p(j) = ∑p(i) x SNN_ij
neighbors
B - Cellrank trajectory inference overview
features
cells
Feature Matrix
KNN graph
probability
fate A
fate B
pseudotime
infer fate
direct edges
cells
cells
Transition matrix
C - Sample density
early stem
density
0.15
LT−HSC
ST−HSC
MPP
LK (sampled)
0.0
0.1
0.2
0.3
0.4
0.5
pseudotime
D - Proliferation versus pseudotime
cycling
1.00
0.75
0.50
0.25
0.00
proliferation onset
pseudotime
E - Lineage probability trends
p(lymph)
p(mono)
p(neut)
p(baso)
p(ery)
p(mk)
pseudotime
F - Pseudotemporal order of events (t1/2)
lineage onset
neut
mono
mk
lymph
ery
baso
pseudotime
early stem
lineage
proliferation
G - Pseudotemporal order of events (tsuperlinear)
early stem
lineage
proliferation
H - STAR visualisation of lineage association
mono
lymph
neut
ery
baso
mk
pseudotime (rank)

Figure 4.1

Figure 4.1 ***(previous page).*** **A**, Schematic of the subsampling strategy used in trajectory inference. Left: lineage probability and pseudotime are computed on the subsampled dataset to reflect the physiological proportions of haematopoietic progenitors. Right: results are projected on the full dataset based on the values of neighbouring cells in the Shared Nearest Neighbour (SNN) dataset. **B** Overview of the cellRank algorithm proposed in [105] with a pseudotemporal kernel: The undirected KNN graph computed on the original matrix is directed using pseudotemporal estimate. The resulting transition matrix is used to infer fate probabilities. **C**, Density of LSK population over pseudotime. A local minimum in cell density indicates an early metastable state. **D**, Proliferative trend over pseudotime fitted using monotonic splines. The half-saturation pseudotemporal coordinate is indicated by the dotted red line. **E**, Bifurcating trends in lineage probability inferred using cellRank. As pseudotime increases, the range of probability values expands. Upward (brown) and downward (grey) trends were fit using monotonic splines, and the half-saturation constant of the upward function is indicated by the dotted brown line. **F**, Joint visualisation of pseudotemporal events detected in the previous panels. The blue-yellow stripes indicate upward lineage probabilities. Underneath, the early stem state, proliferation, and lineage onset values are juxtaposed, revealing the order of these events (grey strip). **G**, same as **F**, but event coordinates were detected using the superlinearity method. **H**, Circular visualization of the lineage probability simplex using the algorithm proposed in [26].

4.4 Mapping probability values to lineage tracing data validates trajectory inference

Despite its appealing theoretical framework, probabilistic trajectory inference is not directly interpretable as differentiation probability, as it represents a simple autonomous model of cellular differentiation that does not account for proliferation, cell death, and extrinsic signalling that is known to play a role in haematopoietic differentiation. For this reason, before seeking further biological insights from cellRank output, I validated the inference outcome using transcriptional evidence from RNA-seq studies featuring lineage tracing or transplantation evidence that reported bias in lineage decisions associated with different subsets. Briefly, I used singleR [106] to map the multi-omic data to the reference datasets, thus compared the fate probabilities with the biases reported in the relevant publication (Fig. 4.2A). The datasets selected for this task are:

- **Drissen et al. 2016** [96], in which authors reported a marked lineage bias in the myeloid output of preGM and GMPs dependent on the expression of *Gata1* and validated their findings using lineage tracing. Gata1^{+} GMPs and preGMs are more likely to produce basophils, while Gata1^{-} progenitors are biased towards monocyte and neutrophil fates;

- **Pietras et al., 2015** [76], where MPPs are further subdivided based on surface marker expression of Flk2 (encoded by *Flt3*), CD150 and CD48 into MPP2 (biased towards megakaryocytic-erythroid lineages), MPP3 (myeloid bias), and MPP4 (lymphoid bias).

Findings are validated through transplantation;

- **Morcos et al., 2022** [54], presented in Chapter 2, this work used lineage tracing and mitotic tracking to unveil heterogeneity in HSC bias based on surface expression of EPCR and Sca-1: while Sca-1^{lo} LT-HSCs are less likely to produce lymphoid progeny, their megakaryocytic output is enhanced.

As shown in Fig.4.2, the probabilities inferred using cellRank agree with data from the reported studies. It should be noted that lineage bias across the different mapped subsets is also detected in LSK populations, whose pseudotemporal placement occurs mainly before the definitive bifurcation of lineage probability. This implies that, even in the early region of the landscape, subtle bias can be assigned to different subsets of progenitor cells based on co-variability patterns of fate probability.

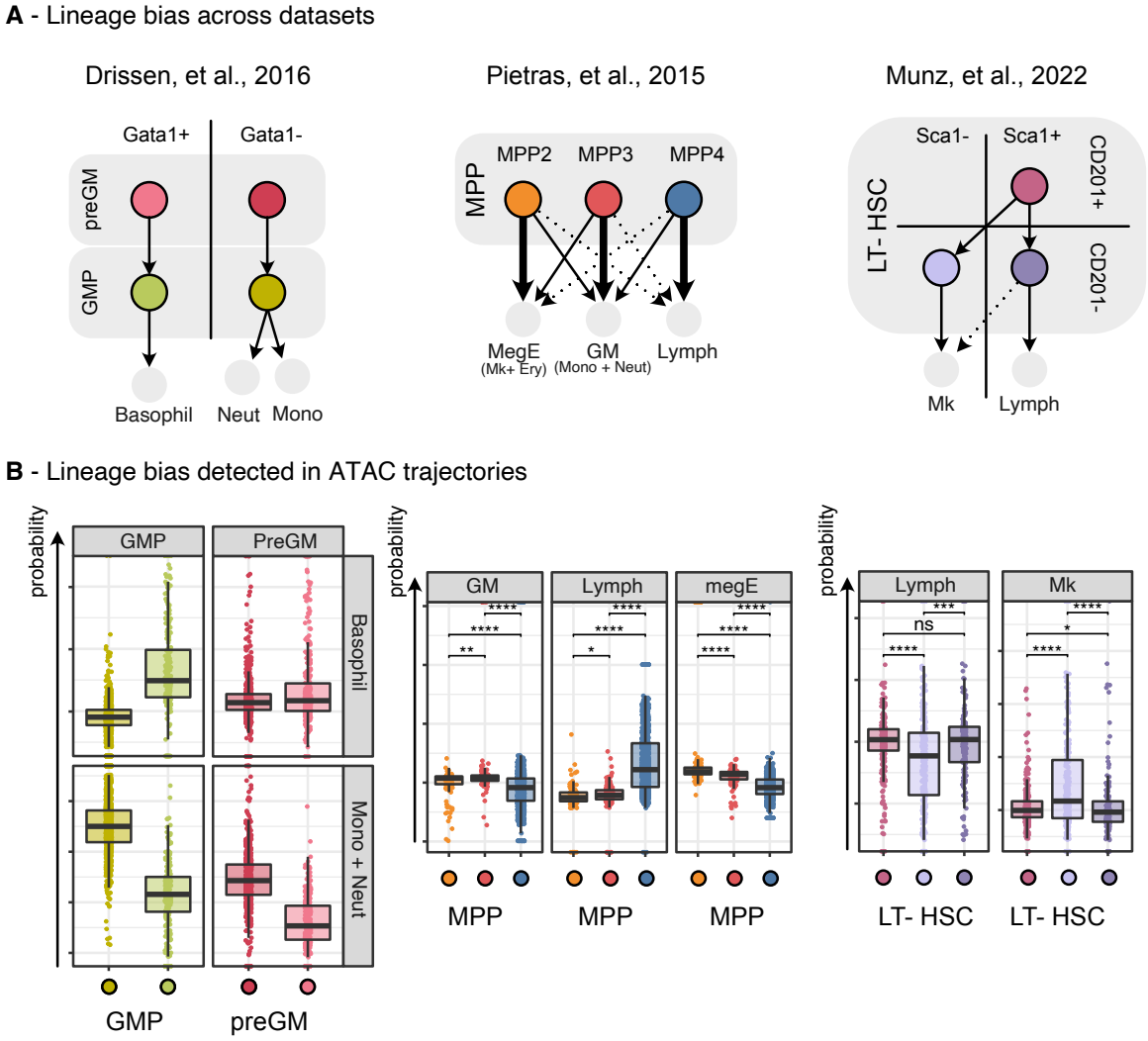


Figure 4.2. **A**, Schematic visualisation of the lineage bias reported across selected papers. **B**, Lineage probability for cells mapped to the progenitor subsets using transcriptional information. Mapped populations are grouped into boxplots and the significance of pairwise comparison using Wilcoxon test is reported.*ns*: non significant, *: p-value < 0.05, **: p-value < 1e-03, ***: p-value < 1e-04, ****: p-value < 1e-05

4.5 Discrete potential model uncovers the topology of haematopoietic differentiation.

Cells with similar properties are typically viewed as a subpopulation. Hence, discretization of the dataset into partitions is desirable as it offers an intuitive view of what kind of progenitors are observable in terms of lineage output. To obtain such description, I designed a two-step method to infer branches starting from probability values:

1. For each detected lineage, I set a probability threshold computed on low pseudotime cells to label cells as lineage-potent (Fig.4.3A, see Methods);

2. During the iterative refinement step, I successively define higher probability thresholds based on the median probability of oligo-potent progenitors to filter out cells with probability values higher than the initial threshold but lower than the progenitor cells (Fig.4.3B, see Methods). In this way, I set adaptive thresholds to classify branching events that occur for higher pseudotemporal values.

The iterative labelling procedure assigns a set of potential fates available to each cell. To visualise them, I used PHATE [16], a visualisation method based on stochastic modelling that is suitable for the representation of continuous differentiating systems (Fig.4.3C). The lineage potential model yields a set of multipotent cells that can differentiate into any mature lineage, generating a natural identification of these cells with the ”stem branch” from which all successive branches originate. Downstream of the stem cell branch, a large number (n=61) of potential combinations are detected in the data, making it impractical to use these subsets for further analyses. To solve this problem, I grouped lineages based on the statistical association among lineages, evaluated using a multi-set intersection test statistic [107]. As shown in Fig.4.3D, most potential combinations are detected in much higher numbers than chance alone would entail. In principle, it is possible that the significance of lineage pairing is solely a consequence of the stem cell branch, in which all lineages co-occur. In fact, removing multipotent cells and repeating the statistical test yielded a much lower number of significant associations (4.3E). By iteratively removing the cells of the highest potency and testing for downstream associations (see Methods), I obtained a significantly reduced set of significant associations (Fig.4.3F), exhibiting the following features:

- Erythroid and megakaryocytic, monocyte, and neutrophil fates are strongly associated. To a lower extent, the monocyte and lymphoid fates are paired too;

- Two prominent 3-fate associations are detected, assigning a highly significant pairing of basophil fate with both myeloid and megE branches;

- A higher-order association is only detected for the non-lymphoid lineages, indicating early branching of the lymphoid cells.

The iterative testing procedure provides a quantitative measure of the hierarchical features of haematopoietic differentiation. Although previous single-cell studies have proposed a model in which unipotent progenitors emerge independently from a 'cloud' of HSPCs [95], the significant associations between the fates obtained here strongly support a more classical view of haematopoiesis, in which three main branches (lymphoid, myeloid, and megakaryocytic erythroid) emerge from a branch of the tip stem. On the other hand, in a strict, tree-like differentiation topology, unipotent progenitors are expected to be only linked to one upstream compartment. This is not the case either, as pairing monocyte and basophil lineages with multiple other branches is incompatible with a tree structure. Interestingly, basophil / mass fate has recently been associated with a *Gata1*-dependent pathway shared with megakaryocytic-erythroid lineage [96] or with other myeloid lineages [108]. Since the goal of the presented analysis was to group lineages in the most informative way, I opted to incorporate basophil fate in the myeloid branch, given the higher significance of the association.

In summary, combining an iterative procedure to define potentials with statistical testing yields a quantitative picture of the hierarchy within the haematopoietic system.

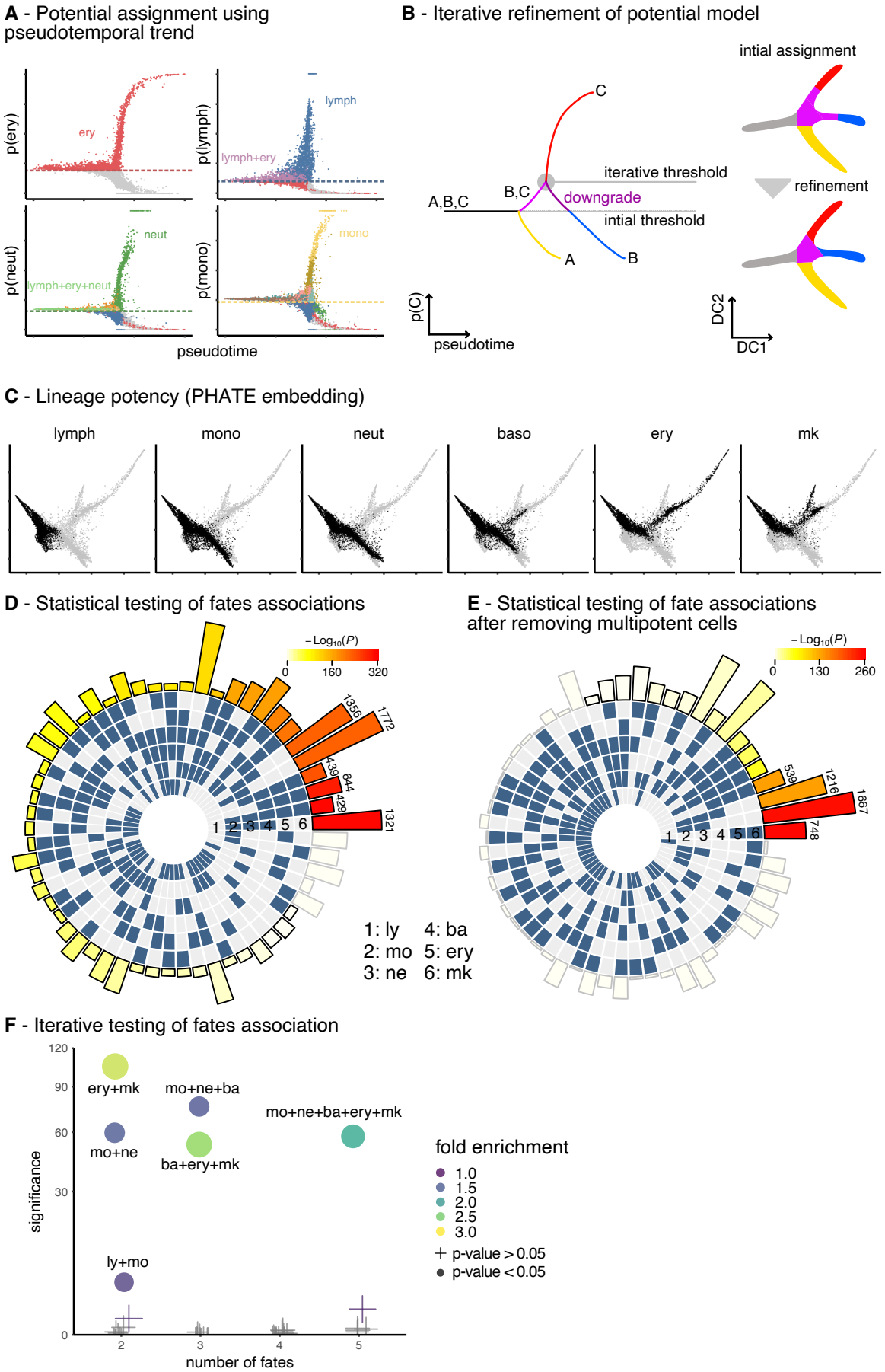

A - Potential assignment using pseudotemporal trend
p(ery)
ery
p(lymph)
lymph
lymph+ery
p(neut)
neut
lymph+ery+neut
p(mono)
mono
pseudotime
B - Iterative refinement of potential model
intial assignment
C
iterative threshold
A,B,C
B,C
downgrade
intial threshold
A
B
refinement
p(C)
pseudotime
DC2
DC1
C - Lineage potency (PHATE embedding)
lymph
mono
neut
baso
ery
mk
D - Statistical testing of fates associations
−Log10(P)
0
160
320
1356
1772
439
644
429
1321
1 2 3 4 5 6
1: ly
2: mo
3: ne
4: ba
5: ery
6: mk
E - Statistical testing of fate associations after removing multipotent cells
−Log10(P)
0
130
260
539
1216
1667
748
1 2 3 4 5 6
F - Iterative testing of fates association
120
90
60
30
0
significance
ery+mk
mo+ne+ba
mo+ne+ba+ery+mk
mo+ne
ba+ery+mk
ly+mo
2
3
4
5
number of fates
fold enrichment
1.0
1.5
2.0
2.5
3.0
p-value > 0.05
p-value < 0.05

Figure 4.3

Figure 4.3 ***(previous page).*** **A**, Initial assignment of lineage potential based on pseudotemporal probability trend. For each lineage, cells whose probability exceeds the baseline probability threshold (dashed line) are annotated as lineage-potent, starting with the erytrhoid potential (top left panel). The addition of a potential classification for lymphoid potential (top right panel) creates potential combinations, indicated by colour. As neutrophil and monocyte potentials are added (bottom panels), multiple lineage combinations are created in the data. **B**, Schematic representation of iterative refinement of the potential model: downstream of {A,B,C}-potent cells, a branching event generates the unipotent branch {A}. Due to the normalisation of the probability, the baseline probability for fates B and C in the remaining cells is artificially increased. To compensate for this, a new threshold is selected based on the probability of {B,C}-potent cells (purple), which is used to reassign potential in downstream cells, resulting in loss of B potential in the dark purple cells. *Right*: Depiction of the net effect of the iterative refinement of the potential model on a dimensionality reduction embedding. **C**, PHATE embedding coloured by lineage potential for each fate. **D**, Visualisation of multi-set intersection significance results. Fate combinations are displayed in the central tiles and ordered on the basis of the p-value, while the height of the bars is proportional to the number of cells in the intersection. Significant combinations (p-value < 0.05) are indicated with black bar outlines. **E**, Same as **D**, but comparisons are made after removal of tip stem cells. **F**, Results of iterative testing of fate combinations: 5-fate combinations were tested after removal of 6-potent cells, 4-fate combinations after removal of 5-potent cells, etc., resulting in six fate combinations that are significantly enriched in the dataset.

4.6 Branching model detects LT-HSC split, choice between lymphoid commitment and proliferation

Building on the findings of the previous analysis, I defined discrete partitions of the data in terms of putative differentiation output, and compared them with sorted populations, proliferation estimates, and chromatin clusters. I categorised cells based on their potential to develop along any of the three identified branches, namely lymphoid, myeloid, and megakaryocytic-erythroid (megE). In this way, I obtained a simplified partition of the data into potential combinations, including lympho-myeloid and myeloid-megE intermediate progenitor stages, recapitulating previous studies that identified populations exhibiting similar patterns of lineage bias using transplantation and lineage tracing data [96, 109], (4.4). Comparing the inferred branches with the original gates used to sort the cells reveals a clearer understanding of branch emergence within phenotypic compartments:

- Although a small portion of MPPs are labelled as stem cells (17%), the majority are already classified as having restricted lineage potential in one of the other branches;
- A more significant proportion of ST-HSCs belong to the stem cell branch (56%), while another considerable subset falls into either the lymphoid or lympho-myeloid branch (29%);
- Most LT-HSCs are labelled as stem cells (65%), but a substantial number of LT-HSCs (33%) have lost their lymphoid potential and are classified as myeloid-megE. This

observation supports the reported lineage bias of Sca-1^{lo} HSCs discussed in Chapter 2.

The resolution of the branching model can be enhanced by dissecting each of the major branches into their constitutive lineages and their relative combinations, as shown in Fig.4.4C. Pseudotemporal analysis of this higher-resolution version of the branching model shows how the unipotent myeloid and megE branches emerge for higher pseudotemporal values (Fig.4.4D).

The outcome of the branching model can be used to examine the proliferative trends in the dataset in further detail. Consistent with the previous result, branches with lower pseudotime are less proliferative, with the oligopotent branches presenting intermediate values between stem cells and committed branches. Interestingly, the lymphoid branch constitutes an exception to this trend, given its low proliferative score (Fig. 4.4E). Next, I directly compared proliferation and lineage emergence by determining differentiation trajectories for each of the identified endpoints and measuring the portion of proliferating cells in each pseudotemporal bin. As shown in Fig.4.4F, proliferation and lineage emergence are generally linked, while the lymphoid trajectory shows lower cell cycle activity. Interestingly, prior analysis of lineage/proliferation onset placed the emergence of the lymphoid lineage in pseudotemporal proximity of the proliferation onset event, suggesting that the initial decision made by stem cells involves choosing between lymphoid commitment and proliferative activation. The inverse relationship between lymphoid potential and proliferation appears to be rooted in the molecular regulation of the key lymphoid transcription factor *Tcf3*, as its knockout induces hyperproliferation and loss of lymphoid potential [110].

Lastly, I compared the branching model with the chromatin clusters identified in the previous chapter. Specifically, I calculated the proportion of cells that exhibit potency in any of the described lineages. As demonstrated in Fig.4.4G and Fig.4.4H, clusters detected in the progenitor gate are predominantly unipotent, while some clusters display a myeloid and lymphoid-myeloid potential. Interestingly, clusters 1 and 7, which exhibit similar transcriptomes but distinct chromatin structures, differ in erythroid, megakaryocytic, and neutrophil potential, implying a subtle bias toward these lineages that cannot be identified using transcriptional information alone.

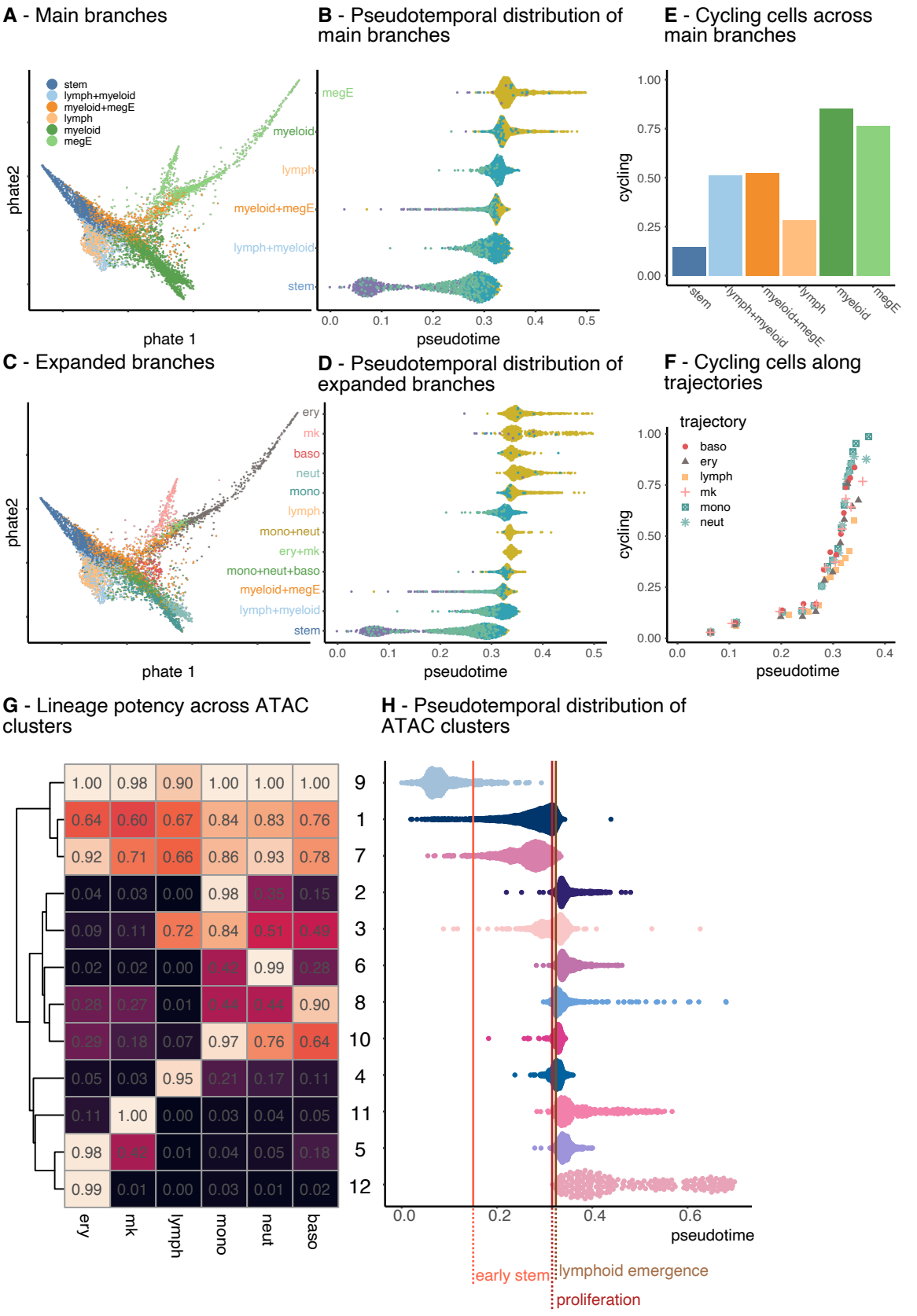

A - Main branches
B - Pseudotemporal distribution of main branches
E - Cycling cells across main branches
stem
lymph+myeloid
myeloid+megE
lymph
myeloid
megE
phate2
phate 1
pseudotime
cycling
C - Expanded branches
D - Pseudotemporal distribution of expanded branches
F - Cycling cells along trajectories
ery
mk
baso
neut
mono
mono+neut
ery+mk
mono+neut+baso
trajectory
G - Lineage potency across ATAC clusters
H - Pseudotemporal distribution of ATAC clusters
early stem
lymphoid emergence
proliferation

Figure 4.4

Figure 4.4 ***(previous page).*** **A,C**, Visualization of main (A) and expanded (C) branches obtained merging potentials based on fate co-occurrence testing. **B,D**, Pseudotemporal distribution of sorted populations across main (B) and extended (D) branches reveals that a significant portion of LT-HSCs, ST-HSCs and MPPs have limited potential. **E**, Portion of cycling cells in each of the main branches. **F**, Portion of cycling cells across multiple trajectories. Cells are binned based on pseudotime and the portion of cycling cells is quantified for each bin within each lineage. **G**, Proportion of lineage-potent cells for each of the ATAC clusters detected in Chapter 2. **H**, Pseudotemporal distribution of ATAC clusters.

4.7 Discussion

A much-used representation used to describe differentiating cells is the Waddington epigenetic landscape [35], in which a differentiating system is depicted as a system of valleys and hills along which cells roll down from a multipotent to a committed stage autonomously. In the haematopoiesis system, the starting point of this descent is represented by long-term HSCs; a population characterised by a low differentiation rate (few differentiation events per cell per year [102]), balanced by equally rare self-renewing divisions. These kinetic rates indicate that stem cells are in a relatively stable state, represented as a valley in the Waddington landscape at the very top of the epigenetic landscape. Accordingly, previous trajectory inference work has theorised that such metastable states would manifest in single-cell omics datasets as dense regions along differentiation trajectories [18, 111]. Puzzlingly, single-cell sequencing experiments featuring the top haematopoietic compartments have yielded a continuum of cellular states [99, 112], in which clusters are largely overlapping and not clearly separated [58, 100]. These studies sample LSK populations more densely to gain enough statistical power to identify relevant subpopulations within the LSK gate at the cost of losing information relative to cell density along the maturation paths. Conversely, unbiased sampling of LS^{-}K cells inevitably causes the comparatively large number of heterogeneous progenitors to determine the features selected for downstream analysis, thus effectively "drowning" stem cell heterogeneity in favour of more prominent branching events [98, 113]. In this chapter, I showed how, by combining feature selection on the LSK-dense dataset and trajectory inference on its unbiased subsample, the divide between the tip HSCs and their progeny can indeed be observed as a low-density interval at low pseudotemporal coordinates, thus delineating a single-cell chromatin landscape that accommodates the slow kinetic aspects of haematopoiesis discovered in lineage tracing experiments [11, 102, 114].

Downstream of this initial state, differentiation progresses towards terminal differentiation in a series of bifurcating events that combine to originate intermediate progenitor states. The computational inference of such intermediate states requires clustering [21, 22, 97] or constraining the available topologies [18, 80]. While previous work has identified such states by defining a probability threshold based on stem cells [108], this approach is limited

in the presence of successive bifurcations. Here, an iterative thresholding method was introduced to define branches that could, in principle, be tested against various acyclic topologies. Using this approach, I defined intermediate progenitor states, evaluated their statistical significance, and thus quantified the amount of hierarchy in the haematopoietic system. The outcome of this analysis shows a strong hierarchical association of monocyte and neutrophil, megakaryocytic, and erythroid fates, as often reported in the literature, thus speaking against a 'cloud' model in which lineages emerge individually from a lowly primed set of HSPCs. Nevertheless, pairing monocyte and basophil fates with multiple other lineages suggests that the topology cannot be strictly defined as a tree, in which each leaf descends exactly from one branch. The work presented in this chapter supports an acyclic-directed graph as the most likely topology of the haematopoietic system using unbiased, data-driven partitions inferred using trajectory inference. Importantly, rather than being a consequence of the clustering algorithm and resolution chosen to discretise the data, this outcome emerges from the significance of potential combinations directly inferred on trajectories. By further refining the branching model to extract lineage-specific trajectories, I observed that downstream of the previously reported early stem cell state, the first molecular decision taken by stem cells is likely to be between lymphoid commitment and proliferative activation. This observation aligns with the findings presented in Chapter 2, where proliferation rates are largely different across the Sca-1-defined branches and is grounded on molecular regulation of the lymphoid lineage, whose emergence is compromised in response to proliferative activation caused by genetic alteration [110] or immunological challenges [76, 115].

However, the procedure proposed here has some limitations. In the first place, the threshold used to classify the potency of cells is arbitrarily set at a fixed percentage (90%) of the baseline probability shown in early stem cells. The dependency of computational pipelines on arbitrary hyperparameters is a prevalent issue in most single-cell methods.

A more fundamental limitation of the work presented here involves the relationship between time and its computational counterpart, pseudotime. Despite its naming, the pseudotime simply measures the distance between a root cell and any other cell on a manifold. Assuming that states close to the root one are reached following a proximity order, one can consider pseudotime as a monotonic function of physical time. This assumption is no longer valid for branching processes, since the speed at which successive states along different branches are reached might differ, implying that comparing pseudotime between cells placed after branching events is not meaningful. More generally, in a multi-dimensional manifold, it is not guaranteed that different directions will be covered at similar speeds. Despite recent advances in pairing pseudotime with physical time using RNA velocity, the latter is riddled with technical and conceptual limitations that limit its applicability to systems characterised by higher kinetic rates, rendering it unsuitable for haematopoiesis [116, 117].

The drawbacks described can help contextualise some of the findings described in this chapter. For example, I showed how the lymphoid lineage is the first to emerge in pseudotemporal order, indicating an early differentiation event. On the contrary, lineage tracing evidence has repeatedly demonstrated that label induced in LT-HSCs of the tip reaches lymphoid-biased cells last [102, 118]. Taken together, these findings suggest that the lymphoid fate is specified early within the LSK compartment, but actual differentiation is slow, due to the high self-renewing capacity of lymphoid-biased progenitors. Interestingly, a population characterised by lymphoid bias and a near complete self-renewal of embryonic origin has recently been described in [119], although its kinetic and dynamic properties are yet to be proven to be distinct from their mature counterparts, LMPPs. Lymphoid emergence is a prime example of how single-cell omics technologies and fate mapping can be integrated to gain deeper insight into differentiation systems.

Notwithstanding the listed caveats, obtaining the pseudotemporal ordering of branching events is paramount in defining the intermediate states between multi- and uni-potent stem cells. In future developments of this work, data from paired transcriptional and lineage tracing datasets ([27, 108, 120]) will be mapped onto the multi-omic landscape and used to reconstruct the kinetics that unfold over the state landscape of steady-state haematopoiesis.

Chapter 5

Molecular regulation of haematopoietic differentiation

5.1 Introduction

In the previous chapter, I used trajectory inference and mathematical modelling to quantitatively describe lineage emergence in haematopoiesis. In this chapter, I describe the molecular interactions that contribute to the generation of the haematopoietic landscape, focussing on the interaction between the transcriptional and chromatin layers.

Multilineage priming is a key concept in the regulation of cell differentiation. The term was initially used to describe the early co-expression of genes expressed by competing branches in HSPCs [121, 122]. These observations lead to a permissive model of commitment, in which the initial opening of gene and enhancer loci for all possible fates is followed by restriction of alternative fates during commitment, as suggested by the detection of open chromatin loci for conflicting lineages prior to commitment in MPPs [43, 44, 123]. Alternatively, lineages can be established by *de-novo* opening of chromatin regions that regulate a specific fate. Current estimates of the balance between de-novo establishment and restriction of alternative fates have been limited by the reliance on bulk analysis [43, 44, 123] or clustering methods [43, 124], which confines the analysis to discrete populations. More recently, a looser definition of lineage priming has emerged in single-cell sequencing studies. In recent scRNA-seq studies [95, 100], priming refers to the potential to infer cellular fate based on the expression of a few key markers. However, a study employing single-cell transcriptomics paired with lineage tracing indicates that transcriptional data alone contain little information on early lineage commitment [27].

In current work [74, 124], lineage priming refers to the possibility that information about fate decisions is available in a subset of epigenetic or transcriptional features before they can be observed phenotypically. Now, single-cell multiomic data allow us to empirically test

how transcriptional and epigenetic layers are temporally related: pseudotemporal ordering which can be compared against transcriptional or chromatin features to establish the order of molecular differentiation events. In [48], a systematic comparison of gene expression and correlated ATAC peaks resulted in the detection of substantial delay between the two layers in stem cells differentiating in the hair follicle of mice. In the first part of this chapter, differential expression and accessibility analysis are used to quantify how RNA and chromatin are regulated in the establishment of mature lineages. Next, I use two complementary methods to quantify the amount of lineage priming that can be observed in the chromatin features with respect to RNA.

Mechanistically, the interplay between chromatin and gene expression is mediated by transcription factors (TFs). These proteins bind to specific DNA sequences, thereby facilitating or alternatively inhibiting the recruitment of RNA polymerase to target genes. Additionally, transcription factors often interact with chromatin remodelling complexes to compact or decompact chromatin regions, influencing gene expression. The role of individual transcription factors such as GATA-1, PU.1, and C/EBPα has been extensively studied [125], revealing their importance in lineage commitment and cell differentiation. However important the singular contribution, the diversity of TF regulation involves interaction, through regulation of TF themselves and the regulation of common targets. A natural way to encode these complex interactions is through gene regulatory networks (GRNs), which provide a schematic representation of interactions between transcription factors, epigenetic modifiers, and other regulatory elements. In this framework, the links between a TF and a gene are often weighted based on whether the relationship is activatory or inhibitory. Understanding these networks is essential to elucidate the mechanisms underlying normal and pathological haematopoiesis. Advances in single-cell sequencing technologies have facilitated more comprehensive analyses of these GRNs, offering new perspectives for targeted therapeutic interventions. Despite these recent technological advancements, scRNA-seq alone has proved so far ineffective in reconstructing causal networks, mainly due to the difficulty in disentangling correlations from causation in transcriptional datasets [50]. The availability of paired transcriptional and chromatin assays has led to improved GRN inference strategies that proved more effective in extracting significant links between TFs and downstream targets [126]. In the second part of this chapter, I use a recently developed GRN inference tool on the haematopoietic dataset to investigate the core aspects of haematopoietic differentiation and identify the drivers of lineage commitment in haematopoiesis.

5.2 Molecular analysis of commitment uncovers hierarchical regulation of lineage markers

Initially, I sought to obtain a molecular description of commitment across lineages using differential expression analysis on differentiation-relevant features (see Methods). I performed differential expression testing between stem cells and tips of the detected lineages for both upregulation and downregulation of RNA and ATAC features (Fig. 5.1A) Given the intermediate position of the basophil lineage and its low cell numbers, I excluded it from this section of the analysis). It has to be noted, however, that while the myeloid and megE lineages are clearly separated in the progenitor population, the cells with the highest lymphoid signal are likely to be phenotypic LMPPs, whose commitment to the lymphoid lineages is not yet complete [127, 128]. Accordingly, RNA features are mostly differential in megE and GM lineages in both directions, with a consistent portion of markers shared between the two branches. Over-representation analysis of these features indicates that translation is enhanced, likely due to increased cell cycle activity required for amplification, while stem cell features are downregulated (Fig. 5.1B). Differential Availability Region (DAR) analysis is dominated by downregulatory events, as reported in [123], favouring a permissive model for commitment. Interestingly, I detected the most opening events in the lymphoid lineage, of which a substantial amount is shared with the GM branch. On the other hand, this branch displays minimal closing compared to MegE and GM branches. Given the lower commitment of LMPPs with respect to their myeloid and megE counterparts, this analysis indicates that, while lymphoid gene opening is robustly established in lymphoid MPPs, the potential for at least GM lineages is retained (as discussed in [127, 129, 130]), as suggested by the paucity of downregulation/closing events. The large amount of closing events shared with the GM lineages is likely linked to the loss of megE competence, already detected at the LMPP level ([76]).

Next, I sought to investigate the pseudo-kinetical aspects of opening and closing by fitting monotonic functions for each of the detected markers and detecting their half-saturation constant (see Methods).

As shown in Fig.5.1C, both upregulation and downregulation tend to follow a hierarchy, in which genes/regions co-regulated across branches are affected earlier, while lineage-specific ones are affected later. Moreover, by aggregating regulation pseudo-series by assay and direction of regulation, consistent precedence of downregulation with respect to upregulation is shown in Fig.5.1D. However, this analysis shows no clear succession between ATAC and RNA regulation, as chromatin and transcriptional features are not linked. In summary, systematic analysis of differentiation features favours a predominantly permissive model of lineage commitment, in which features are hierarchically regulated to origi-

nate the unipotent branches in the data. In the next section, I pair RNA and ATAC features to investigate lineage priming between these layers.

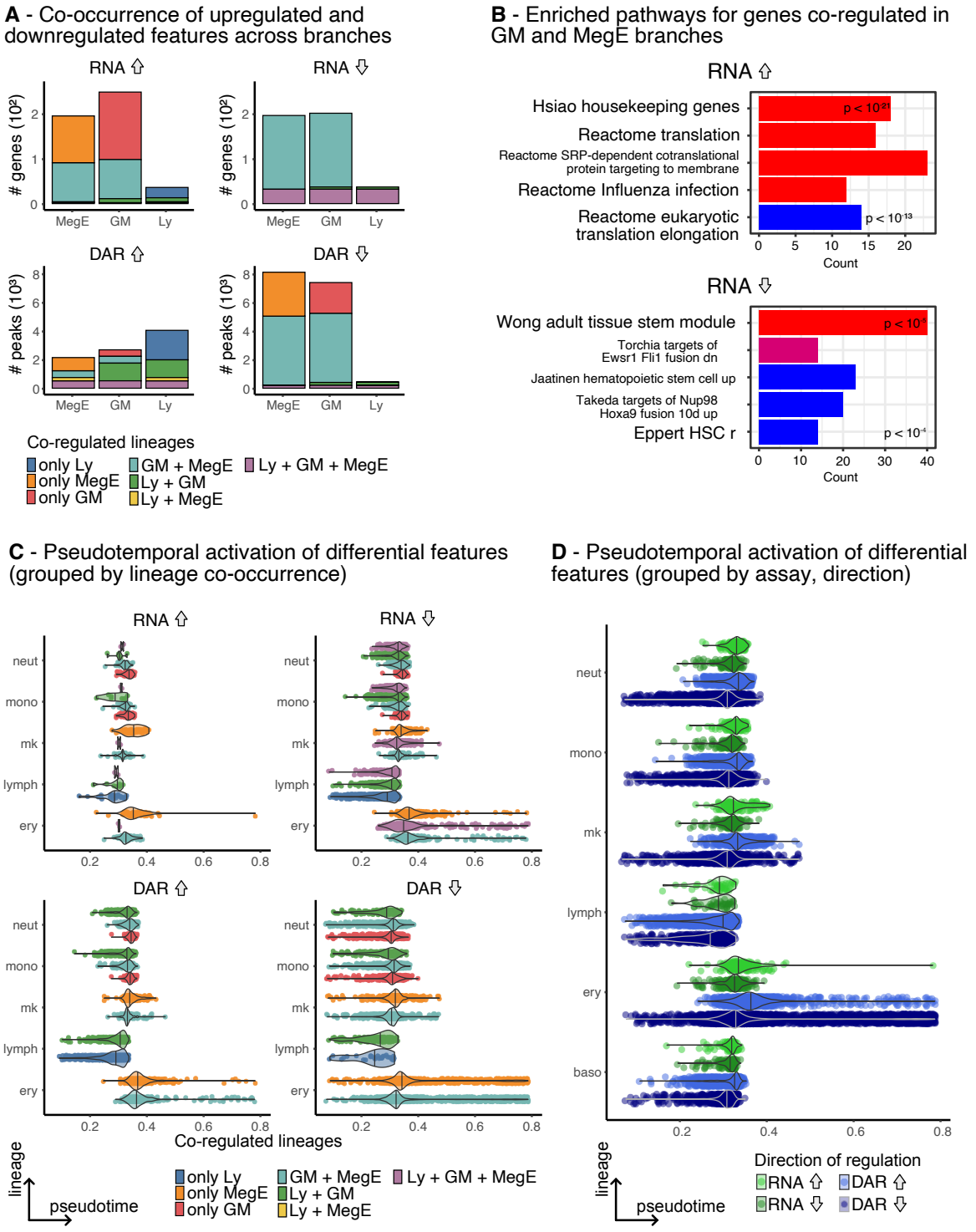


Figure 5.1. **A**, Overlap between differential features across major branches in RNA and ATAC (DAR stand for Differentially Accessible Regions). GM and MegE lineages share a large number of regulated features in both assays. **B**, Over-representation analysis of gene sets in the Curated Msigdb category for RNA markers shared between GM and MegE branches. The x-axis indicates the number of genes in each category, color indicates significance. **C**, Each dot in the violin plots indicates the half-saturation pseudotime for differential features, stratified by assay and direction of regulation. In most cases, shared features are up/down-regulated prior to the assay-specific ones. **D**, Comparison of half-saturation pseudotime across assays and direction. The typical sequence of regulation events is: closing of stem regions in ATAC, downregulation of stem markers expression, the opening of lineage-specific peaks, upregulation of lineage-specific RNA markers.

5.3 Priming analysis unveils chromatin priming around the promoter region

Next, I paired RNA and ATAC features to measure whether a consistent delay could be detected between the two layers. To do this, it is crucial that transcriptional and chromatin features are paired. To achieve feature pairing, I used two approaches that combine RNA expression with chromatin availability in both the gene body region and in distal correlated peaks:

- The Gene Score Metric (GSM), as implemented in ArchR [131], quantifies chromatin accessibility around the gene body. It aggregates accessibility across the entire gene body, applies distance-dependent exponential weighting for potential distal regulatory elements, and sets boundaries to exclude neighbouring genes. While in scATAC-seq analysis this metric is typically used as a proxy for gene expression, in this work I used it to measure the accessibility of the promoter and gene body regions, thus comparing it to actual RNA measurements to quantify lineage priming (Fig. 5.2A);

- DORCs, or Domains of Regulatory Chromatin, are peaks correlated to gene expression. To associate a DORC score to a specific gene, the counts in all peaks significantly correlated with the target gene expression are summed. DORCs have been reported to play a role in the priming of active chromatin states, with chromatin in DORC-regulated genes becoming accessible prior to induction of the corresponding gene expression [48]. Although in its default implementation, all peaks are retained, in my analysis I focused on more distal regions (i.e., peaks distant at least 10^3 bp) to provide a complementary measurement to the Gene Score (Fig. 5.2A) that focuses on putative enhancers.

For both GSM and DORC computation, I restricted the analysis to upregulated lineage markers described in the previous section, while downregulated markers are included in the stem category, as most of them are shared across lineages (Fig. 5.1A).

Once these two methods are used to extract chromatin features linked to RNA expression, the pseudotemporal ordering of cells along trajectories is used to generate pseudotemporal series (or pseudoseries) that capture the expression of features as differentiation progresses downstream of tip stem cells (Fig. 5.2A). If such series display monotonic behaviour, the half-saturation constant computed on the smoothed data can be used to quantify the priming of chromatin features relative to RNA, as described in [48]. Quantification of pseudotemporal activation delay requires that the considered pseudoseries display a monotonic behaviour. To discriminate between monotonic and non-monotonic pseudoseries, I

compared two regularised models used to smooth the data: an unconstrained b-spline and a monotonic spline using positive coefficients. The log-likelihood of the competing models is similar for monotonic series, while nonmonotonic trends show a higher likelihood for the unconstrained b-spline model (Fig. 5.2B, see Methods). Non-monotonicity is detected mostly in RNA pseudotime series, especially in basophil, erythroid and megakaryocytic lineages (Fig. 5.2C). Features that are labelled as non-monotonic are excluded from the quantification of priming, as half-saturation constants are not interpretable in these cases.

DORC detection returned an extensive list of genes potentially regulated by nearby chromatin peaks for each lineage (Fig. 5.2D), whose aggregated expression was tested for priming.

I quantified lineage priming in monotonic trends as previously described. Overall, the distribution of the priming delay is spread around zero. Remarkably, the median priming values indicate a slight precedence of RNA expression over DORC features (Fig. 5.2E). While this observation requires further analysis, it is worth noting that the detection of DORC features is based on correlation, and thus opening or closing of peaks that are annotated as putative enhancers might be caused, rather than caused, by upregulation or down-regulation of lineage markers. In other words, detected peaks could be correlated with RNA expression because they are placed lower in the regulatory cascade that unfolds during differentiation.

Conversely, GSM priming measurements show that, in general, the opening of the chromatin around the gene body precedes the expression of RNA in all branches except the stem, which includes genes down-regulated during differentiation (Fig. 5.2F). Of particular interest is the large amount of priming detected in the lymphoid lineage, whose priming dynamics have been discussed in [132].

Overall, while priming using DORCs cannot be readily interpreted due to the lack of directionality between chromatin opening and RNA expression detected using correlation, GSM analysis offers evidence of a small but consistent amount of lineage priming in chromatin with respect to RNA.

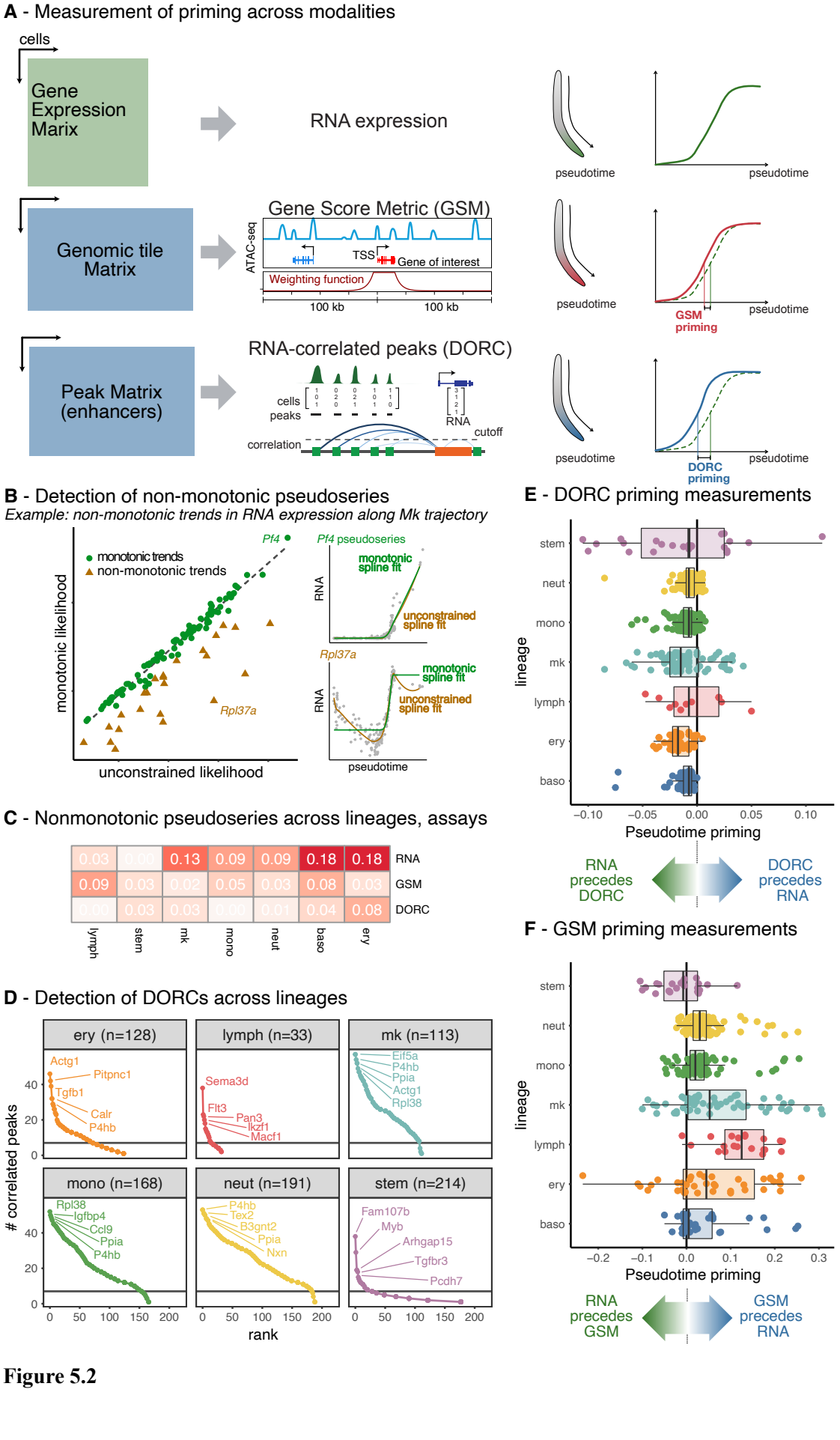

A - Measurement of priming across modalities
cells
Gene Expression Marix
RNA expression
pseudotime
pseudotime
Genomic tile Matrix
Gene Score Metric (GSM)
ATAC-seq
TSS
Gene of interest
Weighting function
100 kb
100 kb
pseudotime
pseudotime
GSM priming
Peak Matrix (enhancers)
RNA-correlated peaks (DORC)
cells
peaks
RNA
correlation
cutoff
pseudotime
pseudotime
DORC priming
B - Detection of non-monotonic pseudoseries
Example: non-monotonic trends in RNA expression along Mk trajectory
monotonic trends
non-monotonic trends
monotonic likelihood
unconstrained likelihood
Pf4
Rpl37a
Pf4 pseudoseries
monotonic spline fit
unconstrained spline fit
Rpl37a
monotonic spline fit
unconstrained spline fit
RNA
RNA
pseudotime
C - Nonmonotonic pseudoseries across lineages, assays
0.03 0.00 0.13 0.09 0.09 0.18 0.18 RNA
0.09 0.03 0.02 0.05 0.03 0.08 0.03 GSM
0.00 0.03 0.03 0.00 0.01 0.04 0.08 DORC
lymph
stem
mk
mono
neut
baso
ery
D - Detection of DORCs across lineages
ery (n=128)
Actg1
Pitpnc1
Tgfb1
Calr
P4hb
lymph (n=33)
Sema3d
Flt3
Pan3
Ikzf1
Macf1
mk (n=113)
Eif5a
P4hb
Ppia
Actg1
Rpl38
mono (n=168)
Rpl38
Igfbp4
Ccl9
Ppia
P4hb
neut (n=191)
P4hb
Tex2
B3gnt2
Ppia
Nxn
stem (n=214)
Fam107b
Myb
Arhgap15
Tgfbr3
Pcdh7
correlated peaks
rank
E - DORC priming measurements
stem
neut
mono
mk
lymph
ery
baso
lineage
−0.10 −0.05 0.00 0.05 0.10
Pseudotime priming
RNA precedes DORC
DORC precedes RNA
F - GSM priming measurements
stem
neut
mono
mk
lymph
ery
baso
lineage
−0.2 −0.1 0.0 0.1 0.2 0.3
Pseudotime priming
RNA precedes GSM
GSM precedes RNA

Figure 5.2

Figure 5.2 ***(previous page).*** **A**, Schematic representation of lineage priming metrics: gene expression is directly compared against pseudotime, while ATAC reads are used to generate both Gene Scores, indicative of accessibility in the gene body region, and DORCS, summarizing accessibility in peaks correlated with gene expression. **B**, *Left*: Pseudotemporal series are fitted using both a monotonic and an unconstrained model. The Likelihoods of gene-specific models are compared to detect non-monotonic trends to exclude from the analysis. *Right*: Example of a monotonic pseudoseries (*Pf4*), in which monotonic and unconstrained models completely overlap, with a non-monotonic one (*Rpl37a*), where the unconstrained model outperforms the monotonic one. **C**, Quantification of non-monotonic pseudoseries across assays and lineages reveals that RNA is generally less monotonic. In particular, 18 % of basophile and erythroid markers show a nonmonotonic behaviour. **D**, Results of the DORC detection pipeline. For each lineage, the genes whose expression is correlated with a high number of peak accessibility are labelled. **E, F**, Difference between half-saturation pseudotime between RNA and DORC (**E**) or GSM (**F**) monotonic trends is reported for all lineages. In most cases, RNA regulation precedes DORC dynamics and follows GSM opening/closing.

5.4 Gene Regulatory Network inference resolves regulatory heterogeneity

Next, I used a gene regulatory network inference framework to establish the regulatory link between RNA and ATAC features that are rooted in molecular interactions between transcription factors and their targets. I used cellOracle, a method that leverages both ATAC and RNA to infer GRN [53]. This method comprises different steps: first, for each gene, the co-accessibility of its transcription starting site (TSS) and nearby peaks (within a 10^6bp window) is computed. The co-accessible peaks are then tested for TF motif enrichments, yielding a set of potential TF-target interactions. In the last refinement step, the RNA expression of each target gene is predicted using a regularised linear model with candidate TF expression as input. Regression coefficients that are significantly different from zero are retained and represent the links in the GRN model (5.3A). The resulting network links transcription factors to putative target genes, together with a signed coefficient that indicates the inferred strength of activation or inhibition. To detect heterogeneity in how target genes are regulated by transcription factors, the last step involving refinement through regularised regression is performed separately for each of the sub-populations in the data. In this work, I used the high- and low-resolution branches defined in the previous chapter to investigate the dynamic aspects of haematopoietic regulation. I selected 127 transcription factors and 920 target genes based on unbiased feature selection using mean-variance relationships and added a few further transcription factors whose role in haematopoiesis has been consolidated in previous studies [133–135] (see Methods).

Methods that infer gene regulatory networks on scRNA-seq data alone are often confounded by the detection of an excessive amount of links between transcription factors and their targets. As shown in Figure 5.3B, the inclusion of chromatin information and regular-

isation increases the sparsity of the inferred networks, prunes a large number of redundant interactions, effectively reducing the number of links by roughly an order of magnitude. As shown in Fig. 5.3C, the majority of the significant links across branches are positive, suggesting that activation is slightly more prevalent than inhibition in the inferred networks.

To gain an overview of the features of the inferred network, positive links were embedded in the stem cell network using a force-directed layout (Fig. 5.3D). In this representation, nodes represent genes, which are coloured based on the branch in which they are expressed maximally. Surprisingly, while most genes reflect the colour of the stem branch, multiple genes expressed in different branches are present.

To analyse this observation, I grouped genes based on the lineage of maximal expression and aggregated the regulatory interactions between lineage genes across the main branches (see Methods). As shown in Figure 5.3E, regulation is not exclusive to genes expressed in the branch under consideration, but involves multiple lineages simultaneously. The regulatory difference across branches lies in the switch from overall activating to inhibitory interactions between lineages. For example, the interactions between lymphoid and neutrophil genes are overall positive in the stem cell branch, but turn negative in the lympho-myeloid progenitors and downstream compartment, indicating how lineage-specific factors inhibit opposing lineages as differentiation progresses. Similarly, the initial co-activation between monocyte and neutrophil markers turns inhibitory in the myeloid branch, in which the two lineages become mutually exclusive. A similar downward trend is observed between neutrophil and basophil, lymphoid, and monocytic lineages (although the overall regulation in the latter never turns fully inhibitory). By contrast, interactions between basophil, erythroid and megakaryocyte branches are inhibitory from the start, with growing intensity as differentiation proceeds downwards.

Next, I set to further resolve regulatory interactions by inferring regulatory networks for the high-resolution branches detected in the previous chapter, resulting in twelve regulatory networks featuring thousands of significant links among TFs and target genes. To interpret this large amount of data, I reasoned that the regulatory importance of a gene or a transcription factor in a network is encoded in its centrality (here measured using eigenvector centrality [82]). Each of the initially selected features has a different centrality score across the inferred networks, resulting in a feature matrix that can be deconvoluted using PCA. In Fig.5.4A, original features are projected on PC2 and PC3, displaying how similarity in regulation can be broadly grouped into three major groups involving stem and lymphoid, myeloid and megakaryocyte-erythroid branches. Notably, while the statistical tests based on trajectory inference presented in the previous chapter suggested slightly closer association of basophils with the myeloid lineages, this analysis groups the basophil lineage with megakaryocytic and erythroid cells. This association is confirmed when computing overall

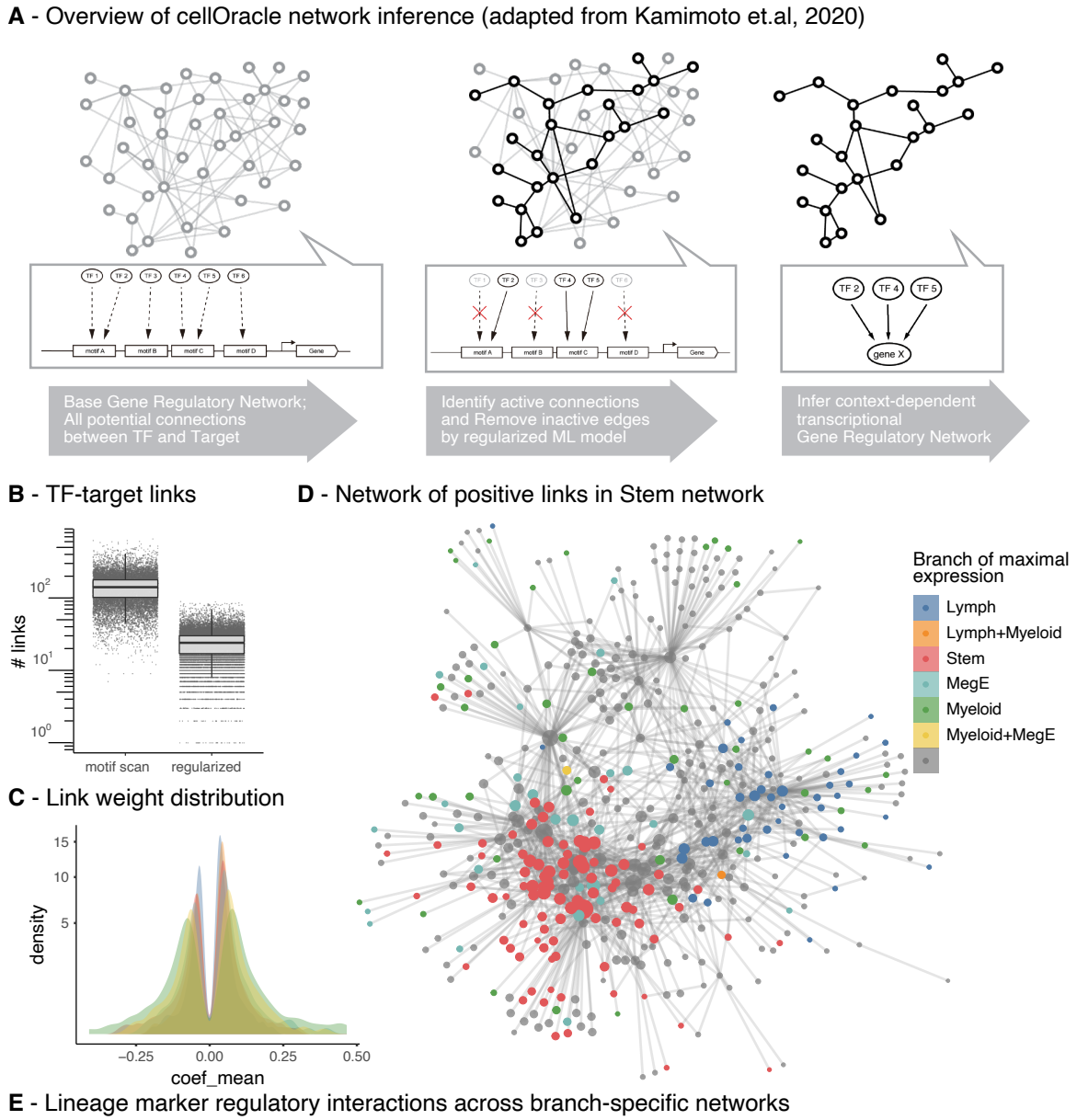

A - Overview of cellOracle network inference (adapted from Kamimoto et.al, 2020)
Base Gene Regulatory Network;
All potential connections
between TF and Target
Identify active connections
and Remove inactive edges
by regularized ML model
Infer context-dependent
transcriptional
Gene Regulatory Network
gene X
B - TF-target links
links
motif scan
regularized
C - Link weight distribution
density
coef_mean
D - Network of positive links in Stem network
Branch of maximal expression
Lymph
Lymph+Myeloid
Stem
MegE
Myeloid
Myeloid+MegE
E - Lineage marker regulatory interactions across branch-specific networks
Stem (Ly + My + MegE)
Ly + My
My + MegE
My
Ly
MegE
Stem
Lymph
Mono
Neut
Baso
Ery
Mk
regulation weight
activation
0.00
inhibition

Figure 5.3

Figure 5.3 *(previous page)*. **A**, Overview of cellOracle pipeline. *Left*: Potential connections between TF and target genes are detected based on motif analysis of co-accessibility patterns between TSSs and ATAC peaks to generate a base GRN. *Centre*: Regularized ML models are used to prune connections based on the expression of transcription factors. *Right*: Regularized regression is performed in each partition to obtain context-dependent GRNs. **B**, Number of links detected in the base GRN and in the final regularized network. Transcriptional-dependent regularization reduces the number of links by a factor of 10. **C**: Distribution of link weights in branch-specific GRNs displays higher prevalence of positive regulation across branches. **D**, Force-directed embedding of positive links in the stem network, in which each node (gene) has been coloured based on the branch of maximal expression. **E**. Overall interactions between lineage markers in each of the branch-specific networks. Each panel refers to a different branch, while each dot represents the aggregate interaction between lineage markers of row and column lineages. The size and colour of each dot indicate the overall strength and direction of inferred regulatory interactions.

correlation across the branches (Fig. 5.4B), in which the three major groups are reproduced by using hierarchical clustering.

At the gene level (Fig. 5.4A, right panel), most genes occupy the centre of the landscape, while known lineage markers are detected in correspondence of the myeloid (*Cebpb*, *Elane*, *Spi1*), stem/lymphoid (*Hlf*, *Satb1*) and megE (*Gata1*, *Apoe*, *Fli1*) branches.

Of particular interest is the centrality of transcription factors, whose expression causes the regulatory heterogeneity so far described. I selected TFs whose centrality is highly dynamic across branches, thus repeated the hierarchical clustering analysis (Fig. 5.4C). While the resulting clusters are similar to the previous descriptions, it is clear here how the Gata factors are the main drivers of megE and basophil branches, while the stem cell state is dependent on a handful of Klf factors. Finally, myeloid lineages are strongly enriched in *Spi1* (encoding PU.1) and *Cebpb*.

A more detailed regulatory analysis can be achieved by comparing gene centralities in a pairwise fashion, as shown in Figure 5.5. In particular, this comparison outputs *Gata1* as the primary driver in the emergence of megE and basophil branches from stem cells, while *Gata2* is also differentially central when these lineages are juxtaposed with lymphoid and myeloid cells. Interestingly, the separation between basophils and megE is imputed to genes (*Auts2*, *Nrf1*) whose role in this branching point is not studied yet, offering interesting candidates for further investigation. On the other hand, lymphoid emergence is attributed to *Hlf* and *Klf* factors, whose interactions and involvement in haematopoiesis have been documented [136, 137], although the similarity in the binding motifs of such factors makes it difficult to pin which specific ones are involved in lymphoid emergence.

Finally, I set to analyse the stem cell subsets detected using trajectory inference and cluster mappability analysis (Chapter 3) and trajectory inference (Chapter 4) to gain a deeper understanding of the regulatory roots of the detected heterogeneity. First I focused on regulatory differences between early stem cells and their immediate multi-potent progeny. Dif-

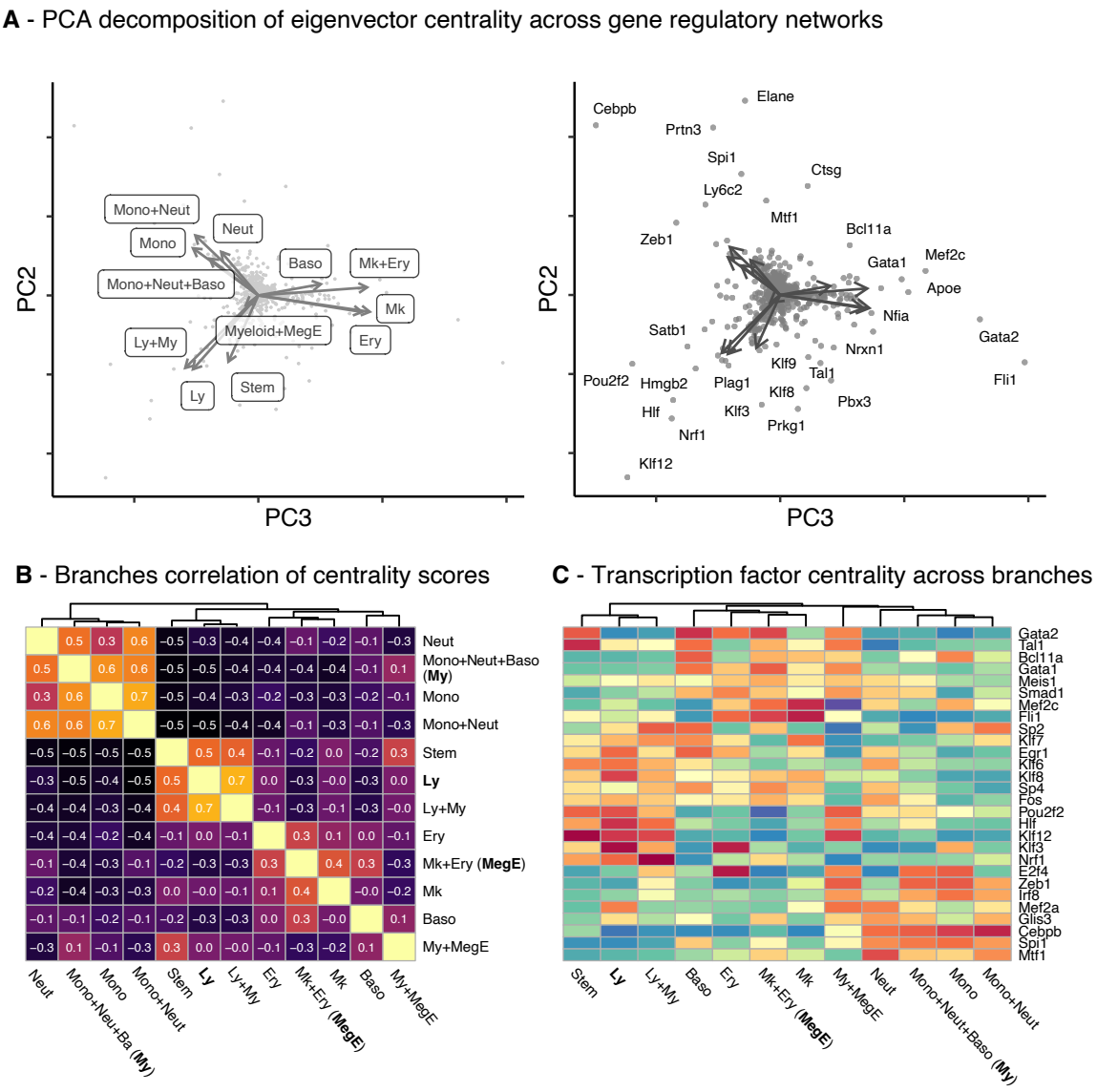


Figure 5.4. **A**, PCA decomposition of gene centralities across gene networks computed on the expanded branch model. *Left*: Branches projection on PC2 and PC3 results in three groups of branches: myeloid, stem+lymphoid, and megE+baso, highlighting the regulatory similarities across branches. *Right*: Gene projections on principal components unveil how regulatory heterogentiy is largely dependent on a handful of lineage markers marked by large principal component projections. **B**, Pearson correlations between gene centrality scores across branches are hierarchically clustered, confirming the branch groups resulting from PCA decomposition. **C**, Hierarchical clustering of TF centrality scores highlights regulatory determinants of regulatory heterogeneity across branches.

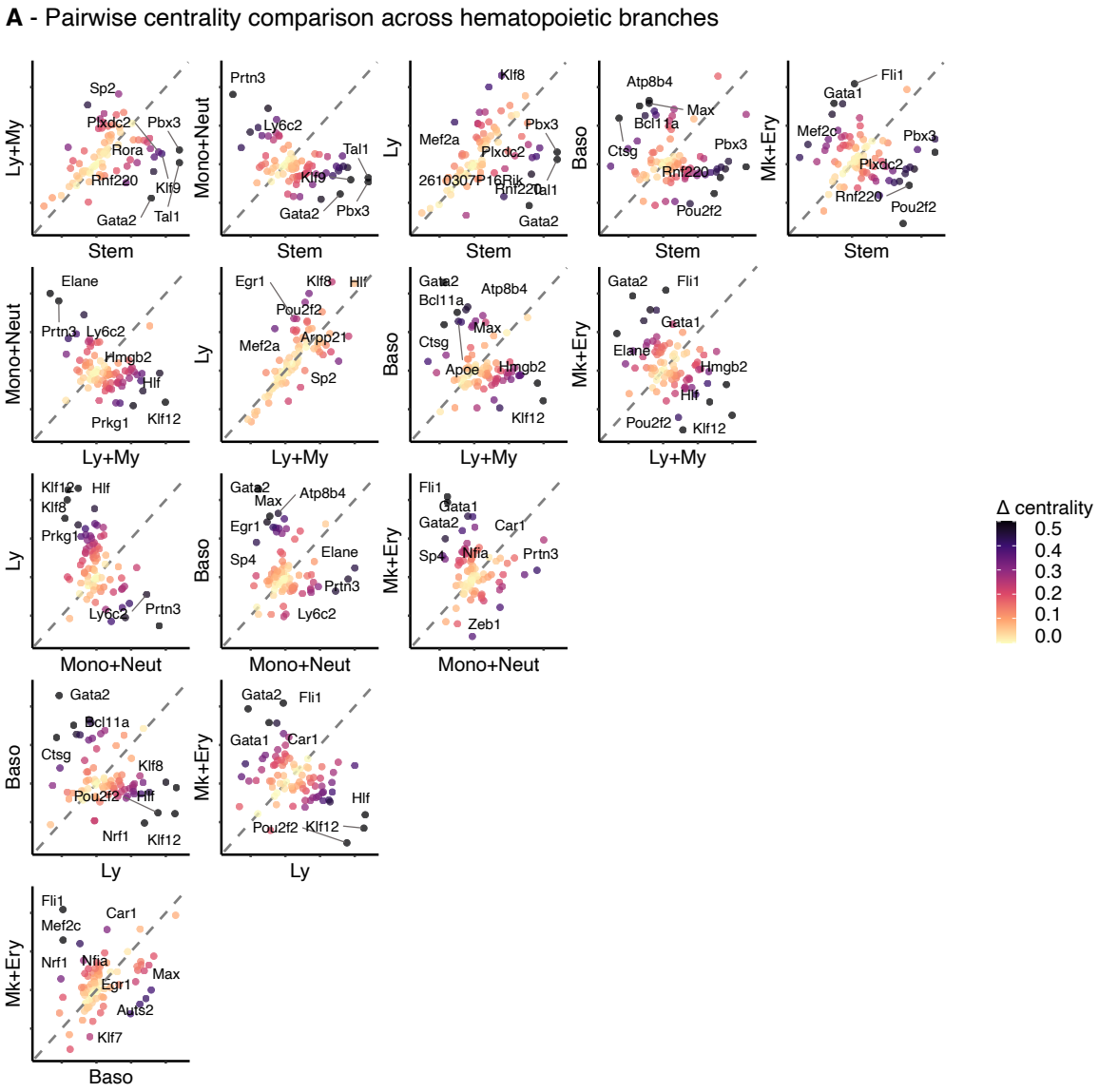


Figure 5.5. Pairwise comparison of gene centrality scores across selected branches.

ferential centrality analysis between the subsets yielded genes whose centrality is dynamically regulated during the transition from early stem cells to multipotent progenitors (5.6A). The early stem cell state is marked by high centrality of crucial transcription factors. In particular, *Egr1* has been assigned a role in keeping HSCs from proliferating and differentiating [138], similarly loss of *Klf6* results in a more proliferative and differentiated state similar to ageing [139]. *Nrxn1* and *Ncam2* were first described in a neural context, but the former acts as an inhibitor of proliferation in HSCs [140], while the latter encodes and adhesion molecule that could play a role in physical interaction with the niche [141]. At the other end of the spectrum, exit from the primitive state is associated with the regulation of *Elf1*, a critical regulator of haematopoietic activation, involved in responsiveness to interferon signalling [142], positive regulation of *Meis1* [143], and co-regulation of myeloid and lymphoid development together with *Fli1* and *Ets1* [144]. Intriguingly, the same transcription factor is the main driver of the regulatory identity of the ATAC-specific cluster detected in Chapter 3, together with *Pou2f2*, *Nfia*, and *Mef2c*, whose role has been described in regulation of the emergence of multiple branches [61].

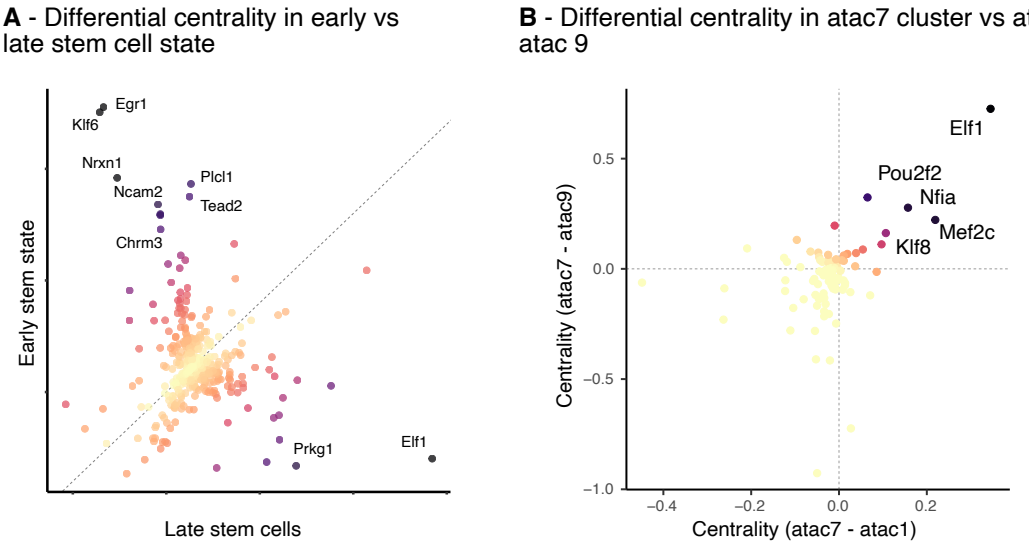


Figure 5.6. **A**, Pairwise comparison of gene centrality between early and late stem cell subsets. **B** Pairwise comparison of differential gene centrality between cluster atac7 and neighbour clusters atac1 and atac9. Each axis contains the difference in gene centrality between cluster 7 and the reference ones.

5.5 Discussion

In the present chapter, I investigated the molecular dynamics of haematopoiesis with a focus on the interrelationship between RNA and ATAC features. This exploration elucidates the cis-regulatory dynamics that underpin lineage determination. First, I investigate how haematopoietic commitment is achieved at the chromatin level. Two possible mechanisms have been proposed: de novo opening of chromatin regions for specific lineages or initial openness followed by progressive restriction of alternative fates (permissive model). Previous estimates of the proportion between the two mechanisms estimated between 50 % [43] and 70% [123] relying on a bulk-level analysis of chromatin data. Here I used the scATAC-seq dataset to confirm that the permissive model is the dominant one in differentiation. Moreover, leveraging the pseudotemporal ordering enabled by single-cell resolution, I established that the two mechanisms are likely to follow an order in which the closing of stem cell peaks precedes de-novo opening of lineage-specific regions, with a large overlap between the two regimens. A similar transient state, in which downregulation of stem features and upregulation of lineage-specific ones coexist, is described in [130, 145] for the LMPP population at the transcriptional level, but here is shown for chromatin features and all lineages. The in-depth analysis of lineage markers sheds further insight into the combinatorial nature of haematopoietic commitment. Following up on the previous findings on pseudotemporal onset of proliferation and lineage commitment, here I showed that while myeloid and megakaryocytic-erythroid lineages upregulate metabolic markers related to translation, the lymphoid branch shows little differences with respect to the stem cell branch. This metabolic activation is accompanied, by several closing events on the chromatin level, while the lymphoid branch follows the opposite trend, opening lymphoid and myeloid-specific regions of DNA. This observation confirms the previously discussed notion that lymphoid specification is not, at this level, accompanied by the exclusion of myeloid fates [128]. This plasticity in lymphoid commitment is likely to be an important element in response to challenges, during which lymphoid MPPs are temporarily re-directed towards the myeloid branch to meet the increased demand for innate immune cells.

However, this is not the sole interpretation of the observed behaviour. It has to be noted how the lymphoid end of the haematopoietic hierarchy here is phenotypically within the MPP compartment, and thus less differentiated than its myeloid-erythroid counterparts. It is possible that the observed metabolic activation and chromatin closing events occur more downstream in the lymphoid maturation process, in the CLP compartment [132]. In addition, the effects described here do not account for the presence, within the lympho-myeloid MPP compartment, of the recently described [119] embryonic MPP population that provides a lifelong contribution to the production of mature lymphoid and myeloid populations:

the features observed in the lymphoid compartment could be a distinctive signature of this eMPPs subset on the chromatin level.

Overall, I show here how commitment at the chromatin level follows a hierarchical order in which shared markers are regulated first, while lineage-specific ones are affected later, and downregulation of stem markers happen before the upregulation of lineage-specific ones.

In the following analysis, I used mathematical modelling to quantify chromatin priming with respect to RNA priming using both promoter and gene body accessibility as defined in [131] and DORCs as defined in [48], which focus on more distal elements. While analysis using the GSM uncovered a consistent amount of chromatin priming, especially in lymphoid and mk-ery lineages, the results from the DORC analysis seem counterintuitive, as in most cases RNA upregulation precedes DORC opening, in stark contrast with what described in [48]. As mentioned in the results paragraph, it is worth noting that in a model in which regulation is a multi-step process where protein and chromatin interact in successive steps, it is not guaranteed that peaks correlated to RNA expression are situated higher in the causal (and thus temporal) chain that triggers transcription.

In the last section of this chapter, I used cellOracle [53] to infer branch-specific GRNs and extrapolate putative causal links between TFs and target genes and extract dynamical features of haematopoietic differentiation. In this case too, I observed a highly dynamic regulatory landscape, in which differentiation is achieved by hierarchical steps of initial co-activation, followed by mutual inhibition of lineage-specific markers. By inferring GRNs on the high resolution branches defined in the previous section, I obtained a low-dimensional representation of the regulatory landscape in haematopoiesis. Interestingly, the basophil lineage, which was previously associated with both myeloid and megakaryocytic-erythroid lineages, is here closer to the latter, although retains typical myeloid-associated factors, such as *Spi1*. In general, this analysis showed how the haematopoietic diversity is can be recapitulated by differential regulation by a limited number of transcription factors.

Chapter 6

Discussion

Running alongside biotechnological advancement, our understanding of haematopoiesis has evolved to incorporate evidence from different modalities. As discussed in the introduction chapter, each technology influences the theoretical framework used to describe the system. For example, the transition from FACS-based analysis to single-cell sequencing technologies fuelled a shift in the description of the haematopoietic hierarchy, from homogeneous discrete populations to a landscape of continuous states. In this thesis, I integrated lineage tracing, single-cell RNA sequencing, and single-cell ATAC sequencing data to gain a more unbiased description of haematopoietic commitment.

Each of the previous chapters includes a Discussion section, thus topics discussed here pertain to topics that involve multiple chapters.

6.1 The topology of the haematopoietic system

A key topic that I addressed in this thesis is the topology of the haematopoietic system. Throughout the previous chapters, I used different methods based on the dataset at hand to extract salient aspects of the differentiation hierarchy. The term "topology" itself underwent a fundamental and yet perhaps overlooked evolution during the advent of single-cell sequencing technologies: historical studies based on FACS and transplantation yielded a tree-like model in which links between populations represented likely differentiation routes. On the other hand, the renewed attention on molecular heterogeneity brought by scRNA-seq shifted the emphasis on proximity relations rather than progeny, given how similar or overlapping in the transcriptional landscape populations appear [26, 58]. However, this observation could be a mere consequence of the visualization algorithms used to embed the data in a two-dimensional layout: even when used on FACS data (usually featuring between 5 and 10 markers in their feature space), the same visualization methods (e.g. UMAP, t-SNE and PHATE) return embeddings in which the gate-defined populations are highly overlapping, due to the nonlinear nature of such methods [146]. Moreover, the interpretation of such plots as a map of states that cells follow while undergoing differentiation is easiliy prone to oversimplification: higher-dimension dynamics forced into a two-dimensional representation result in chaotic behaviour [147], i.e. the high-dimensional trajectories that cells undergo during differentiation are mapped to a discontinuous series of "jumps" in the two-dimensional embedding. Across this thesis, I used trajectory inference methods that account for the higher dimensionality of the data [22, 24] and combined them with lineage tracing data to robustly reconstruct differentiation pathways in haematopoietic datasets.

In the first chapter, this approach was facilitated by the availability of index-sorted data that allowed to combine and compare lineage tracing data and transcriptional heterogeneity. The multiomic dataset shown in the later chapters does not contain labelling propagation evidence, but I leveraged its transcriptional layer to map trajectory inference results onto fate mapping datasets.

Analysis of both datasets robustly confirmed the differentiation pathways leading to megakaryocyte differentiation from tip stem cells:

- A subpopulation of HSCs can be labelled as tip multipotent stem cells: in Chapter 1 this subset was defined based on surface expression of EPCR, while in Chapter 4 pseudotime estimation yielded a locally dense cell-state with the lowest pseudotime;
- Downstream, a subset of HSCs is myelo-erythroid restricted and generates MkPs. In Chapter 1, the subset was detected using stratification of LT-HSCs based on Sca-1 expression, while in Chapter 4 it was detected using the iterative lineage potential model on trajectory inference data.

However, while direct differentiation from HSCs to MkPs does not contradict the results of Chapter 4, the multiomics dataset does not explicitly support the two alternative paths as separate. Notably, the MPP2 subset, reported to have the highest MegE bias among the MPP subsets [76], is not explicitly accounted for in the mathematical model used in [148]. While this population cannot be the source of the fastly labelled $CD48^{-/lo}$ MkPs (Supp. Figure 8e in [148]), they are likely to represent an intermediate pathway between direct HSC differentiation and progressive restriction through CMP and MEP. In other words, the two distinct pathways might be opposite ends of a spectrum of possible input into the MkPs that span HSCs, MPP2 and MEPs population. If that were the case, it is reasonable to expect that there are multiple adjacent "entry points" that are similar both in terms of surface marker expression (aside from Sca-1, MkPs and HSCs can virtually have the same surface marker profile [69], thus resulting hard to distinguish solely based on chromatin similarity).

Moreover, flux estimates between populations available in [148] indicate that the daily flow of cells from Sca1-/lo HSCs to MkPs (10^3 cells/day) is lower than the one from CMPs to MkPs by a factor of 4.3, thus resulting in a pathway whose low cell density can compromise direct observation on snapshot sequencing data.

The emergence of basophil/mast branch from HSPCs is another example of a tolpological aspect of haematopoiesis that benefits from the integration of lineage tracing and single-cell omics technologies. Initially grouped with the neutrophil branch due to their morphological and functional similarity, recent in-vitro and fate mapping evidence highlighted how basophil and mast cell development are dependent on *Gata1* expression [96,149]. Moreover, single-cell sequencing studies have highlighted the proximity between basophils and megE branches using either tree-like hierarchical clustering of single-cell transcriptomes [113] or two-dimensional embeddings [58]. On the other hand, several other studies link basophils transcriptomes to both meyloid and megE branches based on high-dimensional modularity analysis of clustering solutions [22], and trajectory inference constrained to tree topologies (Supp. Fig. 6d in [80]). Moreover, recent fate mapping evidence suggests abundant input from a myeloid GMP-like cluster [108]. These two lines of evidence are contradictory only if haematopoietic differentiation is interpreted as a tree-like system, in which each mature lineage has only one progenitor: using a clustering-free trajectory inference approach paired with statistical testing, I was able to detect the link between basophil emergence and both myeloid and megE branches. Nonetheless, gene regulatory network analysis clearly groups basophils closer to the megE branch, indicating that *Gata1*-driven modulation of gene expression is likely to represent a necessary step along the path of basophil maturation.

Another interesting aspect emerging from this analysis is the interplay between landscapes generated through single-cell sequencing and their dynamical behaviour. Most single-cell sequencing studies correctly place MkPs close to HSCs. On the other hand, transcriptional (and epigenetic) analysis also places the lymphoid region of the system close to LT-HSCs, but all available lineage tracing evidence shows how lymphoid cells receive label very slowly. One can speculate on what causes this dissonance: the computational steps that produce trajectory inference and pseudotemporal estimates are built on a series of steps that can bias the analysis. In particular, manifold distance registered by pseudotime is influenced by the features selected, the normalization and the dimensionality reduction methods: all these factors can affect the final result. But more importantly, the kinetic rates that regulate the flux between populations cannot directly be read using single-cell molecular assays. For example, lymphoid MPPs are characterised by a high self-renewal tendency, contrary to their MPP2 counterpart, as shown by the integrated approach followed in [108].

6.2 The interplay between transcriptional and chromatin landscape

The earliest analyses of multiomic datasets were based on unpaired data, in which distinct biological replicates would be used for different assays. This procedure is cheaper but creates the necessity to successively align modalities, a task that can only be performed by assuming that the same topology underlies the two modalities. In general, some pipelines are focused on integrating the two layers to create a unique embedding that incorporates features from each modality. In this thesis, I compared transcriptional and chromatin layers without integrating the two layers, but rather emphasizing the subtle differences between the modalities and their biological significance.

A positive control for the detection of differential structures across layers is the cell cycle signal: its prominence in RNA over ATAC is well documented [74] and confirmed in this dataset. The availability of chromatin information, which is roughly independent of cell cycle but can, at the same time, produce a fine-resolution description of the differentiation landscape, allowed me to unveil how cell cycle is related to differentiation (for a description of why this is not possible on scRNA-seq data alone, see Chapter 4). From the outcome of the Chapter 4 analysis, I could conclude that exit from the tip state in HSCs does not immediately imply cell cycle activation, but rather an intermediate state in which stem features are lost, followed by proliferation and onset of lineage programs.

In Chapter 3, a direct comparison of clustering solutions on both datasets revealed that transcriptional profiling is more sensitive to cytokine signalling, while chromatin landscape

can detect a cluster enriched in Ctcf motifs which has no transcriptional signature. Classification of these cells using the iterative potential assignment on t.i. results revealed that they might be slightly biased towards the myelo-erythroid end of the differentiation spectrum, hinting at a lineage-primed cell cluster that cannot be detected using transcriptional analysis.

Chapter 5 contains a more systematic analysis of lineage priming. The lymphoid lineage shows a consistent signature of massive priming: while it's the most similar to stem cells transcriptionally, differentially accessible region analysis has shown that massive chromatin re-organization has already taken place at this stage. The single-cell resolution of this multiomic dataset has allowed, in addition, to obtain pseudotemporal series of promoter availability and compare it side by side with the transcriptional onset of lineage markers, revealing chromatin priming in the promoter lineage for all the detected lineages. Alternatively, the detected changes that accompany lymphoid differentiation may originate from the embryonically-derived MPP biased subset described in [119]. However, aside from its embryonic origin, the eMPP subset has a kinetic and molecular behaviour that is indistinguishable from HSC-derived MPP4, raising the question of whether its independent origin has any further implications in differentiation.

6.3 Future directions

Molecular decisions inside single cells depend on extremely diverse mechanisms, that involve chromatin conformation, epigenetic modification, transcription dynamics, post-transcriptional effects, and protein degradation. Aside from extrinsic regulation, all these layers provide a causal contribution in shaping the differentiation path that single cells undergo. Unsurprisingly, thousands of such mechanisms have been investigated in the literature, and their impact on the haematopoietic system assessed. It has now become clear that it is unlikely that one particular interaction has a hierarchical superiority over the others, but decisions are an emergent property that needs a profound knowledge of the internal state of a cell to dissect.

In this thesis, I integrated multiple data modalities to gain a deeper description of such cellular states. However, inferring causal mechanisms that direct cell behaviour from snapshot data requires the adoption of strong assumptions that might limit the applicability of the findings.

For this reason, a desirable future direction for this project is the integration of single-cel barcodes from published datasets [27, 28] through transcriptional mapping, similar to what is shown in Chapter 4 to validate trajectory inference.

However, it is important to note that the results displayed in this thesis were obtained on a single multi-omic dataset. Two more replicates have so far been generated and findings presented in this thesis on the topics of lineage priming, haematopoietic topology and molecular regulation of commitment will be tested against biological replicates before publication.

Methods

In this Appendix I provide a more detailed account of the bioinformatic methods used to obtain results presented throughout the thesis. When previously published methods were used, the relevant publications are referenced. In the following text, methods are ordered based on the Chapter and section in which they first appear. To facilitate reading, functions from specific R libraries are indicated using the *package::function* syntax. Upon publication, the code used to produce results related to the multi-omic dataset will be published as a repository. To facilitate

Chapter 2

Methods for Section 2.3

Preprocessing of scRNA-seq data

The function *isOutlier* from the R package *scran* was used to detect and exclude quality control outliers (library size, detected genes, ERCC, and percentage of mitochondrial counts). Given that one plate used for Smart Seq2 exhibited a large number of poor-quality cells, detection of low-quality cells in this plate was performed by comparison with QC values from the remaining plates. Normalisation was performed using the pooling method (*scran::computeSumFactors* function) followed by scaling within each plate to account for heterogeneous coverage (*batchelor::multiBatchNorm*). After observing poor mixing of plates in dimensionality reduction plots, batch effects were corrected using a linear model (*limma::removeBatchEffect* method). Feature selection was performed using variance modelling on spike-in transcripts (*scran::modelGeneVarWithSpikes* blocked by plate) and selection of genes whose variance could not be fully ascribed to technical variation (biological variance >0 and FDR< 0.05,), thus excluding cell cycle genes from the Gene Ontology database. Selected features were used to compute PCA (*scater::runPCA*).

Embedding using PHATE

PHATE embedding was computed using the phateR package on the first ten principal components.

Scoring of transcriptional lineage sets

Lineage sets were obtained from [60] and scored using the *Seurat::AddMosuleScore* function.

Normalization of surface markers expression

Surface markers expression was transformed using log10 transformation with a pseduocount of 1.

Methods for Section 2.4

Gating on surface marker expression

Sca-1, CD48, and CD201 (EPCR) gating was performed by identifying the minimum in the bimodal distribution of normalised surface marker expression in the considered population. In cases where the distribution did not follow a bimodal distribution, I selected thresholds that reflected the proportions of cells in each gate in the FACS experiments.

Computation of diffusion pseudotime

To compute diffusion pseudotime, I computed Diffusion Maps (*destiny::DiffusionMap*) on 10 PCs and used the cell with lowest PHATE1 coordinate as root, as it correlated with expression of stem markers.

Computation of PAGA graph

PAGA was performed using the python package *scanpy* package on the knn (k =15) neighbourhood graph computed on the first 10 PCs.

Methods for Section 2.6

Gene sets pertinent to specific lineages were retrieved from the Molecular Signatures Database (Msigdb; [63]) via the MsigdbR package, targeting the cell type-specific collection (C8) through lineage-associated keyword searches. Overrepresentation analysis was subsequently conducted using the clusterProfiler::enricher function. Similarly, Gene Set Enrichment Analysis (GSEA) was implemented using the *clusterProfiler::GSEA* function.

Chapter 3

Methods for Section 3.2

Cell calling and pre-processing

Each of the four samples was separately analysed using the R *DIEM* package [150] to identify debris and empty doublets based on the percentage of mitochondrial genes and *Malat1*. Droplets whose debris score exceeded the default threshold of 0.5 were classified as debris. Additionally, the package *DropletUtils* [151] was used to obtain an independent estimate of empty droplets using the *emptyDrops* function with an FDR threshold of 0.001. Droplets classified as non-empty by both methods were retained for further processing.

Cell quality filter was based on both transcriptional (library size, mitochondrial reads) and chromatin (number of fragments, TSS enrichment) metrics. Cells that exceeded sample-specific thresholds determined by the overall metric distribution were retained for downstream analysis.

The retained cells displayed a high amount of ambient RNA signal, likely due to residual cytoplasmatic material during sequencing. To remove it, I used the *celda::decontX* function [152] on each sample.

Normalization, feature selection and dimensionality reduction (scRNA-seq)

Ambient-corrected RNA counts were normalised using the pooling method from the *scran* package [77]. Feature selection (n = 1000) was performed excluding genes involved in cell cycle according to Gene Ontology (GO:0007049) and genes with high contribution to the ambient RNA profile using the *Seurat::FindVariableFeatures* function.

Normalization, feature selection, and generation of feature matrices (scATAC-seq)

The tile matrix was obtained using the *ArchR* package [131] and reduced using the *iterativeLSI* function. Clustering was computed using a resolution of 0.5 on the LSI matrix and Peaks were called using the *addReproduciblePeakSet* function (n = 193554), thus the peaks obtained were used to obtain a Peak Matrix using *ArchR::addPeakMatrix*.

To obtain a Motif Matrix, I used the *addMotifAnnotations* and *addDeviationsMatrix* from the *ArchR* package using the cisbp database motifs [153].

Peaks were normalised using *Signac::RunTFIDF* to normalise peak counts and selected using *Signac::FindTopFeatures*(min.cutoff ='q50') to identify top accessible peaks (n = 96797). SVD reduction (analogous to PCA for scRNAseq data) was obtained using the *RunSVD* function, followed by the exclusion of the first component due to the high correlation with the number of fragments.

Methods for Section 3.3

Transcriptional cell cycle annotation was obtained using the *Seurat::CellCycleScoring* function, while *Seurat::ScaleData* with the G2M and S scores as variables to regress yielded a linearly corrected feature matrix. To produce the embeddings deriving from linear cell cycle correction shown in Fig. 2C, PCA and UMAP (on 20 PCs) were computed using default Seurat functions.

Next, principal component analysis (*Seurat::RunPCA*) and Independent component analysis (*Seurat::RunICA*) were performed on the selected features of the normalised expression matrix. Components from both reductions that were maximally correlated to G2M and S scores were compared (Fig. 2D) and, following the observation of the highest mappability of cell cycle scores to ICs, the two components with the highest correlation to proliferation scores were removed before computation of the UMAP embedding, effectively using 18 ICs to compute the UMAP coordinates.

To compute assortativity scores, Diffusion Maps were computed for each of the cell cycle regression pipelines (no regression, linear regression, ica-based regression) and compared to the ATAC embedding. On each of these Diffusion Maps embeddings, a neighborhood graph (k=30) on the first 20 diffusion components was computed (*Seurat::FindNeighbors* (annoy.metric = 'cosine', k.param = 30)) and assortativity quantified using the *igraph::assortativity_nominal* function.

Methods for Section 3.4

Regression of diffusion components using complementary assay using Random Forests

As a preprocessing step, individual RNA and ATAC diffusion components were scaled between 0 and 1, and outliers outside 0.005 and 0.995 quantiles were trimmed to these quantiles, as diffusion maps are known to assign extreme coordinates to few outliers, potentially hindering the precision of the regression pipeline. Next, the *randomForest::randomForest* function was used to train one model for each RNA component using all ATAC components as predictor variables. The same procedure was repeated by inverting the roles and using RNA to predict each ATAC component.

Clustering of Diffusion Maps embedding

Clustering on diffusion maps was performed using the Leiden algorithm (*Seurat::FindClusters* (algorithm = 4)) and a resolution of 0.3 and 0.7 for RNA and ATAC respectively. Cluster markers were obtained using *scran::findMarker* on normalised RNA and TF motif matrices with the following arguments: pval.type = 'all', direction = 'up', test.type = 'wilcox'.

Modularity analysis for RNA and ATAC clusters

Modularity between clusters was computed on the previously computed nearest neighbour graphs using *bluster::pairwiseModularity*(as.ratio = TRUE) and normalised by dividing for the modularity of single clusters. The difference in modularity across assays was quantified as the difference between the log2 transformed normalised modularity scores across assays.

Purity analysis and marker detection

Neighbourhood purity was computed using *bluster::neighborPurity(k=30)* on the Diffusion Map embedding of the relevant assay. To obtain modality-specific markers and motifs, the *scran::findMarkers* function was used, restricted to clusters with a high number of cross-modality neighbours (i.e. rna12 cells displayed a high number of ATAC neighbours in rna1 and rna3 and it was compared against them, atac7 was compared to atac1 for the same reason).

Chapter 4

Methods for Section 4.2

Downsampling strategy

The proportions between cells in the LT-HSC, ST-HSC, MPP, and LK compartments were computed based on the FACS experiment. Next, cells in the sequencing dataset were sampled according to these proportions. Given that preserving the original proportions was limited by the number of cells in the LK gate, the analyses that did not require an estimate of cell numbers in the LK gate were performed using the same strategy only on LSK populations to retain more cells.

Projection on full dataset using shared nearest neighbourhood graph

Projection of pseudotime and cellRank differentiation probabilities onto the full dataset was performed by assigning the missing metadata using a weighted mean over neighbours in the Shared Nearest Neighbor (SNN) graph (k=20) in which the weight is assigned based on the proportion of shared neighbours.

Methods for Section 4.3

CellRank trajectory inference

The subsampled dataset was exported into adata format to enable cellRank inference in a Python (3.9.12) environment. First, a knn graph (k=15) on LSI components was computed (*scanpy::pp.neighbors* function). The cellRank pipeline can use different kernels to compute the transition matrix between cellular states. Given the low reliability of velocity estimates in haematopoietic datasets, I opted for a pseudotemporal kernel *(cellrank.tl.kernels:: PseudotimeKernel*). Next, to compute diffusion pseudotime (*scanpy::tl.dpt* function), the root cell was selected as the one with the highest projection on the diffusion component maximally correlated to stem score. The transition matrix was then computed using the *cellrank:: compute_transition_matrix* function. Next, among the macrostates obtained using the *compute_schur* (using 20 components) and *compute_macrostates* (n_states =12) functions, 8 were selected as terminal states based on the maximal expression of lineage markers. Of note, while one of the endpoints in proximity of the basophil state was excluded due to low enrichment of lineage markers, two terminal states displayed high monotonic scores, and thus were both retained. The function *compute_absorption_probability* was then used to compare terminal differentiation probabilities. The absorption probabilities for the two monocyte states were summed.

Monotonic spline fitting for trend estimates (proliferation and lineage probability)

Prior to spline fitting, a dataframe containing the relevant quantities was generated:

- **Proliferative trend**: cells were grouped in bins with an equal number of observations using the *ggplot2::cut_number*(n=50) function and counted the portion of cycling cells (inferred phase ≠ G1) in each bin.
- **Lineage probability (upwards)**: First, cells whose probability for the considered lineage was lower than the threshold used for the potential model were excluded. Next, lineage-potent cells were binned using the *ggplot2::cut_number*(n=50) function. For each of the bins, the 0.9 quantile was used to robustly summarise upward probability trend estimates and exclude outliers.
- **Lineage probability (downwards)**: The same procedure for the upwards case was repeated with no cells excluded and the 0.05 quantile to summarise each bin.

A basis of integral splines with 7 knots is generated in correspondence of the following quantiles of the x range: {1/16, 1/8, 1/4, 1/2, 3/4, 7/8, 15/16} using *splines2::iSpline*. Integral splines are built as integral of b-splines, thus non-decreasing by construction. To

obtain a monotonic fit, I used the *penalised* package to constrain the coefficients to be non-negative. In particular, the *optL2* function was used to fit splines using ridge regression (which features a penalisation term proportional to the square sum of the coefficients) whose penalisation parameter was chosen using 10-fold cross-validation. To fit monotonically decreasing trends (i.e. in the downwards probability trend case) the metric of interest was multiplied by -1 and fitted using the described method.

The half-saturation constant was computed by predicting the smoothed value on the uniformly sampled pseudotemporal range and selecting the first value whose prediction exceeded half of the half-saturation value $y_{1/2} = min(f(x)) + 0.5 \cdot (max(f(x)) - min(f(x))$.

The super-linearity constant was computed by computing incremental ratios on the uniformly sampled pseudotemporal range and reporting the first one that exceeded 1.

Methods for Section 4.4

SingleR mapping to published datasets

For each of the comparisons, highly variable features (n=300) that are not annotated as proliferation markers or ambient genes were retained and the *singleR::SingleR function* was used to annotate cells from the query dataset using reference labels. Cells with high label uncertainty, marked as the ones in which the score difference (delta) between the top and second label was lower than an arbitrary threshold (set as the first quartile of the delta distribution), were excluded from figure 4.2.

Methods for Section 4.5

Iterative model for branch assignment

Initial potential assignment For each of the detected endpoints (lymphoid, monocyte, neutrophil, basophil, erythroid, megakaryocyte) a threshold for differentiation probability was set as 90% of the mean differentiation probability of early pseudotime (pt < 0.15) cells (see Fig.4.3A). Cells whose lineage probability is higher than the threshold are assigned potential for that lineage.

Iterative refinement The initial potential assignment outputs for each cell a combination of lineage potentials (e.g. mk + ery + mono + neutrophil). Downstream of the stem branch (in which all 6 potentials are present), cells have a lower number of fates available (e.g. cells that don't have lymphoid potential have 5 possible fates left). I reasoned that, once a fate is no longer available, the threshold for qualifying a cell as potent in a lineage A

or not is better computed on the lowest-potent progenitor that includes A as an attainable fate. For example, the threshold to assign neutrophil potential should not be computed on mean probability of stem cells, but rather the mean probability of cells with mono, neut, baso potential. The reason for this adjustment is rooted in the probabilistic assignment of differentiation probability in cellRank and the normalisation of probabilities to sum to 1: this means that cells with a low number of fates available have, by construction, higher probabilities for all the reachable lineages. In the schematic example shown in Fig 4.2B, the branching event generating the uni-potent branch A yields an increase in differentiation probability towards B and C due to the normalisation of probability.

To compensate for this artefact, I generated a loop that, for each oligopotent branch downstream of stem cells, detects downstream lineage potential combinations (e.g., the {B,C,D} branch has {B,C}, {B,D}, {C,D} as downstream combinations) and sets a new probability threshold based on the average probability for each fate available in the oligopotent branch:

Algorithm 1 Cell Fate Analysis

```
for nfates in [6, 5, 4, 3] do
    Identify unique root_branches                    ▷ e.g. "mk.ery.baso"
    root_branches ← unique values of cellrank_potential_i for cells with 'nfates'
    for each root_branch in root_branches do
        Split root_branch into available_fates
        available_fates ← split root_branch by '.'
        Identify downstream cells for the current root_branch
        downstream ← cells with nfates-1 & corresponding to available_fates
        for lin in available_fates do
            new_threshold ← 0.9 × mean(lin probability in root_branch)
            cells_to_downgrade ← downstream[lin probability < new_threshold]
            potential[cells_to_downgrade] ← potential[cells_to_downgrade] - lin
        end for
    end for
end for
```

PHATE embedding

PHATE embedding was obtained using the *phateR::phate* function (knn=30, t=10, decay=100) on the LSI components of the ATAC layer 2 to 30 (LSI1 was removed due to high correlation with number of fragments).

Multi-set intersection testing

The significance of multi-set intersection was quantified using the *supertest* function from the *SuperExactTest* package in R. Next, multi-set intersection testing was repeated after removing the multi-potent branch, to remove its influence on the significance of association between lineages.

However, this is not sufficient, as the null hypothesis of a hyper-geometric distribution includes an expected amount of multi-potent cells, thus removing all multi-potent cells would skew the results in the opposite way, i.e. lower the significance of overlaps artificially. Thus, in practice, unbiased estimation of overlap significance requires to "add back" some multi-potent cells in agreement with the prediction of the null model.

Let's consider K independent lineages and N cells in total. The goal of this computation is to compute the number of multi-potent cells to ad back (termed s) such that it fulfills the null hypothesis of lineage independence. For each fate k, n_k potent cells are present in the dataset. It follows that the probability of observing K-potent cells after adding s multipotent cells is: $\prod_{k=1}^{K} \frac{n_k+s}{N+s}$. This quantity needs to be equated to the frequency of added multipotent cells: $\frac{s}{N+s}$. The resulting equation was solved by generating a range of integers from 1 to N and selecting the best approximate solution. This procedure was implemented in R and used to adjust the significance in all the comparisons shown in Fig.4.3F.

Branching model

The main branching model was obtained by grouping lineages according to the maximum association detected in the previous stage: myeloid includes monocyte, neutrophil and basophil fates, megE includes megakaryocytic and erythroid, and lymphoid is unaltered.

In the expanded branching model, the main branches were retained unless a cell only had potentials belonging to the same branch, in which case the explicit potential combination was used to assign the branch. For example, a cell with monocyte and lymphoid potential retained the lympho-myeloid classification, while a cell with monocyte and neutrophil potential was assigned the mono+neut expanded branch.

Definition of lineage trajectories

Trajectories for each lineage were defined as the set of all cells that have lineage potential for the selected fate according to the iterative model.

Chapter 5

Methods for Section 5.2

Detection of differentiation-relevant features

First, variable genes were selected separately for each trajectory using *Seurat:: FindVariableFeatures* and selecting genes with variance stabilised mean>0.1 and stabilised variance >1, then excluding mitochondrial genes.

From the union of the variable features, differentiation-relevant features were detected using *scran:: findMarkers* (arguments: pval.type='all', test.type = 'wilcox', direction = 'up') on the expanded branches partition, restricted to stem cells and terminal branches. The procedure was repeated using direction = 'down' to detect downregulated lineage markers.

Due to the impossibility of selecting ATAC peaks based on mean-variance relationships, all ATAC peaks were retained for differential accessibility testing, using *scran:: findMarkers*(test.type = 'binom') to account for the discrete counts in the Peak matrix.

Over-representation analysis of shared GM and megE markers

Over-representation analysis of genes upregulated or downregulated in both GM and megE lineages was performed using the *msigdbr* package to import curated gene sets (category C2). The over-representation test was performed using *clusterprofiler::enricher* using the union of variable features used for differential testing as gene universe.

Detection of monotonic trends for individual markers

For each of the selected markers, pseudotemporal trends were computed as described in Section 6.3, using knn-pseudoaggregates in place of pseudotemporal bins to reduce noise in feature expression.

Methods for Section 5.3

Computation of DORCs

DORCs were computed separately for each of the trajectories. To increase the robustness of correlation estimates, pseudobulk knn-graph aggregates were generated using the Milo package in R [154] and used for the remainder of the DORC analysis.

In order to allow to compute correlations on a wider window, ATAC features were selected using a novel pipeline developed by Jonas Metz, not presented in this thesis. Feature selection on ATAC Peaks allowed to significantly shrink the feature space (from $2 \cdot 10^5$ to $2 \cdot 10^3$).

DORC were computed using the *Signac::LinkPeaks* function with the following parameters: distance = 10^9, min.distance= 10^3, score_cutoff =0.1. The minimal distance parameters were set to this value to reduce the contribution of ATAC reads in the gene body region to DORCs.

Detection of non-monotonic pseudotime series

To distinguish monotonic from non-monotonic pseudotemporal series, monotonic spline fitting as described in 6.3 was paired with a non-monotonic counterpart, in which coefficients were non-constrained. For each of the pseudotemporal series in each lineage, I computed the difference between the log-likelihood of monotonic and non-monotonic fit. Outliers in this distribution (for which $loglikelihood_{monotonic} << loglikelihood_{unconstrained}$ were detected using the *scater::isOutlier* function (nmads=3) and excluded from the priming delay statistics.

Methods for Section 5.3

Inference of Gene Regulatory Networks using cellOracle

The inference of gene regulatory networks was performed by replicating the pipeline described in the cellOracle package documentation:

1. Computation of co-accessibility scores Co-accessibility between peaks was computed in R using *ArchR::addCoAccessibility*(maxDist = 1e06) on the Peak matrix, generating a co-accessibility edge list.

2. Preprocessing of peak data (Python) First, transcription starting sites (TSS) are detected using the *motif_analysis* module from the *cellOracle* package. Next, only links involving TSSs are retained for downstream processing. The resulting list of peaks was annotated using *get_tss_info* such that each peak is paired with the correlated TSS. Next, peaks with low co-accessibility scores (r<0.75) are filtered out.

3. Enrichment of TF motifs in peaks (Python)

   In this phase, the list of peaks generated in the previous step is scanned for TF motifs. This is achieved by accessing the gimmemotifs database (gimme.vertebrate.v5.0.) using the *scan* method of the TFinfo class in the motif analysis (fpr=0.02). The enrichment results are filtered to retain only highly enriched motifs (enriched score >10). The result can be represented as a dataframe with a row for each peak and as one column for each TF whose motifs are enriched in any of the examined peaks. This dataframe constitutes the base GRN, whose links will be pruned in the final step.

4. Feature selection prior to final GRN inference

   In this step, unbiased feature selection of 3000 genes (excluding ribosomial, mitochondrial and ambient genes) was paired with a selection of 150 transcription factors based on an unbiased selection of highly variable transcription factors (vst.mean >0.1, vst.variance.standardized > 1) augmented with a set of transcription factor whose role in haematopoiesis has been described in the literature, for which the variability requirements for inclusion in the selection were relaxed (vst.mean > 0.05, vst.variance.standardized > 0.05).

5. Branch-specific inference of GRN (Python)

   Following the indication of the cellOracle documentation, knn imputation (*.knn_imputation*(k =83)) based on pca neighbours (n_pcs = 20) was used to obtain more robust estimates of gene expression and peak accessibility. Links between TF and target genes were obtained using a regularised linear regression method (bagging ridge) as described in [53], in which target gene expression is regressed as the weighted sum of signed contributions from putative TF regulators using the *get_links* function from cellOracle package. I performed a sensitivity analysis for the regularisation parameter and concluded that it has little relevance in the resulting networks, thus selected a value of alpha = 500. This procedure is repeated for each main branch and extended branch as described in section 6.3 to obtain stage-specific GRNs. Next, top links (n = 2000) are retained for downstream processing. Centrality measures across networks are computed using the *get_network_score* function of the network object.

Aggregation of regulatory interactions across branches

To obtain aggregate interaction scores across genes expressed in different branches, I utilised the lineage markers computed in Section 6.3 and computed the average weight links for interactions between genes of any pair of lineages for each of the inferred networks.

PCA of gene centralities

The eigenvector centrality metrics of the networks obtained using the extended branching model were loaded into an R session and merged by gene, resulting in a matrix with genes as rows and eigenvector centrality in each network as columns. PCA on this matrix was performed using *stats::prcomp*(center=T, scale=T). After noticing that the gene projections on the first principal component were correlated to mean gene expression (not shown), the centrality matrix was reconstructed by excluding the first component exploiting the properties of PCA decomposition: given that the PCA decomposition of the original matrix $X = AB$,

where A contains projections on samples onto principal components, while B contains the loadings of features onto PCs. To reconstruct X while regressing out the first component, it's sufficient to compute $X' = A'B'$, in which A' and B' are the same matrices as in the original decomposition in which the first row (A') and column (B') have been removed.

Corrected centrality matrices were used to compute correlation of centrality scores using Pearson correlation (Fig. 5.4B), hierarchical clustering of TF centralities (Fig 5.4C; using the pheatmap::pheatmap function) and pairwise centrality comparison (Figure 5.5).

List of Abbreviations

LT-HSC: Long-term hematopoietic stem cell
ST-HSC: Short-term hematopoietic stem cell
MPP: Multipotent progenitor
LMPP: Lympho-myeloid primed progenitor
CMP: Common myeloid progenitor
GMP: Granulocyte-macrophage progenitor
CLP: Common lymphoid progenitor
MEP: Myeloid-erythroid progenitor
HSPC: Hematopoietic stem and progenitor cell
HPC: Hematopoietic progenitor cell
MyP: Myeloid Progenitor
EryP: Erythroid Progenitor
preGM: myeloid-restricted pre-granulocyte-macrophage progenitor
GM: Granulocyte-macrophage
MegE: Megakaryocyte-erythroid
Mk: Megakaryocyte
MkP: Megakaryocyte Progenitor
Ery: Erythroid
scRNA-seq: Single-cell RNA sequencing
scATAC-seq: Single-cell assay for transposase-accessible chromatin with sequencing
GEX: Gene expression
FACS: Fluorescence-activated cell sorting
TF: Transcription factor
GRN: Gene Regulatory Network
LSK: Lin− Sca-1+ c-Kit+ cells
PCA: Principal component analysis
ICA: Independent component analysis
LSI: Latent semantic indexing
DAR: Differentially accessible region
DEG: Differentially expressed gene
GSM: Gene score metric
DORC: Domain of regulatory chromatin

Bibliography

[1] Jason Cosgrove, Lucie S. P. Hustin, Rob J. de Boer, and Leïla Perié. Hematopoiesis in numbers. *Trends in Immunology*, 42(12):1100–1112, December 2021. Publisher: Elsevier.

[2] Alexander Maximow. Untersuchungen über Blut und Bindegewebe. *Archiv für mikroskopische Anatomie*, 73(1):444–561, December 1908.

[3] J. E. Till and E. A. McCulloch. A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. *Radiation Research*, 14(2):213–222, 1961. Publisher: Radiation Research Society.

[4] J. W. Visser, J. G. Bauman, A. H. Mulder, J. F. Eliason, and A. M. de Leeuw. Isolation of murine pluripotent hemopoietic stem cells. *The Journal of Experimental Medicine*, 159(6):1576–1590, June 1984.

[5] Mark J. Kiel, Omer H. Yilmaz, Toshihide Iwashita, Osman H. Yilmaz, Cox Terhorst, and Sean J. Morrison. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell*, 121(7):1109–1121, July 2005.

[6] Sidinh Luc, Natalija Buza-Vidas, and Sten Eirik W. Jacobsen. Biological and molecular evidence for existence of lymphoid-primed multipotent progenitors. *Annals of the New York Academy of Sciences*, 1106:89–94, June 2007.

[7] David G. Kent, Michael R. Copley, Claudia Benz, Stefan Wöhrer, Brad J. Dykstra, Elaine Ma, John Cheyne, Yongjun Zhao, Michelle B. Bowie, Yun Zhao, Maura Gasparetto, Allen Delaney, Clayton Smith, Marco Marra, and Connie J. Eaves. Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. *Blood*, 113(25):6342–6350, June 2009.

[8] Brad Dykstra, David Kent, Michelle Bowie, Lindsay McCaffrey, Melisa Hamilton, Kristin Lyons, Shang-Jung Lee, Ryan Brinkman, and Connie Eaves. Long-term prop-

agation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell*, 1(2):218–229, August 2007.

[9] Patricia Jensen and Susan M. Dymecki. Essentials of Recombinase-Based Genetic Fate Mapping in Mice. *Methods in molecular biology (Clifton, N.J.)*, 1092:437–454, 2014.

[10] Katrin Busch, Kay Klapproth, Melania Barile, Michael Flossdorf, Tim Holland-Letz, Susan M. Schlenner, Michael Reth, Thomas Höfer, and Hans Reimer Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542–546, February 2015. Publisher: Nature Publishing Group.

[11] Kristina B. Schoedel, Mina N. F. Morcos, Thomas Zerjatke, Ingo Roeder, Tatyana Grinenko, David Voehringer, Joachim R. Göthert, Claudia Waskow, Axel Roers, and Alexander Gerbaulet. The bulk of the hematopoietic stem cell population is dispensable for murine steady-state and stress hematopoiesis. *Blood*, 128(19):2285–2296, November 2016.

[12] Joseph N. Pucella, Samik Upadhaya, and Boris Reizis. The Source and Dynamics of Adult Hematopoiesis: Insights from Lineage Tracing. *Annual Review of Cell and Developmental Biology*, 36(1):529–550, 2020. _eprint: https://doi.org/10.1146/annurev-cellbio-020520-114601.

[13] Weike Pei, Thorsten B. Feyerabend, Jens Rössler, Xi Wang, Daniel Postrach, Katrin Busch, Immanuel Rode, Kay Klapproth, Nikolaus Dietlein, Claudia Quedenau, Wei Chen, Sascha Sauer, Stephan Wolf, Thomas Höfer, and Hans Reimer Rodewald. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, 548(7668):456–460, August 2017. Publisher: Nature Publishing Group.

[14] Alejo E. Rodriguez-Fraticelli, Samuel L. Wolock, Caleb S. Weinreb, Riccardo Panero, Sachin H. Patel, Maja Jankovic, Jianlong Sun, Raffaele A. Calogero, Allon M. Klein, and Fernando D. Camargo. Clonal analysis of lineage fate in native hematopoiesis. *Nature*, 553(7687):212–216, January 2018.

[15] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. Publisher: John Wiley & Sons, Ltd.

[16] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing

structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, December 2019. Publisher: Nature Publishing Group.

[17] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, August 2023. Publisher: Public Library of Science.

[18] Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, October 2016. Publisher: Nature Publishing Group.

[19] Sophie Tritschler, Maren Büttner, David S. Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J. Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12):dev170506, June 2019.

[20] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, April 2014. Number: 4 Publisher: Nature Publishing Group.

[21] Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, June 2018.

[22] F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9, March 2019. Publisher: BioMed Central.

[23] Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'er. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*, 37(4):451–460, April 2019.

[24] Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. CellRank for directed single-cell fate mapping. *Nature Methods*, 19(2):159–170, February 2022. Number: 2 Publisher: Nature Publishing Group.

[25] Sam Watcham, Iwo Kucinski, and Berthold Gottgens. New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. *Blood*, 133(13):1415–1426, March 2019.

[26] Lars Velten, Simon F. Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P. Hennig, Christoph Hirche, Christoph Lutz, Eike C. Buss, Daniel Nowak, Tobias Boch, Wolf-Karsten Hofmann, Anthony D. Ho, Wolfgang Huber, Andreas Trumpp, Marieke A. G. Essers, and Lars M. Steinmetz. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19(4):271–281, April 2017. Number: 4 Publisher: Nature Publishing Group.

[27] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479):eaaw3381, 2020. Publisher: American Association for the Advancement of Science.

[28] Weike Pei, Fuwei Shang, Xi Wang, Ann-Kathrin Fanti, Alessandro Greco, Katrin Busch, Kay Klapproth, Qin Zhang, Claudia Quedenau, Sascha Sauer, Thorsten B. Feyerabend, Thomas Höfer, and Hans-Reimer Rodewald. Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding. *Cell Stem Cell*, 27(3):383–395.e8, September 2020.

[29] Nina Cabezas-Wallscheid, Florian Buettner, Pia Sommerkamp, Daniel Klimmeck, Luisa Ladel, Frederic B. Thalheimer, Daniel Pastor-Flores, Leticia P. Roma, Simon Renders, Petra Zeisberger, Adriana Przybylla, Katharina Schönberger, Roberta Scognamiglio, Sandro Altamura, Carolina M. Florian, Malak Fawaz, Dominik Vonficht, Melania Tesio, Paul Collier, Dinko Pavlinic, Hartmut Geiger, Timm Schroeder, Vladimir Benes, Tobias P. Dick, Michael A. Rieger, Oliver Stegle, and Andreas Trumpp. Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell*, 169(5):807–823.e19, May 2017.

[30] J. L. Curry, J. J. Trentin, and N. Wolf. Hemopoietic spleen colony studies. II. Erythropoiesis. *The Journal of Experimental Medicine*, 125(4):703–720, April 1967.

[31] Megan Scudellari. The innate debate over HSCs. *Nature Reports Stem Cells*, pages 1–1, August 2009. Publisher: Nature Publishing Group.

[32] Michael A. Rieger, Philipp S. Hoppe, Benjamin M. Smejkal, Andrea C. Eitelhuber, and Timm Schroeder. Hematopoietic Cytokines Can Instruct Lineage Choice. *Science*, 325(5937):217–218, July 2009. Publisher: American Association for the Advancement of Science.

[33] Sandrine Sarrazin, Noushine Mossadegh-Keller, Taro Fukao, Athar Aziz, Frederic Mourcin, Laurent Vanhille, Louise Kelly Modis, Philippe Kastner, Susan Chan, Estelle Duprez, Claas Otto, and Michael H. Sieweke. MafB Restricts M-CSF-Dependent Myeloid Commitment Divisions of Hematopoietic Stem Cells. *Cell*, 138(2):300–313, July 2009. Publisher: Elsevier.

[34] Vionnie W.C. Yu, Rushdia Z. Yusuf, Toshihiko Oki, Juwell Wu, Borja Saez, Xin Wang, Colleen Cook, Ninib Baryawno, Michael J. Ziller, Eunjung Lee, Hongcang Gu, Alexander Meissner, Charles P. Lin, Peter V. Kharchenko, and David T. Scadden. Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell*, 167(5):1310–1322.e17, November 2016.

[35] C. H. Waddington. *The strategy of the genes; a discussion of some aspects of theoretical biology*. Allen & Unwin, London, 1957. OCLC: 1491605.

[36] Alexandra Avgustinova and Salvador Aznar Benitah. Epigenetic control of adult stem cell function. *Nature Reviews Molecular Cell Biology*, 17(10):643–658, October 2016. Number: 10 Publisher: Nature Publishing Group.

[37] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, December 2013. Number: 12 Publisher: Nature Publishing Group.

[38] Cecilia Pessoa Rodrigues, Maria Shvedunova, and Asifa Akhtar. Epigenetic Regulators as the Gatekeepers of Hematopoiesis. *Trends in genetics: TIG*, pages S0168–9525(20)30251–1, October 2020.

[39] Jason D. Buenrostro, M. Ryan Corces, Caleb A. Lareau, Beijing Wu, Alicia N. Schep, Martin J. Aryee, Ravindra Majeti, Howard Y. Chang, and William J. Greenleaf. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*, 173(6):1535–1548.e16, May 2018. Publisher: Cell Press.

[40] Jason D Buenrostro, M Ryan Corces, Caleb A Lareau, Beijing Wu, Alicia N Schep, Martin Aryee, Ravindra Majeti, Howard Y. Chang, and William J. Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548.e16, May 2018.

[41] David Lara-Astiaso, Ainhoa Goñi-Salaverri, Julen Mendieta-Esteban, Nisha Narayan, Cynthia Del Valle, Torsten Gross, George Giotopoulos, Tumas Beinortas, Mar Navarro-Alonso, Laura Pilar Aguado-Alvaro, Jon Zazpe, Francesco Marchese, Natalia Torrea, Isabel A. Calvo, Cecile K. Lopez, Diego Alignani, Aitziber Lopez, Borja Saez, Jake P. Taylor-King, Felipe Prosper, Nikolaus Fortelny, and Brian J. P. Huntly. In vivo screening characterizes chromatin factor functions during normal and malignant hematopoiesis. *Nature Genetics*, 55(9):1542–1554, September 2023. Number: 9 Publisher: Nature Publishing Group.

[42] Constanze Bonifer and Peter N. Cockerill. Chromatin priming of genes in development: Concepts, mechanisms and consequences. *Experimental Hematology*, 49:1–8, May 2017.

[43] David Lara-Astiaso, Assaf Weiner, Erika Lorenzo-Vivas, Irina Zaretsky, Diego Adhemar Jaitin, Eyal David, Hadas Keren-Shaul, Alexander Mildner, Deborah Winter, Steffen Jung, Nir Friedman, and Ido Amit. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–949, August 2014. Publisher: American Association for the Advancement of Science.

[44] Elinore M. Mercer, Yin C. Lin, Christopher Benner, Suchit Jhunjhunwala, Janusz Dutkowski, Martha Flores, Mikael Sigvardsson, Trey Ideker, Christopher K. Glass, and Cornelis Murre. Multilineage Priming of Enhancer Repertoires Precedes Commitment to the B and Myeloid Cell Lineages in Hematopoietic Progenitors. *Immunity*, 35(3):413–425, September 2011. Publisher: Cell Press.

[45] Anjun Ma, Adam McDermaid, Jennifer Xu, Yuzhou Chang, and Qin Ma. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends in Biotechnology*, 38(9):1007–1022, September 2020.

[46] Sarah A. Vitak, Kristof A. Torkenczy, Jimi L. Rosenkrantz, Andrew J. Fields, Lena Christiansen, Melissa H. Wong, Lucia Carbone, Frank J. Steemers, and Andrew Adey. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, 14(3):302–308, March 2017. Number: 3 Publisher: Nature Publishing Group.

[47] Longqi Liu, Chuanyu Liu, Andrés Quintero, Liang Wu, Yue Yuan, Mingyue Wang, Mengnan Cheng, Lizhi Leng, Liqin Xu, Guoyi Dong, Rui Li, Yang Liu, Xiaoyu Wei, Jiangshan Xu, Xiaowei Chen, Haorong Lu, Dongsheng Chen, Quanlei Wang, Qing Zhou, Xinxin Lin, Guibo Li, Shiping Liu, Qi Wang, Hongru Wang, J. Lynn Fink, Zhengliang Gao, Xin Liu, Yong Hou, Shida Zhu, Huanming Yang, Yunming Ye,

Ge Lin, Fang Chen, Carl Herrmann, Roland Eils, Zhouchun Shang, and Xun Xu. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nature Communications*, 10(1), 2019. Publisher: Springer US.

[48] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.e20, November 2020.

[49] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, August 2023. Number: 8 Publisher: Nature Publishing Group.

[50] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, 2020.

[51] Jonas Simon Fleck, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, Malgorzata Santel, Zhisong He, J. Gray Camp, and Barbara Treutlein. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 621(7978):365–372, September 2023. Number: 7978 Publisher: Nature Publishing Group.

[52] Vinay K. Kartha, Fabiana M. Duarte, Yan Hu, Sai Ma, Jennifer G. Chew, Caleb A. Lareau, Andrew Earl, Zach D. Burkett, Andrew S. Kohlway, Ronald Lebofsky, and Jason D. Buenrostro. Functional inference of gene regulation using single-cell multi-omics. *Cell Genomics*, 2(9):100166, September 2022.

[53] Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, February 2023. Number: 7949 Publisher: Nature Publishing Group.

[54] Mina N.F. Morcos, Congxin Li, Clara M. Munz, Alessandro Greco, Nicole Dressel, Susanne Reinhardt, Katrin Sameith, Andreas Dahl, Nils B. Becker, Axel Roers, Thomas Höfer, and Alexander Gerbaulet. Fate mapping of hematopoietic stem cells reveals two pathways of native thrombopoiesis. *Nature Communications*, 13(1):1–13, August 2022. Publisher: Nature Publishing Group.

[55] Irving L. Weissman and Judith A. Shizuru. The origins of the identification and isolation of hematopoietic stem cells, and their capability to induce donor-specific transplantation tolerance and treat autoimmune diseases. *Blood*, 112(9):3543–3553, November 2008.

[56] Alejandra Sanjuan-Pla, Iain C. Macaulay, Christina T. Jensen, Petter S. Woll, Tiago C. Luis, Adam Mead, Susan Moore, Cintia Carella, Sahoko Matsuoka, Tiphaine Bouriez Jones, Onima Chowdhury, Laura Stenson, Michael Lutteropp, Joanna C.A. Green, Raffaella Facchini, Hanane Boukarabila, Amit Grover, Adriana Gambardella, Supat Thongjuea, Joana Carrelha, Paul Tarrant, Deborah Atkinson, Sally Ann Clark, Claus Nerlov, and Sten Eirik W. Jacobsen. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*, 502(7470):232–236, August 2013. Publisher: Nature Publishing Group.

[57] Simon Haas, Jenny Hansson, Daniel Klimmeck, Dirk Loeffler, Lars Velten, Hannah Uckelmann, Stephan Wurzer, Áine M. Prendergast, Alexandra Schnell, Klaus Hexel, Rachel Santarella-Mellwig, Sandra Blaszkiewicz, Andrea Kuck, Hartmut Geiger, Michael D. Milsom, Lars M. Steinmetz, Timm Schroeder, Andreas Trumpp, Jeroen Krijgsveld, and Marieke A.G. Essers. Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors. *Cell Stem Cell*, 17(4):422–434, October 2015. Publisher: Cell Press.

[58] Joakim S. Dahlin, Fiona K. Hamey, Blanca Pijuan-Sala, Mairi Shepherd, Winnie W. Y. Lau, Sonia Nestorowa, Caleb Weinreb, Samuel Wolock, Rebecca Hannah, Evangelia Diamanti, David G. Kent, Berthold Göttgens, and Nicola K. Wilson. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood*, 131(21):e1–e11, May 2018.

[59] Hideyuki Oguro, Lei Ding, and Sean J. Morrison. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell*, 13(1):102–116, July 2013. Publisher: Cell Press.

[60] Alejo E. Rodriguez-Fraticelli, Caleb Weinreb, Shou Wen Wang, Rosa P. Migueles, Maja Jankovic, Marc Usart, Allon M. Klein, Sally Lowell, and Fernando D. Camargo. Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature*, 583(7817):585–589, July 2020. Publisher: Nature.

[61] Christos Gekas and Thomas Graf. CD41 expression marks myeloid-biased adult hematopoietic stem cells and increases with age. *Blood*, 121(22):4463–4472, May 2013. Publisher: American Society of Hematology.

[62] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, December 2014. Publisher: BioMed Central Ltd.

[63] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005. Publisher: National Academy of Sciences.

[64] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. Publisher: NIH Public Access.

[65] Stefano Rossetti, André T. Hoogeveen, and Nicoletta Sacchi. The MTG proteins: Chromatin repression players with a passion for networking. *Genomics*, 84(1):1–9, July 2004. Publisher: Genomics.

[66] Anna H. Schuh, Alex J. Tipping, Allison J. Clark, Isla Hamlett, Boris Guyot, Francesco J. Iborra, Patrick Rodriguez, John Strouboulis, Tariq Enver, Paresh Vyas, and Catherine Porcher. ETO-2 Associates with SCL in Erythroid Cells and Megakaryocytes and Provides Repressor Functions in Erythropoiesis. *Molecular and Cellular Biology*, 25(23):10235, December 2005. Publisher: Taylor & Francis.

[67] Nickolas Steinauer, Chun Guo, and Jinsong Zhang. Emerging Roles of MTG16 in Cell-Fate Control of Hematopoietic Stem Cells and Cancer. *Stem Cells International*, 2017, 2017. Publisher: Hindawi Limited.

[68] Shu Sun, Chen Jin, Jia Si, Ying Lei, Kunying Chen, Yueli Cui, Zhenbo Liu, Jiang Liu, Meng Zhao, Xiaohui Zhang, Fuchou Tang, Matthew T. Rondina, Yueying Li, and Qian-fei Wang. Single-cell analysis of ploidy and the transcriptome reveals functional and spatial divergency in murine megakaryopoiesis. *Blood*, 138(14):1211–1224, October 2021. Publisher: American Society of Hematology.

[69] Thomas Höfer and Hans-Reimer Rodewald. Differentiation-based model of hematopoietic stem cell functions and lineage pathways. *Blood*, 132(11):1106–1113, September 2018.

[70] Yan Liu, Xinyi Zuo, Peng Chen, Xiang Hu, Zi Sheng, Anli Liu, Qiang Liu, Shaoqiu Leng, Xiaoyu Zhang, Xin Li, Limei Wang, Qi Feng, Chaoyang Li, Ming Hou, Chong Chu, Shihui Ma, Shuwen Wang, and Jun Peng. Deciphering transcriptome alterations in bone marrow hematopoiesis at single-cell resolution in immune thrombocytopenia. *Signal Transduction and Targeted Therapy*, 7(1):1–18, October 2022. Publisher: Nature Publishing Group.

[71] Jin Wang, Jiayi Xie, Daosong Wang, Xue Han, Minqi Chen, Guojun Shi, Linjia Jiang, and Meng Zhao. CXCR4high megakaryocytes regulate host-defense immunity against bacterial pathogens. *eLife*, 11, July 2022. Publisher: eLife Sciences Publications Ltd.

[72] Cuicui Liu, Dan Wu, Meijuan Xia, Minmin Li, Zhiqiang Sun, Biao Shen, Yiying Liu, Erlie Jiang, Hongtao Wang, Pei Su, Lihong Shi, Zhijian Xiao, Xiaofan Zhu, Wen Zhou, Qianfei Wang, Xin Gao, Tao Cheng, and Jiaxi Zhou. Characterization of Cellular Heterogeneity and an Immune Subpopulation of Human Megakaryocytes. *Advanced Science*, 8(15):2100921, August 2021. Publisher: John Wiley & Sons, Ltd.

[73] Elinore M. Mercer, Yin C. Lin, Christopher Benner, Suchit Jhunjhunwala, Janusz Dutkowski, Martha Flores, Mikael Sigvardsson, Trey Ideker, Christopher K. Glass, and Cornelis Murre. Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity*, 35(3):413–425, September 2011.

[74] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.e20, November 2020. Publisher: Cell Press.

[75] Kenji Kamimoto1, Christy M Hoffmann1, and Samantha A Morris1. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*, page 2020.02.17.947416, April 2020. Publisher: Cold Spring Harbor Laboratory.

[76] Eric M. Pietras, Damien Reynaud, Yoon A. Kang, Daniel Carlin, Fernando J. Calero-Nieto, Andrew D. Leavitt, Joshua A. Stuart, Berthold Göttgens, and Emmanuelle Passegué. Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell*, 17(1):35–46, July 2015. Publisher: NIH Public Access.

[77] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, April 2016.

[78] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, December 2019.

[79] Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, May 2015. Publisher: American Association for the Advancement of Science.

[80] Huidong Chen, Luca Albergante, Jonathan Y. Hsu, Caleb A. Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, Alexander N. Gorban, Daniel E. Bauer, Martin J. Aryee, David M. Langenau, Andrei Zinovyev, Jason D. Buenrostro, Guo-Cheng Yuan, and Luca Pinello. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nature Communications*, 10(1):1903, April 2019. Number: 1 Publisher: Nature Publishing Group.

[81] Marie-Ming Aynaud, Olivier Mirabeau, Nadege Gruel, Sandrine Grossetête, Valentina Boeva, Simon Durand, Didier Surdez, Olivier Saulnier, Sakina Zaïdi, Svetlana Gribkova, Aziz Fouché, Ulykbek Kairov, Virginie Raynal, Franck Tirode, Thomas G. P. Grünewald, Mylene Bohec, Sylvain Baulande, Isabelle Janoueix-Lerosey, Jean-Philippe Vert, Emmanuel Barillot, Olivier Delattre, and Andrei Zinovyev. Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. *Cell Reports*, 30(6):1767–1779.e6, February 2020.

[82] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, February 2003. arXiv:cond-mat/0209450.

[83] Eva M. Fast, Audrey Sporrij, Margot Manning, Edroaldo Lummertz Rocha, Song Yang, Yi Zhou, Jimin Guo, Ninib Baryawno, Nikolaos Barkas, David Scadden, Fernando Camargo, and Leonard I. Zon. External signals regulate continuous transcriptional states in hematopoietic stem cells. *eLife*, 10:e66512, December 2021.

[84] Léonard Hérault, Mathilde Poplineau, Adrien Mazuel, Nadine Platet, Élisabeth Remy, and Estelle Duprez. Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. *BMC Biology*, 19(1):19, February 2021.

[85] Dominic D. G. Owens, Giorgio Anselmi, A. Marieke Oudelaar, Damien J. Downes, Alessandro Cavallo, Joe R. Harman, Ron Schwessinger, Akin Bucakci, Lucas Greder, Sara de Ornellas, Danuta Jeziorska, Jelena Telenius, Jim R. Hughes, and Marella F. T. R. de Bruijn. Dynamic Runx1 chromatin boundaries affect gene expression in hematopoietic development. *Nature Communications*, 13(1):773, February 2022. Number: 1 Publisher: Nature Publishing Group.

[86] Ying Li, Ziwei Liao, Huacheng Luo, Aissa Benyoucef, Yuanyuan Kang, Qian Lai, Sinisa Dovat, Barbara Miller, Iouri Chepelev, Yangqiu Li, Keji Zhao, Marjorie Brand, and Suming Huang. Alteration of CTCF-associated chromatin neighborhood inhibits TAL1-driven oncogenic transcription program and leukemogenesis. *Nucleic Acids Research*, 48(6):3119–3133, April 2020.

[87] Qian Qi, Li Cheng, Xing Tang, Yanghua He, Yichao Li, Tiffany Yee, Dewan Shrestha, Ruopeng Feng, Peng Xu, Xin Zhou, Shondra Pruett-Miller, Ross C. Hardison, Mitchell J. Weiss, and Yong Cheng. Dynamic CTCF binding directly mediates interactions among cis-regulatory elements essential for hematopoiesis. *Blood*, 137(10):1327–1339, March 2021.

[88] Naoya Takayama, Alex Murison, Shin-Ichiro Takayanagi, Christopher Arlidge, Stanley Zhou, Laura Garcia-Prat, Michelle Chan-Seng-Yue, Sasan Zandi, Olga I. Gan, Héléna Boutzen, Kerstin B. Kaufmann, Aaron Trotman-Grant, Erwin Schoof, Ken Kron, Noelia Díaz, John J. Y. Lee, Tiago Medina, Daniel D. De Carvalho, Michael D. Taylor, Juan M. Vaquerizas, Stephanie Z. Xie, John E. Dick, and Mathieu Lupien. The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. *Cell Stem Cell*, 28(3):488–501.e10, March 2021.

[89] Tae-Gyun Kim, Sueun Kim, Soyeon Jung, Mikyoung Kim, Bobae Yang, Min-Geol Lee, and Hyoung-Pyo Kim. CCCTC-binding factor is essential to the maintenance and quiescence of hematopoietic stem cells in mice. *Experimental & Molecular Medicine*, 49(8):e371, August 2017.

[90] Anjun Ma, Adam McDermaid, Jennifer Xu, Yuzhou Chang, and Qin Ma. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends in Biotechnology*, 38(9):1007–1022, 2020. Publisher: Elsevier Ltd.

[91] Triantafyllos Chavakis, Ioannis Mitroulis, and George Hajishengallis. Hematopoietic progenitor cells as integrative hubs for adaptation to and fine-tuning of inflammation. *Nature immunology*, 20(7):802–811, July 2019.

[92] Christian M Schürch, Carsten Riether, and Adrian F Ochsenbein. Interferons in hematopoiesis and leukemia. *OncoImmunology*, 2(6):e24572, June 2013. Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/onci.24572.

[93] Thomas Höfer, Katrin Busch, Kay Klapproth, and Hans-Reimer Rodewald. Fate Mapping and Quantitation of Hematopoiesis In Vivo. *Annual Review of Immunology*, 34(1):449–478, 2016. _eprint: https://doi.org/10.1146/annurev-immunol-032414-112019.

[94] Nicola K. Wilson and Berthold Göttgens. Single-Cell Sequencing in Normal and Malignant Hematopoiesis. *HemaSphere*, 2(2):e34, March 2018.

[95] Lars Velten, Simon F. Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P. Hennig, Christoph Hirche, Christoph Lutz, Eike C. Buss, Daniel Nowak, Tobias Boch, Wolf Karsten Hofmann, Anthony D. Ho, Wolfgang Huber, Andreas Trumpp, Marieke A.G. Essers, and Lars M. Steinmetz. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19(4):271–281, March 2017. Publisher: Nature Publishing Group.

[96] Roy Drissen, Natalija Buza-Vidas, Petter Woll, Supat Thongjuea, Adriana Gambardella, Alice Giustacchini, Elena Mancini, Alya Zriwil, Michael Lutteropp, Amit Grover, Adam Mead, Ewa Sitnicka, Sten Eirik W. Jacobsen, and Claus Nerlov. Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nature Immunology*, 17(6):666–676, May 2016. Publisher: Nature Publishing Group.

[97] Shobana V. Stassen, Gwinky G. K. Yip, Kenneth K. Y. Wong, Joshua W. K. Ho, and Kevin K. Tsia. Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nature Communications*, 12(1):5528, September 2021. Number: 1 Publisher: Nature Publishing Group.

[98] Myriam L. R. Haltalli, Samuel Watcham, Nicola K. Wilson, Kira Eilers, Alexander Lipien, Heather Ang, Flora Birch, Sara Gonzalez Anton, Chiara Pirillo, Nicola Ruivo, Maria L. Vainieri, Constandina Pospori, Robert E. Sinden, Tiago C. Luis, Jean Langhorne, Ken R. Duffy, Berthold Göttgens, Andrew M. Blagborough, and Cristina Lo Celso. Manipulating niche composition limits damage to haematopoietic stem cells during Plasmodium infection. *Nature Cell Biology*, 22(12):1399–1410, December 2020.

[99] Sonia Nestorowa, Fiona K. Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K. Wilson, David G. Kent, and Berthold Göttgens.

A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8):e20–e31, August 2016. Publisher: American Society of Hematology.

[100] Amir Giladi, Franziska Paul, Yoni Herzog, Yaniv Lubling, Assaf Weiner, Ido Yofe, Diego Jaitin, Nina Cabezas-Wallscheid, Regine Dress, Florent Ginhoux, Andreas Trumpp, Amos Tanay, and Ido Amit. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nature Cell Biology*, 20(7):836–846, July 2018. Number: 7 Publisher: Nature Publishing Group.

[101] Danilo Pellin, Mariana Loperfido, Cristina Baricordi, Samuel L. Wolock, Annita Montepeloso, Olga K. Weinberg, Alessandra Biffi, Allon M. Klein, and Luca Biasco. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nature Communications*, 10(1):2395, June 2019.

[102] Katrin Busch, Kay Klapproth, Melania Barile, Michael Flossdorf, Tim Holland-Letz, Susan M. Schlenner, Michael Reth, Thomas Höfer, and Hans-Reimer Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542–546, February 2015. Number: 7540 Publisher: Nature Publishing Group.

[103] M. Ryan Corces, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, Jonathan K. Pritchard, Anshul Kundaje, William J. Greenleaf, Ravindra Majeti, and Howard Y. Chang. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, 48(10):1193–1203, October 2016.

[104] Mikael N. E. Sommarin, Parashar Dhapola, Fatemeh Safi, Rebecca Warfvinge, Linda Geironson Ulfsson, Eva Erlandsson, Anna Konturek-Ciesla, Ram Krishna Thakur, Charlotta Böiers, David Bryder, and Göran Karlsson. Single-Cell Multiomics Reveals Distinct Cell States at the Top of the Human Hematopoietic Hierarchy, April 2021. Pages: 2021.04.01.437998 Section: New Results.

[105] Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. CellRank for directed single-cell fate mapping. *Nature Methods*, 19(2):159–170, January 2022. Publisher: Nature Publishing Group.

[106] Dvir Aran, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P. Naikawadi, Paul J. Wolters, Adam R. Abate, Atul J. Butte,

and Mallar Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172, February 2019. Number: 2 Publisher: Nature Publishing Group.

[107] Minghui Wang, Yongzhong Zhao, and Bin Zhang. Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports*, 5(1):16923, November 2015. Number: 1 Publisher: Nature Publishing Group.

[108] Iwo Kucinski, Joana Campos, Melania Barile, Francesco Severi, Natacha Bohin, Pedro N. Moreira, Lewis Allen, Hannah Lawson, Myriam L. R. Haltalli, Sarah J. Kinston, Dónal O'Carroll, Kamil R. Kranc, and Berthold Göttgens. A time and single-cell resolved model of hematopoiesis, September 2022. Pages: 2022.09.07.506735 Section: New Results.

[109] Sidinh Luc, Natalija Buza-Vidas, and Sten Eirik W. Jacobsen. Delineating the cellular pathways of hematopoietic lineage commitment. *Seminars in Immunology*, 20(4):213–220, August 2008.

[110] Qi Yang, Lela Kardava, Anthony St. Leger, Kathleen Martincic, Barbara Varnum-Finney, Irwin D. Bernstein, Christine Milcarek, and Lisa Borghesi. E47 controls the developmental integrity and cell cycle quiescence of multipotential hematopoietic progenitors. *Journal of immunology (Baltimore, Md. : 1950)*, 181(9):5885–5894, November 2008.

[111] Malte D Luecken and Fabian J Theis. Current best practices in single□cell RNA□seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), June 2019. Publisher: EMBO.

[112] Simon Haas, Jenny Hansson, Daniel Klimmeck, Dirk Loeffler, Lars Velten, Hannah Uckelmann, Stephan Wurzer, Áine M. Prendergast, Alexandra Schnell, Klaus Hexel, Rachel Santarella-Mellwig, Sandra Blaszkiewicz, Andrea Kuck, Hartmut Geiger, Michael D. Milsom, Lars M. Steinmetz, Timm Schroeder, Andreas Trumpp, Jeroen Krijgsveld, and Marieke A.G. Essers. Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors. *Cell Stem Cell*, 17(4):422–434, 2015. Publisher: Cell Press.

[113] Betsabeh Khoramian Tusi, Samuel L. Wolock, Caleb Weinreb, Yung Hwang, Daniel Hidalgo, Rapolas Zilionis, Ari Waisman, Jun R. Huh, Allon M. Klein, and Merav Socolovsky. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, 555(7694):54–60, March 2018. Number: 7694 Publisher: Nature Publishing Group.

[114] Jianlong Sun, Azucena Ramos, Brad Chapman, Jonathan B. Johnnidis, Linda Le, Yu-Jui Ho, Allon Klein, Oliver Hofmann, and Fernando D. Camargo. Clonal dynamics of native haematopoiesis. *Nature*, 514(7522):322–327, October 2014. Number: 7522 Publisher: Nature Publishing Group.

[115] A.-K. Fanti, K. Busch, A. Greco, X. Wang, B. Cirovic, F. Shang, T. Nizharadze, L. Frank, M. Barile, T.B. Feyerabend, T. Höfer, and H.-R. Rodewald. Flt3- and Tie2-Cre tracing identifies regeneration in sepsis from multipotent progenitors but not hematopoietic stem cells. *Cell stem cell*, 30(2), 2023.

[116] Volker Bergen. RNA velocity—current challenges and future perspectives. (7:e10282 | 2021), 2021.

[117] Melania Barile, Ivan Imaz-Rosshandler, Isabella Inzani, Shila Ghazanfar, Jennifer Nichols, John C. Marioni, Carolina Guibentif, and Berthold Göttgens. Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biology*, 22(1):197, July 2021.

[118] Jianlong Sun, Azucena Ramos, Brad Chapman, Jonathan B. Johnnidis, Linda Le, Yu Jui Ho, Allon Klein, Oliver Hofmann, and Fernando D. Camargo. Clonal dynamics of native haematopoiesis. *Nature*, 514(7522):322–327, October 2014. Publisher: Nature Publishing Group.

[119] Sachin H. Patel, Constantina Christodoulou, Caleb Weinreb, Qi Yu, Edroaldo Lummertz da Rocha, Brian J. Pepe-Mooney, Sarah Bowling, Li Li, Fernando G. Osorio, George Q. Daley, and Fernando D. Camargo. Lifelong multilineage contribution by embryonic-born blood progenitors. *Nature*, 606(7915):747–753, June 2022. Number: 7915 Publisher: Nature Publishing Group.

[120] Weike Pei, Fuwei Shang, Xi Wang, Ann Kathrin Fanti, Alessandro Greco, Katrin Busch, Kay Klapproth, Qin Zhang, Claudia Quedenau, Sascha Sauer, Thorsten B. Feyerabend, Thomas Höfer, and Hans Reimer Rodewald. Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding. *Cell Stem Cell*, 27(3):383–395.e8, September 2020. Publisher: Cell Stem Cell.

[121] M. Hu, D. Krause, M. Greaves, S. Sharkis, M. Dexter, C. Heyworth, and T. Enver. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes & Development*, 11(6):774–785, March 1997. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution:

Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[122] Gerard Brady, Filio Billia, Jennifer Knox, Trang Hoang, Ilan R. Kirsch, Evelyn B. Voura, Robert G. Hawley, Rob Cumming, Manuel Buchwald, Kathy Siminovitch, Neil Miyamoto, Guido Boehmelt, and Norman N. Iscove. Analysis of gene expression in a complex differentiation hierarchy by global amplification of cDNA from single cells. *Current Biology*, 5(8):909–922, August 1995.

[123] Eric W. Martin, Jana Krietsch, Roman E. Reggiardo, Rebekah Sousae, Daniel H. Kim, and E. Camilla Forsberg. Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells. *Epigenetics & Chromatin*, 14(1):2, January 2021.

[124] Anna Maria Ranzoni, Andrea Tangherloni, Ivan Berest, Simone Giovanni Riva, Brynelle Myers, Paulina M. Strzelecka, Jiarui Xu, Elisa Panada, Irina Mohorianu, Judith B. Zaugg, and Ana Cvejic. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell*, 28(3):472–487.e7, March 2021.

[125] David M. Bodine. Introduction to the review series on transcription factors in hematopoiesis and hematologic disease. *Blood*, 129(15):2039, April 2017.

[126] Pau Badia-i Mompel, Lorna Wessels, Sophia Müller-Dott, Rémi Trimbour, Ricardo O. Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, 24(11):739–754, November 2023. Number: 11 Publisher: Nature Publishing Group.

[127] Robert Månsson, Anne Hultquist, Sidinh Luc, Liping Yang, Kristina Anderson, Shabnam Kharazi, Suleiman Al-Hashmi, Karina Liuba, Lina Thorén, Jörgen Adolfsson, Natalija Buza-Vidas, Hong Qian, Shamit Soneji, Tariq Enver, Mikael Sigvardsson, and Sten Eirik W. Jacobsen. Molecular Evidence for Hierarchical Transcriptional Lineage Priming in Fetal and Adult Stem Cells and Multipotent Progenitors. *Immunity*, 26(4):407–419, April 2007.

[128] Robert Månsson, Sasan Zandi, Eva Welinder, Panagiotis Tsapogas, Nobuo Sakaguchi, David Bryder, and Mikael Sigvardsson. Single-cell analysis of the common lymphoid progenitor compartment reveals functional and molecular heterogeneity. *Blood*, 115(13):2601–2609, April 2010.

[129] Jörgen Adolfsson, Robert Månsson, Natalija Buza-Vidas, Anne Hultquist, Karina Liuba, Christina T. Jensen, David Bryder, Liping Yang, Ole Johan Borge, Lina A.M. Thoren, Kristina Anderson, Ewa Sitnicka, Yutaka Sasaki, Mikael Sigvardsson, and Sten Eirik W. Jacobsen. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: A revised road map for adult blood lineage commitment. *Cell*, 121(2):295–306, April 2005. Publisher: Cell Press.

[130] Fatemeh Safi, Parashar Dhapola, Sarah Warsi, Mikael Sommarin, Eva Erlandsson, Jonas Ungerbäck, Rebecca Warfvinge, Ewa Sitnicka, David Bryder, Charlotta Böiers, Ram Krishna Thakur, and Göran Karlsson. Concurrent stem- and lineage-affiliated chromatin programs precede hematopoietic lineage restriction. *Cell Reports*, 39(6):110798, May 2022.

[131] Jeffrey M. Granja, M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3):403–411, March 2021. Number: 3 Publisher: Nature Publishing Group.

[132] Rachael A. Nimmo, Gillian E. May, and Tariq Enver. Primed and ready: understanding lineage commitment through single cell analysis. *Trends in Cell Biology*, 25(8):459–467, August 2015.

[133] Chauncey J. Spooner, Jason X. Cheng, Elisabet Pujadas, Peter Laslo, and Harinder Singh. A recurrent network involving the transcription factors PU.1 and Gfi1 orchestrates innate and adaptive immune cell fates. *Immunity*, 31(4):576–586, October 2009.

[134] Judith Schütte, Huange Wang, Stella Antoniou, Andrew Jarratt, Nicola K Wilson, Joey Riepsaame, Fernando J Calero-Nieto, Victoria Moignard, Silvia Basilico, Sarah J Kinston, Rebecca L Hannah, Mun Chiang Chan, Sylvia T Nürnberg, Willem H Ouwehand, Nicola Bonzanni, Marella FTR de Bruijn, and Berthold Göttgens. An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *eLife*, 5:e11469, February 2016. Publisher: eLife Sciences Publications, Ltd.

[135] Victoria Moignard, Iain C. Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J. Calero-Nieto, Sarah Kinston, Anagha Joshi, Rebecca Hannah, Fabian J. Theis, Sten Eirik Jacobsen, Marella F. de Bruijn, and Berthold Göttgens. Characterization of transcriptional networks in blood stem and progenitor cells us-

ing high-throughput single-cell gene expression analysis. *Nature Cell Biology*, 15(4):363–372, April 2013. Number: 4 Publisher: Nature Publishing Group.

[136] Grigorios Georgolopoulos, Nikoletta Psatha, Mineo Iwata, Andrew Nishida, Tannishtha Som, Minas Yiangou, John A. Stamatoyannopoulos, and Jeff Vierstra. Discrete regulatory modules instruct hematopoietic lineage commitment and differentiation. *Nature Communications*, 12(1), December 2021. Publisher: Nature Research.

[137] Martin Wahlestedt, Vasileios Ladopoulos, Isabel Hidalgo, Manuel Sanchez Castillo, Rebecca Hannah, Petter Säwén, Haixia Wan, Monika Dudenhöffer-Pfeifer, Mattias Magnusson, Gudmundur L. Norddahl, Berthold Göttgens, and David Bryder. Critical Modulation of Hematopoietic Lineage Fate by Hepatic Leukemia Factor. *Cell Reports*, 21(8):2251–2263, November 2017.

[138] Irene M. Min, Giorgio Pietramaggiori, Francis S. Kim, Emmanuelle Passegué, Kristen E. Stevenson, and Amy J. Wagers. The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell*, 2(4):380–391, April 2008.

[139] Alejandro Roisman, Emmalee R. Adelman, Hsuan-Ting Huang, Dean Wade, Daniel Bilbao, and Maria E. Figueroa. Loss of KLF6 Recapitulates Molecular and Functional Changes Associated with Aging in Human Hematopoietic Stem and Progenitor Cells. *Blood*, 134(Supplement_1):447, November 2019.

[140] John Kinzfogl, Giao Hangoc, and Hal E. Broxmeyer. Neurexophilin 1 suppresses the proliferation of hematopoietic progenitor cells. *Blood*, 118(3):565–575, July 2011.

[141] Guillermo López-Ruano, Rodrigo Prieto-Bermejo, Teresa L. Ramos, Laura San-Segundo, Luis Ignacio Sánchez-Abarca, Fermín Sánchez-Guijo, José Antonio Pérez-Simón, Jesús Sánchez-Yagüe, Marcial Llanillo, and Ángel Hernández-Hernández. PTPN13 and β-Catenin Regulate the Quiescence of Hematopoietic Stem Cells and Their Interaction with the Bone Marrow Niche. *Stem Cell Reports*, 5(4):516–531, September 2015.

[142] Leon Louis Seifert, Clara Si, Debjani Saha, Mohammad Sadic, Maren De Vries, Sarah Ballentine, Aaron Briley, Guojun Wang, Ana M. Valero-Jimenez, Adil Mohamed, Uwe Schaefer, Hong M. Moulton, Adolfo García-Sastre, Shashank Tripathi, Brad R. Rosenberg, and Meike Dittmann. The ETS transcription factor ELF1 regulates a broadly antiviral program distinct from the type I interferon response. *PLOS Pathogens*, 15(11):e1007634, November 2019.

[143] Ping Xiang, Chaoyu Lo, Bob Argiropoulos, C. Benjamin Lai, Arefeh Rouhi, Suzan Imren, Xiaoyan Jiang, Dixie Mager, and R. Keith Humphries. Identification of E74-like factor 1 (ELF1) as a transcriptional regulator of the Hox co-factor MEIS1. *Experimental hematology*, 38(9):798–808.e2, September 2010.

[144] Josette-Renée Landry, Sarah Kinston, Kathy Knezevic, Ian J. Donaldson, Anthony R. Green, and Berthold Göttgens. Fli1, Elf1, and Ets1 regulate the proximal promoter of the LMO2 gene in endothelial cells. *Blood*, 106(8):2680–2687, October 2005.

[145] Serena Belluschi, Emily F. Calderbank, Valerio Ciaurro, Blanca Pijuan-Sala, Antonella Santoro, Nicole Mende, Evangelia Diamanti, Kendig Yen Chi Sham, Xiaonan Wang, Winnie W. Y. Lau, Wajid Jawaid, Berthold Göttgens, and Elisa Laurenti. Myelo-lymphoid lineage restriction occurs in the human haematopoietic stem cell compartment before lymphoid-primed multipotent progenitors. *Nature Communications*, 9(1):4100, October 2018. Number: 1 Publisher: Nature Publishing Group.

[146] Shadi Toghi Eshghi, Amelia Au-Yeung, Chikara Takahashi, Christopher R. Bolen, Maclean N. Nyachienga, Sean P. Lear, Cherie Green, W. Rodney Mathews, and William E. O'Gorman. Quantitative Comparison of Conventional and t-SNE-guided Gating Analyses. *Frontiers in Immunology*, 10, 2019.

[147] Laleh Haghverdi and Leif S. Ludwig. Single-cell multi-omics and lineage tracing to dissect cell fate decision-making. *Stem Cell Reports*, 18(1):13–25, January 2023. Publisher: Elsevier.

[148] M.N.F. Morcos, C. Li, C.M. Munz, A. Greco, N. Dressel, S. Reinhardt, K. Sameith, A. Dahl, N.B. Becker, A. Roers, T. Höfer, and A. Gerbaulet. Fate mapping of hematopoietic stem cells reveals two pathways of native thrombopoiesis. *Nature Communications*, 13(1), 2022.

[149] Roy Drissen, Supat Thongjuea, Kim Theilgaard-Mönch, and Claus Nerlov. Identification of two distinct pathways of human myelopoiesis. *Science Immunology*, 4(35), 2019. Publisher: Sci Immunol.

[150] Marcus Alvarez, Elior Rahmani, Brandon Jew, Kristina M. Garske, Zong Miao, Jihane N. Benhammou, Chun Jimmie Ye, Joseph R. Pisegna, Kirsi H. Pietiläinen, Eran Halperin, and Päivi Pajukanta. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Scientific Reports*, 10(1):11019, July 2020. Number: 1 Publisher: Nature Publishing Group.

[151] Jonathan A. Griffiths, Arianne C. Richard, Karsten Bach, Aaron T. L. Lun, and John C. Marioni. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature Communications*, 9(1):2667, July 2018. Number: 1 Publisher: Nature Publishing Group.

[152] Shiyi Yang, Sean E. Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima, and Joshua D. Campbell. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, 21(1):57, March 2020.

[153] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, September 2014.

[154] Emma Dann, Neil C. Henderson, Sarah A. Teichmann, Michael D. Morgan, and John C. Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, February 2022.

Acknowledgements

From looking at tree models of haematopoiesis, it appears as if there's an element of symmetry in the choices that cells undergo. Differentiating into a megakaryocyte is like committing to a lymphocyte, albeit in a mirror. During my years of PhD, I fell from the symmetrical heavens of Physics to find that, when it comes to biology, evolution is a lousy programmer. It copy-pastes the same lines over and over, it has no regard for consistency, and overlapping bits of code act as a safeguard against dangerous bugs. But how come when I copy-paste code in the wrong place I get nonsensical results, while evolution gets the adaptive immune system? Seems unfair to me.

At the same time, during the first months of education as a researcher, I was initially horrified and then delighted to discover that scientific discovery is far muddier and more dynamic than its textbook counterpart.

Starting with the people who accompanied me during these years of wonder and frustration, I'd like to first mention my supervisor, Thomas Höfer. Coming from a Physics background, he immediately spotted the perplexities he might also have experienced and promptly helped me deal with them. I would like to thank Thomas for his patience and advice during these years, for valuing my research output and showing me how knowing the limitations and clauses of one's work makes it more valuable rather than the opposite. I learned that the job of a researcher in my field goes beyond chewing on theoretical and computational considerations, but needs to produce a synthesis that has a clear biological implication and that, in communication, making a message shorter will make it clearer. In that spirit, I'll move on.

Unfortunately, the mundane parts of the PhD are not punctuated by newly found wisdom or results. Plunging into some dataset to come back with some gold nuggets results most of the time in a handful of dirt, not the easiest treasure to bring home. I'd like to thank the office mates who shared this experience with me daily: Tamar, Anna, and Maurice. Day in and day out, sharing the occasional frustration or complication of one's work over coffee has frequently made the difference between a terrible day and an I'll-try-again-tomorrow day. A special acknowledgement to Jonas, who has proven an invaluable companion in brainstorming new ideas, delineating shared projects, and complaining about the messiness

of single-cell sequencing data.

Finally, a heartfelt thank you to all the people who joined the Theoretical System Biology group throughout the years, whose talent and drive have provided a constant source of inspiration towards the scientist I'd like to be.

At this point, I'd like to spend a few words for the people who likely scrolled directly to this part, glancing at the previous bunch of pages and not quite sure of what they are supposed to mean. At times, I wondered myself.

I met Carolina on a sunny afternoon on a bench that overlooks the Neckar River. The shops and bars were closed due to pandemic restrictions, and we could not do much else than stroll along the Neckarwiese. We quickly fell in love, and got married on a sunny afternoon last August. Before and after that day, our life together is a blissful sequence of invaluable, loving, ordinary moments. We have plenty of fun and exciting events, yes, but the happiness that I gather from simply spending time with you, be it window shopping, padel swinging, or improbable dance-offs are the texture that make my time joyous. It amazes me how you met me during one of the most challenging times in my life and offered unwavering hope for things to turn out better, making it a point to get excited about future plans, however uncertain and shifting. Now that I'm finishing up this thesis, I can't wait to make up for all the months of long faces that you endured with me and to be excited together about what awaits our family in the next years.

Next comes my birth family. My father says that children are like kites, at one point you have to let them fly away. And like kites, Marco, Mattia and I flew away from our hometown when we were about eighteen, to chase our corners of the sky, leaving our parents looking up from an empty house. But that does not amount to flying away. Throughout the last ten years, we experienced awe and sheer unpreparedness for the sky at large. During each challenge and each difficult time, we felt the gentle pull of the warmest support that we could ask for, be it a phone call to vent, a pleasant anecdote, or the recount of the latest shenanigans that everyone has been up to. However difficult, coming and going each year from and to my hometown made me aware of the subtle but pervasive joy of having a loving family, like water that fish swim into without really knowing it's there. Thank you, Mom and Dad, for the unrelenting love and support. And thank you, Mattia and Marco, for being the bright and wide kites that I forever look up to.

Lastly, I count myself lucky to have found, repeatedly and reliably, incredible friends who helped me through breakups, moving couches through narrow doors, finding the strength to go through another light-less winter, lit my days to no end with lightheartedness and delight.

Thank you to Simo, Lucie, Christoph and Mareike for our time together at our Neuenheim home, where I learned how truly countless the paths to self-realisation are, and how

being carefree is a wondrous balance of courage and presence. To this day, nothing beats pizza and JBL speaker on a makeshift terrace on a 10 square meters rooftop. My life is richer for that, and I'll never stop feeling lucky to have lived it with you.

Thank you to Gianmario, Mattia, and Stefano for our time together in our virtual CyberSalotto, a beacon of motivation and accountability during the darkest times, but mostly a glorious haven for all kinds of non-sensical banter.

Thank you to Chiara, Mario, Simona, Domenico, Nunzia, Sebastian, Cecilia, Enrico-detto-Jullo, Federica, and Ricardo for creating the loudest, funniest, and most exhilarating moments in the last couple of years. In the everlasting quest for a home away from home, you turned my long study stay into a colourful, fulfilling life.

Thank you to Leonsito and Federico for boundless happiness: without a reason, without a tooth, without a thought.