Delineating Megakaryocyte Lineage Commitment with Isoform Resolved Single Cell Transcriptomics

By

Anita Luisa Ahlert Scoones

Thesis submitted to the University of East Anglia for the Degree of *Doctor of Philosophy*

May 2023

Declaration

I hereby declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. Specific details of work done in collaboration are given at the start of relevant chapters. The contents of this thesis are not substantially the same as any that has been submitted or is being submitted for a degree. The total length of the main body of this thesis including figure legends is 94,905 and therefore does not exceed the limit of 100,000 words.

Anita Scoones.

May 2023.

Valeu a pena? Tudo vale a pena
Se a alma não é pequena.

–Fernando Pessoa

Abstract

Revised models of megakaryocyte (Mk) commitment from haematopoietic stem cells (HSCs) are emerging, supported by increasing evidence that heterogeneity in the HSC pool enables rapid platelet replenishment through a direct commitment to the Mk lineage (Haas *et al.*, 2015; Roch, Trachsel and Lutolf, 2015; Grover *et al.*, 2016). Platelet-biassed differentiation potential in a subset of HSCs corresponding to high expression of von Willebrand Factor (vWF) may support this direct commitment (Sanjuan-Pla *et al.*, 2013; Shin *et al.*, 2014), but the cellular and molecular transitions underpinning this process, and how they vary with age and under stress, remain to be elucidated.

In this thesis, single-cell transcriptomics was used to explore the continuum of differentiation between HSC and Mk, in both steady state and in response to stresses including platelet depletion and ageing. Full-length scRNA-seq of mouse bone marrow-derived Lin^- c-Kit^+ $Cd150^+$ (LK CD150+) cells was employed, hypothesising that this gating strategy would enable unbiased capture of HSCs, Mk progenitors as well as any intermediate cell types and states. The use of full-length scRNA-seq, generated from both short- and long-read sequencing platforms, enabled the analysis of both gene and isoform expression during megakaryopoiesis. To explore the plasticity of the system under stress, this thesis outlines the cellular and transcriptomic changes along this trajectory in response to acute platelet depletion as well as normal ageing.

scRNA-seq data from a total of 2,016 LK Cd150+ cells was generated, and pseudotime analysis used to confirm that this population captures the entire trajectory of early Mk and erythroid differentiation. Upon acute platelet depletion, global and cell-type specific changes in gene expression were observed, including an increased proportion of HSCs exiting quiescence, and the expansion of Mk progenitor populations with distinct expression profiles marked by upregulation of markers involved in Mk maturation compared to steady state control. Taken together, these data suggest an ‘emergency’ response is triggered to counteract the threat of acute thrombocytopenia at multiple levels of megakaryopoiesis, involving the activation of stem cells and the generation of novel intermediate cell types.

Differential expression programmes were identified between young and old Mk progenitors for genes related to proliferation, plasma membrane and inflammation as well as altered frequencies of HSC and progenitor populations, consistent with previous reports that show an expansion of HSCs with age.

This thesis also explores differential isoform expression arising from alternative splicing, revealing heterogeneity across HSCs in key genes for HSC and Mk function, highlighting the importance of isoform-level interrogations of single cells in megakaryopoiesis. Overall, this work adds to the understanding of the mechanisms which enable the first steps with which HSCs commit to the Mk lineage, representing an important resource for further insights into ageing, stress and plasticity in haematopoiesis.

Acknowledgements

With the writing finally done, it’s hugely gratifying to be able to thank all those who helped make this work possible. First and foremost, I would like to thank my supervisor Dr Iain Macaulay who has supported me beyond all my expectations. His guidance, mentorship and unequalled understanding throughout the last four years have made these not only my most successful but the happiest years of my education. I would like to extend this to my co-supervisor Dr Wilfried Haerty also for his guidance and mentorship. Both have provided me with invaluable support and offered me many professional opportunities throughout my time as their student. I could not have wished for two better supervisors, and thank them both for teaching that although it takes wisdom and expertise to be a good mentor, it's patience and compassion that will make you a great one. I will forever aspire to lead by their example.

I also offer heartfealt thanks to all members of the Macaulay and Haerty groups, past and present, for their support academically and otherwise. Ashleigh Lister, Laura Mincarelli and Edyta Wojtowicz - three wonderful women who were vital in my training during my first months and taught me so much about the challenges and good qualities that it takes to be a successful woman in science. A special thanks to Dr David Wright, who particularly in the last year has been crucial to my progress in the last leg of my PhD, as well as been an incredible source of support. And I also extend my thanks to Charlotte Utting and Lydia Pouney for their assistance in generating data towards this thesis, and Silvia Ogbeide and Yash Bancil for the invaluable words of encouragement. All of this work would have been impossible without such fantastic colleagues, it has been a joy to work with you all.

The works contained within this thesis would also not have been possible without contributions by Dr Stuart Rushworth and members of the Rushworth group, I thank Stuart for the provision of animals and Dr Jayna Mistry for their assistance with all mouse work. I would also like to acknowledge the members of Genomic Pipelines at the Earlham Institute who have contributed to sequencing all the samples presented in this thesis and have always gone above and beyond to answer my questions and help.

My deepest gratitude goes to my fellow students at the Earlham Institute who offered me a continuous source of support throughout my PhD. I could always depend on this cohort for advice and I thank them for always making me see the bright side and smile even at late hours after long hours working.

I would finally like to thank my family, Mum and Dad who have lovingly supported me in the pursuit of my career and instilled perseverance and determination in me throughout my life. To my brother William, who continues to inspire me to be a better person every day. To my partner Danilo, for being by my side as I completed my PhD. I thank you for your support, patience and love over these four years. And last but not least to my wonderful dogs Lucy, Dougal, Polly, Ninna and Pepe who are my greatest treasures and have provided endless distraction when most needed, as well as emotional support in the form of cuddles.

Contents

Abbreviations

AML	Acute myeloid leukaemia
AS	Alternative splicing
BM	Bone marrow
CCA	Canonical correlation analysis
CCS	Circular consensus sequencing
CCS	Circular consensus sequencing
CD	Cluster of differentiation
cDNA	Copy/Complementary DNA
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
DAPI	4’,6-Diamidino-2-Phenylindole
DC	Dendritic cell
DEGs	Differential Expressed Genes
DPBS	Dulbecco’s Phosphate Buffered Saline
dU	Deoxy uracil
EB	Elution buffer
Ery	Erythrocyte
FACS	Fluorescence-activated cell sorting
FCS	Foetal calf serum
FMO	Fluorescence minus one
FSC	Forward Scatter
FSM	Full splice match
G&T-seq	Genome and transcriptome sequencing
gDNA	Genomic DNA
GEM	Gel Beads-in-emulsion
GM	Granulocyte-macrophage
GMP	granulocyte-monocyte progenitor
GRE	Gene regulatory element
GSEA	Gene set enrichment analysis
HiFi	High Fidelity
HSC	Haematopoietic stem cell
HSPC	Haematopoietic stem and progenitor cell
HSPC gate	Lin- Sca1+ c-Kit+ sorting gate

HT	High-throughput
ID	Identification
ISM	Incomplete splice match
KLF	Kruppel-like factor
LCR	Locus control region
Lin	Lineage
LMPP	Lymphoid multipotent progenitor
LR	Long reads
LSK	Lin- Sca-1+ c-Kit+ phenotype
LT	Low-throughput
Ly-HSC	Lymphoid biassed haematopoietic stem cell
LT-HSC	Long-term haematopoietic stem cell
MEP	Megakaryocyte-erythroid progenitor
Mk	Megakaryocyte
Mk-HSC	Megakaryocyte biassed haematopoietic stem cells
MkPs	Megakaryocyte progenitors
MPN	Myeloproliferative neoplasms
MPP	Multipotent progenitors
Mye-HSC	Myeloid biassed haematopoietic stem cell
mRNA	Messenger RNA
NGS	Next Generation Sequencing
NIC	Novel in catalogue
NK	Natural Killer Cells
NNC	Novel not in catalogue
P	Progenitor
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate-buffered saline
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PreGM	Pre-granulocyte-macrophage progenitor
PreMegE	Pre-megakaryocyte-erythrocyte progenitor
QC	Quality control
qRT-PCR	Quantitative real-time polymerase chain reaction
RBC	Red blood cells
RNA-seq	RNA sequencing

RT	Reverse transcription
scATAC-seq	Single-cell assay for transposable-accessible chromatin
scRNA-seq	Single-cell RNA sequencing
S-read	Segmented HiFi read
ST-HSC	Short term haematopoietic stem cell
SR	Short reads
SSC	Side Scatter
t-SNE	t-distributed stochastic neighbour embedding
TF	Transcription factor
TPO	Thrombopoietin
TSO	Template switch oligo
TSS	Transcription start site
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique oligonucleotide molecular identifier
ZMW	Zero-mode waveguides

List of Figures

List of Tables

Chapter 1:

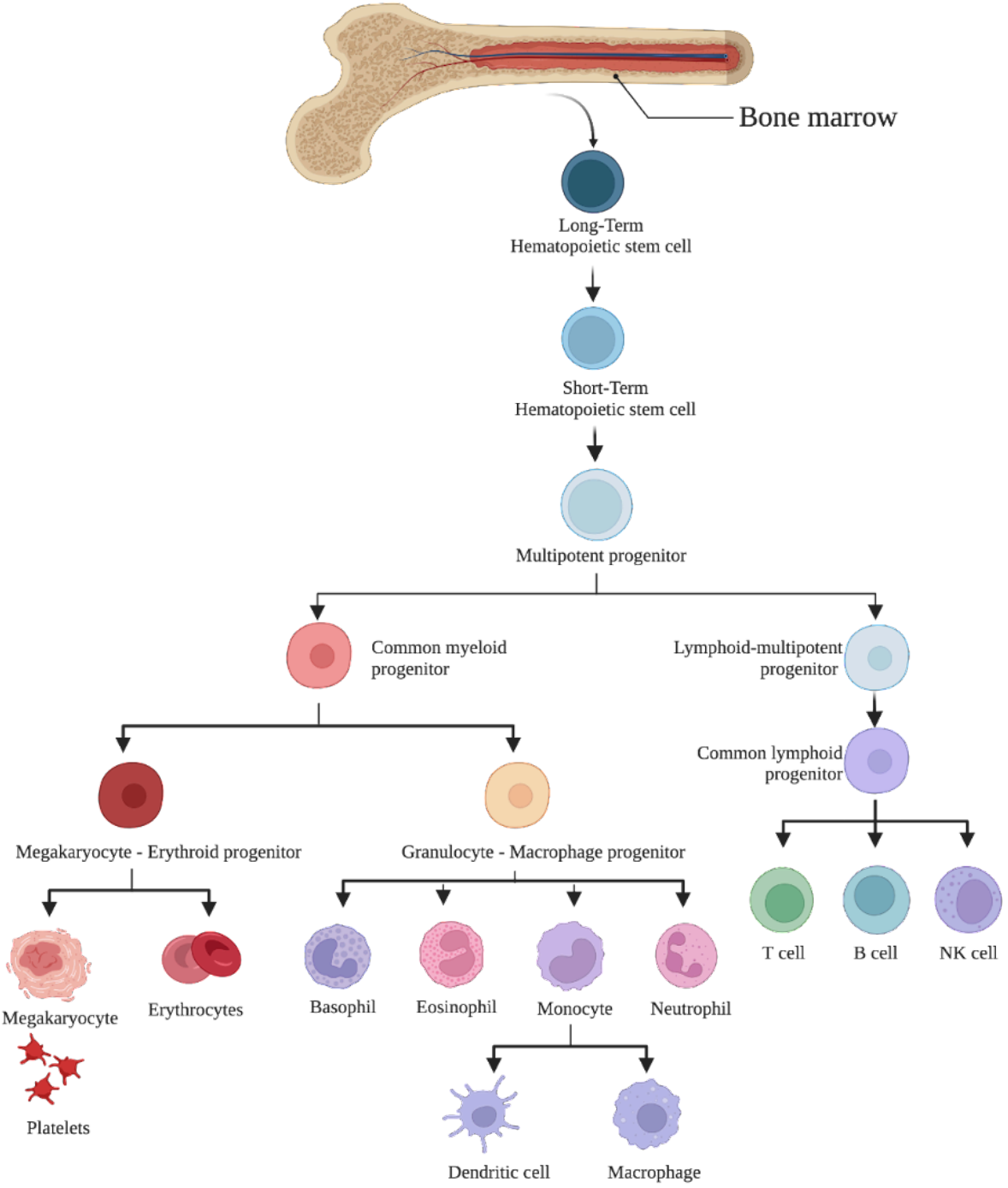
Introduction

Preface

Haematopoiesis is the process in which the heterogeneous cell populations that comprise blood are formed and replenished throughout an organism's lifetime. It is an essential process for organism survival, with its primary functions being the transport of oxygen throughout the body by red blood cells (erythrocytes), the generation of the white blood cells of the immune system (leukocytes) that are critical for fighting infections, and the vital formation of blood clots by platelets (thrombocytes) to stop bleeding (Orkin, 2000). Haematopoiesis is classically depicted in a hierarchical fashion, with haematopoietic stem cells (HSCs) at the apex giving rise first to progenitors and then to increasingly lineage-restricted precursors (Figure 1.1).

Given in part by the accessibility of blood cells relative to other tissues, and the well-established experimental approaches for studying haematopoiesis over the last 50 years it is arguably one of the most thoroughly studied cellular systems. With over 10 blood cell fates and its generation of more than 300 billion cells daily, it has served as a paradigm for understanding heterogeneous cellular systems, stem cell biology and function and how when behaving aberrantly they contribute to disease, oncogenesis and ageing (Orkin and Zon, 2008). The lack of proliferative ability and limited life span of most mature blood cells - with estimates suggesting the production of 1.5×10^6 blood cells every second in an adult human - necessitates their constant replenishment and requires extensive homeostatic regulatory mechanisms to keep up with the high turnover rate (Fliedner *et al.*, 2002; Orkin and Zon, 2008). This control primarily resides with HSCs, the first tissue-specific stem cell to be isolated and routinely used clinically in the treatment of a variety of blood cell diseases (Delaney, Gutman and Appelbaum, 2009; Munoz *et al.*, 2014; Piemontese *et al.*, 2015). It is also mediated at the level of a subset of more committed progenitors, that exhibit high proliferative capacity within the haematopoietic system (Pietras *et al.*, 2015).

Haematopoiesis is divided into three 'waves' during the mammalian lifespan, each characterised by the type of cells that are produced, the location of haematopoiesis, and the growth factors and cytokines that regulate the process (Figure 1.2) (Galloway and Zon, 2003). The first site of haematopoiesis in human and murine ontogeny occurs in the yolk sac, where the transitory primitive wave takes place and gives rise to primitive erythrocytes and macrophages to facilitate tissue oxygenation in early embryonic development. The second wave of haematopoiesis occurs in the foetal liver and spleen, it is characterised by the production of definitive erythrocytes (Ery), granulocytes, and monocytes until ultimately the system migrates to the adult bone marrow (BM) where all types of blood cells are produced



[1]

Figure 1.1. The classic hierarchy model of the haematopoietic system. A schematic representing the classical dogma of the haematopoietic system, with HSCs at the apex of the hierarchy.

[1] Created with BioRender.com

throughout the organism's lifetime, comprising the third and final wave of haematopoiesis (Ivanovs *et al.*, 2011, 2014).

Our understanding of haematopoiesis has undergone profound shifts over the last 50 years as a consequence of methodology and technological developments, particularly within genomics and advances in the field of single-cell biology. This chapter covers the evolution of our current understanding of haematopoiesis, with particular emphasis on lineage commitment towards the platelet lineage. Detailing key research and the implications of single-cell technologies in the field of megakaryocyte (Mk) lineage commitment, whilst highlighting gaps in our knowledge within this field.

[2]

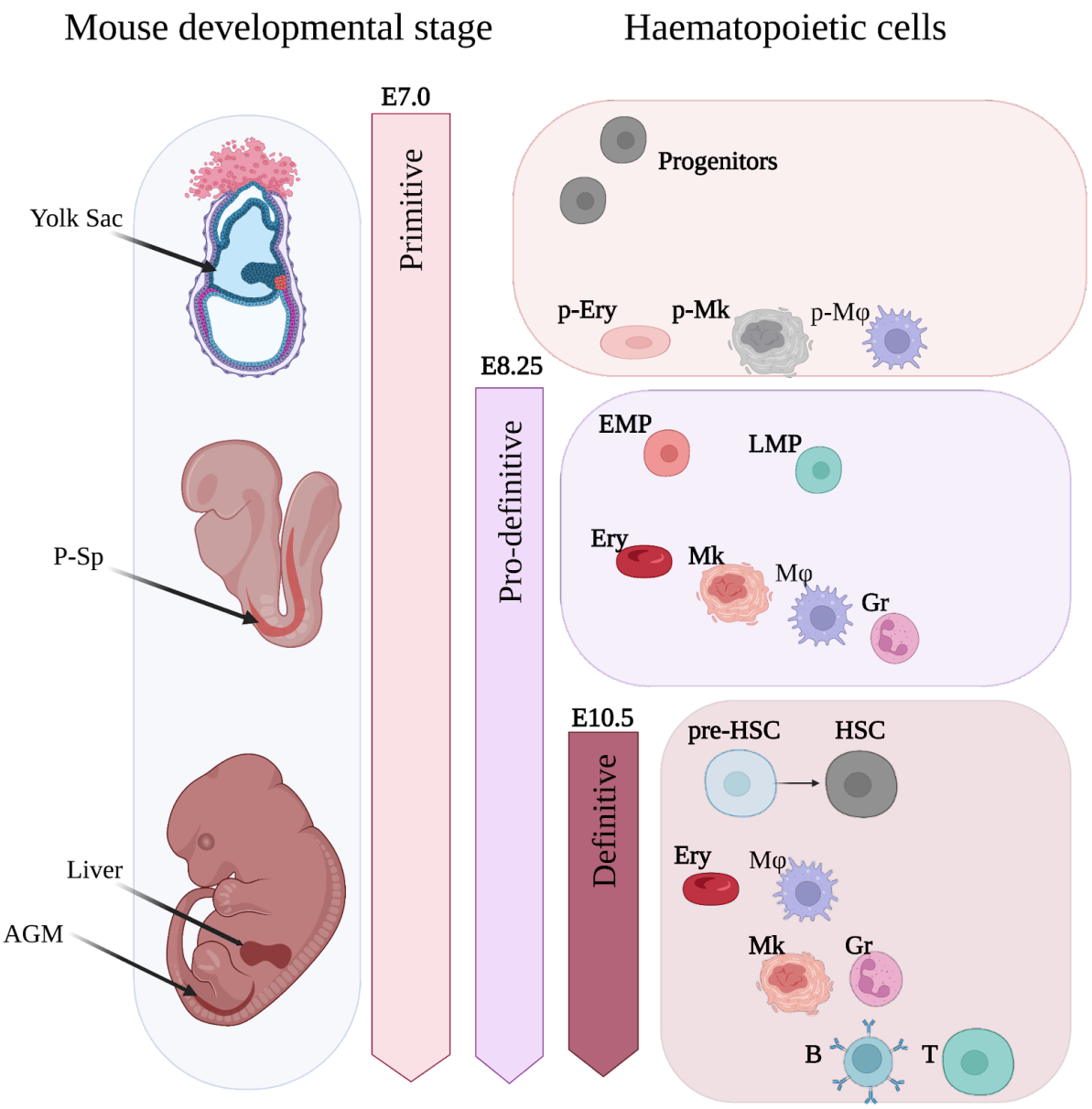


Figure 1.2. The three waves of haematopoiesis. The first wave begins around embryonic day 7 (E7.0) and is called the primitive haematopoiesis which gives rise to p-Ery, p-Ml and p-Mφ cells. The second wave, called pro-definitive, starts at E8.25 where EMPs begin to emerge. The third wave, termed definitive, sees the generation of both haematopoietic stem (HSC) and their resulting progenitor cells.

[2] Created with BioRender.com

1.1.1. The classic hierarchical model of haematopoiesis

The classic model of haematopoiesis is credited to research led by Dr Ernest McCulloch and Dr James Till at the University of Toronto in the early 1960s, who first described a small number of homogenous HSCs existing atop a cellular hierarchy, capable of ultimately giving rise to all mature blood cell types through a cascade of differentiation steps (Figure 1.1). A series of experiments helped define HSCs by their ability to differentiate into all blood cell lineages as well as a self-renew and enter a non-proliferative quiescent state - which remain key defining properties of all stem cells (Till and McCulloch, 1961; Becker, McCulloch and Till, 1963; Siminovitch, McCulloch and Till, 1963; Wu *et al.*, 1967, 1968).

Over the following two decades, significant progress has been made in the field of HSC research, thanks to advancements in techniques such as fluorescence-activated cell sorting (FACS) and magnetic-activated cell sorting (MACS). These technologies have played a crucial role in the isolation of HSCs and the identification of progenitors necessary for rebuilding the blood hierarchy. A major contributing factor to these advancements has been the increased availability of monoclonal antibodies, which exhibit high specificity for specific antigens. This development has greatly facilitated the isolation and functional evaluation of different cellular subsets. As a result, researchers have been able to discover key markers that are still utilised today to identify specific cell populations (Weissman and Shizuru, 2008).

One notable example of these includes Sca-1 (van de Rijn *et al.*, 1989), which was found to separate BM cells into approximately 25% Sca1-positive and 75% Sca1-negative populations. Importantly, only the Sca-1-positive cells exhibited consistent clonal and in vivo reconstitution capabilities (Spangrude, Heimfeld and Weissman, 1988). Today, we recognise Sca-1 as a key stem cell antigen used as a marker for the identification of primitive HSCs. Its expression decreases as cells differentiate, and it plays a vital role in maintaining the bone marrow HSC compartment throughout an individual's life (Chatterjee *et al.*, 2010). Together with the expression of c-kit (CD117) and the low expression of several surface markers associated with mature blood lineages (collectively known as Lin-), the Lin- Sca-1+ cKit+ (LSK) profile is used in FACS enrichment for HSCs and the most immature blood progenitors. Later, Nakauchi and colleagues identified that ~30%–40% of the CD34- LSK cells have HSC potential (Osawa *et al.*, 1996), followed by the discovery of numerous other markers such as CD150 over subsequent years (Kiel *et al.*, 2005; Balazs *et al.*, 2006; Wagers and Weissman, 2006; Benveniste *et al.*, 2010). These discoveries improved HSC enrichment enabling HSC

fractionation thus making molecular and behavioural analyses within the HSC compartment feasible.

With such developments, the classical model of haematopoiesis was further expanded to subcategorise HSCs into long-term (LT) or short-term (ST) subsets. LT-HSCs are largely quiescent, capable of both symmetric and asymmetric division, but critically are defined by their ability to completely reconstitute the haematopoietic system over several months post-irradiation (Morrison and Weissman, 1994). ST-HSCs are also capable of multilineage repopulation, however unlike LT-HSCs, they exhibit finite self-renewal capacity (Morrison *et al.*, 1997). Within the same bone-marrow compartment are multipotent progenitors (MPPs), capable of transiently providing multilineage reconstitution (Morrison *et al.*, 1997). Because HSCs themselves only rarely divide (Bradford *et al.,* 1997, Cheshier *et al.,* 1999) and are limited in numbers, these MPPs, which are more numerous and harbour differential proliferative potentials, also serve to maintain a primary level of homeostatic control (Passegue *et al.,* 2005). Together these cells exist at the top of a hierarchy of increasingly lineage-restricted progenitors towards each of the blood lineages, with HSCs at the conceptual apex. The fundamental aspect of this model of haematopoiesis is that at each stage of differentiation, cells undergo binary fate choices in discrete stages, whereby cells within the same conceptual stage are uniformly lineage committed and have an equal propensity to differentiate towards unipotency (Figure 1.3, left).

1.1.1. Paradigm shifts from discrete differentiation to a continuum of haematopoietic lineage commitment

The separation of cells on the basis of surface markers, followed by bioassays of developmental potential have identified progenitors committed to lymphoid or myeloid programmes (common lymphoid (CLP) and myeloid (CMP) progenitors, respectively (Kondo, Weissman and Akashi, 1997; Akashi *et al.*, 2000). The classical model of haematopoiesis proposed that this first lineage bifurcation separated progenitors with myeloid vs lymphoid potential, however, it was subsequently demonstrated that the earliest bifurcation occurred immediately downstream of multipotent HSCs, in MPPs, whereby cells with lympho-myeloid but no Mk-Ery potential diverged from cells with combined Mk-Ery and myeloid potential (Adolfsson *et al.*, 2001, 2005; Månsson *et al.*, 2007; Luc *et al.*, 2008). Supporting this, the comparison of CLPs and a pre-granulocyte-macrophage (PreGM) cell type of predominant myeloid potential revealed they share more similarity than either cell type has to cells of the Ery lineage (Pronk *et al.*, 2007). Together these results suggest a lymphoid and myeloid branch (LMPP) precedes before the Ery lineage, and as cells become increasingly lymphoid-restricted (losing granulocyte-monocyte (GM) potential) generate CLPs, which lack myeloid potential but can rapidly produce natural killer (NK) cells, B cells and T cells (Kondo, Weissman and Akashi, 1997).

Although the classical model still serves as a valuable paradigm, improved cell purification combined with large-scale single-cell transcriptomics and lineage tracing experiments have provided deeper insights into the transcriptomic landscape of haematopoiesis (Laurenti and Göttgens, 2018). Further complexities within haematopoiesis that were not explained through the classic model were revealed. One notable finding is the discovery of substantial functional and molecular variability among cells exhibiting similar cell surface marker phenotypes. This variability indicates that cells with seemingly identical characteristics possess distinct gene expression profiles and functional attributes; including differential self-renewal capacities and multiple routes of commitment towards distinct lineages (Notta *et al.*, 2016).

Importantly, it became apparent that HSCs exhibit significant heterogeneity in properties key to stem cell function. Studies within the HSC compartment showed that individual HSCs exhibit differential reconstitution patterns, cell cycling kinetics and self-renewal durability, with numbers of mature cells that individual HSCs produce ranging from 1 to up to almost 100% of the recipient's peripheral blood cells (Müller-Sieburg *et al.*, 2002; Dykstra *et al.*, 2007; Beerman *et al.*, 2010; Benveniste *et al.*, 2010; Morita, Ema and Nakauchi, 2010; Wilkinson and Göttgens, 2013). In addition, the ratio of myeloid and lymphoid cells generated can vary

significantly between individual HSCs (Muller-Sieburg *et al.*, 2004; Sieburg *et al.*, 2006; Dykstra *et al.*, 2007; Challen *et al.*, 2010).

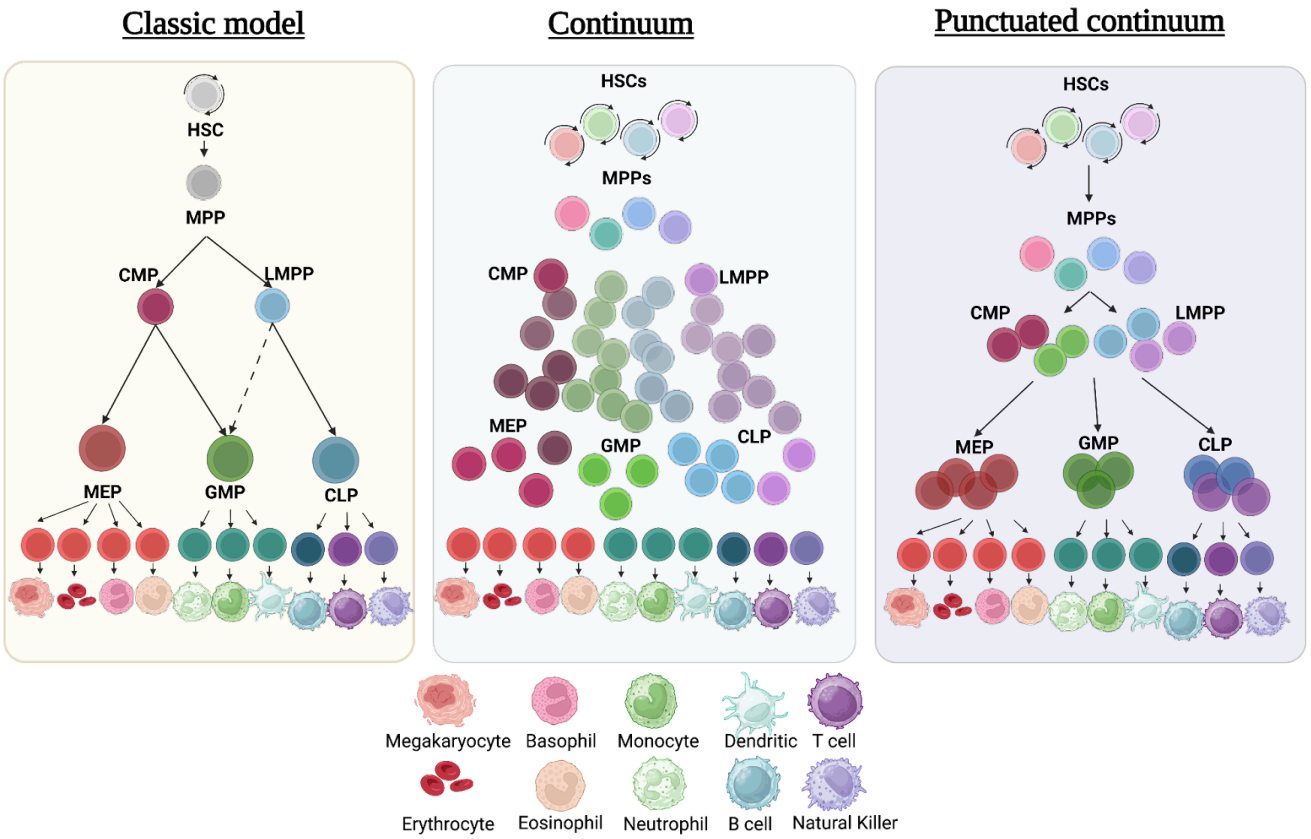
Moreover, advances in functional transplantation assays shed light on alternative lineage commitment pathways, further expanding our knowledge of the intricate cellular processes involved in haematopoiesis. Naik *et al*. showed that nearly 50% of the LMPP compartment is biassed toward dendritic cell (DC) commitment, which was previously thought to arise strictly via the CMP-to-GMP differentiation route (Naik *et al.*, 2013). Similarly, Yamamoto *et al.* identified a HSC subpopulation that could produce one multipotent daughter cell, and one lineage-committed daughter, revealing that some multipotent stem cells can undergo direct lineage commitment (Yamamoto *et al.*, 2013). This work demonstrated the differentiation of Mk-Er progenitors (MEPs) from HSC that arise directly without cell division, 'by-passing' intermediate differentiation stages of progressing through conventional MPPs and CMPs. Moreover, approximately 10% of LT-HSCs were found to express the Mk surface antigen CD41 and contained self-renewing cells with differential patterns of restricted differentiation output (CMP-like 20%, Mk-like 12%, and Mk-E 2%), as well as intermediate and ST-HSCs. This indicated that lineage-restricted progenitors could have long-term repopulating activity, challenging the assumption that only HSCs at the top of the hierarchy could self-renew (Yamamoto *et al.*, 2013)

Broadly, two models to explain the heterogeneity within the HSC compartment and their early downstream progenitors were formed. Either there still exists a mixture of distinct cell types, but they consist of lineage-biassed subsets; or cells are in fact all equivalent and the heterogeneity is a result of stochastic fate choices and/or differences in extrinsic influences such as niche signals (Schroeder, 2010). While it is clear that extrinsic signals do play a significant role in differentiation behaviours (Zhang *et al.*, 2003; Florian *et al.*, 2012; Vas *et al.*, 2012; Mirantes, Passegué and Pietras, 2014; Mead *et al.*, 2017), research has demonstrated consistent lineage bias that is heritable across cell divisions - suggesting lineage bias is an intrinsic property in some stem cells that can be inherited by their offspring (Dykstra *et al.*, 2007; Kent *et al.*, 2009; Weinreb *et al.*, 2020). To date, lineage-biassed HSPCs have been identified for the lymphoid lineage (Ly-HSCs) (Muller-Sieburg and Sieburg, 2008), myeloid lineage (Mye-HSCs) (Muller-Sieburg *et al.*, 2004; Beerman *et al.*, 2010; Challen *et al.*, 2010), and Mk lineage (Sanjuan-Pla *et al.*, 2013; Grover *et al.*, 2016; Carrelha *et al.*, 2018; Rodriguez-Fraticelli *et al.*, 2018).

As single-cell transcriptomics became possible, there was a unique opportunity to assay cells of the haematopoietic system simultaneously, and hierarchically organise them based on their

unique transcriptomic profiles. Assessments of populations of cells that previously were thought to be homogeneous and analyses of differentiation continuity at single-cell resolution transformed haematopoietic research. The identification of differential functional properties within immunophenotypically defined cell types challenged the assumption that 'marker-pure' subsets (ie. cells isolated based on surface marker expression) are synonymous with 'functionally-pure' subsets. The transcriptional continuity of single-cell transcriptomics promoted a paradigm shift in the field, contributing to a revised model of haematopoiesis as a continuum of differentiation rather than a series of discrete cell types where multiple differentiation routes towards unipotency exist (Figure 1.3, centre).

The continuum model addressed some of the oversimplifications of the classic model of haematopoiesis, however, gene expression alone arguably may be insufficient to distinguish discrete cell populations and ignores other relevant aspects of cellular commitment. Coupling gene expression with the assessment of different subpopulations has defined distinct functional groups (Pietras *et al.*, 2015). The perturbation of key master regulators of haematopoiesis, such as haematopoietic transcription factors (TFs) has identified clear transition points that occur over the continuous transcriptomic landscape, suggesting the presence of punctuated transitions (Figure 1.3, right) (Giladi *et al.*, 2018; Laurenti and Göttgens, 2018; Liggett and Sankaran, 2020).



[3]

Figure 1.3. Revisions to classic haematopoiesis model. The classic model (left) describes differentiation through a series of discrete branching populations of increasing lineage restriction until reaching unipotency, where HSCs and progenitor populations display uniform lineage potential and make binary fate decisions. The continuum of differentiation model (centre) describes sub-populations of HSCs and progenitors with differential propensities towards specific lineages, and differentiation as a stochastic and continuous rather than step-wise process. The punctuated continuum model (right) describes heterogeneous subpopulations of varying lineage biases existing within distinct cell states (Liggett and Sankaran, 2020)adapted from (Liggett and Sankaran, 2020).

[3] Created with BioRender.com

One of the key concepts in haematopoiesis is the process of lineage commitment, which refers to the mechanism by which a cell becomes restricted to producing a single progenitor or mature blood cell type. This occurs through a multistep process, first involving the decision by HSCs for either self-renewal versus differentiation followed by the decision for commitment to distinct lineage fates. A cell is considered committed to a specific lineage once it has acquired irreversible cell-type specific characteristics and activated signalling pathways that regulate the expression of genes involved in the development and function of that particular cell type. These characteristics and pathways typically, though not always, reflect the cell's developmental history and dictate its future fate. The molecular mechanisms, cellular interactions and timing of lineage commitment are fundamental to the regulation of blood production in homeostasis and in disease.

1.2.1 Regulation of cell fate decisions

The study of lineage specification has been greatly facilitated by *in vitro* colony forming assays, which have provided valuable insights into the cellular and molecular mechanisms underlying this process. The limitations of these assays from a physiological standpoint are minimised by combining *in vivo* strategies, such as single-cell transplantation and label-retention assays, that allow the tracking of haematopoietic lineage restriction at different stages of commitment (Perié *et al.*, 2015; Upadhaya *et al.*, 2018; Lee-Six and Kent, 2020; Rodriguez-Fraticelli *et al.*, 2020; Weinreb *et al.*, 2020). These complementary approaches, when combined with functional assays that analyse the properties of the succeeding cell progeny, have greatly enhanced our understanding of the dynamic and complex process of cell fate decisions. As a result, we have gained new insights into the heterogeneity and plasticity of haematopoietic stem and progenitor cells, revealing previously unrecognised levels of complexity and regulation.

In addition, with the advent of high-throughput sequencing technologies, research of gene expression patterns across haematopoietic populations has revealed cell-type specific signatures and led to the identification of key regulators of haematopoietic differentiation. Indeed, cell fate choices are closely associated with changes in gene expression, as different cell types express sets of genes that are responsible for their molecular features and biological functions. By controlling gene expression, cells up- and down-regulate genes that promote differentiation towards specific lineages. Gene expression involves the conversion of DNA into RNA, which may then be translated into a protein by a ribosome. Its regulation occurs through several

mechanisms at multiple levels of the process, including transcriptional control, post-transcriptional regulation, and translational control, which are in turn influenced by a multitude of extrinsic signalling and epigenetic factors. Transcriptional regulation involves the direct or indirect binding of *trans*-acting TFs to specific DNA elements, which can either activate or repress gene expression (Wilson *et al.*, 2010; Palii *et al.*, 2019). Post-transcriptional regulation involves the processing of RNA molecules to produce mature mRNAs, which can be subject to various forms of RNA modifications and stability control. Finally, translational control involves the regulation of protein synthesis by factors that affect the activity of ribosomes and the efficiency of translation initiation. Together, these regulatory mechanisms ensure careful control over gene expression in response to different physiological signals and environmental cues, and underpin the dynamic nature of haematopoietic lineage commitment.

1.2.2. Gene regulatory elements

Gene regulatory elements (GREs) are generally *cis-acting* components of DNA that interact with specific TFs and other regulatory proteins to modulate gene activity (Davidson, 2010). The term GRE encompasses multiple types of elements categorised by their function; including promoters, enhancers, insulators, silencing elements and locus control regions (Maston, Evans and Green, 2006). In the context of haematopoietic lineage specification, GREs are responsible for orchestrating the activation or repression of genes that drive the differentiation of HSCs into specific blood lineages. Their intricate regulatory networks and interactions ensure the progression and balance of different lineages, contributing to homeostatic control of the cellular composition of the haematopoietic system.

Promoters

A promoter is a region of DNA upstream of a gene that promotes the initiation of transcription. By sequence-specific TF and RNA polymerase interactions, promoters facilitate the binding and assembly of the pre-initiation complex. The most common is the TATA box promoter, which induces DNA partial unwinding to facilitate transcription upon binding the TATA-binding protein (Lee and Young, 2000).

Enhancers

Enhancers ensure proper regulation of transcription levels by activating transcription above basal levels, resulting in tissue-specific patterns of gene expression. They regulate the spatio-temporal activity of genes, and are considered to be key determinants of cell identity. Enhancers can be distal or co-localised with promoters, where TF binding resulting in enhancer looping to the promoter or “chromatin hub” interactions are their respective regulatory

mechanisms (Bulger and Groudine, 2011; Ong and Corces, 2011). For example, MYC has been found to be regulated by the BENC cluster of enhancers (Bahr *et al.*, 2018). MYC is an essential TF for HSC and progenitor regulation with implications in haematopoietic malignancies (Delgado and León, 2010). Bahr et al. showed BENC is composed of lineage-specific enhancer modules and the deletion of these modules leads to cell-type-specific downregulation of MYC expression (Bahr *et al.* 2018).

Locus control regions

Locus control regions (LCR) enhance the expression of linked genes at distal chromatin regions in a tissue-specific manner, regulating gene expression through chromatin domain-opening activity (Maston, Evans and Green, 2006). Unlike enhancers, LCRs possess all the properties necessary for opening a chromosome domain and preventing hetero-chromatinisation at ectopic sites (Grosveld *et al.*, 1987). The first identified LCR was in the human β-globin Ery locus, and illustrated cell lineage–specific gene expression regulation based on long-range interactions of various *cis-elements* and chromatin alterations, not exclusively gene-proximal elements (promoters, enhancers, and silencers) (Li *et al.*, 2002).

Silencers

Silencers are sequence-specific elements that confer a negative (i.e., repressing) effect on the transcription of a target gene (Maston, Evans and Green, 2006). Individual TFs that drive the gene expression for cells to develop down one pathway can simultaneously repress pathways to other lineages. For example, high levels of the Ery Kruppel-like factor (KLF) promote erythropoiesis whilst suppressing megakaryopoiesis, in part by repressing the level of FLI-1 (Frontelo *et al.*, 2007)

Insulators

Also known as boundary elements, insulators function to block genes from being affected by the transcriptional activity of neighbouring genes, thus preventing crosstalk between genomic regions by limiting the action of transcriptional regulatory elements to defined domains (Maston, Evans and Green, 2006). For example, a region upstream of the Ankyrin-1 Ery promoter is a barrier insulator *in vivo* in Ery cells, demonstrating both prevention of gene silencing, and occupancy by barrier-associated proteins (Gallagher *et al.*, 2010).

1.2.3. Transcription factors and haematopoietic differentiation

The decision of HSCs to differentiate or self-renew, or of progenitors to differentiate towards particular lineages depends on complex and tightly regulated processes that are contingent on the appropriate and timely execution of specific expression signatures. In particular, TFs are critical gene expression regulators that orchestrate fate specification. TFs are *trans*-acting regulators that function in conjunction with other GREs such as chromatin modifiers and co-factors, as well as other TFs, which enable them to establish transcriptional profiles and cell fates in often a cell-type-specific manner (Wilkinson and Göttgens, 2013). Early studies of haematopoiesis focused on uncovering the specific roles of master regulator TF in the fate decision process, often through genetic ablation of TFs or epigenetic regulators, such as GATA-1, SPI-1 (PU.1), CEBPA, GFI-1, and IKZF1 (Ikaros) in mice to determine their functional roles (Shivdasani *et al.*, 1997; Nichogiannopoulou *et al.*, 1999; Hock *et al.*, 2004; P. Zhang *et al.*, 2004; Chou *et al.*, 2009; Doulatov *et al.*, 2012).

Several critical TFs are known to play major roles in HSPC regulation. Using genome-wide computational analysis of TF binding patterns and functional validation, Wilson *et al.* reported combinatorial interactions for ten key regulators of HSPCs including TAL1, LYL1, LMO2, GATA-2, RUNX-1, MEIS-1, ERG, FLI-1, and GFI-1B. This seminal work revealed the interaction between a heptad of HSPC-associated TFs, including functional links between RUNX1 and other key HSPC regulators (Wilson *et al.*, 2010). *RUNX-1* for example is involved in the maintenance of HSC quiescence, by controlling the balance between self-renewal and differentiation through regulating the expression of genes maintaining the expression of those important for HSC function, such as CD41 and CD42. Its' deletion in adult BM causes the expansion of immature progenitors with a concomitant reduction of LT-HSC activity (Growney *et al.*, 2005). GATA-2 has been shown to promote the self-renewal inducing HSC quiescence by interacting with a network of TFs that specify early lineage commitment (Tipping *et al.*, 2009; Doré *et al.*, 2012; Johnson *et al.*, 2012; Collin, Dickinson and Bigley, 2015).

1.2.4. Chromatin structure and epigenetics

Epigenetics originally referred to heritable features of a cellular phenotype independent of changes in DNA sequence, but has over time evolved to define chromatin-based reactions that regulate DNA-templated processes (Zhao *et al.*, 2023). The interplay between chromatin structure, DNA methylation, and histone modifications contributes to the epigenetic regulation of haematopoiesis (Sashida and Iwama, 2012). These mechanisms involve modifications to the packaging of DNA in chromatin, as well as the establishment and maintenance of epigenetic marks that control gene expression. Chromatin can exist in different states, ranging from closed and condensed (heterochromatin) to open and accessible (euchromatin) (Felsenfeld and Groudine, 2003). The balance between these states is dynamically regulated through chromatin-remodelling during haematopoietic development to allow for appropriate gene expression ie. enabling lineage-specific genes to become accessible while silencing genes associated with alternative lineages (Rodrigues, Shvedunova and Akhtar, 2020).

Epigenetic modifications on DNA and histones also contribute to haematopoietic cell-fate decisions. DNA methylation typically results in the repression of gene expression, while depending on the type, histone modifications can lead to both expression activation and repression. Acetylation is generally associated with gene activation, while certain histone methylations can be either activating or repressive depending on the specific residues and context.

Adelman and colleagues analysed histone modifications and DNA methylation together with RNAseq data of HSCs during ageing and revealed widespread epigenetic changes reporting significant epigenomic deregulation in aged HSCs as compared to young. For instance, 35% of all active enhancers lost H3K27ac (acetylation of the 27th lysine on the H3 histone protein) with age, including enhancers regulating numerous TFs such as RUNX3, FLI1, GATA2, GFI1, HIF1A, and KLF6, as well as epigenetic modifiers – implicating enhancer deregulation as a key factor responsible for HSC loss of function during ageing (Adelman *et al.*, 2017).

Moreover, dissociated TF motif activity variability within immunophenotypically defined populations have been shown to correlate to specific axes of differentiation, for instance, GATA motif activity in HSCs is likely to represent indicators of lineage priming (Buenrostro *et al.*, 2015). Most recently, chromatin mapping in haematopoietic cells found that HSCs and Mks have strongly overlapping chromatin signatures, with open sites corresponding particularly to key Mk TF binding sites implicating that epigenetics in the control of Mk differentiation from HSC (Heuston *et al.*, 2018). We are only now beginning to resolve how the epigenomic landscape is involved in the process of cell fate decisions, with single-cell resolution epigenetic approaches demonstrating promising signs its contributions are significant.

1.3.1 Megakaryocytes and platelets: form and function

Megakaryocytes (Mks) are rare (0.05%-0.1%) terminally differentiated polyploid cells that primarily reside within the bone marrow, whose most prominent function is the production and release of anucleate thrombocytes (platelets) into the bloodstream. Each cell measures up to 100 μm and can generate between 1,000-3,000 platelets (Ebaugh and Bird, 1951; Machlus, Thon and Italiano, 2014). Mks have also been found in other organs such as the lungs, kidney, liver and spleen (Davis *et al.*, 1992; Lefrançais *et al.*, 2017). Unlike other haematopoietic progenitors that require cytokinesis for maturation, Mks undergo multiple rounds of DNA replication without cell division during maturation, resulting in an endomitotic lobulated nucleus in some cases reaching up to 64N before the cell undergoes terminal maturation and platelet release (Figure 1.4) (Noetzli Leila J., French Shauna L. and Machlus Kellie R., 2019). The multi-lobed nucleus is retained in the Mk as the rest of the Mk cell body is transformed into protrusions known as proplatelets (platelet-sized swellings connected by cytoplasmic bridges are released into sinusoidal blood vessels), and then once the entirety of the Mk cell body is used the nucleus is extruded and degraded (Italiano *et al.*, 1999; Machlus, Thon and Italiano, 2014). It is thought Mks are polyploid in order to support the large quantities of mRNA and protein that are packaged into granules and ultimately platelets while still retaining their ability to perform multiple functions.

Platelets are the second most abundant blood cell type (after erythrocytes) at between 150-400×10^6 per mL of blood, with vital roles in maintaining the balance between haemostasis and blood clot formation (thrombosis), as well as roles in inflammation (Grozovsky *et al.*, 2015; Portier and Campbell, 2021). They are small (2–4 μm in diameter), anuclear cells that stay in circulation for 7–10 days in humans before being consumed in clot formation or eliminated by macrophages in the spleen, and to a lesser extent the liver, as part of homeostatic turnover (van der Meijden and Heemskerk, 2019). Platelets are metabolically active; equipt with several functionally active organelles including mitochondria, several types of storage granules and multiple intracellular membrane structures, including endoplasmic reticula, Golgi apparatus, lysosomes, peroxisomes and endosomes (Thon and Italiano, 2012). Though platelets have no nucleus or genomic DNA, they contain mRNA that is transported from Mks during platelet release, after which there can be no further transcription and therefore no further RNA generation. For a long time, the proteome of platelets was considered static, determined by the Mks from which they originate, wherein transcription may be influenced by external stimuli in

the bone marrow microenvironment such as inflammation (Rondina, Weyrich and Zimmerman, 2013). However, more recent data has shown platelets not only contain a pool of Mk-derived mRNAs (Newman *et al.*, 1988; Gnatenko *et al.*, 2003), and are capable of encoding for different proteins (Kieffer *et al.*, 1987; Freedman, 2011; Plé *et al.*, 2012; Bray *et al.*, 2013). Moreover, they also contain the complete machinery for *de novo* protein synthesis, making possible dynamic modifications of protein expression in mature platelets, with activating stimuli shown to induce proteome reorganisation (Bray *et al.*, 2013; Cimmino *et al.*, 2015).

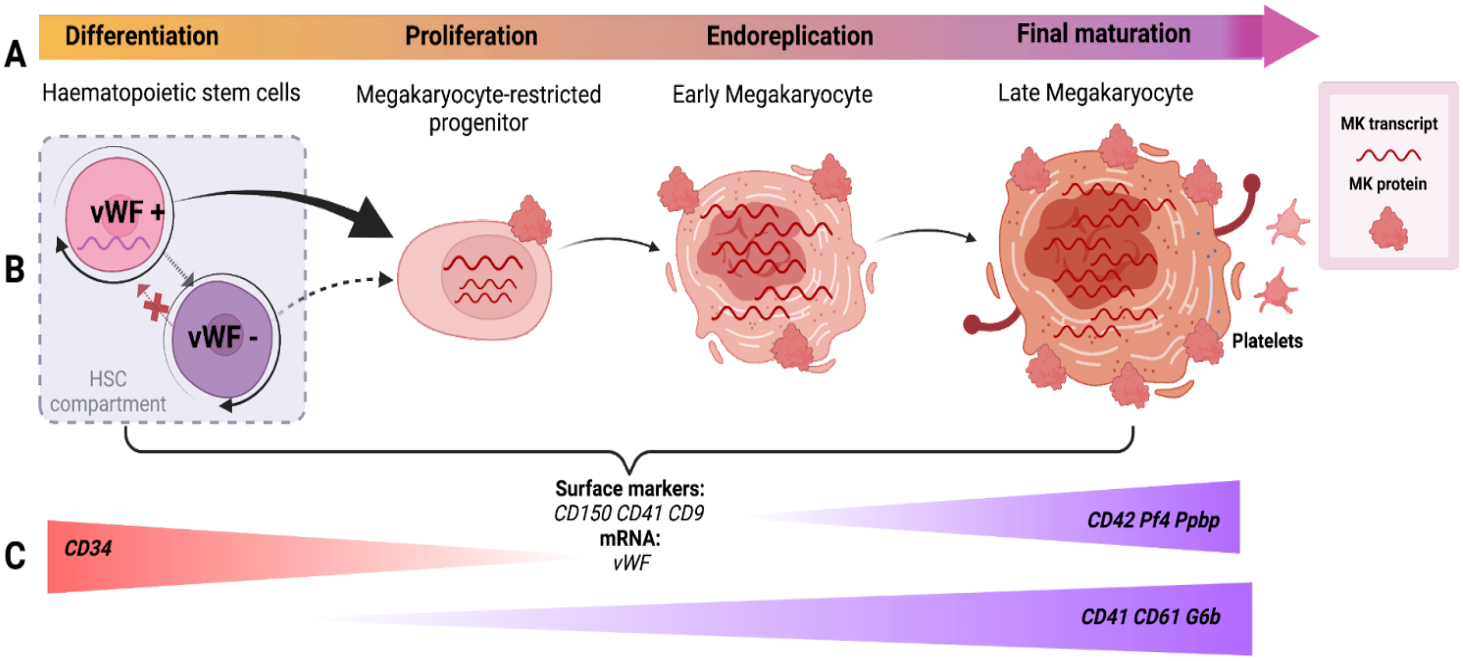


Figure 1.4. Features of megakaryocyte developmental stages. (a) Developmental stages from HSCs differentiating towards Mk lineage commitment (b) vWF+ HSCs are primed for platelet-specific gene expression increasing expression of Mk-associated transcripts and proteins with Mk maturation (c) Markers associated with specific stages of Mk lineage commitment with (adapted from Davizon-Castillo *et al,* 2020).[4]

[4] Created with BioRender.com

1.3.2 Regulation of megakaryopoiesis and the expression of Mk lineage genes

Megakaryopoiesis involves the coordinated interplay of TF-controlled cellular programmes with extracellular cues in supporting niches or as circulating factors. Its regulation occurs across multiple levels by different cytokines, the most important of which is TPO, which is regarded as the key regulator of Mk maturation and proliferation (Noetzli Leila J., French Shauna L. and Machlus Kellie R., 2019).

1.3.2.1. Signalling Pathways in Megakaryopoiesis

Thrombopoietin (TPO) regulates Mk differentiation from HSCs, with all progenitors primed to become Mks expressing the TPO receptor, *Mpl* (Debili *et al.*, 1995). This regulation is the result of a feedback loop; where constitutive TPO production by the liver is sequestered by circulating platelets in an Mpl–dependent manner (Lok *et al.*, 1994; Alexander *et al.*, 1996; de Sauvage *et al.*, 1996; Kuter and Begley, 2002). The reduction in platelet counts leads to increased levels of circulating TPO, which exerts its stimulatory effects on BM HSC increasing Mk (and platelet) numbers (McCarty *et al.*, 1995; Fielder *et al.*, 1997). TPO binding results in the initiation of multiple signalling pathways, notably including the phosphorylation of JAK2 which in turn phosphorylates downstream targets including activation of the TFs STAT3/STAT5 (Miyakawa *et al.*, 1995, 1996; Yamada *et al.*, 1995). This signalling cascade causes downstream activation of Mk-specific TFs and regulation of expression of Mk-specific genes. Kimura *et al.* demonstrated the importance of TPO signalling for Mk function using *Mpl* KO mice, showing a global decrease in HSCs and a drastic reduction in specifically Mks and platelets (Kimura *et al.*, 1998).

Besides the well-established role of TPO signalling in optimal Mk function, there are TPO-independent pathways of megakaryopoiesis that are less understood but still exist. Some patients lacking functional TPO signalling are still capable of platelet production, indicating the presence of alternative mechanisms (van den Oudenrijn *et al.*, 2000). One such TPO-independent mechanism involves the chemokine IGF-1 (insulin-like growth factor-1). IGF-1 promotes the differentiation of CD34+ cells toward the Mk lineage through AKT signalling. Studies have shown that administering IGF-1 *in vivo* increases platelet counts in mice lacking the Mpl receptor and in lethally irradiated mice, suggesting a TPO-independent phenotype (Chen *et al.*, 2018). Previous research has also implicated other signalling effectors in TPO-independent megakaryopoiesis, including interleukin (IL) 1α, CCL5 (C-C motif chemokine ligand 5), and Notch signalling (Mercher *et al.*, 2008; Nishimura *et al.*, 2015; Machlus *et al.*, 2016)).

1.3.2.2. Transcription Factors in Megakaryopoiesis

TFs are involved from foetal HSC specification to adult HSC maintenance and renewal and Mk commitment, regulating differentiation along each lineage decision point from HSC to Mk (Noetzli Leila J., French Shauna L. and Machlus Kellie R., 2019).

At the CMP juncture, myeloid commitment depends on the balance of expression of antagonistic transcription factors GATA1 and PU.1 (Spi-1) which influence skewing towards the MEP and GMP respectively (Rekhtman *et al.*, 1999; Rhodes *et al.*, 2005; Arinobu *et al.*, 2007). It was shown these TFs were capable of upregulating their own expression whilst inhibiting the expression of the other (Tsai, Strauss and Orkin, 1991; Yu *et al.*, 2002; Ferreira *et al.*, 2005; Okuno *et al.*, 2005; Rosmarin, Yang and Resendes, 2005). This work proposed that when both lineage-specific TFs are expressed at low levels, myeloid progenitors remain in a multipotential state until one of the TFs dominates resulting in an autoregulatory loop that drives commitment one way or another (Cantor and Orkin, 2001). This hypothesis is supported by data showing co-expression of *Gata1* and *Spi1* within CMPs (Palii *et al.*, 2019), however, some reports dispute the abrupt binary switching for myeloid fate decisions, suggesting instead that these TFs work to reinforce lineage choice once made (Hoppe *et al.*, 2016).

MEP differentiation to the Mk vs. Ery lineages is coordinated by the time- and dose-dependent expression of various TFs. MEPs express NFE2 (nuclear factor, erythroid 2), SCL, GFI1B (growth factor–independent transcriptional repressor), GATA1, GATA2, KLF1 (Kruppel-like factor), and ETV6, whilst EKLF (erythroid Krüppel-like factor) and c-Myb are exclusive to the Ery lineage and FLI1 and RUNX1 are exclusive for the Mk lineage (Starck *et al.*, 2003; Bouilloux *et al.*, 2008; Doré and Crispino, 2011; Kuvardina *et al.*, 2015).

GATA family TFs are among the most studied haematopoietic TFs, and are known to bind to *cis*-GREs of many Mk- and Ery-specific genes. GATA-1 is essential for RBC development and is expressed during the later stages of haematopoietic differentiation. In megakaryopoiesis, it has been shown to be required for terminal maturation, while GATA-2 on the other hand is required for HSC maintenance as well as further downstream during early megakaryopoiesis. GATA-2 and GATA-1, along with its cofactor FOG1, are expressed in an antagonistic manner in the MEP; GATA-2 promotes megakaryopoiesis at the expense of erythropoiesis, whereas GATA-1 promotes erythropoiesis (Ikonomi, Noguchi, *et al.*, 2000; Ikonomi, Rivera, *et al.*, 2000; Cantor and Orkin, 2002; Galloway *et al.*, 2005).

RUNX-1 also directs MEPs toward Mk fate, suppressing the Ery-specific TF KLF-1. This shift increases the ratio of Mk-specific FLI-1 to KLF-1, ultimately promoting megakaryocyte

differentiation (Bouilloux *et al.*, 2008; Kuvardina *et al.*, 2015). Moreover, Mk-specific TF FLI-1 (which is negatively regulated by ETV6) regulates the expression of Mk receptors GPIX and GPIbα (glycoprotein Ib platelet subunit alpha) (Kwiatkowski *et al.*, 1998, 2000). Finally, During the final stages of Mk maturation, NFE2 is essential. It controls the Mk-specific microtubule component β1-tubulin (Lecine *et al.*, 1998; Schwer *et al.*, 2001). Mice lacking NFE2 exhibit normal Mk proliferation but impaired platelet production, highlighting its critical role in Mk proplatelet formation (Shivdasani *et al.*, 1995; Levin *et al.*, 1999).

1.3.3 Differentiation models of megakaryopoiesis

In the classic haematopoiesis model, after myeloid/lymphoid bifurcation at the MPP, CMPs differentiate into bipotent MEPs which eventually go on to differentiate into unipotent mature Mk progenitors (MkPs) (Figure 1.5A) (Akashi *et al.*, 2000). Over the last decade however, refined experiments have demonstrated this to be an oversimplified model that does not account for the complexities of Mk differentiation/lineage commitment subsequently driving many to explore other possible models.

1.3.3.1. The overlapping signatures of Mk and Ery lineages

The vast majority of cells produced by the BM are of Mk and Ery lineages. Data have consistently shown that the Mk and Ery lineages are closely related, including their largely parallel ontogeny during embryonic development, common regulatory networks and shared expression of lineage-determining transcription factors such as GATA-1 and -2, TAL1, FOG-1 and GFI-1b (Psaila and Mead, 2019). The antagonistic expression of these TFs factors with others has been shown to promote Mk-Ery differentiation while simultaneously repressing myeloid programmes (Chou *et al.*, 2009). In the clinic, the link between Mk and Ery lineages was recognized from observations that erythroleukaemia cell lines and blasts isolated from patients with bi-phenotypic leukaemias can possess features of both Ery and Mk lineages, whereas Ery or Mk with B or T lymphoid characteristics were rarely observed (Matutes *et al.*, 2011). Further illustrating their shared signature, erythropoietin (EPO) treatment was found to stimulate platelet production in addition to erythropoiesis, and the exposure of CD34+ cells to both EPO and thrombopoietin (TPO) was found to increase both Ery and Mk progenitors (McDonald *et al.*, 1987; Papayannopoulou *et al.*, 1996).

The first early clonogenic assays of multipotency found that MEPs gave rise to Mk-only, Ery-only, and Ery-Mk mixed colonies (Debili *et al.*, 1996; Akashi *et al.*, 2000; Klimchenko *et al.*, 2009). These MEPs were distinguishable from CMP and GMP populations through the absence of the surface antigens CD123 or Flt3 (Manz *et al.*, 2002; Adolfsson *et al.*, 2005). However, the majority of cells from MEP subsets were later shown to generate primarily single lineage progeny, with a small minority generating mixed Mk/Ery progenitors (Pronk *et al.*, 2007). The lineage potential of MEPs, initially presented as a bipotent cell type, was largely skewed toward the Ery lineage, with up to 80% of single MEPs producing pure Ery colonies (Manz *et al.*, 2002) - a finding that has been recapitulated in multiple investigations since (Psaila *et al.*, 2016; Miyawaki *et al.*, 2017).

A caveat to studies on the output of MEPs concerns the difference between the maturation of Mk vs Ery lineages. Debili et al. showed dual potential progenitors in liquid culture quickly made lower numbers of Mk than Ery cells, which disappeared after making platelets followed by a much greater numerical expansion of Ery cells (Debili *et al.*, 1991). Moreover, considering Ery-committed cells undergo considerably more cell divisions that Mk-committed cells, and how mature Mks typically survive < 3 weeks in culture it is possible the analysis of cell potential in culture by FACS could be masking the yield of rare Mks; missing Mk potential no longer evident at later time points as used to assess Ery and myeloid progeny (Besancenot *et al.*, 2010; Pop *et al.*, 2010; Sim *et al.*, 2016; Xavier-Ferrucio and Krause, 2018). This lack of effective MEP purification complicated the early studies aiming to establish the role of the MEP in haematopoiesis. Using the more reliable MEP gating strategies, it was later revealed the originally defined MEP compartment is composed of at least three subfractions with distinct gene expression and functional capacities - cells enriched for Mk/Ery output but with residual myeloid differentiation capacity ("Pre-MEP"), and Ery-primed and Mk-primed bipotent fractions ("E-MEP" and "Mk-MEP") (Psaila *et al.*, 2016).

Studies suggest that the MEP fate decision is governed at least in part by the regulation of several TFs factors promoting either Ery and Mk differentiation (Starck *et al.*, 2003; Doré and Crispino, 2011; Bianchi *et al.*, 2015; Sanada *et al.*, 2016). But also upstream of TFs GTPase activating proteins such as Arhgap21 and Rgs18 have been implicated in the MEP fate decision, supporting Ery and suppressing Mk commitment. Acting as an effector for Gfi1, Rgs18 has been shown to regulate downstream signalling through Erk1 to modify the balance of Fli1 and Klf1 (Sengupta *et al.*, 2013; Xavier-Ferrucio and Krause, 2018; Xavier-Ferrucio *et al.*, 2018)

Sufficient evidence exists at the transcriptional, immunophenotypic, and functional levels connecting Mk and Ery differentiation pathways and the existence of a bipotent MEP. While MEPs clearly do exist *in vivo* in both mice and humans, alternative pathways to achieve fate commitment to the Ery and Mk lineages that do not require that they pass through this bipotent MEP stage are now known to exist. For example, the reconstruction of differentiation trajectories across murine cKit+ single cells indicated coupling of Ery and basophilic lineages, and the earlier divergence of Mk cells from multipotent progenitors, with lympho-myeloid differentiation occurred along a separate trajectory - suggesting Mk-Ery lineages are not always intertwined (Tusi et al. 2018). MEPs may therefore represent a transient state and are thus difficult to isolate in contrast to other more well-defined progenitors.

1.3.3.2 Shared features between Mks and HSCs

The conceptual distance between Mk commitment and the HSC compartment became increasingly shorter with the more we came to discover about both cell types. First with the revision to the haematopoietic model describing the Mk-Ery lineage bifurcation from the myelo/monocytic lineage that co-segregates with the lymphoid fate as the first major lineage fate decision in haematopoietic differentiation (Doulatov *et al.*, 2010). Indeed, in spite of the hierarchical distance between Mks as terminally differentiated cells and HSCs as the most undifferentiated cell type, they share several important features (Figure 1.4). These include cell surface molecules, lineage-specific TFs, and specialised signalling pathways - supporting the notion that these cell types are more closely connected than once thought, and the possibility that Mk lineage specification occurs at immature haematopoietic stages (Nishikii, Kurita and Chiba, 2017).

Both Mks and HSCs rely on TPO signalling, critical for platelet production and stem cell maintenance respectively, where HSCs derived from *Mpl*$^{-/-}$ mice showed significantly reduced long-term repopulating capacity (Kimura *et al.*, 1998; Solar *et al.*, 1998; Qian *et al.*, 2007; Yoshihara *et al.*, 2007). Likewise, TPO KO itself reduces HSC function It has also been shown that severe thrombocytopenia (disease of low blood platelet count) is caused by loss of function mutations in *Mpl,* and patients have been found to have a higher risk of bone marrow failure, further implicating the role of TPO in both Mk and HSC function (Huang and Cantor, 2009). Other key Mk cell surface receptors co-expressed by HSCS include CD9 (Karlsson et al., 2013), CD41 (Gekas and Graf, 2013), CXCR4 and CD150 (Kiel *et al.*, 2005). To demonstrate the effect of CD41 on stem cell phenotypes, Gekas and Graf used competitive BM transplantation to show that CD41+ HSCs had a more quiescent phenotype, and the knockout of CD41 in mice resulted in pancytopenia in animals, where platelet and red blood cell and leukocyte numbers were significantly reduced (Gekas and Graf, 2013). High CD150 expression has long been used for the enrichment of HSCs, having been identified as a marker for HSCs with greater self-renewal (Beerman *et al.*, 2010).

Several important TFs play significant roles in both HSCs and Mks. These shared TFs include RUNX-1, GATA-2, EVI-1, TAL1 and PBX1 (Huang and Cantor, 2009). Specifically, deficiencies in RUNX-1 have been shown to reduce HSC numbers and result in abnormalities in Mk nuclei, including hypo-lobulation, low DNA ploidy, and under-developed cytoplasm (Sun and Downing, 2004; Talebian *et al.*, 2007). Similarly, EVI-1 KOs significantly decrease the number of phenotypic HSCs, leading to impaired self-renewal and repopulation capacity. These mice also experience thrombocytopenia and delayed platelet recovery following treatment with the cytotoxic agent 5-fluorouracil (5-FU) (Goyama *et al.*, 2008). These findings

collectively suggest that certain TFs have lineage-specific functions in both HSC and Mk development.

Moreover, Notch signalling has a significant role in both the multipotency of HSCs and the specification of Mk lineage from HSCs in *in vitro* co-culture systems. Notch signalling is a well-established pathway that influences various developmental processes and regulates cell fate decisions (Wilson and Radtke, 2006). Activation of Notch signalling was found to promote increased Mk specification from HSCs in co-culture systems, while inhibition of Notch signalling reverses this effect (Burns *et al.*, 2005; Mercher *et al.*, 2008; Cornejo *et al.*, 2011).

The functional and compositional similarities among Mks, HSCs, and endothelial cells suggest an important connection between HSCs and Mks. Platelets, derived from Mks, serve to repair endothelial lesions and prevent bleeding. This process involves adhesion to exposed subendothelial structures, platelet activation, aggregation, and activation of angiogenesis. Consequently, platelets and endothelial cells share pathways that regulate hemostasis and thrombosis. Additionally, many lineage-specific factors expressed in Mks and HSCs are also present in endothelial cells and/or hemangioblasts, the common precursor of HSCs and endothelial cells during embryogenesis (Choi *et al.*, 1998; Lancrin *et al.*, 2009). Given the similarities in the development of HSCs and endothelial cells, as well as the functional roles shared by Mks and endothelial cells, it is reasonable to propose this strengthens links between Mks and HSCs.

In recent years, the co-localisation of HSCs and Mks within the BM has emerged as another important factor. Both HSCs and Mks reside in the vascular sinusoid regions of the BM, and recent studies have highlighted the significant roles of Mks in supporting the HSC niche (Kiel *et al.*, 2005). Bruns *et al.* demonstrated in mice that Mks directly regulate the size of the HSC pool. They found that endogenous HSCs are frequently located in close proximity to Mks in a non-random manner. Their findings were further confirmed by *in vivo* depletion of Mks, which resulted in the loss of HSC quiescence and expansion of functional HSCs. This indicates that terminally differentiated Mks derived from HSCs contribute to the HSC niche (Bruns *et al.*, 2014). Multiple mechanisms have been proposed to explain how Mks directly influence HSC behaviour, including the involvement of CXCL4, TPO, and TGF-beta (Bruns *et al.*, 2014; Nakamura-Ishizu *et al.*, 2014; Zhao *et al.*, 2014).

Extensive research has focused on the shared characteristics between Mks and HSCs to elucidate Mk lineage commitment. As the body of evidence has grown, it became increasingly clear that HSCs exhibit lineage bias towards the Mk fate.

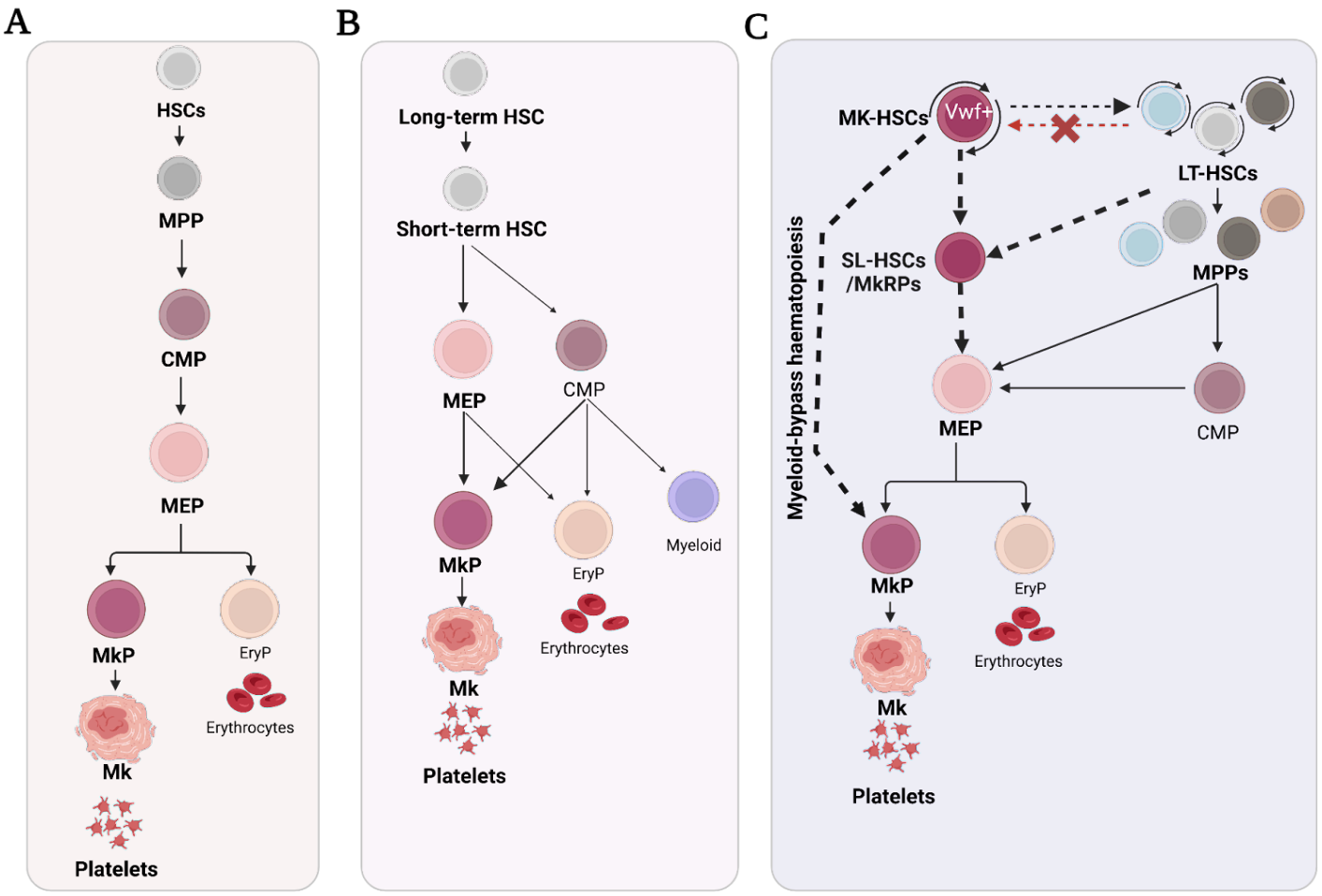
1.3.3.3. Mk-biassed HSCs

The prospective isolation of platelet-biassed LT-HSCs by Sanjuan-Pla and colleagues confirmed the existence of a HSC subset showing strong Mk lineage-biassed reconstitution (with limited lymphoid potential) proving that this lineage is not strictly derived from the CMP to MEP route. This subset exhibits platelet-specific gene expression that is identifiable based on high expression of von Willebrand factor (*vWF*), a key Mk marker protein involved in platelet aggregation (Sadler, 1998; Sanjuan-Pla *et al.*, 2013). Importantly, they were among the first to show a hierarchical relationship between HSC subtypes *in vivo*, where only *vWF+* can give rise to v*WF+* HSCs (Ly-HSCs), and not the opposite, strongly suggesting platelet-primed HSCs exist at the apex of the HSC subtypes (Figure 1.5C) (Sanjuan-Pla *et al.*, 2013). Using a transgenic (vWF)–green fluorescent protein (GFP) mouse model they demonstrated that around half of murine LT-HSCs expressed the Mk gene vWF and with platelet-biassed reconstitution and myeloid but low lymphoid contribution.

Further corroborating Mk-priming in the HSC compartment, cKit^{high} LT-HSCs were found to exhibit Mk-biassed potential. Researchers showed that HSCs with differential lineage output correlated with cKit cell surface expression, where cKit^{low} HSCs were more quiescent than cKit^{high} HSCs and displayed lower Mk-reconstitution potential (Shin *et al.*, 2014).

Soon after, several groups demonstrated Mk-lineage bias using single-cell *in vitro* methods, confirming a subpopulation of HSCs exhibit strong Mk bias (Mk-HSCs) but also revealing even direct differentiation towards Mk unipotent progenitors without cell division (Yamamoto *et al.*, 2013; Nishikii *et al.*, 2015; Roch, Trachsel and Lutolf, 2015)). Notably, Yamamoto *et al.* demonstrated both the direct commitment from HSCs towards Mk unipotent progenitors, as well as the presence of a functional MEP *in vivo.* Using FACS-sorted candidate HSPCs transplanted into recipients, they showed that reconstituting HSCs generated 1 HSC and 1 MkP showing that Mk unipotency can arise directly from HSCs (Yamamoto *et al.*, 2013). This work further confirmed the existence of Mk-HSCs in addition to MEP-based Mk reconstitution, suggesting that Mk-HSCs possibly serve as an important source of Mks under specific physiological conditions.

Consistent with Yamamoto's findings, a single cell transplantation approach of over 1000 single-cells found approximately 10% of vWF+ LT-HSCs were capable of stable replenishment only Mks, whilst 90% replenished other lineages in addition to Mks. Mk-biassed HSCs sustained multipotency *in vitro* and in secondary transplants and platelets were the only lineage that was invariably reconstituted in 100% of transplants (Carrelha *et al.*, 2018).



[5]**Figure 1.5. Models of megakaryocyte commitment** (a) The classical hierarchical haematopoiesis model: the bifurcation of myeloid/lymphoid lineage first occurs in MPPs, and MkPs eventually generate from MEPs as the progeny of CMPs (b) Alternative model: Based on the identification of LMPPs from short-term HSCs, with almost no Mk/ Ery lineage potential where the bifurcation of Mk lineage first occurs during differentiation from HSCs (c) Proposed model from recent literature: An immunophenotypic-defined HSC population contains a functionally heterogeneous, Mk-biassed subpopulation of HSCs that directly gives rise to MkPs and bypass the MEP stage.

[5] Abbreviations: HSC, haematopoietic stem cell; MPP, multipotent progenitor; CMP, common myeloid progenitor; Ery, erythroid; LT-HSC, long-term haematopoietic stem cells; ST-HSC, short term HSC; MEP, Mk/Ery progenitor; Mk, megakaryocytes; MkRP, Mgk-repopulating progenitors; SL-MkP, stem-like Mk committed progenitor, Mk HSC, Mk-biassed *Vwf*+ HSCs.

Moreover, whole transcriptome single-cell analysis of the HSC compartment identified a subpopulation expressing a strong signature of Mk-specific genes (Grover *et al.*, 2016). Together these data indicate whilst lineage biases have been seen across other blood cell lineages, the platelet lineage is the only cell type that has been found to exclusively comprise 100% progeny in some HSCs, suggesting that HSC Mk-priming may be a phenomenon exclusive to the Mk lineage (Ceredig, Rolink and Brown, 2009).

Lineage-tracing experiments using transposon tagging for clonal tracing the fates of progenitors and HSCs have also provided evidence for Mk-biassed HSCs. Rodriguez-Fraticelli *et al*. found approximately 50% of Ery clones shared transposon tags with myeloid cells whilst comparatively very few Mk clones were shared with Ery cells, which would have been predicted under the premise a shared MEP is the predominant differentiation route in haematopoiesis. This work suggested instead a shared origin for Ery and myeloid lineages, but a largely separate pathway for Mk differentiation (Rodriguez-Fraticelli *et al.*, 2018). Together these data prompted important revisions to the existing model of megakaryopoiesis, of platelet-primed HSCs at the apex of the hierarchy with multiple and even direct routes towards Mk commitment (Figure 1.5C).

With indications overwhelmingly in agreement that multiple routes for Mk generation exist, it seems reasonable to presume this may be a protective mechanism linked to the blood demand in emergency cases such as infections or acute bleeding. Under this hypothesis, Haas *et al.* used an LPS infection system and found Mk repopulating progenitors (Mk-RPs) within the Lin- cKit+ CD150+ CD48- compartment, in addition to Mk-HSCs. These unipotent progenitors upregulated CD41 expression directly giving rise to Mks by-passing any other intermediate progenitors of the Mk differentiation axis (Haas *et al.*, 2015).

The evolving understanding of megakaryopoiesis, specifically the direct differentiation of Mks from Mk-HSCs, presents new opportunities to study this process more thoroughly. It is crucial to understand the branching points of the haematopoietic system, as it has significant clinical implications. This understanding can improve our knowledge of how the body responds to haematopoietic stresses like ageing and infection, as well as the initiation and progression of leukaemia and other myeloproliferative neoplasms (MPNs). Moreover, by studying the regulation of Mk commitment pathways, we can identify therapeutic targets for disorders such as anaemia or thrombocytopenia, which are associated with abnormal Mk function. Finally, research suggests changes in Mks and platelets during disease can exacerbate inflammation by affecting the BM environment. Therefore, it is essential to investigate Mks not only as platelet-producing cells but also as cells involved in maintaining the BM niche

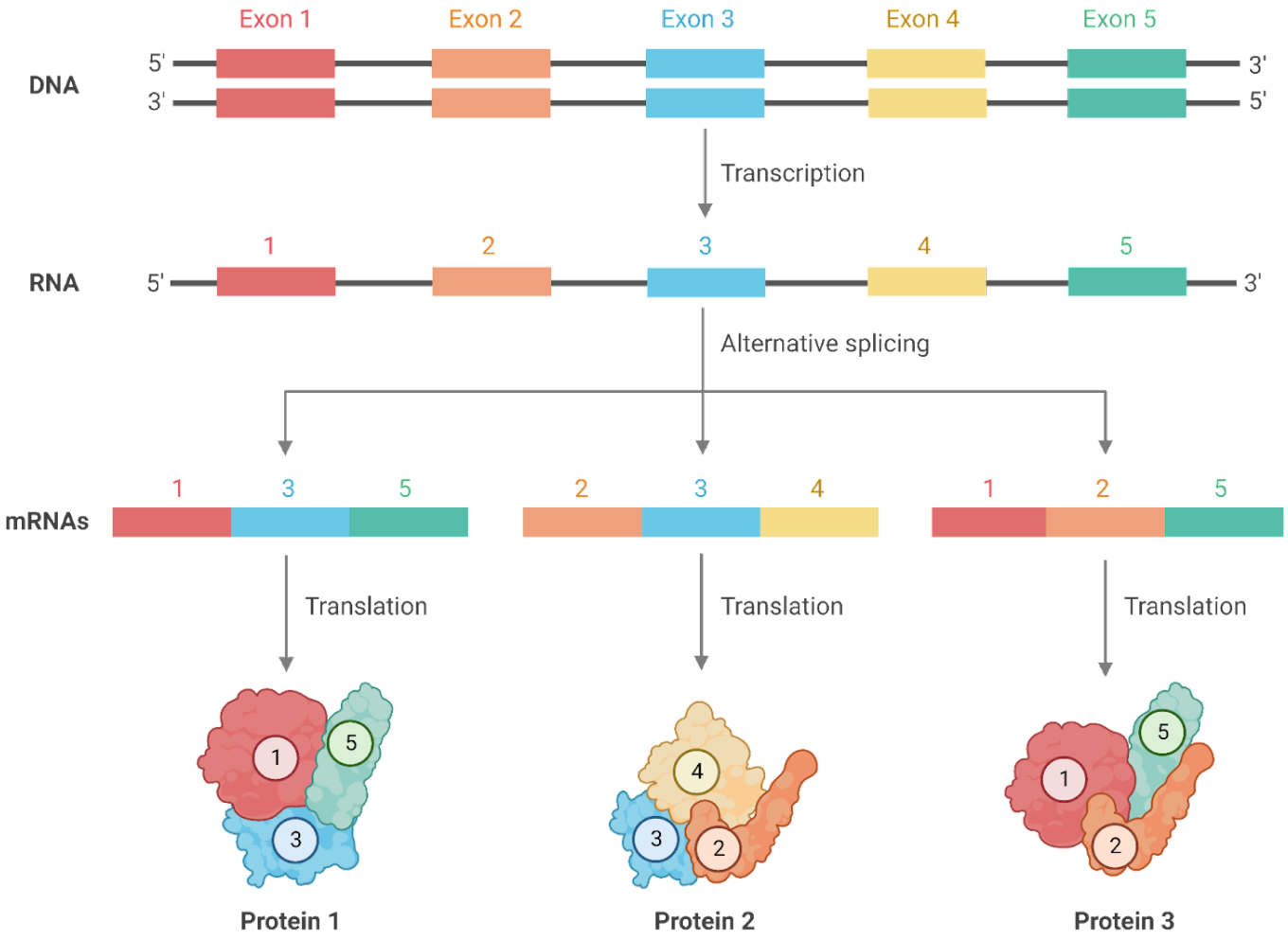
1.4 Alternative splicing in lineage commitment

In the 1950s and ’60s, the study of bacterial genetics first opened the way toward understanding life as the genetically encoded interactions of macromolecules and the shift to seeing the world through the lens of molecular biology (Stent 1968). In these early days of modern molecular biology, scientists almost exclusively worked with bacterial systems, which were easy to grow and manipulate in the laboratory. Although the concept of “genetic information” had been rapidly and widely adopted at this point, no one was clear about what exactly genetic information might consist of. The central dogma of molecular biology - that DNA is transcribed into RNA and RNA translated into proteins - was first publicly proposed in 1957 by Francis Crick as part of a Society for Experimental Biology Symposium on the Biological Replication of Macromolecules held at University College London (Crick, 1958). Ultimately the “flow of information” and the concept of RNA as an intermediate between DNA and protein resolved the link between base sequences of nucleic acids and those of amino acids in a protein (Crick, no date). At the time ribosomes were known only as microsomes, and their function was uncertain; messenger RNA was still undreamt of—it would be properly identified only in the summer of 1960, and the discovery was not published until the following year (Brenner, Jacob and Meselson, 1961; Gros *et al.*, 1961; Cobb, 2015). We now know coding regions of the genome, defined as nucleic acid sequences that when transcribed into mRNA can ultimately be translated into polypeptides, are characterised as genes.

1.4.1 Alternative Splicing

Molecular biology progressed and with it, our understanding of genomics developed. The next major challenge in understanding the process of gene expression was to prove the conversion of DNA into protein. Several groups in the late 1970s studying the relationship between cytoplasmic RNA and the DNA structure of adenovirus using electron microscopy first identified a series of RNA molecules late in viral infection containing sequences from noncontiguous sites in the viral genome that they termed "mosaics" (Berget, Moore and Sharp, 1977; Chow *et al.*, 1977). This heterogeneous nuclear RNA discovered by Sharp, Berget and Moore turned out to be the evidence of a process coined as RNA splicing (Berget, Moore and Sharp, 1977). This landmark discovery described for the first time the process by which DNA information is transcribed into mRNA. When a gene is copied into RNA, it contains a long, 'jumbled message' of nucleotide sequences called exons and introns (Gilbert, 1978). While exons compose a message with specific instructions from a particular gene, they are separated by introns that are interspersed in the RNA, rendering the message incomprehensible. Similar to a message that needs to be decoded because of a sentence with extra letters, introns must be excised from pre-mRNA to form a coherent message consisting only of coding exons (codons) as a single-stranded molecule called a mature mRNA transcript - the genetic information used to synthesise proteins in the cytoplasm. This discovery revealed that the expression of genes and their protein products involves a series of complex processing stages. Subsequently, introns were found in many other viral and eukaryotic genes, one of the first being those for haemoglobin and immunoglobulin (Darnell, 1978; Early *et al.*, 1980).

Before knowledge of RNA splicing, the consensus was that all organisms have the same gene structure as bacteria, which lack introns. Bacterial RNA transcripts are mostly collinear - meaning there is one-to-one correspondence of bases between the gene and the mRNA transcribed from the gene - following this hypothesis that 'one gene = one protein', when taking the human transcriptome with a predicted ~20,000 genes as a reference, there should be ~20,000 canonical human proteins. RNA splicing demonstrated that eukaryotic cells, with their discontinuous genes, are far more complex than bacterial cells. It also debunked the dogma that one gene produces one mRNA, and all mRNAs from a gene produce one protein. The completion of the Human Genome Project in 2003 confirmed a vast discrepancy between the number of annotated protein-coding genes and the number of observed human polypeptides violating this hypothesis (International Human Genome Sequencing Consortium, 2004). It quickly became clear that the expression of genes and their protein products relies on many more levels of regulation. In eukaryotic transcription, RNA polymerases in the nucleus copy DNA that is composed of exons (amino-acid coding regions) and introns (non-amino-acid



[6]

Figure 1.6. Alternative splicing during transcription results in different mRNA isoforms and protein products.

[6] Created with BioRender.com

coding regions) 5’ to 3’ generating complementary messenger RNA (mRNA) products. The purpose of transcription is to process the DNA such that intronic regions are excised from the resulting mRNA transcript, leaving behind only the coding exons (codons) as part of the single-stranded molecule for export into the cell’s cytoplasm.

Alternative splicing is the summation of multiple mechanisms in every eukaryotic cell including differential usage of splice sites, transcription start sites (TSS) and polyA sites that enable the expression of multiple unique isoforms for each gene (Wang *et al.*, 2008) (Figure 1.6). Consequently, the proteins translated from alternatively spliced mRNAs will contain differences in their amino acid sequence and, often, in their biological functions. It is estimated that up to 95% of human multi-exon genes undergo alternative splicing to encode proteins with different functions (Figure 1.7). In fact, splice isoforms can have even opposing functions and there are many instances whereby a splice isoform acts as an inhibitor of canonical isoform function, thereby adding an additional layer of regulation to important processes.

Exon skipping

Exon skipping is a splicing event where specific exons are excluded or skipped during mRNA processing, leading to the removal of particular sequences from the final mRNA transcript. Mutations within an exon or its adjacent splice sites can disrupt normal splicing, resulting in the exclusion of the affected exon during mRNA processing. As a consequence, the protein produced from the aberrantly spliced mRNA may be incomplete or truncated, leading to functional impairment or loss.

Intron retention

Intron retention occurs when an intron is not properly removed from the pre-mRNA during splicing and is retained in the final mRNA transcript. This results in the inclusion of intronic sequences within the mature mRNA, which can have functional implications.

Alternative 5’ donor site

This occurs when an alternative site within an exon or intron is used as the starting point for the splicing process. For example, in T cells, the CD45 (Ptprc) gene exhibits AS with different 5' donor sites, giving rise to CD45 isoforms with distinct expression patterns and functions, thereby impacting the structure and function of the protein (Hermiston, Xu and Weiss, 2003)

Alternative 3’ acceptor site

This occurs when an alternative site within an exon or intron is used as the ending point (acceptor site) for the splicing process. The utilisation of alternative 3' acceptor sites can result in the inclusion of additional exonic sequences or the skipping of exons. A mutation upstream of GATA1, a crucial TF for Ery and Mk lineage development, was identified in two unrelated

patients with a unique form of dyserythropoietic anaemia. Researchers revealed the mutation reduced normal splicing of this region of GATA1 and promoted an intron retention event of 15 nucleotides involving an alternative splice acceptor site (Abdulhay *et al.*, 2019).

These diverse splicing mechanisms offer cells the capacity to generate a broad spectrum of protein isoforms, thereby enhancing protein diversity and functional complexity. Dysregulation of AS can have far-reaching consequences for cellular functions and is closely linked to various diseases. An in-depth investigation of AS and its influence on protein expression and function holds great potential for advancing our comprehension of intricate biological processes and disease mechanisms in hematopoiesis at the single-cell level. By unravelling the intricacies of AS, we can gain valuable insights into the molecular underpinnings of haematopoietic differentiation, lineage specification, and disease progression.

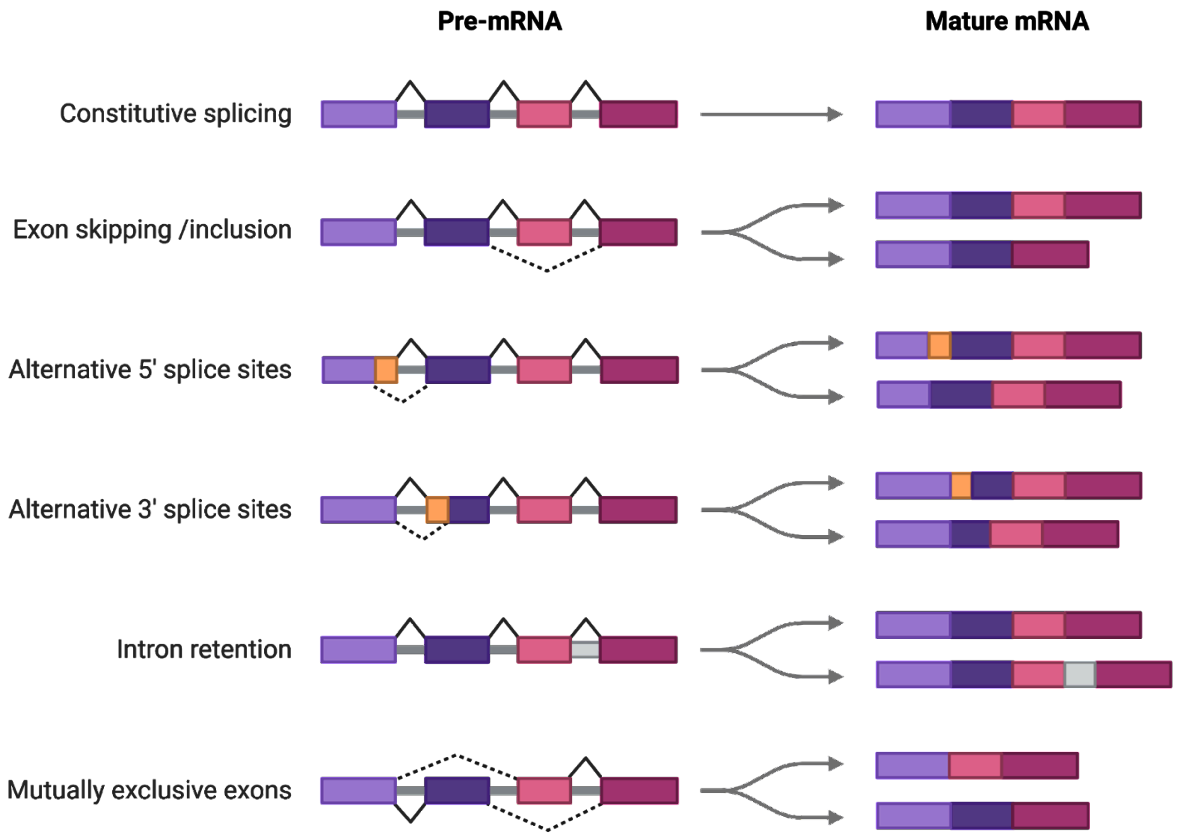


Figure 1.7. Mature mRNA products that arise through different types of splicing of pre-mRNA.

1.5 Single-cell biology

The demand for sufficiently sensitive methods to investigate molecular-level heterogeneity within tissues was an evident universal challenge in biology. Traditional gene expression analyses often obscured critical differences in gene expression between cells, hindering the understanding of divergent functions. The use of bulk cell populations provides information on cell population averages, making the assumption of homogeneity within a population and, depending on the experimental question, often fails to capture the true state within the sample (Figure 1.8). The emergence of single-cell resolution approaches, combined with computational tools necessary for their interpretation, has significantly enhanced our capacity to comprehend complex cellular systems.

For decades, stem and progenitor cells were characterised based on their behaviour using the traditional haematopoietic single cell assays such as *in vitro* colony-forming assays, other clonal assays and transient *in vivo* repopulation (Kondo, Weissman and Akashi, 1997; Akashi *et al.*, 2000; Adolfsson *et al.*, 2005; Månsson *et al.*, 2007; Pronk *et al.*, 2007). But how properties of the assays may influence the readout and therefore the conclusions that can be reached is an important consideration when interpreting generated data. For instance, HSC transplantation assays evaluate repopulation ability but may not reflect the cells' contribution to haematopoiesis in their native, steady-state setting (Wilson *et al.*, 2008; Bernitz *et al.*, 2016). Lineage potential and fate can vary greatly between in vivo and in vitro conditions (Carrelha *et al.*, 2018). Likewise, progenitor cell populations, even when identified by specific markers, may exhibit functional heterogeneity, and individual cells do not necessarily adopt all possible lineage outcomes *in vitro* (Rieger *et al.*, 2009). Therefore, while early models of haematopoiesis undoubtedly served as useful guides for genetic perturbation studies, they also have revealed technical limitations.

Cell type and state are influenced by various molecular aspects, including genomics, epigenomics, transcriptomics, proteomics, and metabolomics, which are in turn shaped by intrinsic and extrinsic factors (Mincarelli *et al.*, 2018). Single-cell transcriptome, epigenetic and transplantation analysis, barcoding, and *in vitro* clonal functional assays have helped to resolve molecular heterogeneity that population-based strategies could not capture. Consequently, many annotations have been redefined through the lens of single-cell transcriptomics, enabling exploration of molecular heterogeneity.

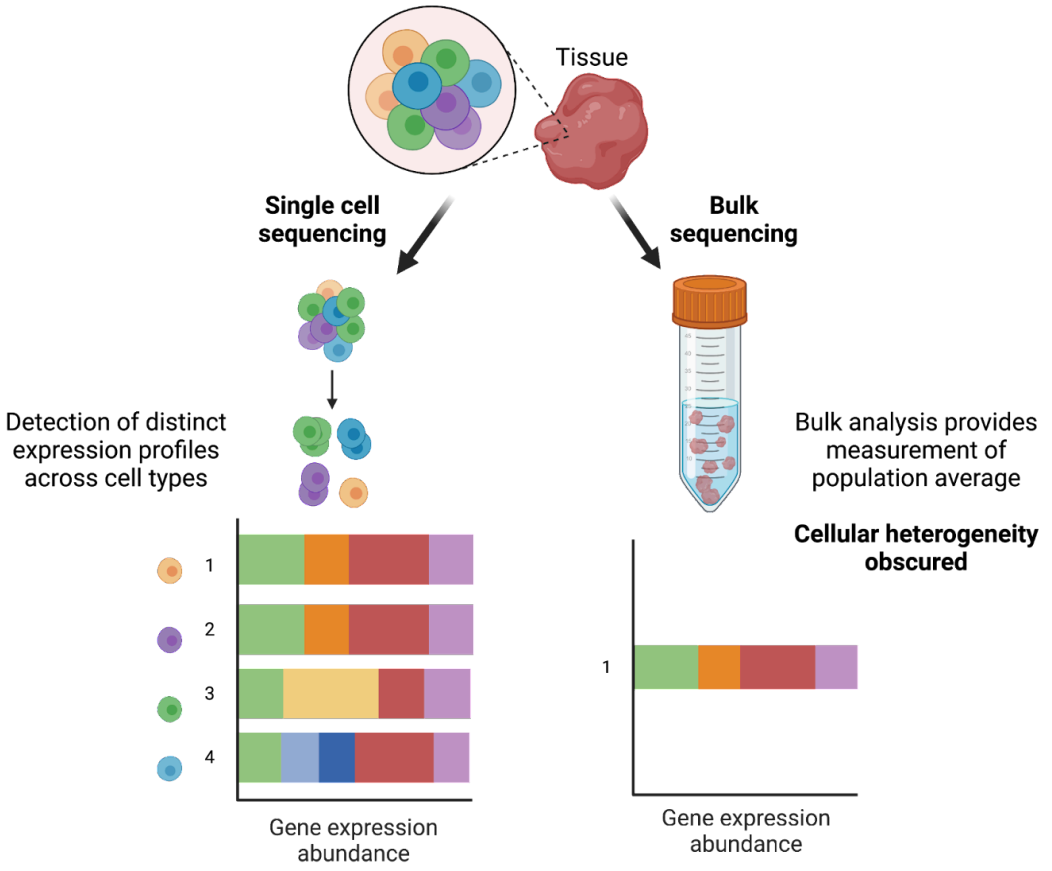


Figure 1.8. Schematic of single-cell analysis compared to traditional bulk sample analysis. Single-cell assays enable tissue profiling at single-cell resolution, where distinct expression profiles of individual cells can be appreciated. Single-cell approaches allow for the detection of unique signatures between cell types where genes that are lowly expressed at the tissue level and often obscured can be captured and studied. [7]

[7] Created with BioRender.com

Single cell technologies

Key to the study of single cells is the capacity to effectively isolate cells to enable analysis of their unique molecular identity. Numerous methods to achieve this are available, and broadly can be categorised into manual isolation (using micropipettes or micromanipulation), FACS based isolation, combinatorial indexing and microfluidic isolation (Mincarelli *et al.*, 2018). The approach selected is dependent on the specific biological question, as well as the practical considerations and experimental objectives for a given experiment.

Combinatorial indexing involves using a two-step barcoding strategy combined with FACS-sorting enabling each sequencing read to be assigned to an individual cell. This is a popular strategy to increase throughput without the need for microfluidics. Microfluidic based isolation, in which cells are captured in individual droplets or nanowells for processing, is well suited to maximising throughput and minimising the reagent cost per cell. Cells are co-encapsulated in droplets with uniquely barcoded oligodT primers, enabling cDNA to be pooled and sequenced in parallel, with reads assigned to individual cells based on their barcode. Such approaches (such as the popular 10x Genomics platform) enable 3′ or 5’ transcript counting from thousands of cells in parallel.

Plate-based scRNA-seq strategies use the physical single cell separation and combine full-length, PCR-based cDNA amplification with tagmentation-based Next Generation Sequencing (NGS) library preparation to generate single cell libraries. These whole transcript methods (such as Smart-seq2) that sequence reads from all regions of an RNA are superior in their ability to enable high sequence coverage often leading to higher numbers of genes detected per cell, while also capturing sequence variation (SNVs, UTRs, and alternative splicing) within the transcriptome (Picelli *et al.*, 2014). Hence such assays are most often applied in experiments studying rare cell types and/or genes due to insufficient coverage provided from high-throughput strategies (Figure 1.9).

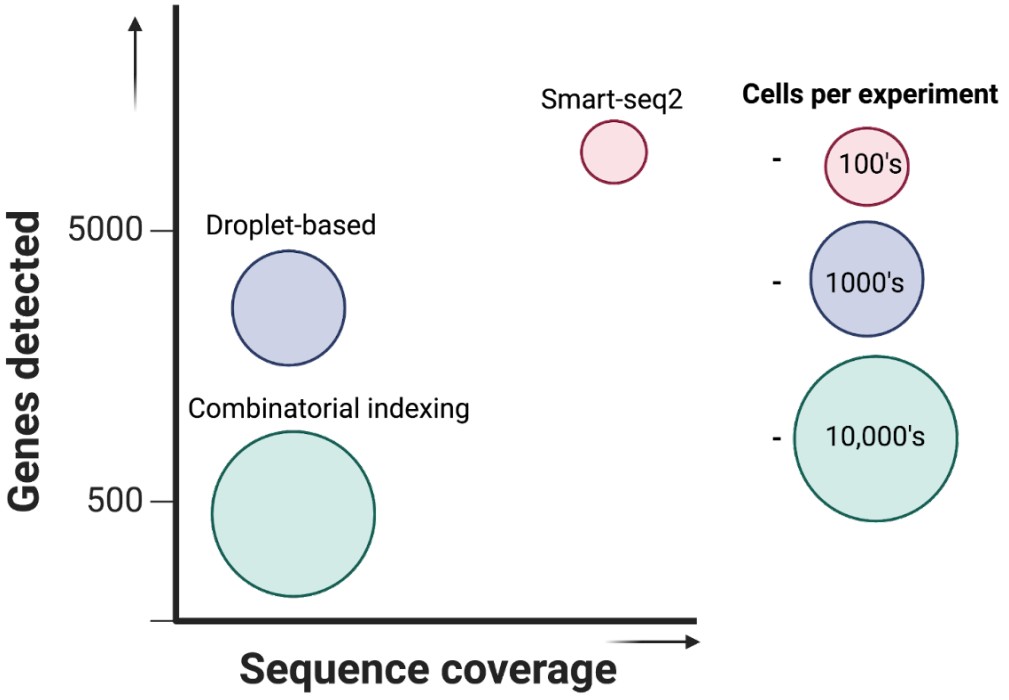


Figure 1.9. Single-cell technologies and their coverage and number of genes detected. [8]

[8] Created with BioRender.com

Isoform resolved single cell sequencing

scRNA-Seq is now well established and has been successful in identifying cell types and novel cell trajectories, revealing gene expression differences across cell types. Up to now, single-cell transcriptomic analyses have largely focused on gene-level expression, where gene expression represents the aggregation of isoforms originating from the same single gene. Short read (SR) scRNA-Seq studies using full-length methods have uncovered significant cell-to-cell variation in isoform expression (Shalek *et al.*, 2013; Marinov *et al.*, 2014; Yap and Makeyev, 2016; Song *et al.*, 2017). However, due to SR constraints these studies largely focused on changes in exon usage and/or splice junctions, leaving the true complexity of isoform expression in and between single cells unresolved.

Coupling long-reads (LRs) with single cell sequencing is able to provide the missing isoform information and has the potential to once again revolutionise transcriptomics (Figure 1.10). Full-length isoform sequencing enables the characterisation of all aspects of isoforms including exon-skipping, alternative 3' and 5' splice sites, intron retention and alternative transcription start and end sites (Wen, Mead and Thongjuea, 2020). LR RNA-sequencing is able to leverage on the full-length cDNA generated from library preparations through direct sequencing of transcripts without prior fragmentation. Examples of studies combining LRs with single-cell technology were all performed on less than ten cells using either Oxford Nanopore Technology (ONT) (Byrne *et al.*, 2019) or PacBio sequencing (Macaulay *et al.*, 2015; Karlsson and Linnarsson, 2017), and were the first to demonstrate the potential of applying LR sequencing into single cell studies.

Later demonstrations of LR scRNA-Seq across a more substantial number of cells utilised the PacBio ScISOr-Seq method, complemented by SR scRNA-Seq (Gupta *et al.*, 2018). Isoforms from over one thousand single cells were sequenced, but a small median number of reads (270) and genes (129) per cell were captured (Gupta *et al.*, 2018). More recently, deeper per-cell profiling identified differential isoform expression in 395 genes across cell types, including 76 high confidence novel isoforms (Joglekar *et al.*, 2021). They demonstrated differential isoform expression due to a single cell type changing its isoform expression pattern, providing critical insight into the relative importance of cell-types, and cell composition in defining splicing patterns (Joglekar *et al.*, 2021).

These studies highlighted an important trade-off between per-cell depth (important to enable isoform comparisons) and number of cells sequenced; where sequencing large cell quantities with current LR scRNA-Seq technologies either results in low per-cell read depth or high experimental cost. Attractive solutions to this limitation were developed which combine the

power of both SR and LR sequencing to counteract each technologies' inherent limitations. Mincarelli *et al.* leveraged the 10X Genomics cell barcode to integrate SR and LR data to single cells sequenced across both (Mincarelli *et al.*, 2023), whilst another strategy instead implemented a sub-sampling strategy where only ~10–20% of the cells from a 10x Genomics experiment are sub-sampled for ONT sequencing (Tian *et al.*, 2021). This approach uses SRs to identify the cell barcodes and UMIs as well as providing a broader view of the cell-types present, while LR scRNA-Seq enabled insights into isoform usage at the single cell level (Tian *et al.*, 2021; Mincarelli *et al.*, 2023).

The combination of single cell and LR technologies is still in its infancy, and with the developments in throughput and/or accuracy of LR, subsampling and/or matched SR scRNA-Seq may eventually become unnecessary. So far, experiments into coupling LRs with single cell sequencing have only reinforced the relevance and importance to the study of isoform heterogeneity and AS patterns. In the same way bulk-level gene expression was found to be often insufficient to appreciate gene expression heterogeneity, it has become evident that bulk isoform-sequencing does not delineate the heterogeneity across AS profiles. Therefore the technological development of LR scRNA-seq approaches is a major focus of current research to provide novel insights into the splicing landscape of cells and tissues.

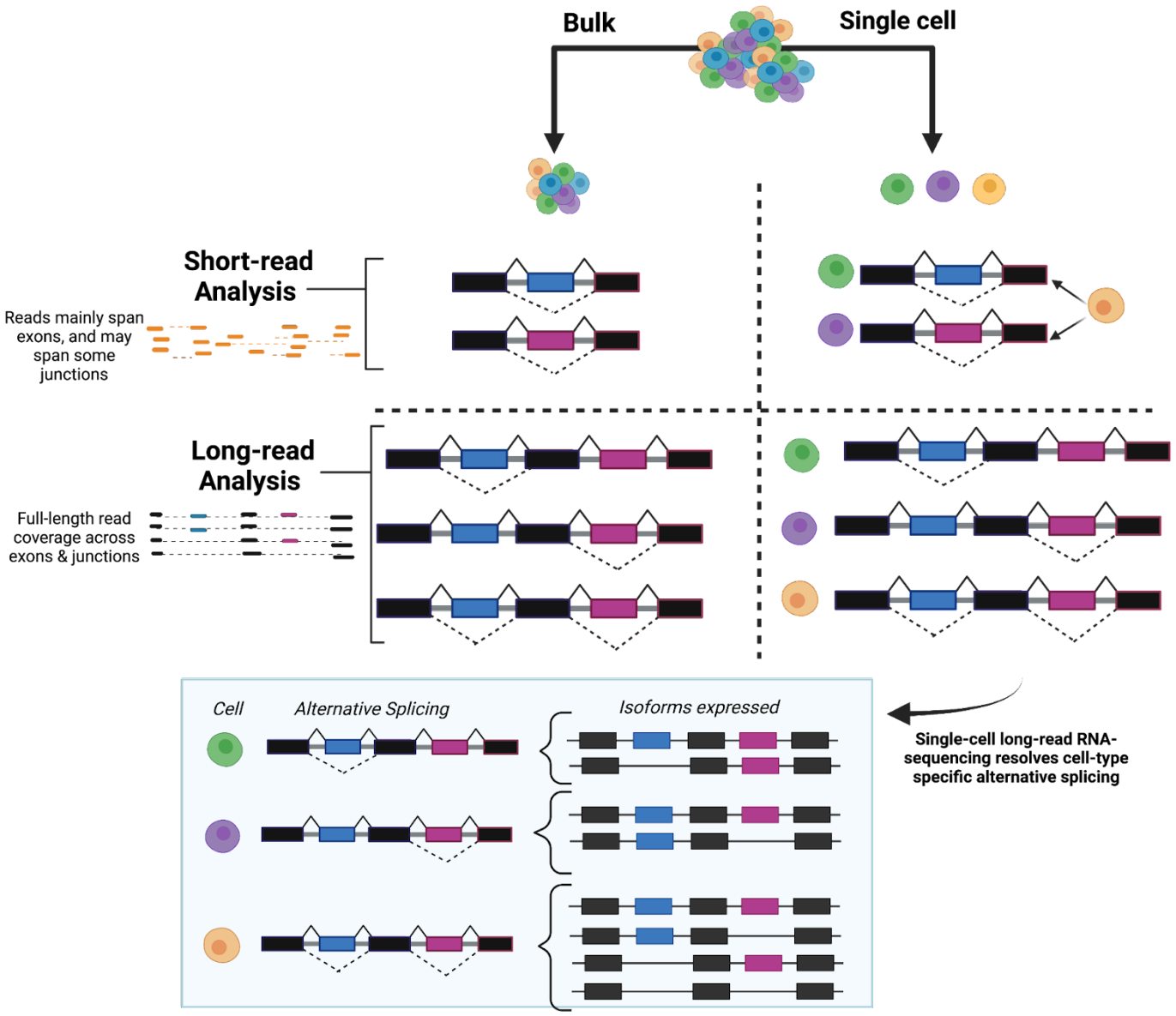


Figure 1.10. Unique insights gained through single-cell alternative splicing analysis using short- and long-read RNA-sequencing. Top-left: Bulk short-read RNA-sequencing is unable to resolve cell of origin for alternative splicing event. Top-right: Single-cell short-read RNA-sequencing is able to distinguish cells of origin for each alternative splicing event. With the exception of cells with coordinated alternative splicing events (yellow) where it would be inferred that there are two isolated alternative splicing events. Bottom-left: Bulk long-read RNA-sequencing is able to distinguish isolated and coordinated alternative splicing events but is unable to assign the events to the cell of origin. Bottom-right: Single-cell long-read RNA-sequencing is able to distinguish isolated and coordinated alternative splicing events as well as assign the events to the cell of origin (adapted from Wen, Mead and Thongjuea, 2020).

The aim of this project was to characterise the transcriptional landscape of the Mk lineage at single-cell resolution. Mks primarily serve for the lifelong continuous production of platelets (thrombopoiesis), with additional roles beyond platelet production across other physiological processes including HSC quiescence, inflammation, immunity, and bone metastasis. Contrary to initial theories, haematopoietic stem cells and downstream haematopoietic progenitors exhibit differential propensities towards distinct lineages. Specifically, the classic route of Mk differentiation from HSCs via increasingly lineage-restricted intermediate progenitors has been challenged with the identification of direct commitment from multipotent HSCs as the first lineage bifurcation. Moreover, in non-steady state haematopoietic conditions such as inflammatory stress or acute platelet depletion (thrombocytopenia), novel stem-like progenitors have been identified within the stem cell compartment exclusively restricted to the Mk lineage suggesting alternate pathways of HSC-Mk commitment exist revealing important gaps in our understanding of Mk commitment. Dysfunction in Mk differentiation including both excessive proliferation or deficient generation of megakaryocytes has long-established implications in the development of haematological disorders. Thus, the specific emphasis of this thesis was to provide a comprehensive landscape of the transcriptional signatures across cells of the Mk lineage at high resolution.

Characterisation of cells along the Mk lineage was achieved by employing scRNA-seq with Smart-seq2 to FACS-sorted cells. Cells were isolated with a broad gating strategy based on high Cd150 expression; shown to be expressed across HSCs through to committed Mk progenitor cells. The hypothesis behind isolating haematopoietic stem and progenitors with this approach was that this would enable the capture of cells at different stages of commitment to the Mk lineage, including intermediate cell states that may have been missed in the existing literature that instead employed cell-type specific canonical gating strategies based on individual cell-type marker expression. This thesis outlines Mk commitment trajectories under steady-state haematopoiesis and the transcriptomic signatures that arise under haematopoietic stressors including ageing and acute thrombocytopenia.

Bioinformatic analysis of single-cell transcriptomic profiles produced a comprehensive roadmap of cell types along the Mk lineage, including Mk progenitor sub-populations with differential signatures with age and activated upon stress. Pseudotemporal ordering of cells enabled the delineation of HSC-Mk commitment trajectories, identifying genes whose pattern of expression correlates with Mk commitment. This includes known Mk-associated genes affirming their role in Mk fate commitment and differentially expressed genes (DEGs) that have not been implicated previously in Mk lineage specification.

Finally, this project sought to explore the heterogeneity in isoform expression in genes with important roles in Mk function and commitment. Alternative splicing leads to the expression of multiple isoforms from individual genes that often play important and in some cases even opposite roles in cell function. The inherent technical challenges involved in the study of single cells at the isoform level have meant that historically this has been overlooked when studying lineage fate decisions in haematopoiesis. Technological development of long-read sequencing from low-input samples and novel strategies to increase the throughput of single-cell long-read data, this project sought to also evaluate isoform expression in genes key for Mk function. Haematopoietic stem cells characterised based on their short-read transcriptomic signatures were profiled with long-read sequencing to compare the additional power long-read sequencing provides when seeking to evaluate isoform expression. Strategies for concatenation of 10X scRNA-seq cDNA were tested and applied to haematopoietic cells demonstrating the feasibility and advantages of combining gene and isoform level measurements from single cells to heterogeneity within the haematopoietic compartment

Chapter 2:

Materials & Methods

Preface: This chapter describes the methods implemented to generate data for Chapters 3 to 5. Relevant sections from these methods are referenced throughout the experimental chapters along with the specific experimental approach used. All experiments were performed by Anita Scoones unless otherwise stated in each respective section.

2.1 Materials

2.1.1 Equipment

BD FACS Aria Fusion-	BD Biosciences
Beckman Coulter Biomek FXP-	Beckman Coulter
Agilent 2100 Bioanalyzer system-	Agilent Technologies Inc
Centrifuge 5430 R-	Eppendorf
Countess 3 Automatic Cell Counter-	Thermo Fisher Scientific Inc.
BD FACS Melody-	BD Biosciences
EVOS XL Core Imaging System-	Thermo Fisher Scientific Inc.
Mosquito Liquid Handling Robot-	SPT Labtech
Qubit Fluorometer-	Thermo Fisher Scientific Inc.
Bio-Rad C1000 Touch Thermal cycler-	Bio-Rad Laboratories Inc
Digital general purpose water bath-	VWR International, LLC
Femto Pulse system-	Agilent Technologies Inc
10X Genomics Magnetic Separator-	10X Genomics
ThermoMixer C-	Eppendorf
10X GenomicsVortex Adapter-	10X Genomics
10X Genomics Chromium controller-	10X Genomics
EasySep™Magnet-	Stem Cell Technologies

2.1.2 Buffers and Solutions

No	Name	Reagents
Tissue culture		
1	Platelet counting media -	DPBS, 2mM EDTA, 0.1% BSA, 0.02 U/ml Apyrase, 0.0001M PFI2
2	Dissection media -	DPBS 1X, 5% FCS, 2mM EDTA
3	FACS sorting media -	DPBS 1X, 4% FCS
Smart-seq2* *Volumes for 100x samples		
4	Smartseq2 lysis buffer -	1 µl of SUPERase RNase Inhibitor (Invitrogen) 19 µl of 0.2% (vol/vol) Triton X-100 (Sigma).
5	Smartseq2 annealing mix -	10 µl of 100 µM oligo-dT (see oligonucleotide sequence section 2.1.4: 1) 100 µl of 10 mM dNTP mix (ThermoFisher) 80 µl of nuclease-free water
6	Smartseq2 reverse transcription mix -	50 µl of Superscript II RT (Invitrogen) 200 µl of 5X Superscript II First Strand Buffer (Invitrogen) 50 µl of 100 mM DTT (Invitrogen) 25 µl of SUPERase RNAse Inhibitor 200 µl of 5M Betaine (Sigma) 6 µl of 1 M MgCl2 (Ambion) 10 µl of 100 µM TSO (see oligonucleotide sequence section 2.1.4: 2) 29 µl of nuclease-free water
7	Smartseq2 pre-amplification mix -	1250 µl of 2X KAPA HiFi HotStart Ready Mix (KAPA Biosystems) 25 µl of 10 µM IS PCR primer (see oligonucleotide sequence section 2.1.4: 3) 225 µl of nuclease-free water
10X Genomics* *Volumes for 1x sample		
8	LT 10X GEM mix - (HT mix volumes)	18.8 µl (41.25 µl) RT Reagent B (10X PN: 2000165, PN: 2000435) 2.4 µl (5.15 µl) Template Switch Oligo (10X PN: 3000228) 2 µl (4.25 µl) Reducing Agent B (10X PN: 2000087) 8.7 µl (12.95 µl) RT Enzyme C (10X PN: 2000102, PN: 2000436)
9	10X Elution solution -	98 µl EB Buffer (Qiagen) 1 µl 10% Tween 20 (Bio Rad) 1 µl Reducing Agent B (10X PN: 2000087)
10	10X cDNA amplification mix -	50 µl Amplification Mix (10X PN: 2000103, PN: 2000440) 15 µl cDNA primers (10X PN: 2000089)

11	10X Fragmentation, end repair and A-tailing mix -	5 µl Fragmentation buffer (10X PN: 2000091) 10 µl Fragmentation enzyme (10X PN: 2000104, PN: 2000090) 25 µl of EB buffer (Qiagen)
12	10X ligation mix -	20 µl 10X ligation buffer (10X PN: 2000092) 10 µl DNA ligase (10X PN: 220131, PN: 220110) 20 µl Adaptor oligos (10X PN: 2000094)
	PacBio Iso-Seq* * Volumes for 1x sample	
13	Repair and A-tailing mix -	8 µl Repair buffer (PacBio 102-182-700) 4 µl End repair mix (PacBio 102-182-700) 2 µl DNA repair mix (PacBio 102-182-700)
14	Ligation mix -	30 µl Ligation mix (PacBio 102-182-700) 1 µl Ligation enhancer (PacBio 102-182-700)
15	Nuclease mix -	5 µl Nuclease buffer (PacBio 102-182-700) 5 µl Nuclease mix(PacBio 102-182-700)
	PacBio MAS-seq* *Volumes for 1x sample	
16	TSO PCR mix -	25 µl MAS PCR mix (2X) (PacBio 102-692-800) 5 µl MAS capture primer Fwd (PacBio 102-693-300) 5 µl MAS capture primer Rev (PacBio 102-693-900)
17	MAS PCR primer mix -	125 µl Nuclease-free water 212.5 µl MAS PCR mix (2X) (PacBio 102-692-800)
18	MAS primer digestion mix -	1.5 µl MAS adapter A Fwd (PacBio 102-695-800) 1.5 µl MAS adapter Q Rev (PacBio 102-695-900) 20 µl MAS ligation additive (PacBio 102-696-400)
19	MAS ligase mix -	10 µl MAS Ligase buffer (PacBio 102-693-100) 10 µl Ligase (PacBio 102-693-000)
20	DNA damage repair mix -	6 µl Repair buffer (PacBio 102-696-100) 2 µl DNA repair mix (PacBio 102-696-000)
21	Nuclease treatment mix -	5 µl Nuclease buffer (PacBio 102-696-300) 5 µl Nuclease mix (PacBio 102-696-200)

2.1.3 Kits & General Reagents

No	Kit name	Product number	Source
1	SPRIselect Reagent Kit	B23318	Beckman Coulter, Inc
2	Buffer EB	19086	Qiagen
3	Glycerin (glycerol), 50% (v/v) Aqueous Solution	3290-32	Ricca Chemical Company
4	Low TE Buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA)	12090-015	Thermo Fisher Scientific
5	Nuclease-free Water	AM9937	Thermo Fisher Scientific
6	KAPA Library Quantification Kit for Illumina Platforms	KK4824	KAPA Biosystems
7	High Sensitivity DNA Kit	5067-4626	Agilent
8	Qubit 1x dsDNA HS Assay Kit	Q32854	Thermo Fisher Scientific
9	SMRTbell cleanup beads	PacBio 102-158-300	Pacific Biosciences
10	MAS-Seq for 10x 3’ concatenation kit	PacBio 102-407-900	Pacific Biosciences
11	Chromium Next GEM Single Cell 3′ LT Kit v3.1	PN-1000325	10X Genomics
12	Chromium Next GEM Single Cell 3' HT Kit v3.1	PN-1000370	10X Genomics
13	Dual Index Kit TT Set A	PN-1000215	10X Genomics
14	SMRTbell prep kit 3.0	PacBio 102-182-700	Pacific Biosciences
15	PacBio Elution Buffer	PacBio 101-633-500	Pacific Biosciences
16	EasySep Mouse Haematopoietic Progenitor Cell	19856	Stem Cell Technologies
17	Trypan Blue Solution, 0.4%	15250061	Thermo Fisher Scientific
18	Femto Pulse gDNA 165 kb analysis kit	FP-1002-0275	Agilent
19	AMPure XP beads	A63882	Beckman Coulter, Inc

2.1.4 Oligonucleotide sequences

No	Oligo name	Sequence (5'-3')	Source
1	Oligo-dT30V *	AAGCAGTGGTATCAACGCAGAGTAC(T30)VN	IDT
2	TSO oligo (LNA)	AAGCAGTGGTATCAACGCAGAGTACATrGrG	Qiagen
3	IS PCR *	AAGCAGTGGTATCAACGCAGAGT	IDT
4	10X cDNA primer (Fwd)	CTACACGACGCTCTTCCGATCT	10X Genomics
5	10X cDNA primer (Rev)	AAGCAGTGGTATCAACGCAGAG	10X Genomics
6	10X Adapter Oligos	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC TCTAGCCTTCTCG	10X Genomics
7	Illumina P5 (Dual Index Plate TT set A)	AATGATACGGCGACCACCGAGATCTACAC	10X Genomics
8	Illumina P7 (Dual Index Plate TT set A)	AAGCAGAAGACGGCATACGAGAT	10X Genomics

Smart-seq2* (Picelli *et al.*, 2014)

2.2 Methods

2.2.1 Sample collection

Mice

C57BL/6 female mice were obtained from the University of East Anglia under a United Kingdom Home Office project licence held by Dr. Stuart Rushworth. Mice were bred and maintained in individually ventilated cages in conditions complying with the *Code of Practice for the Housing and Care of Animals Bred, Supplied or Used for Scientific Purposes*. All experiments as part of this thesis were performed in accordance with the regulations set by the United Kingdom Home Office and the *Animal Scientific Procedures Act* (1986), where euthanasia was performed in compliance with *Schedule 1* of the act.

2.2.1.1 Platelet depletion experiment

A total of 6 mice were used for experiments presented in Chapter 3. Mice were subject to platelet depletion through intravenous injection of anti-GPIb antibody (Emfret Analytics #R300). Dr Jayna Mistry performed tail vein injections under the guidance of Dr Stuart Rushworth. Antibody-induced thrombocytopenia was performed in four mice by injections of anti-GPIb suspended in 200 µl sterile PBS at 2 µg/g average body weight (platelet depleted samples). Mice's body weight ranged between 18-24 g ie. body weight average of 22 g was used for injections. Two mice were injected with an equal dose of IgG control, also suspended in 200 µl sterile PBS total volume (control samples). Mice were housed in separate cages and 24 hours post-injection whole-blood samples were collected prior to euthanization by cardiac puncture to confirm platelet depletion by FACS analysis.

To prepare blood samples for platelet quantification, whole blood in Microvette EDTA tubes (Sarstedt 16.444) was centrifuged very briefly to collect peripheral blood at the bottom of the tubes. 6 µl of whole blood from control samples was mixed with 114 µl of platelet counting mix (x20 dilution) (see section 2.1.3. Buffers and solutions: 1), and 5 µl of platelet depleted samples was mixed with 5 µl SPHERO beads and 300 µl of PBS. Next 2 µl of CD41-PC7 antibody was added to all samples, followed by brief vortexing and then incubation at 4°C for 15 minutes. After incubation, 300 µl of platelet counting mix was added to samples, pipette mixed, and 5 µl was transferred into new tubes with SPHERO beads at a 1:1 ratio (5 µl). Platelet and bead quantification for each sample was performed on the FACS Melody by gating with size (forward scatter, FSC) granularity (side scatter, SSC) and CD41 expression - ensuring both platelets and beads were visible but distinct on FSC and SSC plots. This was performed in

triplicates per sample, along with unstained controls and CD41-PC7 stained beads to ensure proper voltages for PC7. Data acquisition was performed on a total of 500,000 of events/cells.

2.2.1.2 Ageing experiment

A total of 6 mice were used for experiments presented in Chapters 4 and 5. Three 8-10 weeks of age (young) and three of 72 weeks of age (aged) mice, preserved in accordance with the regulations set by the *United Kingdom Home Office and the Animal Scientific Procedures Act* (1986).

Human PBMCs

Sample preparation of human PBMCs was performed by Charlotte Utting and Lydia Pouncey. Two vials of frozen human peripheral blood mononuclear cells (PBMCs) were obtained from Stem Cell Technologies (cat: #70025) and stored in -80°C until used. Cells were isolated from peripheral blood (PB) leukapheresis samples using density gradient separation and/or red blood cell lysis using Institutional Review Board (IRB)-approved consent forms and protocols. Samples were thawed in a water bath at 37°C with gentle shaking and used immediately. The total volume per vial was measured by aspirating all of the contents of each vial. To determine total cell counts per sample, 20 µl of each sample was mixed at 1:1 ratio with Trypan blue and cell counts obtained using the Countess 3 automated cell counter. Cell counting was performed three times per sample to obtain an average concentration from three readings. After cell counting, the remaining volume per vial was transferred to a 50 mL falcon tube. To capture any cells left in the vial, each vial was rinsed with an equal volume of PBS, gently swirled, and added to the respective falcon for each sample. To wash cells, each falcon was supplemented with PBS media until 20 mL and then pelleted in a centrifuge at 300g for 10 minutes at room temperature. After centrifugation the supernatant was discarded and each pellet was gently resuspended again in 20 mL of PBS for a second wash. Using the same conditions, cells were pelleted one final time in a centrifuge. After this, most of the supernatant was discarded only this time leaving behind approximately 2 mL of media, and cells were gently resuspended by flicking each tube. To obtain a final count of each cell suspension after wash steps, cells were re-counted using Trypan blue on the Countess 2 instrument. Cells were also studied under a microscope at x20 magnification to visually inspect the quality of cells.

2.2.2 Mouse bone marrow dissection

Bone marrow was isolated from mouse spine and hind legs (femora, tibiae, tibiofemoral joints and ilia) following schedule one. Clean bones were crushed using sterile pestles and mortars containing dissection medium (see section 2.1.3. Buffers and solutions: 2). Cell suspensions obtained from each mouse were filtered through 70 µM filters into 50 ml polypropylene falcon tubes (Eppendorf) and then pelleted by centrifugation at 300 g for 5 minutes at 4°C.

2.2.3 Red blood cell depletion of bone-marrow samples

After discarding the supernatant, pellets containing bone-marrow cells were resuspended in 3 ml of dissection medium, and 4 ml of ammonium chloride solution (NH_4Cl) was added for red blood cell depletion. The influx of NH_4Cl into erythrocytes causes cellular swelling and rupture enabling white blood cells to be isolated from whole blood marrow. Samples were incubated on ice for 10 minutes and the centrifugation step was repeated, where after removing the supernatant containing red blood cells, pellets were resuspended in 5 ml of PBS.

2.2.4 Enrichment of haematopoietic stem and progenitor cells from bone marrow

To quantify cell suspension concentrations and determine cell viability after red blood cell depletion, samples were diluted, stained with Trypan blue and manually counted with a cell counting chamber. To deplete lineage-positive cells, first cells were incubated with Mouse Haematopoietic Progenitor Cell Isolation Cocktail and Rat Serum for 15 mins at 4°C in the dark, then mixed with streptavidin-coated magnetic particles (RapidSpheres™) and incubated at room temperature for a further 10 mins (EasySep™ Mouse Haematopoietic Progenitor Cell Isolation Kit). Samples were supplemented with PBS to the recommended volume (2.5 ml) and incubated at room temperature for 3 mins on a magnet for column-free immunomagnetic separation (EasySep™ Magnet). In this step, unwanted cells are targeted for removal with biotinylated antibodies directed against non-haematopoietic stem cells and non-progenitor cells (Table 2.1) and streptavidin-coated magnetic particles. The depleted cell suspensions were poured into new tubes, the above step was repeated once more to capture any remaining desired cells from the original suspension, combining the first and second fractions into a single cell suspension tube per sample.

2.2.5 Cell staining for fluorescence-activated cell sorting of bone-marrow samples

After lineage depletion, cells were counted as described above and then resuspended in an equimolar antibody cocktail containing primary antibodies for 30 mins at 4°C in the dark (Table 2.2). After incubation, cells were centrifuged to remove excess antibodies and resuspended into approximately 1 ml cell suspension of cell sorting media (see section 2.1.3 Buffers and solutions: 3)

Table 2.1. Lineage cocktail panel

Cell surface Marker	Clone	Manufacturer	Catalogue No
CD4	REA604	Miltenyi Biotec	130-118-692
CD8a	53-6.7	BioLegend	100711
CD5	REA421	Miltenyi Biotec	130-106-205
Gr-1	REA810	Miltenyi Biotec	130-112-302
CD11b (Mac-1)	REA592	Miltenyi Biotec	130-113-810
CD45R (B220)	REA755	Miltenyi Biotec	130-110-851
Ter119	REA847	Miltenyi Biotec	130-112-914

Table 2.2. Antibody panel used for FACS isolation of HSC and early Mk progenitors.

Antigen	Fluorochrome	Clone	Manufacturer	Catalogue No
CD48	Pe-Cy5	HM48-1	Biolegend	103405
CD117 (cKit)	PE-vio770	REA791	Miltenyi Biotec	130-111-695
Sca1	BV786	D7	BD Biosciences	563991
CD150	BV510	TC15-12F12.2	BioLegend	115929
Flt3	APC	A2F10	BioLegend	135309
CD16/32	APC-Cy7	93	BioLegend	101325
CD105	PE	MJ7/18	BioLegend	120407
CD41	FITC	MWReg30	BioLegend	133903
Lineage cocktail (see Table 2.1)	VioBlue	NA	NA	NA

2.2.6 FACS gating strategy for isolating LK Cd150+ single-cells

2.2.6.1 Plate-based scRNA-seq experiments

To ensure a high level of purity for single-cell RNAseq flow cytometry analysis and cell sorting was performed using the BD FACSMelody cell sorter (BD Biosciences, San Jose, California) according to manufacturer's instructions. Viable LK (Lin^-, $cKit^+$) single cells were sorted by gating forward and side scatter for lineage-negative $cKit^+$, and Cd150+ expressing cells. Due to the rarity of stem cells in bone marrow, the gating was adapted for some wells to include the Sca-1 marker, where Sca-1 low (LSK Cd150+) cells were sorted to enrich for more immature HSPCs. To record the levels of the conventional surface markers that separate HSCs and progenitor subpopulations index sorting was utilised for some of the sorted cells.

2.2.6.2 10X Genomics experiments

After processing samples as described above (see sections 2.2.2 - 2.2.5), to generate a sample suspension concentrated enough for FACS sorting a sufficient number of cells for loading the 10X Genomics chip, equal concentrations of each mouse cell suspension were pooled to create a single concentrated cell suspension for cell sorting. Using the same gating strategy as for Smart-seq2 FACS sorts, viable LK (Lin^- $cKit^+$) by gating forward and side scatter for lineage-negative $cKit^+$, and Cd150+ expressing single-cells were sorted. To enrich for more immature HSPCs, after sorting 3600 LK Cd150+ cells, a new gate including the Sca-1 marker was created to sort 800 Sca-1 low (LSK Cd150+) cells into the same well. Cells were sorted into a single well of a 96-well plate containing 7 µl of FACS sorting media (see Buffers & Solutions 2.1.2, 3). Each well consisted of 4,400 cells, with approximately 80% coming from LK Cd150+ and 20% LSK Cd150+ gates. In total, two wells were sorted for this experiment ie. two technical replicates from the same cell suspension for sample loading.

2.2.7 Single-cell RNAseq

Smart-seq2

Full-length RNA-seq libraries from single cells were generated using the Smart-seq2 protocol as previously described (Picelli *et al.*, 2014). Reactions were performed in UV-treated 96-well PCR plates, and all centrifugation steps were performed at 4°C for 1 minute at 500 g unless otherwise stated.

2.2.7.1 Single-cell lysis

Cells were sorted directly into wells containing 2.2 µl Smart-seq2 lysis buffer (see 2.1.3 Buffers and solutions: 4). After sorts were completed, plates were centrifuged and immediately stored at -80°C until processed. All plates were processed within 6 months or less of sorting.

2.2.7.2 Reverse transcription and pre-amplification

Before collecting plates from long-term storage, the annealing and reverse transcription (RT) master mixes were prepared within laminar flow hoods and kept on ice until used (see 2.1.3 Buffers and solutions: 5, 6). Plates containing sorted cells were thawed on ice until fully defrosted and 2.2µl of annealing mix was added to each well. Plates were then sealed and centrifuged to ensure all liquid was collected at the base of wells and incubated for 3 minutes at 72° C. Immediately after incubation plates were placed back onto ice for approximately 2 minutes. For RT, 5.5 µl RT mix was deposited into wells, plates were securely sealed, centrifuged, and transferred into the thermocycler for RT using conditions detailed in Table 2.3. Upon completion of RT, 15 µl of the PCR pre-amplification mix was added to each well (see 2.1.3 Buffers and solutions: 7). Plates were then sealed, centrifuged, and transferred into the thermal cycler for PCR pre-amplification using conditions detailed in Table 2.4.

Table 2.3. Thermal cycling programme for Smartseq2* reverse transcription (RT).

Step	Temperature (C°)	Time (min)	Cycles
RT and template switching	42	00:90:00	1
RNA unfolding	50	00:02:00	10
Completion of RT and template switching	42	00:02:00	
Enzyme inactivation	70	00:15:00	1
Hold	4	∞	1

* (Picelli *et al.*, 2014)

Table 2.4. Thermal cycling programme for Smart-seq2* pre-amplification.

Step	Temperature (C°)	Time	Cycles
Denaturing	98	00:03:00	1
Annealing	98 67 72	00:00:20 00:00:15 00:06:00	21
Extend	72	00:05:00	1
Hold	4	∞	1

* (Picelli *et al.*, 2014)

After pre-amplification, plates containing single-cell cDNA were stored in -20°C, or immediately cleaned up using the Beckman Coulter Biomek FXP using a 1:0.8 ratio of sample to AMPure XP beads. All samples were washed twice with 80% ethanol and the final PCR-purified cDNA was eluted in 22 µl of nuclease-free water, 20 µl of which was transferred to a new 96-well plate.

To check the quality of cDNA libraries, 8-10 single-cell wells were randomly selected (as well as at least one positive and negative control well) from each plate and assayed for both concentration and size distribution on the Qubit fluorometer (Qubit 1X dsDNA HS assay kit) and Agilent 2100 Bioanalyzer system (Agilent High Sensitivity DNA Kit) respectively.

2.2.7.3 Library preparation

All Smart-seq2 libraries were prepared using the Illumina NextEra XT DNA Library Preparation Kit and the Nextera XT 96-Index Kit (384 samples) according to manufacturer's instructions, optimised to 1/12.5 of the volume of the original protocol.

Prior to NextEra, cDNA was first normalised to the recommended input concentration for library preparation. After measuring the average cDNA concentration of single-cell wells for each plate of Smart-seq2 described in the previous step, aliquots of cDNA were transferred into new 96-well plates with a multichannel pipette and diluted in nuclease-free water to generate 0.2ng/ µl dilutions for each sample.

All NextEra liquid handling steps were automated using the Mosquito LV multi-channel liquid handling robot for accurate liquid transfer at low reaction volumes. To enable this, a 384-well 'reagent plate' containing sufficient volumes of all reagents to process four 96-well plates of Smart-seq2 cDNA (+ 10% overage) were aliquoted into specific positions for liquid transfer into sample wells. Robot tips were changed after every liquid handling step to prevent

cross-contamination of samples. First, 400nL of each sample was transferred from four 96-well plates containing Smart-seq2 cDNA normalised to 0.2ng/ µl into wells of a 384-well plate for Nextera reactions. To this, 1.2 µl of Tagmentation mix (see table 2.5) was transferred into all sample wells of the 384 reaction plate, after which plates were immediately sealed, centrifuged to collect liquid at the base of wells, and incubated at 55°C for 10 minutes in a thermal cycler for tagmentation. In this step Nextera transposome tagments cDNA, which is a process that creates short fragments and then tags the DNA with adapter sequences in one step.

Immediately proceeding tagmentation, 400 nL of NT buffer was added to all samples, mixed, and incubated at room temperature for 5 minutes. This neutralises the reaction to prevent over tagmentation. After 5 minutes, 1.2 µl of PCR master mix along with 800 nL of pre-mixed i7 and i5 index primers were added to each sample well across the 384-well plate. Plates were then sealed, centrifuged to collect liquid at the base of wells, and incubated on a thermal cycler with the NextEra Limited-cycle PCR programme (Table 2.6). This step uses the adapters to amplify the DNA whilst also adding index adapter sequences on both ends which enables dual-indexed sequencing of pooled libraries for Illumina sequencing.

Each 96-well plate was pooled to generate a single library per plate by combining equal volumes from each well. The pooled single-cell library underwent manual bead clean-up using a 1:0.8 ratio of sample to AMPure XP beads according to manufacturer's instructions. All samples were washed twice with fresh 80% ethanol and the libraries were eluted in 20 µl of nuclease-free water which was transferred to a new 1 ml Eppendorf tube.

Table 2.5. Reagent master mixes to process 384 samples for Nextera XT library preparation.

NextEra master mix	Reagent	Volume of reagent in master mix (384X samples) (µl)	Volume of mix per 1X sample (µl)
Tagmentation mix	Amplicon tagment mix	59.5	1.2
	Tagment DNA buffer	110.5	
NT buffer	Neutralise tagment buffer	17	0.4
PCR master mix	Nextera PCR mix	23	1.2
Sample index adapters (unique for each sample)	Pre-paired i7 and i5 index adapters	NA	0.8

Table 2.6. Nextera PCR amplification thermal cycling programme.

Step	Temperature (C°)	Time	Cycles
Denaturing	72 98	00:03:00 00:00:30	1
Annealing	95 55 72	00:00:10 00:00:30 00:01:00	12
Extend	72	00:05:00	1
Hold	4	∞	1

2.2.7.4 Library quality control and normalisation

To ensure the pooled libraries (pooled 96-well plates) would be sequenced at equal concentrations to generate an even read distribution for all samples, libraries underwent manual normalisation to the same concentration before volumetric pooling of multiple plates. First, the quality of cDNA libraries was assessed by determination of final fragment size distribution and concentration on an Agilent HS DNA Bioanalyzer and Qubit fluorometer respectively. This step also provided visibility into possible library issues, such as adapter dimers or unexpected library sizes. Libraries within 300-600 bp that met the minimum concentration requirements for sequencing were selected for final pooling.

The quantification of plate pools to enable equimolar pooling of multiple plates for sequencing was performed using quantitative polymerase chain reaction (qPCR) using the Kapa Library Quantification Kit according to manufacturer's instructions. In comparison to Qubit readings, qPCR quantification results in higher accuracy quantification because it only detects molecules with complete sequencing adapters at both ends ie. the only fragments that will successfully generate sequence reads.

qPCR was performed at two dilutions per sample (1/100 and 1/10,000) in triplicates. Six standards of known concentration, also added in triplicates, were used for all assays to generate a standard curve for accurate library quantification. Calculations to obtain final total nanomolar (nM) readings for each plate pool were performed, each pool was diluted to 4 nM in nuclease-free water, and pooled at equal volumes to generate final sequencing-ready libraries, each containing four pooled plates of 96 wells totalling 384 wells per sequencing library.

10X Genomics

This section describes the protocols implemented for 10X Genomics Single Cell 3' library preparation for two sets of experiments, both presented in Chapter 6. The methodology for both experiments follow the same workflow and are outlined below, where details that differ between the two experiments are mentioned within the text.

2.2.7.5 GEM incubation and cDNA generation

Mouse samples (see Materials and Methods 2.2.6.2) were processed using the 10X Genomics Single Cell 3' LT v3.1 kit, and PBMC samples (see Materials and Methods 2.2.1: Human PBMCs) were processed using the 10X Genomics Single Cell 3' HT v3.1 kit according to manufacturer's instructions. To prepare samples for loading at the optimal cell concentration for each experiment, mouse LK Cd150+ cells were FACS sorted directly into the total volume of PBS to be loaded, and PBMCs were prepared by diluting the cell suspension after final cell counting in PBS. ~1000 mouse single-cells for the LT and ~9000 PBMC cells for the HT runs were the target number of cells to capture for each experiment. A sample mixture was generated by combining 10X GEM mix (see mix 2.1.3 Buffers and solutions: 8) and nuclease-free water with the volume of cell suspension containing ~4000 and ~15,000 cells respectively. The 10X Chromium NextGEM chip ('L' 10X PN: 2000414 for LT experiment, 'M' 10X PN: 2000417 for HT experiment) was carefully loaded with the sample mixture, and all unused wells of the chip loaded with 50% glycerol solution. To complete chip assembly, 10X partitioning oil and barcoded gel beads were pipetted into chip wells according to manufacturer's instructions, and the chip loaded onto the 10X Chromium Controller for GEM generation. Upon completion of the GEM generation, the GEM suspension was carefully transferred into a tube strip placed on ice, and then immediately incubated in a thermal cycler with the thermal cycling programme for GEM reverse transcription (RT) (Table 2.7). In this step, the gel bead is dissolved, primers are released, and the co-partitioned cell is lysed to enable the cell lysate and RT reagents to mix and produce barcoded, full-length cDNA from poly-adenylated mRNA.

Table 2.7. Thermal cycling programme for 10X GEM reverse transcription incubation.

Step	Temperature	Time
1	53°C	00:45:00
2	85°C	00:05:00
3	4°C	Hold

2.2.7.6 Post GEM-RT cleanup and cDNA amplification

After GEM RT incubation, 125 µl 10X recovery agent was added to the samples and incubated at room temperature for 2 minutes until biphasic separation. The separated recovery agent was carefully removed, and the remaining aqueous phase containing first-strand cDNA from the post-GEM-RT mixture, which includes leftover reagents and primers, was bead purified using 10X Dynabeads MyOne SILANE (10X PN: 2000048) and 10X magnetic separator according to manufacturer's instructions. Samples were washed twice with 80% ethanol and purified products eluted in 35 µl 10X Elution Solution (see 2.1.3 Buffers and solutions: 9). For amplification of cDNA, samples were mixed with 65 µl of 10X cDNA amplification mix (see 2.1.3 Buffers and solutions: 10), pipette mixed, and incubated in a thermal cycler with the cDNA amplification programme according to manufacturer's instructions (Table 2.8). This generates barcoded, full-length cDNA at sufficient mass for library construction. The amplified cDNA samples were then cleaned up using SPRIselect beads at 0.6:1 SPRIselect reagent to sample ratio, to remove excess amplification reagents. Samples were washed twice with 80% ethanol and eluted in 40 µl of EB buffer (Qiagen). To assess the quality of the final cDNA, the concentration was determined using a Qubit fluorometer and size distribution was determined on a Bioanalyzer High Sensitivity chip.

Table 2.8. Thermal cycling programme for 10X cDNA amplification.

Step	Temperature	Time	Cycles
Denaturing	98°C	00:03:00	1
Annealing	98°C 63°C 72°C	00:00:15 00:00:20 00:01:00	12
Extension	72°C	00:01:00	1
Hold	4°C	∞	1

2.2.7.7 cDNA Fragmentation, End Repair and A-tailing

After ensuring successful cDNA generation, 10 µl of cDNA generated with the LT experiment and 20 µl of cDNA generated with the HT experiment was used to generate sequencing ready libraries. For this, enzymatic fragmentation and size selection are performed to optimise the cDNA amplicon size, where TruSeq Read 2 (read 2 primer sequence) is added via End Repair, A-tailing. First, cDNA was transferred to a new tube strip and mixed with 10X fragmentation end repair and A-tailing mix (see 2.1.3 Buffers and solutions: 11). This was then incubated on a pre-cooled thermal cycler on the fragmentation, end repair and A-tailing programme (Table 2.9). After this, samples were bead purified for double-sided size selection using SPRIselect beads, where samples were mixed at 0.6:1 bead-to-sample ratio, then after 5 minutes and

magnetic separation, the supernatant was transferred to a new tube strip (leaving behind larger fragments) and mixed with SPRIselect reagent at 0.8:1 bead to sample ratio. Samples were washed twice with 80% ethanol and fragments on beads were eluted in 50 µl of EB buffer (while smaller fragments are discarded).

Table 2.9. Thermal cycling programme for 10X fragmentation, end repair and A-tailing.

Step	Temperature	Time
Pre-cool block	4°C	∞
Fragmentation	32°C	00:05:00
End Repair and A-tailing	65°C	00:30:00
Hold	4°C	∞

2.2.7.8 Adaptor ligation and sample index PCR

To prepare final libraries, Illumina P5, P7, i7 and i5 sample indexes are ligated to amplicons through PCR. For this, 50 µl sample post-double-sided size selection was mixed with 50 µl 10X ligation mix (see 2.1.3 Buffers and solutions: 11) and incubated in a thermal cycler for 15 minutes at 20°C. Post ligation, samples were purified using SPRIselect magnetic beads at 0.8:1 bead-to-sample ratio. Samples were washed twice with 80% ethanol and eluted in 30 µl of EB buffer. To add sample index sets, 50 µl of 10X amplification solution (10X PN: 2000103) was added to the sample along with 20 µl of an individual Dual Index TT Set A (10X PN: 3000431) and incubated on a thermal cycler with the sample index PCR programme (Table 2.10). After sample indexing, PCR libraries again underwent double-sided size selection, first at 0.6:1 then 0.8:1 bead-to-sample ratio to remove large and smaller fragments respectively. Samples were washed twice with 80% ethanol and final sequencing-ready libraries eluted in 35µl EB buffer. Libraries were stored at -20°C until sequencing.

Table 2.10. Thermal cycling programme for 10X sample index PCR.

Step	Temperature	Time	Cycles
Denaturing	98°C	00:00:45	1
Annealing	98°C	00:00:20	16
	54°C	00:00:30	
	72°C	00:00:20	
Extension	72°C	00:01:00	1
Hold	4°C	∞	1

2.2.8 PacBio Iso-Seq library preparation

2.2.8.1 Sample selection

cDNA generated from young and aged single-cells for Smart-seq2 (see Methods sections 2.2.7.1 - 2.2.7.2) was used to obtain material for Iso-Seq library preparation. This was achieved using results from clustering short-read scRNA-seq data of cells from the Smart-seq2 experiment in Chapter 4. To do this, after clustering the cell identities (sample names) of cells classified as LT-HSCs and HSCs, annotated based on the expression of canonical markers in the short-read data, were extracted into two lists. Each list consisted of the cell identities of cells from young and old mice, along with the well and plate ID for each cell. The 96-well plates containing purified cDNA generated with Smart-seq2 were removed from -20°C storage, thawed on ice, then centrifuged briefly to collect cell volumes at the bottom of all wells. Using the two lists of cells and their well positions in each plate, two pools were created in separate 1.5 ml Eppendorfs by careful aspiration of 10 µl from each listed HSC well based on mouse age resulting in two mini-bulk pools of cDNA from single-cells from either young (34 cells) or aged mice (46 cells). The resulting two samples were measured for cDNA concentration using a Qubit fluorometer and size distribution was measured on a Bioanalyzer in order to confirm the amount of cDNA available in each pool to proceed with Iso-Seq library preparation (minimum 160 ng).

2.2.8.2 cDNA purification

Iso-seq libraries were prepared using the PacBio SMRTbell prep kit 3.0 (PacBio 102-182-700). After pooling libraries and performing initial quality control, libraries were first purified using Ampure XP beads at a 1:1 bead-to-sample ratio to remove traces of primer fragments from Smart-seq2 cDNA and also to concentrate each pool into the required input volume for SMRTbell prep. Samples were washed twice with freshly prepared 80% ethanol and eluted in 47 µl of low TE buffer. Final cDNA concentration post-purification was measured on a Qubit fluorometer and size distribution was measured on a Bioanalyzer..

2.2.8.3 Repair and A-tailing

Purified cDNA was then mixed with 14 µl Repair and A-tailing mix (see 2.1.3 Buffers and solutions: 13). The reaction of 60 µl was pipette mixed thoroughly, briefly centrifuged to collect all liquid and incubated in a thermal cycler with the Repair and A-tailing programme (Table 2.11).

Table 2.11. Repair and A-tailing thermal cycling programme.

Step	Temperature	Time	Cycles
Repair	37°C	00:30:00	1
A-tailing	65°C	00:05:00	1
Hold	4°C	∞	1

2.2.8.4 Adapter ligation

To ligate PacBio adapters which enable SMRT sequencing, 4 µl of SMRTbell adapter (non-barcoded) was added to each sample from the previous step along with 31 µl ligation mix (see 2.1.3 Buffers and solutions: 14). Each sample was pipette mixed thoroughly and centrifuged briefly to collect all liquid and then incubated on a thermal cycler at 20°C for 30 minutes. After adapter ligation, each sample was bead purified with 124 µl SMRTbell cleanup beads (1.3X) according to manufacturer's instructions. Samples were first mixed with beads, incubated at room temperature for 10 minutes then placed in a magnetic separation rack until beads separate fully from the solution and the supernatant discarded. Beads were washed twice with freshly prepared 80% ethanol, and eluted in 40 µl of elution buffer (PacBio 101-633-500).

2.2.8.5 Nuclease treatment

For nuclease treatment of the final Iso-Seq libraries, 10 µl of nuclease mix (see 2.1.3 Buffers and solutions: 15) was added to each sample, thoroughly pipette mixed and briefly centrifuged to collect all liquid. Samples were then incubated on a thermal cycler at 37°C for 15 minutes. After 15 minutes, samples were finally bead purified with 65 µl SMRTbell cleanup beads (1.3X) according to manufacturer's instructions. Samples were first mixed with beads, incubated at room temperature for 10 minutes then placed in a magnetic separation rack until beads separate fully from the solution and the supernatant discarded. Beads were washed twice with freshly prepared 80% ethanol, and final libraries were eluted in 15 µl of elution buffer (PacBio 101-633-500). Final Iso-Seq libraries were assessed by measuring the concentration and size distribution of each cDNA sample with a Qubit Fluorometer and Bioanalyzer respectively. Final SMRTbell libraries were stored at -20°C until sequencing.

2.2.9 MAS-seq library preparation

Preface: MAS-seq is a newly released library preparation approach for high-throughput long-read transcriptome sequencing by PacBio collaboratively developed with Dr Aziz Al'Khafaji et al. of the Broad Institute (Al'Khafaji *et al.*, 2021). This novel method uses intramolecular multiplexing to maximise the yield of PacBio sequencing of cDNA libraries prepared with a 10x Genomics kit through the concatenation of multiple cDNA molecules into a large library fragment.

MAS-seq library preparation of mouse samples generated by Anita Scoones using the 10X Genomics Single Cell 3' LT v3.1 kit was performed by Dr Eirini Lampraki of PacBio. MAS-seq preparation of PBMC cDNA generated using the 10X Genomics Single Cell 3' HT v3.1 kit was performed by Ashleigh Lister and Lydia Pouncey.

2.2.9.1 cDNA sample input

cDNA products made using the 10X Genomics Single Cell 3' LT v3.1 kit with FACS sorted mouse Lin- cKit+ Cd150+ cells (see 2.2.7.1 - 2.2.7.2), along with cDNA products made using the 10X Genomics Single Cell 3' HT v3.1 kit with PBMC samples (see 2.2.8.1 - 2.2.8.2) were first thawed on ice from -20°C storage. After mixing samples briefly and centrifuging to collect all the liquid all samples were evaluated using a Qubit fluorometer and Bioanalyzer to determine sample concentration and size distribution. Aliquots were taken from each sample, and normalised to 15 ng in EB buffer (Qiagen).

2.2.9.2 TSO artefact depletion

First a PCR was performed to generate biotinylated DNA fragments to enable the removal of TSO priming artefacts generated during 10X cDNA synthesis. To do this, up to 15 µl of each library was combined with 30 µl TSO PCR master mix in a tube strip tube (see 2.1.3 Buffers and solutions: 16), and the overall reaction volume was supplemented with nuclease-free water to total 50 µl. Each sample was pipette mixed and briefly centrifuged to collect all liquid. The tube strip was then incubated on a thermal cycler with the TSO PCR programme (Table 2.12). After the PCR, each cDNA sample was purified using SMRTbell cleanup beads (PacBio 102-158-300) at a 1.5:1 bead-to-sample ratio. Beads were pipette-mixed with samples thoroughly, and incubated for 10 minutes at room temperature before magnetic separation. The supernatant was removed without disturbing beads, and beads were washed twice with 80% ethanol. Purified samples were eluted in 42 µl of EB. Finally, sample concentration was measured on a Qubit fluorometer.

Next, the removal of DNA fragments containing TSO artefacts was performed using MAS capture beads. MAS capture beads (PacBio 102-428-400) were prepared by adding 10 µl of resuspended beads per sample to a tube strip placed on a magnet. Upon magnetic separation, the supernatant was removed and 40 µl of bead binding buffer was gently used to resuspend the beads. This was discarded, and repeated once more off the magnet, where the beads resuspended in 40 µl binding buffer were transferred to PCRs tubes. 40 µl of samples were combined with beads at a 1:1 ratio and mixed carefully with wide bore tips and incubated at room temperature. After 15 minutes, the tube strip was placed on a magnet and the supernatant removed. Beads were washed twice with MAS bead washing buffer, and once with nuclease-free water, before the MAS bead-DNA complex was finally resuspended in 40 µl EB buffer. To cleave the captured DNA products from the MAS beads, 2 µl MAS enzyme was added to samples, pipette mixed and briefly spun to collect liquid. Samples were then incubated on a thermal cycler for 30 minutes at 37°C for TSO artefact removal.

Samples were purified post-TSO artefact removal using SMRTbell cleanup beads at 1.5:1 beads-to-sample ratio as previously described. Beads were washed twice whilst on the magnet with 80% ethanol, and the final cDNA products were eluted in 46 µl EB buffer and transferred to a fresh tube strip. Final cDNA yield post-TSO artefact depletion was measured on a Qubit fluorometer.

Table 2.12. TSO PCR thermal cycling programme.

Step	Temperature	Time	Cycles
Denaturing	98°C	00:03:00	1
Annealing	98°C 65°C 72°C	00:00:20 00:00:30 00:04:00	5
Extension	72°C	00:05:00	1
Hold	4°C	∞	1

2.2.9.3 MAS Primer PCR

To generate DNA fragments containing orientation-specific MAS segmentation adapter sequences, 16 parallel cDNA amplification reactions with MAS primers were performed. This was achieved by first combining 45 µl purified cDNA from the previous step with 337.5 µl MAS PCR primer mix (see 2.1.3 Buffers and solutions: 17) on ice. 22.5 µl of this MAS PCR primer mix containing cDNA was arrayed across 16 tubes per sample being processed. To each of the 16 tubes 2.5 µl of MAS pre-mixed primers were added in the order shown in Table 2.13 to total final volumes of 25µl per tube. Each sample was pipette mixed and briefly spun to collect liquid, then incubated on a thermal cycler on the MAS PCR programme (Table 2.14).

Table 2.13. MAS assay PCR primer sets

Reaction	MAS primer premix set	PacBio Reagent Code
1	MAS primers premix A	102-694-000
2	MAS primers premix B	102-694-100
3	MAS primers premix C	102-694-200
4	MAS primers premix D	102-694-300
5	MAS primers premix E	102-694-400
6	MAS primers premix F	102-694-500
7	MAS primers premix G	102-694-600
8	MAS primers premix H	102-694-700
9	MAS primers premix I	102-694-800
10	MAS primers premix J	102-694-900
11	MAS primers premix K	102-695-000
12	MAS primers premix L	102-695-100
13	MAS primers premix M	102-695-300
14	MAS primers premix N	102-695-500
15	MAS primers premix O	102-695-600
16	MAS primers premix P	102-695-700

Table 2.14. MAS PCR thermal cycling programme.

Step	Temperature	Time	Cycles
Denaturing	98°C	00:03:00	1
Annealing	98°C	00:00:20	9
	68°C	00:00:30	
	72°C	00:04:00	
Extension	72°C	00:05:00	1
Hold	4°C	∞	1

2.2.9.4 Pooling and MAS array formation

The entire volume of all 16 reactions were pooled into a single 1.5ml LoBind Eppendorf tube, and purified with SMRTbell cleanup beads at 1.5:1 beads to sample ratio as previously described. Sample-bound beads were washed twice with 80% ethanol and the final purified product was eluted in 50 µl of EB buffer. The final yield of cDNA post MAS array PCR was measured on a Qubit fluorometer to ensure sufficient material for ligation.

To create single-stranded extensions to enable directional assembly of cDNA segments into a linear array PCR amplified cDNA fragments were treated with MAS enzyme. This was achieved by adding 10 µg of sample from the previous step in 47 µl EB to a 0.2 ml PCR tube with 23 µl of MAS primer digestion and 20 µl MAS ligase mix (see 2.1.3 Buffers and solutions: 18 and 19). The reaction volumes were pipette mixed, briefly spun to collect liquid, and Eppendorfs incubated for 1hr at 42°C for ligation.

After ligation reactions were bead purified again using SMRTbell cleanup beads at 1.2:1 beads to sample ratio as previously described. Sample-bound beads were washed twice with 80% ethanol, and purified ligated cDNA product was eluted in 43 µl EB buffer. MAS array yield post-ligation was measured on a Qubit fluorometer.

2.2.9.5 DNA damage repair and nuclease treatment

In a new PCR strip tube, 5 µg of MAS array in EB buffer was combined with 8 µl of DNA damage repair mix (see 2.1.3 Buffers and solutions: 20), pipette mixed and briefly spun to collect liquid. Samples were incubated at 37°C for 30 minutes. After incubation, reactions were bead purified again using SMRTbell cleanup beads at a 1.2:1 beads-to-sample ratio as previously described. Sample-bound beads were washed twice with 80% ethanol, and purified ligated cDNA product was eluted in 40 µl EB buffer. For nuclease treatment, 10 µl of nuclease treatment mix (see 2.1.3 Buffers and solutions: 21) was added to each sample, pipette mixed, and briefly spun to collect liquid. Samples were then incubated at 37°C for 1 hour on a thermal cycler.

For the final cleanup of MAS libraries, samples were purified with SMRTbell cleanup beads at a 1.2:1 beads-to-sample ratio as previously described. Samples were washed twice with 80% ethanol, and final MAS-seq products were eluted in 20 µl EB buffer. To determine final reaction yield, sample concentration and size distribution were determined using a Qubit fluorometer and Femto Pulse System respectively. SMRTbell MAS libraries were stored at -20°C until sequencing.

2.2.10 Sequencing parameters

2.2.10.1 Smart-Seq2 libraries

All Smart-Seq2 libraries were sequenced on the NovaSeq 6000 using SP flow cells with 150bp PE reads. Each Nextera library of 384 pooled single cells was submitted at approximately 4 nM, sequenced on one SP lane to generate approximately 1M reads per cell on average.

2.2.10.2 10X Genomics libraries

Mouse derived LK Cd150+ libraries from 10X Genomics Single Cell 3′ LT Kit v3.1

Libraries generated from mouse single cells with the 10X Single Cell 3′ LT kit were sequenced on 1 lane of the MiSeq v3 flow cell with 28-10-10-90 configuration to generate 22 million reads per lane for each direction sequenced (3.3 Gb sequencing data).

Human PBMC libraries from 10X Genomics Single Cell 3′ HT Kit v3.1

Libraries generated from human PBMCs with the 10X Single Cell 3′ HT kit were sequenced on 1 lane of the NovaSeq v1.5 with 28-10-10-90 configuration to generate 28K reads per cell for each direction sequenced.

2.2.10.3 Iso-seq libraries

Mouse HSC Iso-Seq libraries

Libraries were sequenced following PacBio recommendations for Iso-Seq libraries. Approximately 5 nM of each library was submitted for sequencing on two SMRT cells of the PacBio Sequel II (8M, v2) with a 30hr movie to generate 3M polymerase reads, and an expected yield of 140 Gb per SMRT cell.

MAS-seq libraries

MAS-seq libraries were sequenced following PacBio recommendations for MAS-seq libraries. Approximately 5 nM of each library was submitted for sequencing on 4 SMRT cells of the PacBio Sequel IIe (8M, v2) with a 30 hr movie. Sequel II Binding kit 3.2 and a 2-hour pre-extension time was used for these libraries as recommended by the manufacturer to generate at least 3.5M polymerase reads per sample.

Table 2.15. Summary of experiments presented in this thesis.

Experiment	Species	Sample	Experimental Condition	Cell isolation approach	Library preparation method	Cells per experiment	Sequencing method	Chapter(s) featured
1	*Mus musculus* (n = 6)	Single-cells (Lin^- $cKit^+$ Cd150+)	Platelet depletion vs Control	FACS	Smart-Seq2 NextEra	1288	Illumina	3
2	*Mus musculus* (n = 6)	Single-cells (Lin^- $cKit^+$ Cd150+)	Young vs old	FACS	Smart-Seq2 NextEra	672	Illumina	4
3	*Mus musculus* (from experiment 2)	cDNA (Smart-Seq2 cDNA)	Young vs old	HSCs pooled based on single-cell clustering of Experiment 2	Iso-Seq SMRTbell prep	NA	PacBio	5
4	*Mus musculus* (n = 2)	Single-cells (Lin^- $cKit^+$ Cd150+)	NA	FACS and 10X Genomics	10X Genomics Single Cell 3' **LT** and MAS-seq	73	Illumina PacBio	5
5	*Mus musculus* (n = 2)	Whole bone-marrow suspension	NA	10X Genomics	10X Genomics Single Cell 3' **LT** and adaptation of HITscIso-Seq	1040	Illumina PacBio	5
6	*Homo sapiens* (n = 2)	Peripheral blood mononuclear cells (PBMCs)	NA	10X Genomics	10X Genomics Single Cell 3' **HT** and MAS-seq	>12,000	Illumina PacBio	5

2.3 Computational analysis

For all computational packages mentioned, versions and citations are detailed Table 2.16.

Smart-seq2 scRNA-seq data pre-processing pipeline

Single-cell RNAseq Smart-seq2 data preprocessing was performed by Anita Scoones using the *ScOmix: Integrated Single-cell analysis pipeline* developed by Matthew Madgwick.

2.3.1 Genome alignment

Raw FASTQ paired-end sequencing reads were aligned using the universal RNA-seq STAR aligner (Dobin *et al.*, 2013) to the mouse reference genome version M23 (GRCm38.p6) obtained from GENCODE (Frankish *et al.*, 2019) using annotations extracted from the GTF file for that genome. Pre-alignment statistics were obtained from the raw sequences using FASTQC (Andrews, 2010). These results were aggregated together with additional post-alignment quality control summary statistics using MultiQC (Ewels *et al.*, 2016) the read quality and alignment scores against the reference genome.

2.3.2 Gene count quantification

BAM files of mapped sequence reads were processed using SamTools (Li *et al.*, 2009). For quantification, reads were annotated (vM23q) with genomic features (as genes, junctions, exons, promoters, gene bodies, genomic bins and chromosomal locations) and counted using the general-purpose featureCounts (Liao, Smyth and Shi, 2014) summarisation program for read count quantification. Gene expression matrices containing the counts of unique RNA molecules that mapped to each gene ID (rows) for each cell (columns) were created for each sample plate of cells sequenced.

10X Genomics scRNA-seq data preprocessing pipeline

10X Genomics data pre-processing was performed by Dr Yuxuan Lan using the CellRanger (https://github.com/10XGenomics/cellranger) analysis pipelines for Chromium single-cell data analysis.

2.3.3 10X Genomics data demultiplexing

First, raw base call (BCL) files containing sequencing data of all the libraries in the sequencing run generated by the Illumina software were demultiplexed using the CellRanger (Zheng *et al.*, 2017) *mkfastq* pipeline - a pipeline adapted for 10X Genomics data that wraps Illumina's *bcl2fastq* software. This generated FASTQ files for each individual library sequenced.

2.3.4 Gene count quantification

Cellranger *count* was then used to generate single cell gene counts for each GEM well that was demultiplexed by CellRanger *mkfastq* using default software settings. The run ID, FASTQ files, sample name and the mouse or human transcriptome references for each experiment respectively were provided as input. *Count* performs read alignment, UMI counting, and secondary analysis (dimensionality reduction, clustering, and visualisation) for each run.

First, the first 16 bases of read 1 containing the 10x barcode, which identifies the partition from which the DNA originates, were extracted from read pairs. Using the 10X barcode whitelist of 737,000 different sequence barcodes error correction was performed, which checked whether all observed barcodes matched any barcode on the whitelist due to sequencing error. After trimming the barcode sequence, the trimmed read pairs were aligned to the reference genome and post-alignment duplicate read pairs were marked using the heuristic that when two read pairs with the same barcode align to the same fragment on the reference genome they are duplicates of each other. Next, to define cell barcodes in partitions containing a cell (each barcode labels a partition but not every partition contains a cell) the distribution of non-duplicate reads with a mapping quality of at least 30 per barcode was calculated. Finally, to compute a coverage profile matrix the read-pair coverage over the genome for each cell barcode using only read-pairs that had mapping quality of at least 30 and were not marked duplicates was performed. The final output included a *web_summary.html* file containing summary metrics and automated secondary analysis results, and the feature, barcode, and count folders containing counts for every gene per cell for downstream analysis.

Bioinformatic data analysis

2.3.5 Single-cell RNAseq quality control

Bioinformatics analysis was conducted in RStudio and processed using the *Seurat* R package (Stuart *et al.*, 2019). Matrices of gene expression counts per cell for each 96-well plate of samples were loaded into Seurat as .tsv files. For quality control, the number of reads mapped to the genome (counts), number of genes detected (features), and the proportion of the reads mapping to the mitochondrial genome per cell were quantified for every cell. Cells with low sequencing depth (less than 1M reads), expressing less than 200 or more than 10,000 features, and a high fraction of reads mapping to the mitochondrial genome (>15%) were excluded from downstream analyses. Prior to merging the matrices into a single aggregated Seurat object, global-scaling normalisation was employed where the feature expression values were normalised for each cell by the total expression and then multiplied by a scale factor of 10,000, then log-transformed using log1p to account for any 0s before normalisation.

2.3.6 Data integration

Datasets were integrated using nonlinear transformation of the underlying data and identified anchors across dimensions 1:50 of the datasets using the *FindIntegrationAnchors* function. Post-normalisation, the top 5000 most highly variable genes were selected using variance-stabilising transformation, calculated prior for each separate object using the *FindVariableFeatures* function (subset of features that exhibit cell-to-cell variation in the dataset most likely to highlight biological signal). These features (genes) identified as most variables across the dataset were then used to complete the integration (*IntegrateData)*.

2.3.7 Cell-cycle annotation and regression

To explore the effects of cell cycle heterogeneity in the data, cells were scored based on canonical cell cycle markers loaded within Seurat (Kowalczyk *et al.*, 2015). After completing the initialisation of the Seurat object, the *CellCycleScoring* function was implemented to assign and store S and G2/M scores in the metadata along with the predicted cell cycle classification (either G2M, S or G1 phase) per cell. To minimise the influence of cell-cycle associated genes in downstream analysis, the difference between the G2M and S phase scores *(CC.difference)* were regressed during data scaling according to Seurat's *Alternative Workflow* for cell-cycle regression. This approach maintains signals separating non-cycling cells and cycling cells, but differences in cell cycle phase among proliferating cells were regressed so as to not dominate marker identification. This was achieved by providing *CC.difference* into the vars.to.regress parameter during linear transformation (data-scaling).

2.3.8 Dimensionality reduction and clustering

Principal component analysis (PCA) was performed to the normalised scaled expression values on the first 50 principal components (PCs). Both Jackstraw and Elbow plots were inspected to determine the appropriate number of top PCs capturing the most variances. Clustering of the processed data was performed using the *FindClusters* function. To visualise and explore the clusters within the dataset, non-linear dimensional reduction with uniform manifold approximation and projection (UMAP) was applied with the same principal components as input to the clustering analysis, placing similar cells together in a low-dimensional space. This superimposes the clusters on the two-dimensional UMAP projection.

2.3.9 Single-cell differential expression analysis for cell type annotation

Wilcoxon Rank Sum test was used through Seurat's *FindAllMarkers* to identify differentially expressed genes (DEGs) for each cluster to serve as marker genes for cell type annotation using canonical lineage/cell type-specific markers published in haematopoiesis datasets/ literature that is cited where relevant within the text. The data were represented by heatmaps and expression plots to demonstrate the unique transcriptional profiles of each cluster for cell type identification.

2.3.10 Pseudotemporal ordering of single-cells

Cells were ordered along a pseudotime trajectory using the *Monocle3* R package (Trapnell *et al.*, 2014; Cao *et al.*, 2019). This applies a reversed graph embedding machine learning strategy to reconstruct a cell trajectory of differentiation using the top DEGs with most biological over-dispersion. Genes to be used for defining pseudotime ordering in the construction of the trajectory were selected by identifying DEGs between cells based on cell types assigned in *Seurat*. This was achieved through two approaches, first the direct conversion of the *Seurat* object into the *Monocle* equivalent cell data set (*cds*) using *as_cell_dataset()* from *Seurat_Wrappers* - a collection of community-provided methods and extensions for Seurat. This approach preserves cell-type annotations and UMAP cell embeddings per cell to recreate the same partitions for pseudotime analysis. Or second, the *Seurat* object was loaded into monocle post single-cell annotation, pre-processed (*pre_process_cds()*) using the default recommended parameters based on the size of the dataset, and reduced into smaller dimensions. *cluster_cells()* cells was applied using default parameters unless otherwise stated in the text to cluster cells and assign them into partitions. This approach adopted by *Monocle* is based on the partition-based graph abstraction (PAGA), a topology-preserving map of single cells (Wolf *et al.*, 2019). Annotations stored within the *cds* object metadata from *Seurat* were then re-assigned to partitions if the partitions identified correlated with the existing clusters. Using these clusters as input, a graph was learned over the existing projected cells with *learn_graph()*. This applies

a principal graph-embedding procedure that is based on the SimplePPT algorithm ultimately creating a path of connecting points across clusters enabling pseudotime ordering (Mao *et al.*, 2015, 2017). To perform pseudotime ordering of cells the cds was processed using the *order_cells()* function, with default parameters, and the resulting trajectory was visualised using *plot_cell_trajectory* (Qiu *et al.,* 2017). To perform differential expression analysis of genes differentially expressed both along pseudotime and across experimental conditions described within results chapters, graph-autocorrelation analysis (*graph_test())* and regression analysis (*fit_models()*) were used respectively.

2.3.11 Differential expression analyses

Differential gene expression testing was achieved using the pseudobulk approach, aggregating data from single cells at the sample level and subsetting based on the cell-type or condition being tested. This enabled tools traditionally created for bulk expression testing to be used as opposed to more recent methods designed for single-cell data, which continue to be out-performed due to inherent technical noise artefacts in single-cell data such as dropout, zero-inflation and high cell-to-cell variability (Hicks *et al.*, 2018; Squair *et al.*, 2021).

Pseudobulk differential gene expression analyses were performed with *edgeR* and *DESeq2* (Robinson, McCarthy and Smyth, 2010; Love, Huber and Anders, 2014). First, the counts from single cells were aggregated using the *AggregateExpression()* function based on pseudobulk annotations (cell type and biological replicate identity). For *edgeR,* raw counts were used as input, where *DGEList()* was used to generate a data object from counts and group identifier extracted from the metadata (cell type and biological replicate split by condition). The *DESeq2,* equivalent to this was achieved using *DESeqDataSetFromMatrix()*. Median-of-ratios normalisation was performed on the count data and normalisation factor values noted across samples.

Next, pre-filtering of genes was performed to retain for downstream analysis only those above a minimum threshold of 10 reads, and present across all samples across experimental conditions. *EdgeR* filtering was achieved with *filterbyExpr()*, retaining genes that had sufficiently large counts in a statistical analysis by determining library size (a numeric vector giving the total count for each library). After pre-filtering, *estimateDisp()* was applied to give the estimate of the common dispersions across samples to determine within-group variability. This calculated the adjusted profile log-likelihood for each gene, of which the square root of the common dispersion was calculated to obtain the coefficient of biological variation. Next, *edgeR glmFit()* was used to fit a negative binomial generalised log-linear model (GLM) to the read counts for each gene, followed by applying *glmLRT()* to conduct likelihood ratio tests for coefficients in the linear model. To extract the most significant DEGs the GLM outputs were filtered for

adjusted *P values (<=0.05)* and ranked by descending absolute log-fold-change. For *DESeq2* analysis, the *DESeq()* function (wraps *estimateSizeFactors*, *estimateDispersions* and GLM fitting as a single default function) was used to achieve differential expression analysis based on the negative binomial distribution. Statistical tests were implemented as default to determine the significance of differential expression across conditions. The Wald test was used to compare the estimated fold-change in expression between the conditions to the estimated standard error of the fold-change. P-values were adjusted for multiple testing using the Benjamini-Hochberg method to control the false discovery rate. Visualisation of results was achieved with *ggplot2* and *EnhancedVolcano* packages.

2.3.12 Functional Analysis

To gain insights into over-represented functional annotations from gene sets such as DEGs across pseudotime, cell types or conditions gene enrichment analysis was performed using the *clusterProfiler() (Wu et al., 2021)* and *Reactome() (Gillespie et al., 2022)* R packages. This involved first converting lists of genes of interest from gene symbols into their respective Entrez gene identifiers from the National Center for Biotechnology Information (NCBI) database (Maglott *et al.*, 2007). The Entrez ID represents tracked, unique identifiers for genes assigned as a stable species-specific integer. Functional annotation based on known associations described across biological databases was performed on gene sets using *enrichGO(), enrichPathway()* and/or *enrichKEGG()* to perform over-representation analysis across gene ontology classes (GO), pathway and/or Kyoto Encyclopedia of Genes and Genomes (KEGG) databases respectively (Kanehisa and Goto, 2000; G. Yu *et al.*, 2012; Yu and He, 2016). Results were visualised using *ggplot2*.

Iso-seq data pre-processing pipeline

2.3.13 Genome alignment

HiFi reads (FASTQ) were aligned to the mouse reference genome version M23 (GRCm38.p6) obtained from GENCODE using annotations presented in the GTF file for that genome using *minimap2* (Li, 2018; Frankish *et al.*, 2020). This generated mapped BAM files for each sample.

2.3.14 Collapsing of redundant isoforms

After transcript sequences were mapped to the mouse reference genome, the software package cDNA cupcake (Tseng, no date) was used to collapse redundant transcripts into unique isoforms. This step identifies groups of isoforms with high sequence similarity, combining them into single clusters. Low-quality isoforms were redacted based on default parameters and any differences between the remaining isoforms within each group were resolved to generate the final consensus isoform representative for that group. This step generated collapsed isoforms unique for each sample in GFF format and secondary files containing information about the number of reads supporting each unique isoform.

2.3.15 Classification of isoforms using SQANTI3

Once the final dataset containing collapsed isoforms per sample was generated, they were classified and annotated using the SQANTI3 software (Tardaguila *et al.*, 2018). SQANTI3 uses a set of defined criteria to classify isoforms based on their transcript structure, coding potential and support from existing/ published annotations. The input required for this included the collapsed isoform data in GFF format, the mouse reference genome (GRCm38.p6) and annotations presented in the GTF file for that genome. Additionally, the polyA motif list (TXT format) and CAGE peak data (BED format) containing annotated transcription start sites (TSS) for mouse obtained from the SQANTI3/data GitHub repository were also provided. This enables additional QC of TSS (RNA degradation at the 5’ end can result in mistaken classification of novel TSS) and detection of transcription terminal sites (TTS) within or proximal to polyA motifs. Isoforms were classified into categories described in Table 2.16.

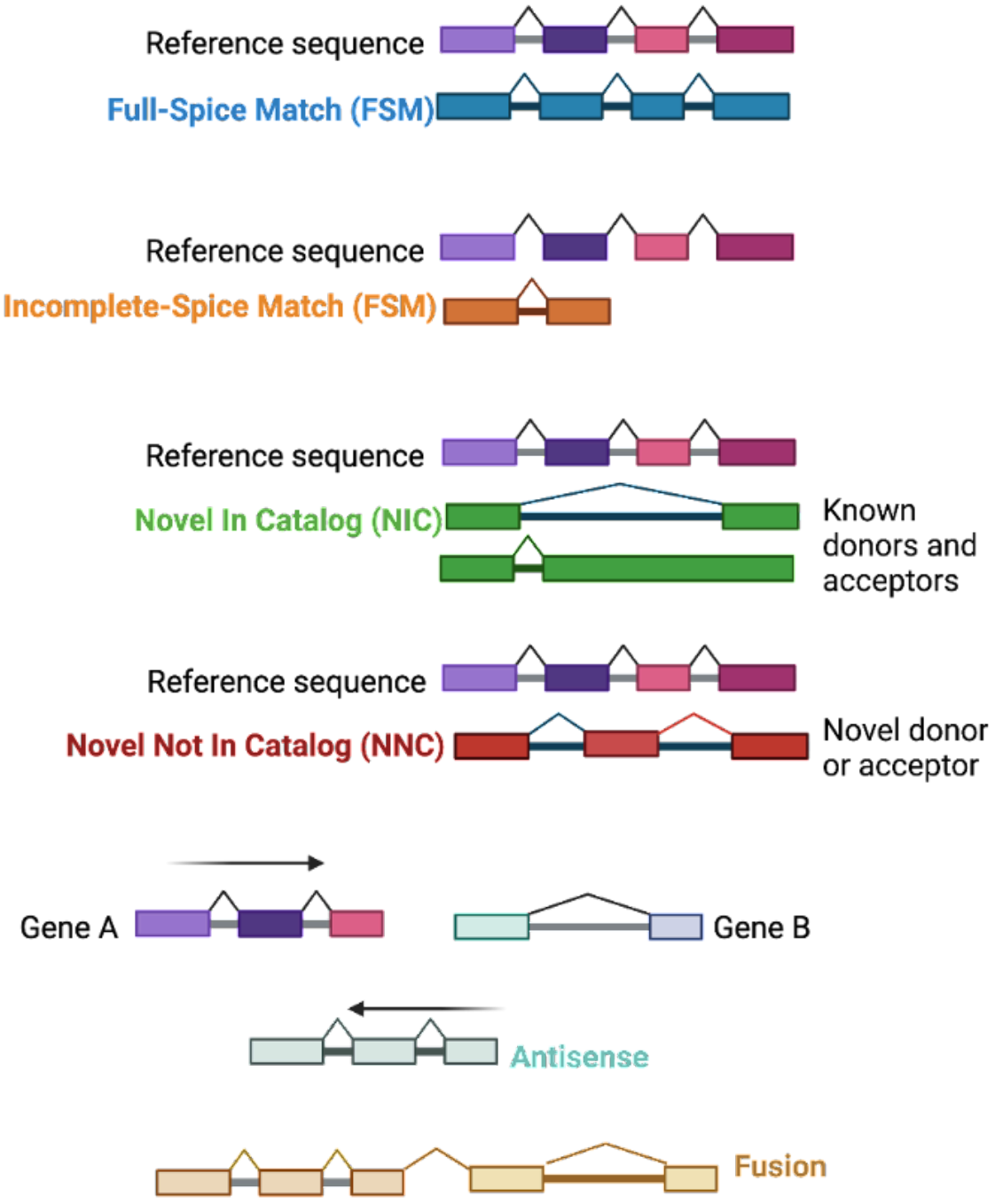
Table 2.16. SQANTI3[9] Isoform classification categories used to annotate isoforms in the data.

Type	Description
Full Splice Match (FSM)	Complete match of all splice junctions in the reference
Incomplete Splice Match (ISM)	Partial match of splice junctions in the reference
Novel in Catalogue (NIC)	Isoform with a novel combination of known splice sites
Novel not in Catalogue (NNC)	Novel isoform with at least 1 novel splice site
Antisense	Does not overlap a same-strand reference gene but is anti-sense to an annotated gene
Genic Intron	The isoform lies entirely within an annotated intron
Genic Genomic	The isoform overlaps introns and exons of the reference
Intergenic	The isoform is within the intergenic region of the reference

2.3.16 Isoform artefact removal

Finally, the output classification file generated from SQANTI3 QC was filtered with the *sqanti_filter.py* python script included in the SQANTI3 workflow. Here, isoforms post-classification were filtered using default parameters defined within the rules argument of the script which searches for artefact isoforms. This step classifies artefact isoforms based on several criteria. Intrapriming at the 3' end in isoforms categorised as FSMs may result in false FSM classification, hence FSM transcripts found to contain >=12 adenines within 20 bp of an annotated TTS are deemed as intrapriming artefact and removed from the data. For all transcripts not defined as FSMs, transcripts are excluded under the following criteria: 3' end is deemed an intrapriming artefact, a junction previously annotated as RT-switching is detected, or if non-canonical junction read coverage does not meet the minimum threshold. This step generated a filtered version of the classifications from SQANTI QC, with an additional QC column indicating whether the isoform was retained or discarded and the reason for its removal.

[9] (Tardaguila *et al.*, 2018) (Figure 2.1)



[10]

Figure 2.1. Schematic illustrating the structural categories used for isoform characterisation based on Structural and Quality Annotation of Novel Transcript Isoforms (SQANTI3) descriptions (Tardaguila *et al.*, 2018).

[10] Created with BioRender.com

MAS-seq data analysis

2.3.18 Deconcatenation of PacBio MAS-seq reads

Data pre-processing was performed using the PacBio SMRT Link software v11.1. First, PacBio Hifi reads (CCS reads with >Q20 sequencing accuracy) were deconcatenated into segmented reads (S-reads) of individual cDNA molecules using the PacBio SMRT read splitter programme *Skera* version 0.1.0. *Skera* was used to split MAS-Seq single-cell reads at adapter positions to generate BAM records for each S-read using adapter sequences provided in FASTA format.

2.3.19 Primer removal

After deconcatenation, the removal of primers from S reads and identification of cell barcodes was performed using the PacBio barcode demultiplexer programme *lima.* Here, unwanted primer sequences and non-barcoded cDNA sequences were removed and reads oriented into 5' - 3' orientation for downstream processing steps.

2.3.20 Cell barcode and UMI extraction and read refinement

Cell barcode and UMI extraction were performed using the PacBio tool isoseq3 *Tag,* by providing the bam output from the *lima* step above together with the 10X 12bp UMI and 16bp 3' barcode lists, generating tagged full-length (FL) reads. In this step, correct UMI and cell barcodes sequences were clipped from reads, and metadata for each successfully tagged read was generated. Next, FL tagged reads were refined using isoseq3 *Refine* by trimming of poly(A) tails with a minimum of 20 bp in length to generate full-length tagged non-concatemer (FLTNC) reads.

2.3.21 Unique molecule identification and deduplication

To improve yield and accuracy during barcode calling in downstream steps, isoseq3 *correct* was used to correct errors in cell barcodes that may have occurred during sequencing. Barcode error correction of FLTNC reads were indexed against a list of candidate cell barcodes taken from the 10x 3' (v3.1) kit, whereby candidates matching the barcode whitelist were marked as real cells, and statistics such as the number of reads matching known barcodes and whether they match cellular barcodes or UMI sequences is annotated. The UMIs correctly identified within the data were then used in isoseq3 *groupdedup* to collapse reads that match the same sequence, correcting for PCR duplicates resulting in only unique molecules.

2.3.22 Genome mapping

Transcripts were mapped to their reference genomes (*Mus musculus* and *Homo Sapiens)* and classified against transcript reference annotation from GENCODE (*Mus musculus* and *Homo sapiens)* using *pbmm2* - a minimap2 SMRT wrapper for PacBio data generating mapped bam files for each cell.

2.3.23 Removal of redundant isoforms and isoform classification

Isoseq3 collapse was used to collapse redundant transcripts based on exonic structures into unique isoforms generating unique isoforms in GFF format along with meta information including the number of reads supporting each unique isoform. classified using *Pigeon.* This characterises isoforms into categories following the SQANTI3 classification categories (see section 2.3.6) generating a classification text file, along with files detailing reads that spanned junctions, and a Seurat-ready input for tertiary downstream analysis.

Table 2.17. Software package versions for all packages used in data analysis for this thesis.

Package	Access/ citation
CellRanger v 3.1.0	(Zheng *et al.*, 2017)
CellLoupeBrowser v 4.1.0	(Zheng *et al.*, 2017)
FlowJo v 10	BD Biosciences
FACSChorus	BD Biosciences
FACSDiva	BD Biosciences
STAR v 2.7.9a	(Dobin *et al.*, 2013)
FastQC v 0.11.7	(Andrews, no date)
MultiQC v 1.5	(Ewels *et al.*, 2016)
minimap2 v 2-2.7	(Li, 2018)
featureCounts v 1.6.4	(Liao, Smyth and Shi, 2014)
Python v 3.6.0	(Van Rossum and Drake, 2009)
Biopython v 1.61	(Cock *et al.*, 2009)
Samtools v 1.15.1	(Danecek *et al.*, 2021)
cDNA cupcake v 29.0.0	(Tseng, no date)
SQANTI3 v 4.2	(Tardaguila *et al.*, 2018)
R v. 4.2.0	R Core Team (2022).
RStudio v 22.07.1	R Core Team (2022)
edgeR package v .3.38.4	(Robinson, McCarthy and Smyth, 2010)
goseq v 1.48.0	(Young, Wakefield and Smyth, 2012)
monocle3 v 1.0.0	(Trapnell *et al.*, 2014)
ggplot2 v 3.3.6	(Wickham, 2016)
Seurat v 4.2.0	(Butler *et al.*, 2018; Stuart *et al.*, 2019; Hao *et al.*, 2021)
ReactomePA v 1.40.0	(Gillespie *et al.*, 2022)
clusterProfiler v 4.4.4	(Wu *et al.*, 2021)
tidyverse v 1.3.2	(Wickham *et al.*, 2019)
reshape2 v 1.4.4	(Wickham, 2007)
SingleCellExperiment v 1.18.0	(Amezquita *et al.*, 2020)
stringr v 1.4.1	(Wickham, 2010)
BiocGenerics v 0.42.0	(Huber *et al.*, 2015)
Iso-Seq (3.8.2)	https://isoseq.how/

Chapter 3:

Single-cell profiling of the trajectory from haematopoietic stem cells to megakaryocyte progenitor in response to stress

Chapter disclosures:

Preprocessing of scRNA-seq libraries (genome alignment, sequencing quality-control, and gene count quantification) were performed by Anita Scoones using the *ScOmix: Integrated Single-cell analysis pipeline* (unpublished) developed by Matthew Madgwick (see Materials and Methods 2.3.1 - 2.3.2).

3.1 Introduction

Distinct subsets of HSCs exist with differential propensities towards the generation of lymphoid or myeloid fates is an accepted phenomenon within haematopoiesis (Müller-Sieburg *et al.*, 2002; Adolfsson *et al.*, 2005; Dykstra *et al.*, 2007; Copley, Beer and Eaves, 2012). One subset in particular marked by high expression of the *Vwf* gene not only has a higher tendency towards myeloid cell differentiation but is distinctly stably biassed towards the platelet lineage, with competitive transplantation assays verifying that *Vwf*+ HSCs have a long-term increased platelet output compared to *Vwf*- HSCs which instead show lymphoid-biassed reconstitution (Sanjuan-Pla *et al.*, 2013). Extensive functional assays interrogating the potent platelet-primed reconstitution pattern of *Vwf*+ HSCs, combined with global gene expression profiling have shown this to be in part regulated by TPO signalling, and exhibit higher expression of Mk-lineage genes including *Vwf,* and *Itga2b (Cd41)* (Gekas and Graf, 2013; Sanjuan-Pla *et al.*, 2013).

The cell surface molecule signalling lymphocytic activation molecule family member 1 (Slamf1) also designated Cd150 - which is expressed in the primitive HSC compartment (Kiel *et al.*, 2005) - has been identified as a specific marker for subfractionation of functionally distinct myeloerythroid precursors with high-clonal Mk capacity both *in vitro* and in transplant assays *in vivo* (Pronk *et al.*, 2007). Although high Cd150 expression is commonly used to isolate HSCs, its expression was revealed to be maintained in cells Mk lineage and of the myeloerythroid compartment including MkPs. Both CMPs and MEPs were divided into Cd150+ and a Cd150- cells, where the Cd150+ cells exhibited high clonal megakaryocyte capacity indicating the potential of utilising this as a marker for cells committed to the Mk fate (Pronk *et al.*, 2007).

The strong evidence for Mk-biassed reconstitution by HSCs (Månsson *et al.*, 2007; Sanjuan-Pla *et al.*, 2013; Shin *et al.*, 2014), and corroborating evidence of Mk-restricted cells identified within the HSC compartment (Yamamoto *et al.*, 2013) in steady-state haematopoiesis prompted the question as to whether this potential bypass of intermediate commitment steps exists as part of a protective mechanism under non-homeostatic conditions. In extensive platelet consumption, such as caused by infection or injury, the haematopoietic system recruits HSCs to generate the necessary platelet precursors to coordinate platelet regeneration in order to restore homeostasis. According to the commitment roadmap of classic haematopoiesis, this would require the hierarchical transition through discrete stages of MPP, CMP, MEP to reach Mk unipotency before Mk maturation. Using single-cell *ex vivo* lineage tracking Haas *et al.* demonstrated in an infection model that inflammation-driven changes within the phenotypic

HSC compartment was detected in a particular subset of cells that exclusively generated mature Mks but are distinct from MkPs (Haas *et al.*, 2015). They showed that this fraction of cells - coined stem-like Mk progenitors (SL-MkPs) - exists as an Mk-committed subset of the wider Mk-primed but multipotent *Vwf*+ HSC population identified by Sanjuan-Pla *et al.*, with uniformly high expression of all Mk genes not only enriched transcriptional priming.

Taken together this research has provided important insights into the cells involved in Mk lineage fate commitment from HSCs, suggesting multiple routes of Mk differentiation exist. However the transcriptional heterogeneity of cells along the Mk lineage, the differences in expression along the Mk differentiation trajectory and how this is impacted by acute platelet depletion is yet to be elucidated at high-resolution. Moreover, although assays have been performed at the single-cell level, they have been applied to discrete populations independently isolated based on cell-surface marker expression, therefore may be missing intermediate stages of Mk fate restriction.

3.1.1 Aims

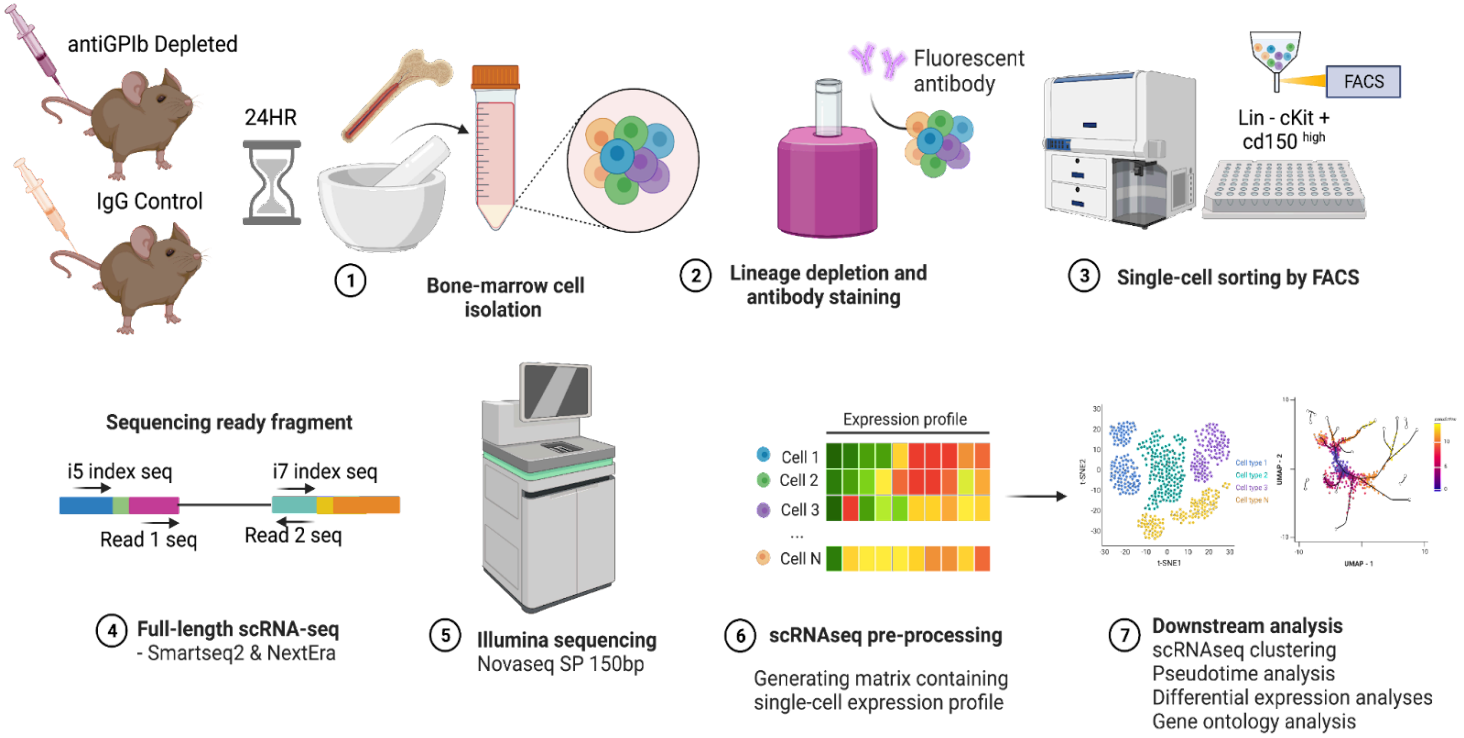
The aims of this chapter were to:

- Capture cells along the trajectory towards Mk commitment by FACS gating LK and LSK Cd150+ cells.
- Apply single-cell transcriptomics using Smart-seq2 to order cells along the continuum of differentiation between HSC and MkP.
- Interrogate the differentiation trajectory in both steady state and in response to platelet depletion.

These aims were addressed by profiling HSPCs from mouse bone marrow using scRNA-seq. Through targeted depletion of platelets in mice, the hypothesis was that this would induce megakaryopoiesis for the replenishment of platelets, prompting changes in the bone-marrow compartment including increased HSC Mk differentiation and consequently the increased expression of genes promoting commitment to Mks. Employing Smart-seq2 on cells captured using a broad gating approach enabled the analysis of cells from HSC to committed MkPs and any intermediate cell types that may represent important transitional stages in Mk differentiation, normally excluded by narrow gating strategies. The scRNA-seq profiles of cells were used to visualise and study the transcriptional heterogeneity of cells in the LK Cd150+ BM fraction, and using pseudotime trajectory analysis single cells were computationally ordered along a differentiation trajectory. This chapter describes the analysis of transcriptional changes along Mk commitment and differential expression signatures associated with a model of acute thrombocytopenia.

3.2 Experimental approach

To investigate the trajectories of the Mk lineage in response to platelet depletion, intravenous injection of anti-GPIb antibody was administered to mice. This treatment induces antibody-induced thrombocytopenia, causing significant and irreversible Fc-independent platelet depletion in mice (Bergmeier *et al.*, 2000; Nieswandt *et al.*, 2000). Four mice were injected with anti-GPIb, while two mice were injected with IgG control suspended in PBS as controls. LK and LSK Cd150+ single-cells were sorted into 96-well plates (Methods 2.2.2 - 2.2.6.1) and processed for Smart-seq2 as previously described (Picelli *et al.,* 2014) (Methods 2.2.7.1-2.2.7.4). A total of 14 96-well plates of scRNA-seq data were generated and sequenced. The resulting single-cell transcriptional profiles were analysed to determine the captured cell types. Differential expression and functional analyses were performed to investigate the effects of platelet depletion Mk differentiation trajectories.



[11]**Figure 3.1.** Schematic workflow of the experimental approaches implemented for Chapters 3.

[11] Created with BioRender.com

3.3 Results

3.3.1 Isolation of haematopoietic stem and megakaryocyte progenitors for scRNA-seq

To capture cells along the Mk trajectory for scRNA-seq, cells were isolated from mouse bone marrow and sorted by single-cell FACS coupled with index sorting (Figure 3.1).

Prior to single-cell sorting, we conducted quantitative measurements to evaluate the efficacy of platelet depletion across all subjects. Blood aliquots obtained via cardiac puncture 24 hours antibody treatment were subjected to flow cytometry analysis targeting the expression of Itga2b (CD41), a platelet-specific glycoprotein, relative to the number of FACS beads introduced into each sample. FACS beads, engineered with a consistent fluorescence intensity, served as internal controls. Following CD41 staining and addition of FACS beads, samples were processed through the flow cytometer to quantify the fluorescence intensity of individual cells. The frequency of CD41+ events relative to FACS bead events was utilised to determine platelet counts in each sample. In platelet-depleted mice, the average frequency of platelets per bead was 0.8, contrasting with the range of 150-151 platelets per bead observed in mice administered IgG control antibody. This discrepancy confirms the successful and significant depletion of platelets in mice treated with anti-CD41 antibody (Figure 3.2).

In addition, the same blood aliquots were analysed using an automated haematology analyser (Alinity HQ, Abbott Laboratories) to assess the specificity of anti-GPIb treatment to platelets, and ensure other blood cell populations remained unaffacted by the experiment. As expected these results show significantly fewer platelets 24 hours post anti-GPIb treatment (13.8 billion cells / L) compared to mice injected with control antibody (626 billion cells / L), with no significant differences observed in the frequency of white blood cells (WBC) or red blood cells (RBC) between platelet-depleted and control mice (see Appendix Supplementary Figure 3.1). These findings confirmed the selective action of the antibody on platelets, successfully depleting platelets without affecting other cellular components of blood.

Broad gates based on c-Kit and Cd150 expression (Lin - cKit + Cd150+) were used to sort LT-HSCs, Mk progenitors and all intermediate cells with high Cd150 expression, which has been shown to be a shared marker that is expressed at all stages of Mk commitment (Pronk *et al.*, 2007). Due to their low frequency compared to other cell types in the LK Cd150+ gating strategy, LT-HSCs were gated (Lin - Sca + cKit + Cd150+) and sorted separately to ensure sufficient coverage of the population. Cells from both platelet-depleted and control mice were sorted into fourteen 96-well PCR plates, with LK Cd150+ and LT-HSC cell enrichment

represented for each plate (Figure 3.3A) for a total of 1288 cells (784 LK Cd150+ and 504 LSK Cd150+ cells respectively, excluding controls). Using this sort layout accounts for possible batch effects that could arise during sorting and downstream library preparation of plates (Figure 3.3B).

Plates were randomised and processed in four batches for Smart-seq2 single-cell RNAseq as previously described (Picelli 2014). After the generation of cDNA from samples, 0.2ng/uL of cDNA was used per sample to generate Illumina sequencing-ready libraries with Nextera. Sample quality was evaluated at several stages before sequencing, including post-cDNA amplification product clean-up and post-Nextera library preparation (Appendix Supplementary Figure 3.2). Plates were pooled at equimolar concentrations to ensure equal read coverage across libraries during sequencing to generate a minimum of 1M reads per cell. A total of four libraries with pooled single cells were sequenced to generate data for this chapter (Table 3.1).

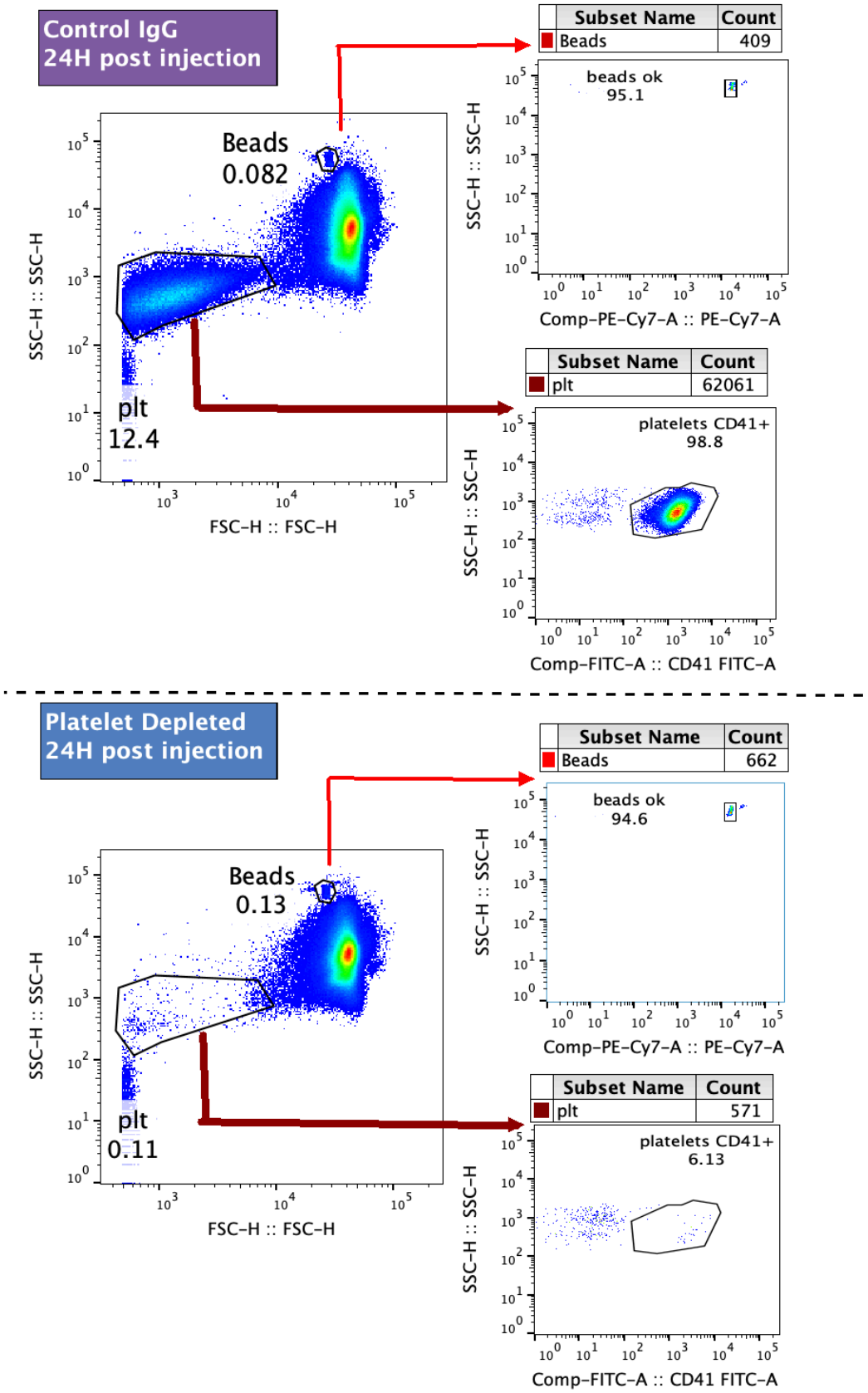
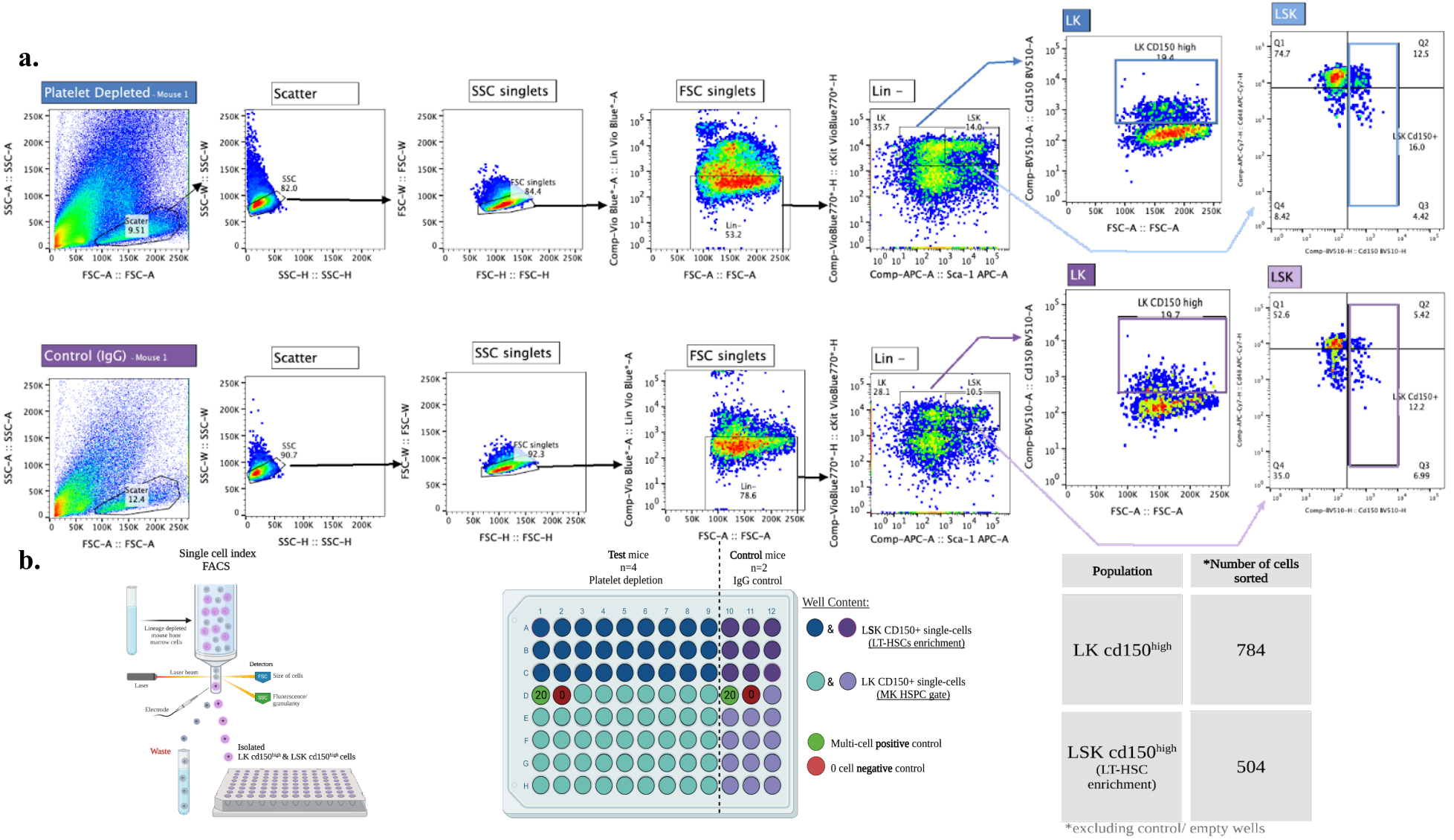


Figure 3.2. Flow cytometry analysis of whole blood collected from mice 24 hours post-injection. Shows either anti-GPIb treatment or isotype control confirms depletion of platelets.



Population	*Number of cells sorted
LK cd150^{high}	784
LSK cd150^{high} (LT-HSC enrichment)	504

*excluding control/ empty wells

Figure 3.3. Isolation of mouse bone-marrow HSCs and Mk-EryPs single cells with index FACS sorting. (a) Flow cytometry sorting gating strategy used for FACS isolation of LK and LSK Cd150+ single cells. (b) Summary of cell isolation experimental design outlining plate sorting strategy and numbers of cells isolated per population.

Table 3.1. Overview of sequencing strategy employed to generate single-cell libraries. [12]

Sequencing Batch	Number of Plates	Plate IDs	Illumina index sets used for pooling	Total number of cells (excluding controls)
1	4	1, 5, 7 & 11	1, 2, 3 & 4	368
2	3	2, 8, & 13	1, 2 & 3	276
3	4	3, 6, 9 & 12	1, 2, 3 & 4	368
4	3	4, 10 & 14	1, 2 & 3	276

[12] Sequences for the four Illumina index sets of 96-well plates used in Nextera library preparation are listed in Supplementary Table 2.1.

3.3.2 Sequencing quality metrics and markers of poor cell viability allows the identification of high-quality samples for downstream analyses

To evaluate the integrity of the sequencing data, the output generated by FASTQC underwent comprehensive analysis using MultiQC. This process generated a detailed report for each single-cell library, presenting statistics on various parameters including read count, read length distribution, GC content, adapter contamination, and duplication rates. Additionally, the percentage of reads uniquely mapping to the mouse genome per cell was assessed (refer to Appendix Supplementary Figure 3.3). During this analysis, few samples exhibited lower-quality indicators such as reduced read quality or the presence of overrepresented sequences. However, all samples were retained as part of the dataset and subsequently processed using featureCounts to generate a feature counts matrix for each batch of libraries.

To remove low-quality cells from the dataset, commonly used metrics to assess single-cell data quality were assessed. These criteria meant samples with fewer than 100,000 total reads per cell, a total number of genes detected below 2,000 per cell, and mitochondrial gene expression content exceeding 15% were excluded from the dataset - leaving a total of 933 high-quality single-cell samples suitable for further analysis (Figure 3.4 and 3.5). Moreover, any samples with preceeding indications of low quality with MultiQC that were not already removed based on these metrics were specifically assessed and individually excluded. The calculation of these QC statistics confirmed that sequencing resulted in an average 1.3M reads and a median of 8,646 genes detected for cells that passed QC, meeting the expected minimum read-depth criteria of ~1M reads per cell (Figure 3.5). The relationship between the number of reads per cell and the number of detected genes follows a trend of proportionality up to a saturation point. Beyond this point, additional reads cease to contribute to the detection of new genes. This phenomenon can be attributed to the random sampling of each mRNA molecule during library preparation and sequencing. As the number of reads increases, there is a higher probability of capturing rare or lowly expressed transcripts. Consequently, this leads to the detection of a greater diversity of genes within the sample. This trend is evident in this data when visualising the relationship between reads and counts per cell where the steep increase in genes detected as counts per cell increases begins to plateau after approximately 1M reads per cell (Figures 3.4 and 3.5). The average number of genes captured from platelet-depleted and control mice were 8,764 and 6,942 respectively, values consistent using Smart-seq2 scRNA-seq gene detection across the literature - ranging between 5-15K genes per cell depending on experimental conditions and quality of input cells (Svensson *et al.*, 2017; Ziegenhain *et al.*, 2017).

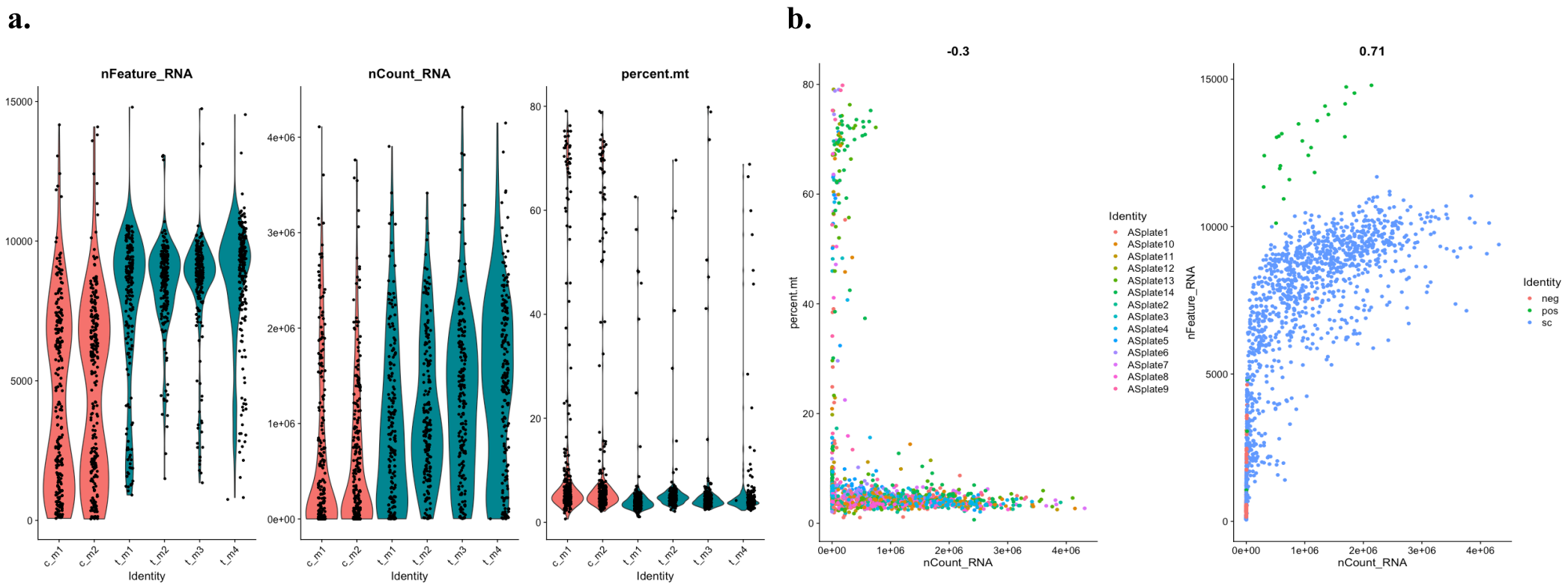


Figure 3.4. Single-cell sample quality control for selecting high-quality cells suitable for further analysis. The top panel shows raw data statistics of 1288 cells. (a) Violin plots showing: (1) number of genes (nFeature), (2) number of reads (nCount) and (3) percentage of reads attributed to mitochondrial genes (percent.mt) per single cell. Violins are coloured by experimental condition (red = **control**, teal = **platelet depleted**) and split by mouse ID (**control** 1 and 2, **platelet depleted** 1 to 4). (b) (1) Correlation between the number of reads and the percentage of mitochondrial content per cell, coloured by plate ID (**plates 1-14**). (2) Correlation between the number of reads and the number of genes detected per cell, coloured by well type (**pos** = multi-cell well, **neg** = empty well, **sc** = single-cell well).

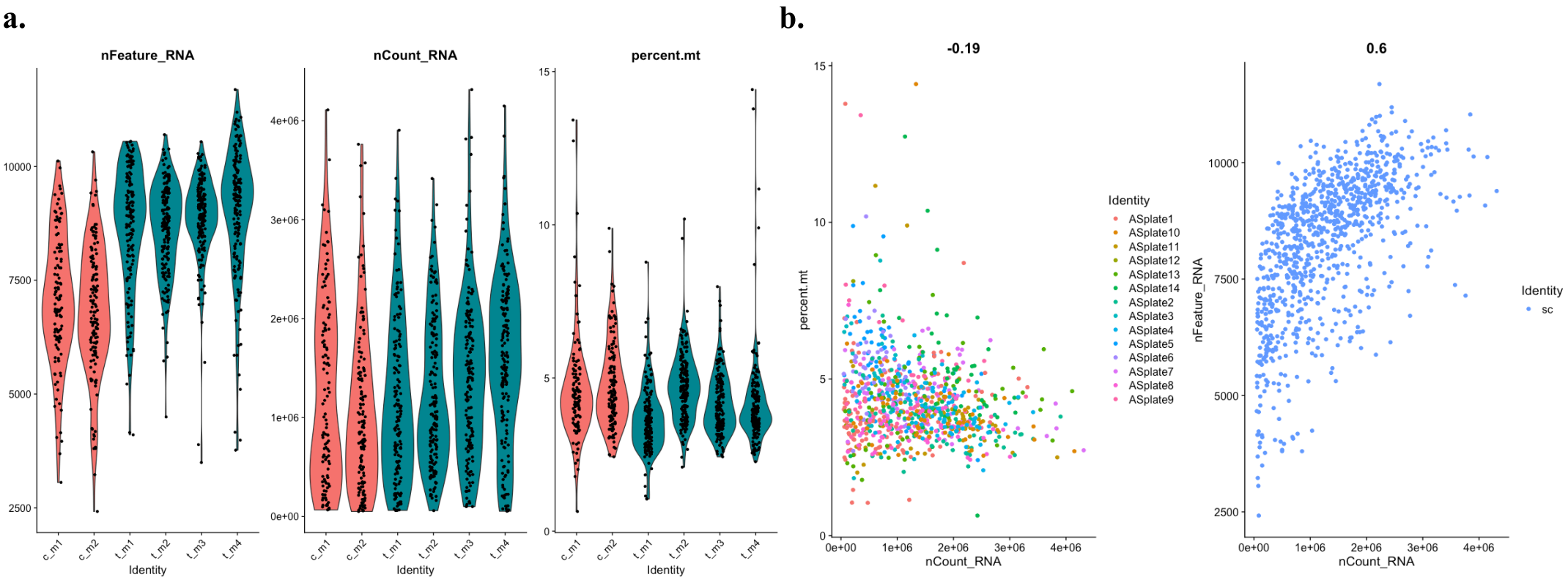


Figure 3.5. Single-cell post-quality control data selection, comprising 933 cells. (a) Violin plots showing: (1) number of genes (nFeature), (2) number of reads (nCount) and (3) percentage of reads attributed to mitochondrial genes (percent.mt) per single cell. Violins are coloured by experimental condition (red = **control**, teal = **platelet depleted**) and split by mouse ID (**control** 1 and 2, **platelet depleted** 1 to 4). (b) (1) Correlation between the number of reads and the percentage of mitochondrial content per cell, coloured by plate ID (**plates 1-14**). (2) Correlation between the number of reads and the number of genes detected per cell, coloured by well type (**sc** = single-cell well).

3.3.3. Batch correction and cell cycle marker regression enable the identification of primary sources of heterogeneity of scRNA-seq data

After the selection of high-quality samples, the data were aggregated into four objects based on batch (1-4), normalised (see Methods section 2.3.5) and the top 5,000 variable features for each object calculated independently. This enabled the identification of repeatedly variable genes to use for the integration of data from across the four batches. Once variable features per batch were calculated, a list of anchors between the batches - or cross-dataset cell pairs that were in a matched biological state - were used to integrate samples back into a single dataset for further analysis. A major challenge in single-cell data integration is the presence of batch effects that arise from technical variations between samples that can significantly influence gene expression measurements, complicating the distinction between biological and technical effects. ScRNA-seq data integration combines data from multiple samples/experiments generating a unified dataset for downstream analysis. Moreover, data integration from multiple samples often aids in the accuracy and resolution of cell type identification and annotation. By combining data from multiple experiments the number of cells available for analysis is larger, which can improve the sensitivity and specificity of cell type identification. Here, the objective of integrating the data based on sequencing batches of randomised plates was to correct for technical differences between scRNA-seq datasets i.e. removal of 'uninteresting' sources of variation such as technical noise or batch effects, as opposed to matching data based on conserved cell types across the treatment conditions. This approach ensured that any differences that may be present as a result of either technical or biological sources of variability between treatment conditions were ignored, and intentionally performed as such so as not to mask the presence of differential expression signatures or cell types not conserved across the two treatment conditions which would be key in data interpretation.

Another common source of variation is the inherently high expression of cell-cycle-associated gene markers within cells, which often has a confounding role during dimensionality reduction and clustering of scRNA-seq data (Kowalczyk *et al.*, 2015; Tirosh *et al.*, 2016). Whilst this is a true source of biological variation and is valuable in understanding cell states within the data particularly in differentiating processes such as haematopoiesis, the dominating effect on gene expression can obscure other important signals during principal component analysis (PCA). To minimise the influence of cell-cycle gene expression variability on downstream analysis, Seurat's cell-cycle scoring approach was employed to first: score cells into either G2M, S or G1 cell-cycle phases based on the expression of canonical cell cycle associated markers (Tirosh *et al.*, 2016), and secondly to: regress the difference between G2M and S phase scores during scaling of the data. This approach maintains variability caused by non-cycling and cycling cells, preserving signals that would likely be important to distinguish stem vs progenitor cells,

whilst regressing differences in cell-cycle phase in cells undergoing proliferation, mitigating the unwanted effects of cell-cycle heterogeneity.

To identify the primary sources of heterogeneity present in the newly integrated dataset, containing cells from all mice and both treatments, the 5000 features of highest variance were recalculated post-data integration. The contribution of these features were then assessed using PCA, where the first principal component (PC) contains the features contributing the most heterogeneity, the second most and so forth. This enabled initial exploration of highly variable gene patterns present in the data, providing early indications of the cells captured within the experiment.

Canonical haematopoietic lineage-specific markers for myeloid (*Mpo, Ctsg, Elane)* , platelet (*Pf4),* erythroid (*Car1)* as well as HSC-specific markers (*Cd74)* were among the genes showing highest standardised variance, providing preliminary confirmation the cell types expected were captured (Figure 3.6A). PC loadings represent the weights of each gene used when calculating the PCs. Positive loadings indicate a positive correlation between the expression of genes and the PC, whilst negative loadings indicate a negative correlation, whereby the larger (either positive or negative) gene loading indicates a strong effect on the specific PC.

To pull-out which genes were contributing to the construction of PCs and identify correlated gene sets prior to clustering or cell type annotation, the loadings of the first 50 PCs were assessed. Given the genes with the highest standard variance are, for the most part, associated with myeloid cell function (Figure 3.6A), unsurprisingly the loadings for the first PC is populated by other myeloid-associated genes including *Mafb* - essential in early myeloid differentiation (Kelly *et al.*, 2000) and *Aif1* - highly expressed in committed subsets differentiating from common myeloid CMPs (Elizondo *et al.*, 2019) (Figure 3.6B). The positive loadings of PC2, which is uncorrelated with the first component and accounts for the next largest variance, include genes related to HSC function *Mpl, Esam, Hlf, Ifitm1* and *3* with negative loadings for myeloid-associated genes.

Altogether, the loadings for the first PCs provided initial insights suggesting variable expression in myeloid-associated genes, as well as HSC and genes previously linked with Mk-Ery commitment. Although the high variability in the expression of myeloid genes could be of biological significance, it is most likely due to the fact that myeloid cells have a more significantly different expression profile to all other cells captured. HSCs along with cells of the Mk and Ery lineages have a higher proportion of genes in common with each other compared to myeloid cells. As the implemented sorting strategy was intended for the isolation of HSCs and cells from Mk-Ery lineages, it is most probable this saturation of

myeloid-associated genes in PC loadings is a product of a small number of myeloid cells that were captured that express a stronger diverging signature compared to the rest of the dataset.

For efficient downstream analysis of single-cell expression signatures and cell clustering, the standard deviation of PCs was calculated to determine the number of components containing the highest sources of heterogeneity. This mathematical optimisation technique where variation is plotted as a function of the number of PCs enables the determination of the minimum number of components that contains the most variation to describe the data, to not only avoid unnecessary computational work-load but most importantly data overfitting and remove potential noise from the dataset. To achieve this a typical approach would be to implement a heuristic approach called the "elbow method" and represent the data in a scree plot. This would capture the PCs which have the largest proportion of variation explained in the data. Here the first seven PCs contain the highest standard variation, and PC fifteen was determined as the cut-off (or "elbow") for downstream clustering analyses - adding further PCs was determined as unnecessary for modelling of the data (Figure 3.6C).

As part of PCA, each cell is given a score in each PC which can be used to project and visualise the distribution of cells in relation to each other based on their gene expression (Figure 3.6D). While PCA is an informative method of dimensionality reduction and pulls out variations in the data, it assumes the data has a linear structure and may therefore miss any non-linear patterns (Lever, Krzywinski and Altman, 2017). Moreover, PCA can only take into consideration two or three principal components at a time, whereas more modern approaches of dimensionality reduction consider all components only plotting them in two dimensions (Maaten and Hinton, 2008; McInnes, Healy and Melville, 2018). scRNA-seq datasets, such as this, are nonlinear in nature, and have a highly-dimensional and complex structure, therefore PCA is not recommended as the sole method to represent the underlying and often hidden structure of the dataset (Becht *et al.*, 2018). As a linear technique that assumes normal data distribution, PCA projections were used for the purposes of preliminary exploration of variance in the data based on experimental conditions, but other methods were used to determine single-cell clusters. PCA confirms no detectable batch effect from the processing of plates - where cells are not separated based on plate ID (1-14) nor differential treatment condition (control or depleted), but it is clear that is insufficient to resolve cell-type clusters (Figure 3.6D).

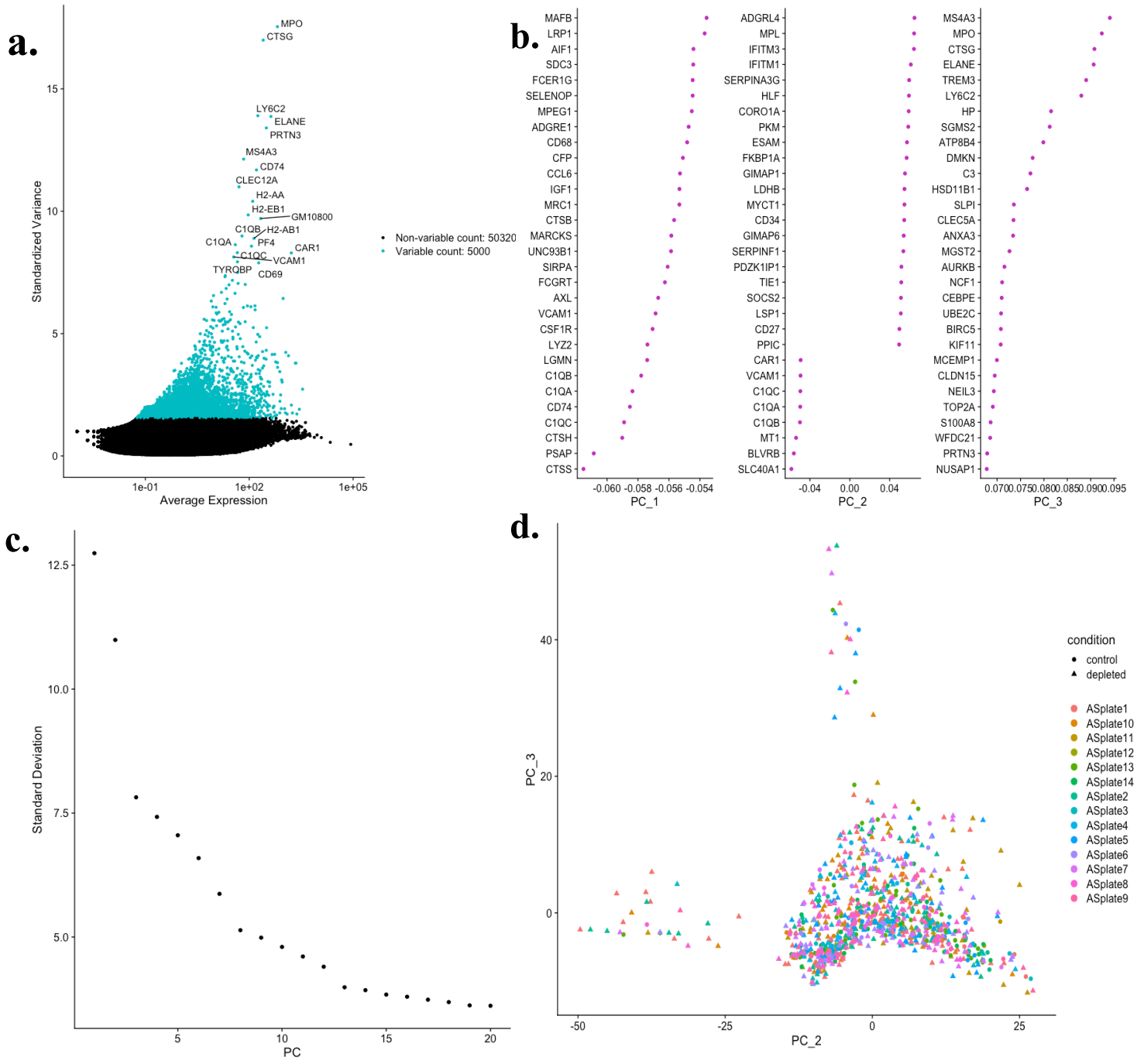


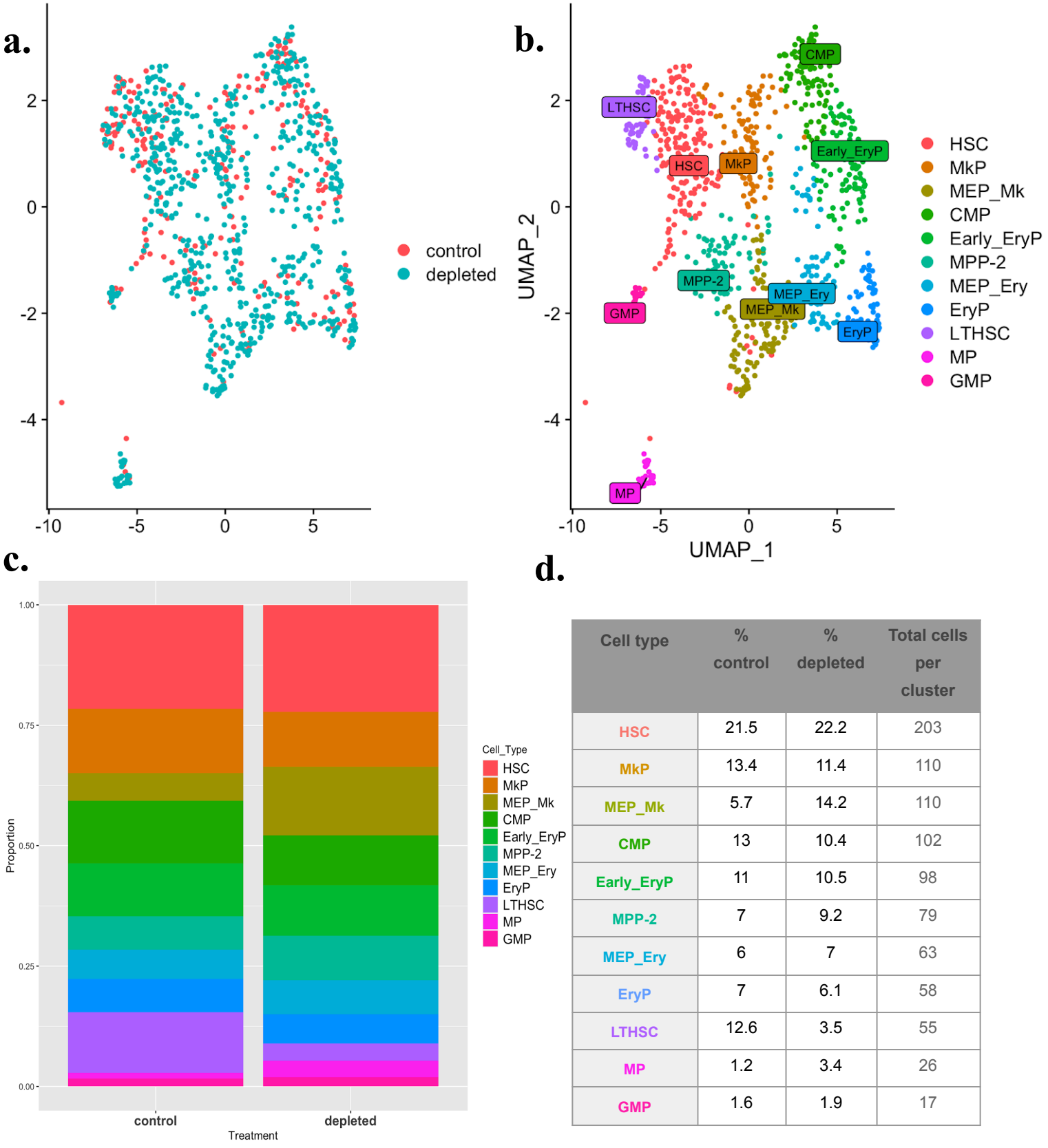
Figure 3.6. Principal component analysis of the dataset post-integration identifies the top principal components containing the most highly-variable genes and confirms no plate- or treatment-induced batch effects. (a) Standard variance of genes against average expression. Teal: 5000 most highly variable genes, with the top 20 gene IDs annotated (b) Loadings for top three principal components (c) Scree plot of standard deviation as a function of principal component number used to determine the number of components to include for downstream clustering (d) PCA projection of all cells coloured by the plate of origin (1-14) where point shapes (circle and triangle) indicate control or depleted treatment conditions.

3.3.4 Unbiased hierarchical clustering reveals transcriptionally distinct populations from LT-HSC towards Mk and Ery progenitor signatures

Unsupervised clustering of the single-cell dataset was performed using Seurat's *FindClusters()* function, where cells were grouped into clusters based solely on their gene expression profiles without the input of cell type classifications or experimental metadata. FindClusters is a shared nearest neighbours method which builds a graph over the data to determine which cluster each cell should be in based on how "similar" its neighbours are. It then recursively applies a modularity optimization technique based on the Louvain algorithm to recursively group cells, merging each of these nodes (cells) into communities (clusters). Clustering the data in this way enabled first the identification of cell populations captured in the experiment without investigator bias influencing the structure of the data, as well as facilitating dimensionality reduction, visualisation and the downstream analysis of the different expression signatures of cells within the dataset. Dimensionality reduction to visualise clusters calculated with a resolution = 1 was performed using uniform manifold approximation and projection (UMAP) (Figure 3.7). The gene expression signatures of the top most highly expressed genes per cluster were visualised using a heatmap (Figure 3.8).

Cells were grouped into 11 clusters, which were individually assigned into cell types by thorough analysis of gene lists of the top marker genes with positive expression within each cluster against all other clusters using previously published literature sources and expression atlases as references (Pronk *et al.*, 2007; Haas *et al.*, 2015; Paul *et al.*, 2015; Pietras *et al.*, 2015; Miyawaki *et al.*, 2017; Dahlin *et al.*, 2018). Cell-type annotations were assigned to clusters onto the integrated dataset, with both treatment conditions combined. This enabled grouping of cell types based on their expression signatures, irrespective of treatment condition. No dominating effect on clustering as a result of platelet depletion was observed, therefore confirming no cluster was composed of cells unique to one condition (Figure 3.7A).

The expression signature of cells in cluster 1 is consistent with a canonical immature HSPC signature with high expression of markers such as *Cd34,* and more HSC-specific *Esam,* and *Hlf* a key regulator in HSC quiescence (Ishibashi *et al.*, 2016; Komorowska *et al.*, 2017). Cluster 9 largely shares an overlapping expression of multiple genes highly expressed in cluster 1 suggesting it is also composed of HSCs, however with low expression of *CD34*, and highest levels of genes including *Procr, Sult1a1,* and *Mpl* - markers of rare and most primitive HSCs, it is consistent with well-established literature stipulating the *Cd34 low* HSC-subset as murine long-term multilineage reconstituting HSCs (LT-HSCs) (Balazs *et al.*, 2006; Gazit *et al.*, 2014; Ali *et al.*, 2017) (Figure 3.9).



Cell type	% control	% depleted	Total cells per cluster
HSC	21.5	22.2	203
MkP	13.4	11.4	110
MEP_Mk	5.7	14.2	110
CMP	13	10.4	102
Early_EryP	11	10.5	98
MPP-2	7	9.2	79
MEP_Ery	6	7	63
EryP	7	6.1	58
LTHSC	12.6	3.5	55
MP	1.2	3.4	26
GMP	1.6	1.9	17

Figure 3.7. Dimensionality reduction and clustering of single cells from control and platelet-depleted mice. (a) UMAP projection of integrated dataset where each point represents a single-cell, coloured based on treatment condition (b) UMAP projection coloured instead by clusters annotated by cell-type (c) Proportion of cells in each cluster per condition (total cells: control = 246, depleted = 675) (d) Numbers of cells per cluster and proportion from total cell numbers per condition.

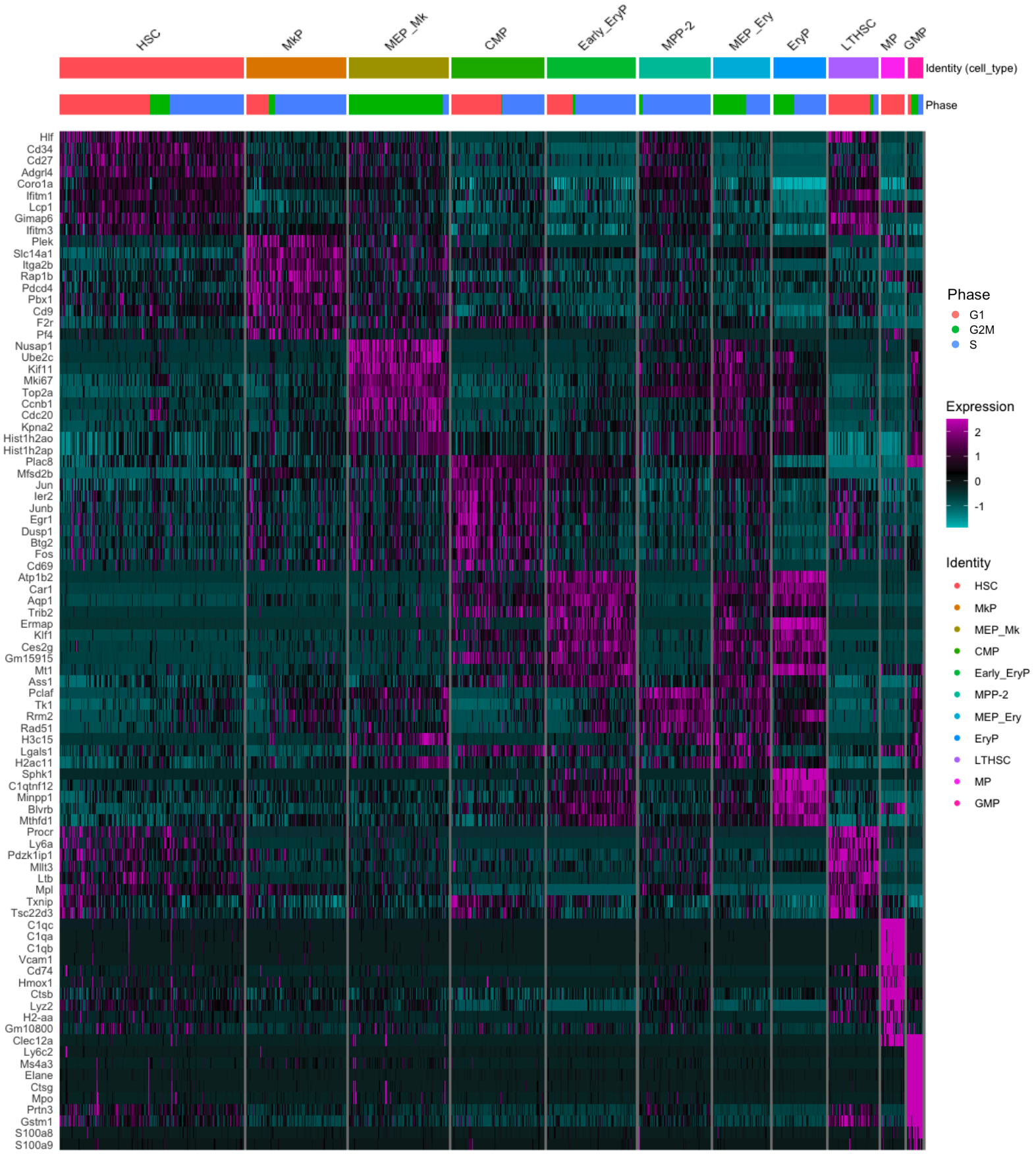


Figure 3.8. Heatmap showing the distribution of expression levels of the top 10 markers per cluster.

Cluster 2 contains MkP cells with high expression of genes canonical to Mk function including *Plek, Pf4 and Itga2b* (CD41), as well as *Mpl* and *Vwf* - notably also expressed in LT-HSCs (Paulus *et al.*, 2004; Chen, Hu and Shivdasani, 2007; Pronk *et al.*, 2007; Lambert *et al.*, 2009; Sanjuan-Pla *et al.*, 2013; Grover *et al.*, 2016) (Figure 3.9). Clusters 5 and 8 exhibit high expression of markers associated with erythrocytes, including "master" transcription factor in erythropoiesis *Gata1,* known to regulate all aspects of erythroid maturation and function at the transcriptional level (Gutiérrez *et al.*, 2020), as well as other erythroid cell markers *Klf1, Epor, Rhd (Siatecka and Bieker, 2011; Watowich, 2011)*. The two clusters are separated based on differential expression of genes related to cell-cycle stage, and genes associated with distinct Ery progenitor maturity stages where cluster 8 contains more cells in the G2M cell cycle phase and high expression in markers for lineage committed erythroid progenitors (Figure 3.10A).

Clusters 10 and 11 respectively showed strong signatures of monocyte-macrophage progenitors (MPs) and granulocyte-macrophage progenitors (GMPs). The three subunits of C1q are *C1qA, C1qB,* and *C1qC*components of the C1 complement activation complex and known mature macrophage markers, along with *Vcam 1* and *Mafb* - key regulators of macrophage are highly expressed in cluster 10 (Kishore and Reid, 2000; Yang *et al.*, 2022). Classic markers for the more immature GMPs *Mpo, Elane,* and *Clec12a* are most highly expressed in cluster 11 (Figure 3.10B and 3.10C).

Altogether there were more cells from depleted than control samples that passed QC (675 and 246 cells respectively). A comparison of cell-type abundance per condition was achieved by calculating the distribution of cells for each condition across clusters as percentages (Figure 3.7C). Cells from both treatment conditions were identified in every cluster, with a few notable differences. A lower proportion of LT-HSCs were captured from platelet-depleted mice than control. Also, platelet-depletion samples contained higher numbers of cells in the MPP-2 and Mk-MEP clusters (Figure 3.7D). These results suggest that platelet depletion induces LT-HSC exit from the HSC compartment, and the expansion of MPP-2 and Mk-MEPs for the differentiation of Mks and ultimately the rescue of platelet levels in the blood.

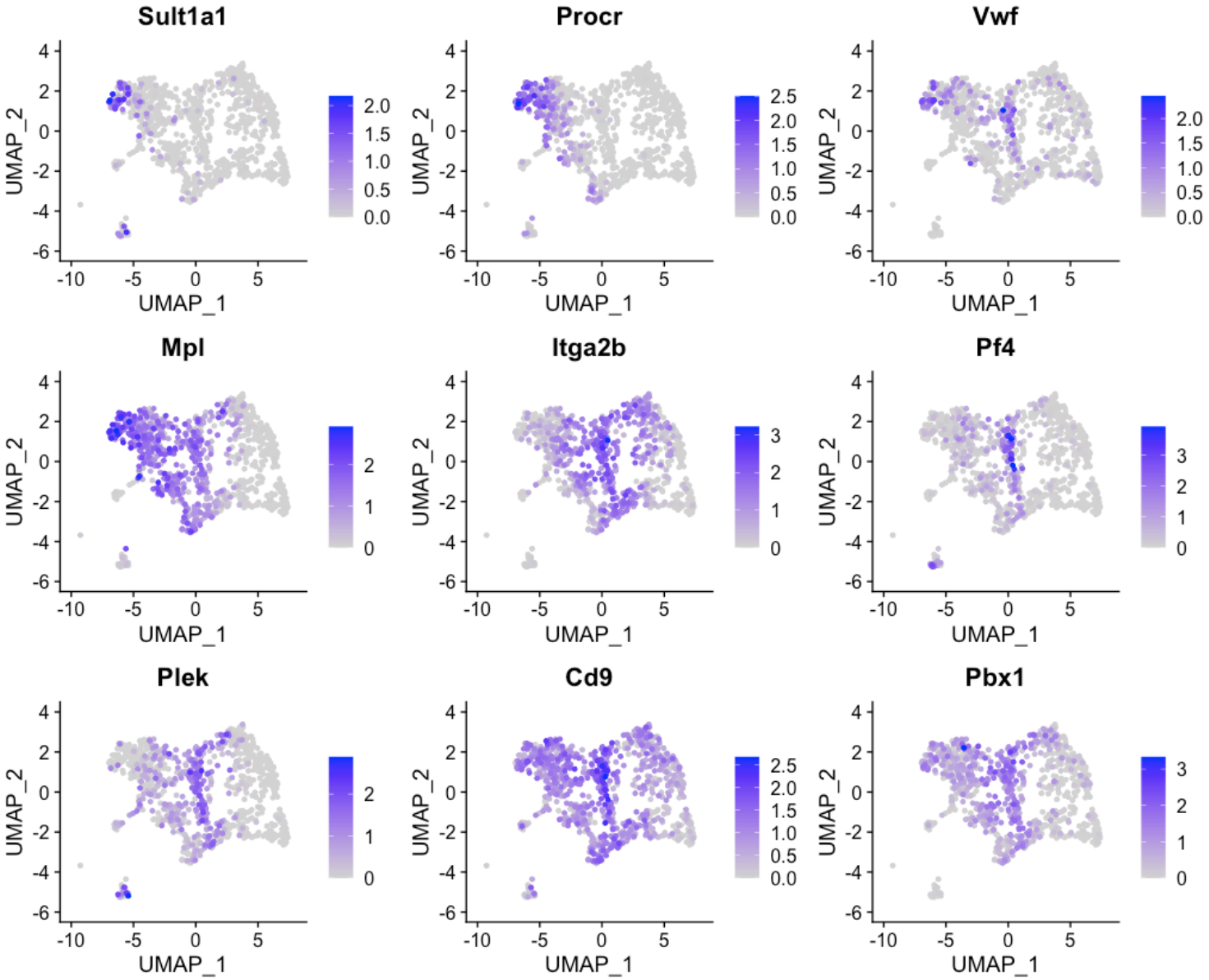


Figure 3.9. Feature expression of markers associated with Mks and HSCs.

Clusters 3 and 7 are also cell populations where genes associated with the G2M cell-cycle phase acted as principal components governing the clustering. This is evident from the high expression of, for example, DNA topoisomerase II alpha (*Top2a),* essential in governing topological states of DNA during transcription and Ubiquitin-conjugating enzyme E2C (*Ube2c)* known for its required role in the destruction of mitotic cyclins during cell cycle progression - both highlighted as positive markers (Figure 3.11). To explore further the relationship of cell-cycle phase across this dataset, the proportion of cells classified as either being in G2M, S or G1 cell cycle phases were compared across all clusters (Figure 3.11). This confirmed that cluster 3 was indeed primarily composed of cells in G2M, undergoing cell division, explaining the high saturation of DNA replication and cell cycle progression genes listed as cluster 3 markers (*Prr11, Pimreg, Nusap1, Hmmr,* and *Mki67)*. Further inspection of cluster 3 markers, aside from those related to cell cycle, shows a positive expression of Mk-associated genes including *Plek, Mpl, Gfi1b,* and *Itga2b*. Cluster 7 in comparison expressed higher levels of Ery-associated markers such as both *Gata1* and *Gata2 and Epor*. Clusters 3 and 7 best recapitulate the expression of, and were annotated as, MEP progenitors which were sub-clustered whereby cluster 3 has higher expression of Mk-lineage genes (Mk-MEP) and 7 higher expression of Ery-lineage genes (MEP Ery) respectively.

Though successful isolation of haematopoietic progenitors with bi-potency towards exclusively Mk and Ery lineages has been demonstrated (Manz *et al.*, 2002; Pronk *et al.*, 2007; Klimchenko *et al.*, 2009; Sanada *et al.*, 2016), gaps in our understanding of the properties of these cells remain and whether these are sufficient to warrant designation as a unique cell-type, or whether these MEPs exist only as a transient state of differentiation. This is enforced by the difficulty of the isolation of MEPs based on a cell-surface repertoire and the challenge of identifying a unique gene expression signature (Xavier-Ferrucio and Krause, 2018). It is, however, both consistent that MEPs can be found within the LSK Cd150+ BM fraction isolated in this experiment, and the expression patterns observed here correlate with previous work studying MEP lineage output (Psaila *et al.*, 2016). Moreover, both exhibit limited expression of myeloid-associated genes (e.g. *Mpo, Elane*) suggesting the cells captured here are specifically primed for either Mk-specific or Ery-specific gene expression (Leonard *et al.,* 1993; Tsai *et al.,* 1994; Lira and Friedman, 1997; Osawa, 2002). When testing clustering of data using a lower resolution, clusters 3 and 7 were grouped into a single population, having a signature most closely resembling CMPs, suggesting that these populations share a similar pattern of expression with one another and have a more immature expression pattern than traditional unipotent progenitors (Notta *et al.*, 2016). Altogether, these data suggest these cells express transcriptomic signatures consistent with our current knowledge of MEPs.

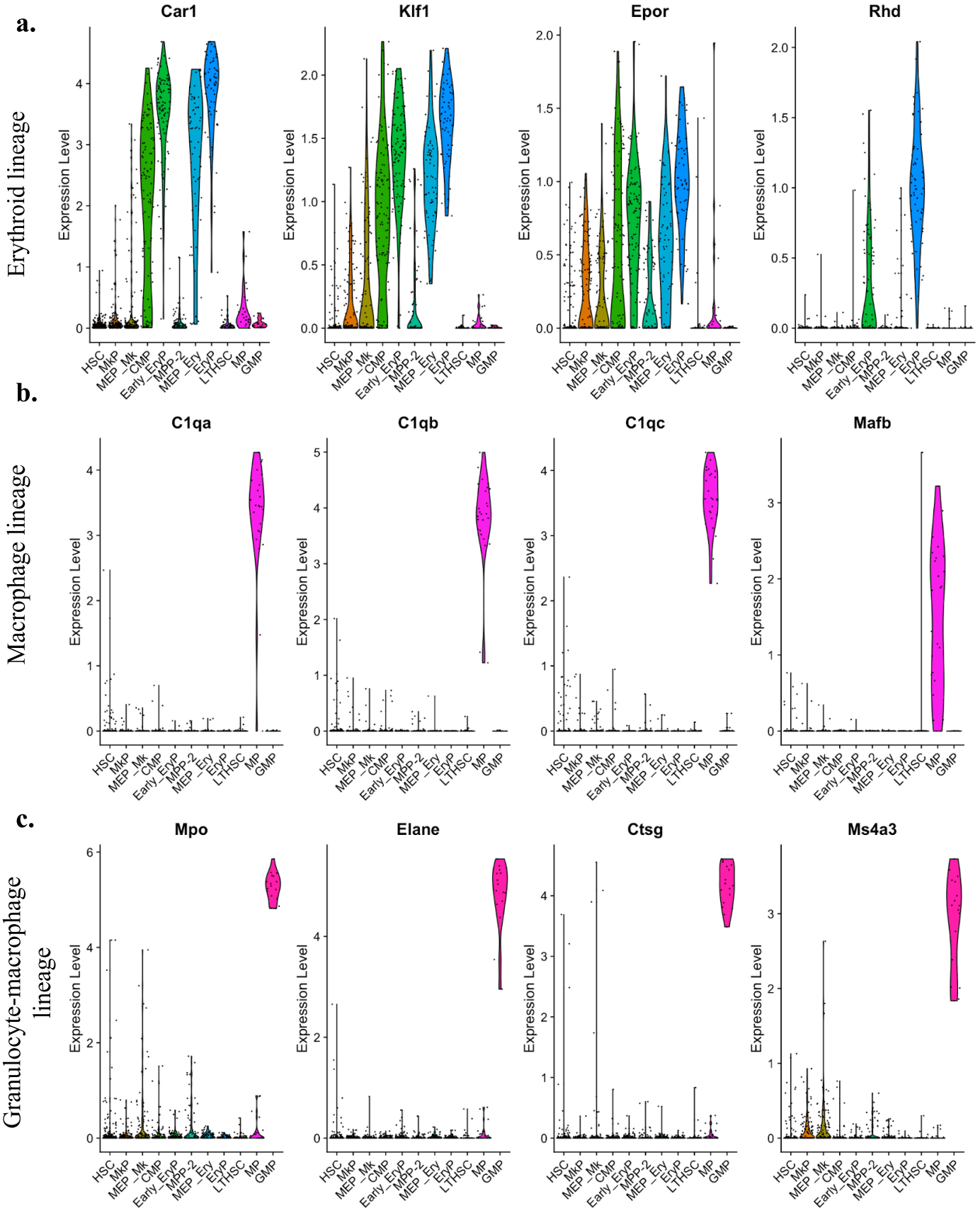


Figure 3.10. Violin expression plots of example canonical markers. (a) erythroid (b) macrophage-monocyte and (c) granulocyte-macrophage progenitors.

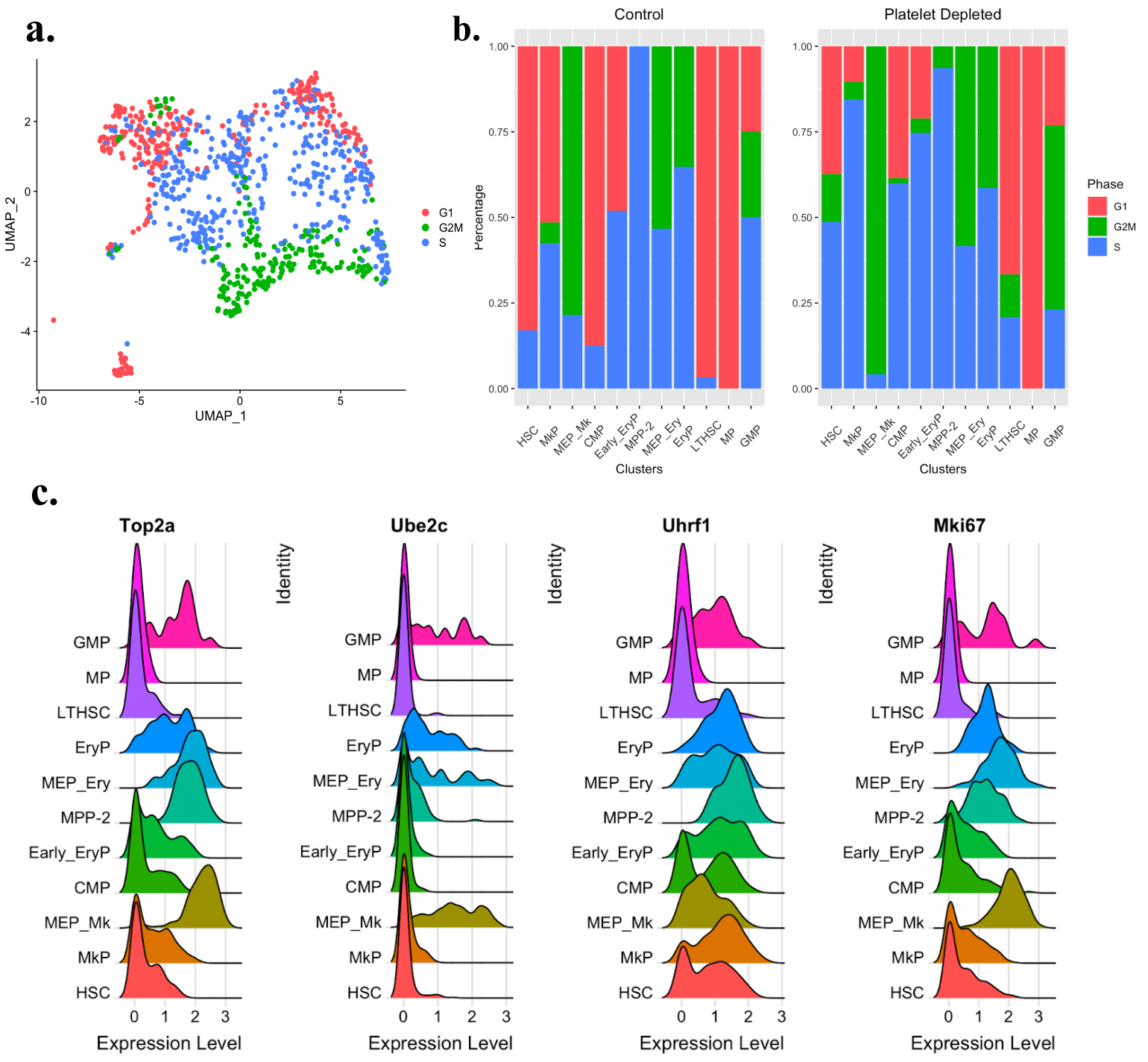


Figure 3.11. Cell-cycle phase analysis and pattern of expression of G2M and S phase markers across clusters. (a) UMAP projection of single cells captured coloured by cell cycle stage **(b)** Proportion of cells in each cell cycle stage per cluster from control (left) and platelet depleted (right) samples **(c)** Expression levels of four cell-cycle associated markers across all clusters.

3.3.5 Transcriptional ordering along pseudotime reveals a differentiation continuum

Clustering is a useful tool to characterise gene expression patterns within a dataset inherently by forcing high-dimensional data from many cells into discrete groups, enabling differential expression analyses between cell populations and across 'meta' parameters including in this case treatment condition. However, clusters do not adequately represent the differentiation continuum occurring in haematopoiesis as the point of differentiation is not taken into account. Particularly in haematopoiesis, critical differences in gene expression between cells can be explained by the temporal position they occupy along the process of lineage commitment. For this reason, tools for trajectory analysis inference from 'snapshot' scRNA-seq data were developed to order cells along a pseudotime metric, recapitulating the dynamic process in question by providing temporal resolution to expression measurements from single cells.

Since 2015, more than 50 algorithms for trajectory inference from scRNA-seq data have been published (Saelens *et al.*, 2019). One of these approaches, packaged in the tool *Monocle*, uses an algorithm to learn the sequence of gene expression changes each must go through as part of a dynamic process, and using the overall trajectory of gene expression changes projects cells in order (Trapnell *et al.*, 2014). Briefly, *Monocle* first employs a differential expression test, similar to *FindVariableFeatures()* of *Seurat,* to identify a number of genes that vary significantly across cells in the dataset. Then it applies independent component analysis for dimensionality reduction, computing a minimum spanning tree to build a trajectory by finding the longest connected path in that tree. Monocle then projects each cell onto the nearest point to them along the longest connected path of the tree, where in case of multiple 'end points' a branched trajectory representing diverging trajectories are constructed (Qiu *et al.*, 2017).

Whilst a wide variety of trajectory inference tools are available, *Monocle* is well-established as one of the methods best suited for pseudotime ordering complex trajectories with multi-branching topologies, and includes functions for differential gene expression analyses along trajectories across multiple criteria (Saelens *et al.*, 2019). To infer a trajectory from this single-cell RNAseq dataset, pseudotime analysis was performed using *Monocle3,* the latest version released by the Trapnell lab. As cells were assigned into clusters using Seurat, the dataset from *Seurat* object was directly converted into the *Monocle* equivalent cell data set (cds), preserving cell-type annotations and UMAP cell embeddings per cell to recreate the same partitions for pseudotime analysis as those presented in section 3.3.4. Using these partitions as input, a graph was learned over the projected cells, creating a path of connecting points across clusters enabling pseudotime ordering. To calculate cell-wise pseudotime the cds was parsed into *order_cells()* calculating a numeric value per cell based on the position in which it lies along pseudospace. This was performed semi-supervised, where the only parameter used for pseudotime ordering was setting the root point of the trajectory as cells in the LT-HSC cluster.

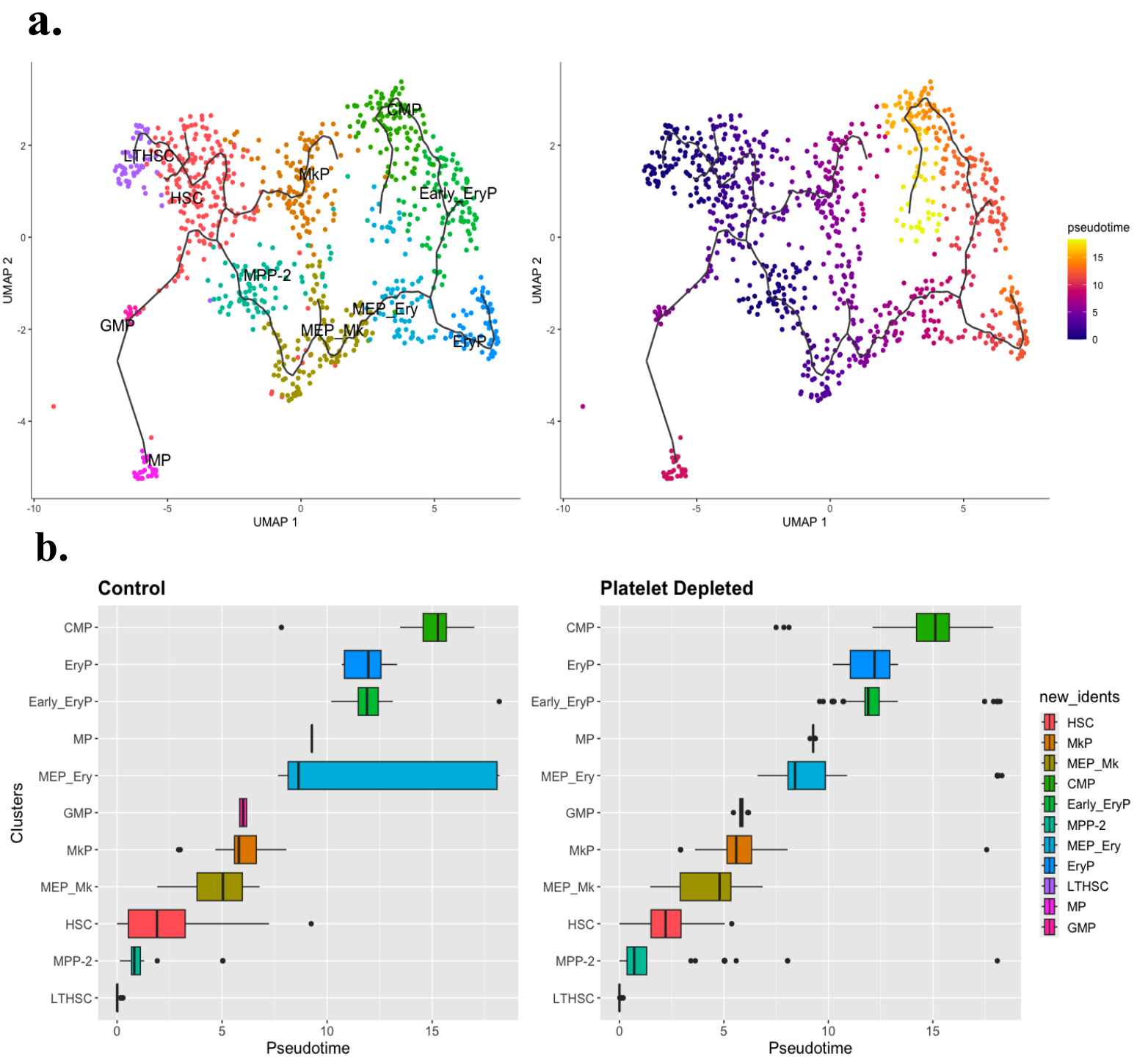


Figure 3.12. Pseudotime analysis of single-cells (a) UMAP projection of single-cells showing the trajectory learned over dataset (black) coloured by cluster (left) and pseudotime values (right). **(b)** Median pseudotime values per cluster along pseudospace per treatment condition.

Pseudotime analysis ordered cells along a trajectory from LT-HSCs to Mk and Ery progenitors, with LT-HSCs having the lowest and EryPs the highest pseudotime values (Figure 3.12). To visualise the order and range of pseudotime values assigned to each cell in all clusters, the median pseudotime value per cluster was plotted across pseudotime (Figure 3.12B), showing the successive order of clusters within the pseudospace. This order corroborates the existing literature suggesting that commitment to the Mk lineage occurs at earlier stages of haematopoiesis as compared to other mature blood cells including the Ery lineage (Haas *et al.*, 2015; Grover *et al.*, 2016; Miyawaki *et al.*, 2017; Psaila and Mead, 2019). This also revealed a small discrepancy between control vs platelet depleted samples, where MkP cells from the depleted object were assigned into an earlier slot along pseudotime than MkPs from control mice.

3.3.6 Differential expression along pseudotime reveals expression dynamics of genes implicated in Mk function

For unbiassed identification of genes within this dataset that are differentially expressed across the constructed trajectory, the *Monocle3* graph_test was applied which draws on the Moran's *I* statistical test, a technique in spatial correlation analysis (Moran, 1950). Moran's *I* is a measure of multi-directional and multi-dimensional spatial autocorrelation which describes whether cells at nearby positions on a trajectory will have similar (or dissimilar) expression levels for all genes tested. The goal of this analysis was to extract genes that vary significantly in expression across cells from low to high pseudotime values, in order to confirm our data recapitulates well-established gene signatures of Mk lineage restriction, but also potentially identify novel genes involved in lineage specification towards Mks. This identified 3,212 genes that vary along pseudotime across the dataset. Many genes identified in this analysis corroborate with existing literature to vary in expression during HSC differentiation along the Mk-Ery lineage. For example, *Mpl, Mecom, Esam, Socs2* and *Hlf* were all identified as significantly differentially expressed early in the trajectory; these are well-established markers for HSCs and only expressed in the most primitive HSPC compartment (Balazs *et al.*, 2006; Kustikova *et al.*, 2006; Qian *et al.*, 2007; Yokota *et al.*, 2009; Balenci *et al.*, 2013). Correspondingly, in late pseudospace, lineage-specific markers were identified as differentially expressed. This includes Ery-specific *Klf1, Epor* and *Car1,* myeloid/ monocyte-specific *Plac8* and *Apoe* and Mk-specific *Itga2b* and *Gp5* - all enriched in cells with high pseudotime values as expected (Figure 3.12) (Debili *et al.*, 2001; Hodge *et al.*, 2006; Klimchenko *et al.*, 2009; Reddi and Belibasakis, 2012; Song *et al.*, 2012).

Conducting differential expression analysis on the entire dataset revealed the global changes along the entire tree end-to-end. However, with MkPs having lower pseudotime values than the EryP, GMP and MP populations captured in this experiment, markers for these populations

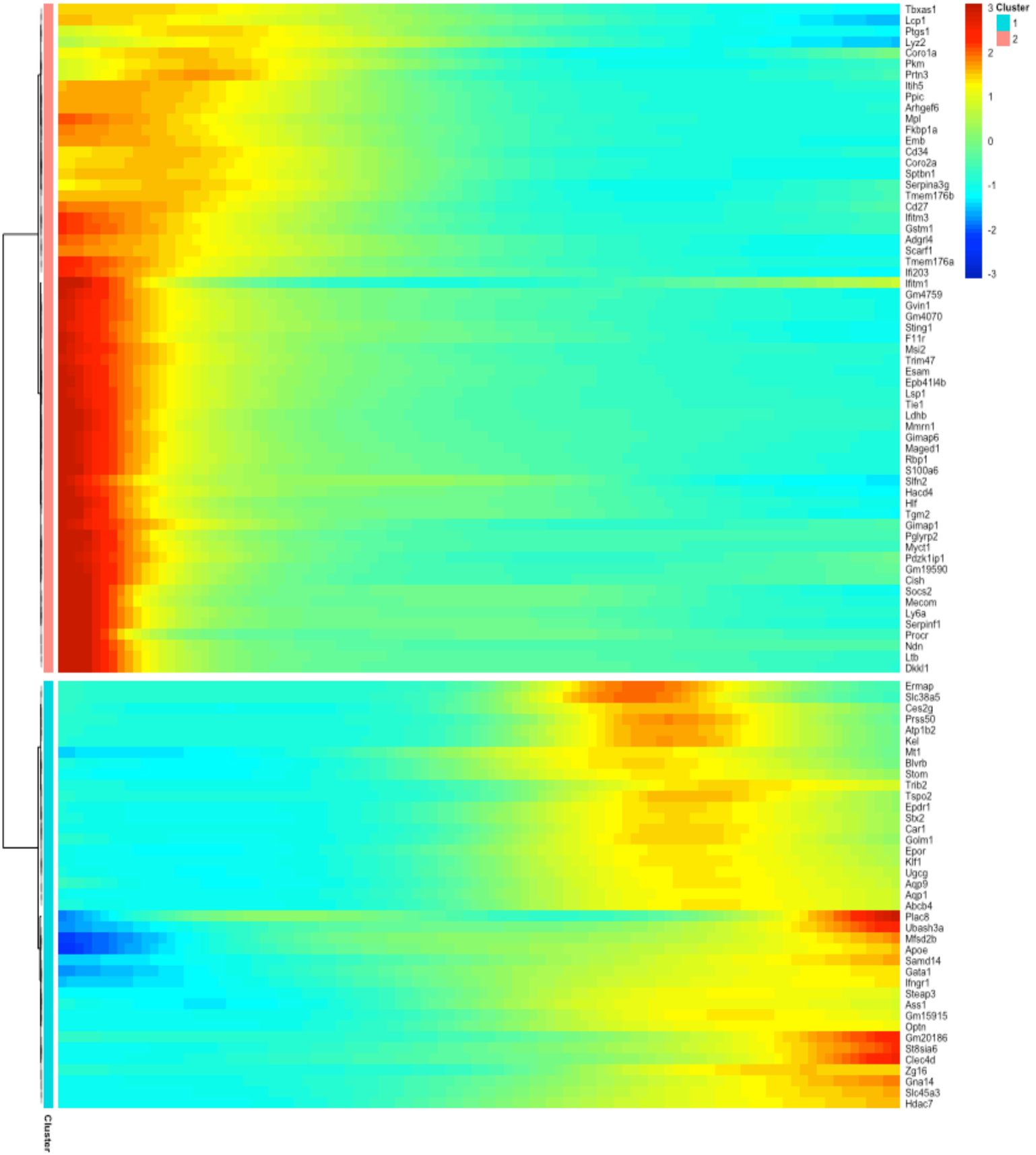


Figure 3.13. Heatmap of top 100 genes differentially expressed along pseudotime. Top genes were selected from those which were statistically significant (sig q < 0.05).

characterised as further along in pseudotime were dominating the list of genes returned as differentially expressed, consequently complicating the changes along the Mk-lineage from being resolved with this approach. This is emphasised in Figure 3.13, where the heatmap is clearly clustered into two compartments based on the two extremes of the pseudotime scale, with comparatively lower detection of genes along intermediate pseudotime states - largely composed of the MkP cells of interest (Figure 3.12). To specifically interrogate the dynamics in the expression of genes along the Mk trajectory, the dataset was subset to include only cells along the Mk lineage. This was achieved by using *Monocle's choose_cells()* function, which enables the specific selection of cells into a new object for further analysis, where only LT-HSCs, HSCs, MPP-2, Mk-MEP and MkP cells were included (Figure 3.14).

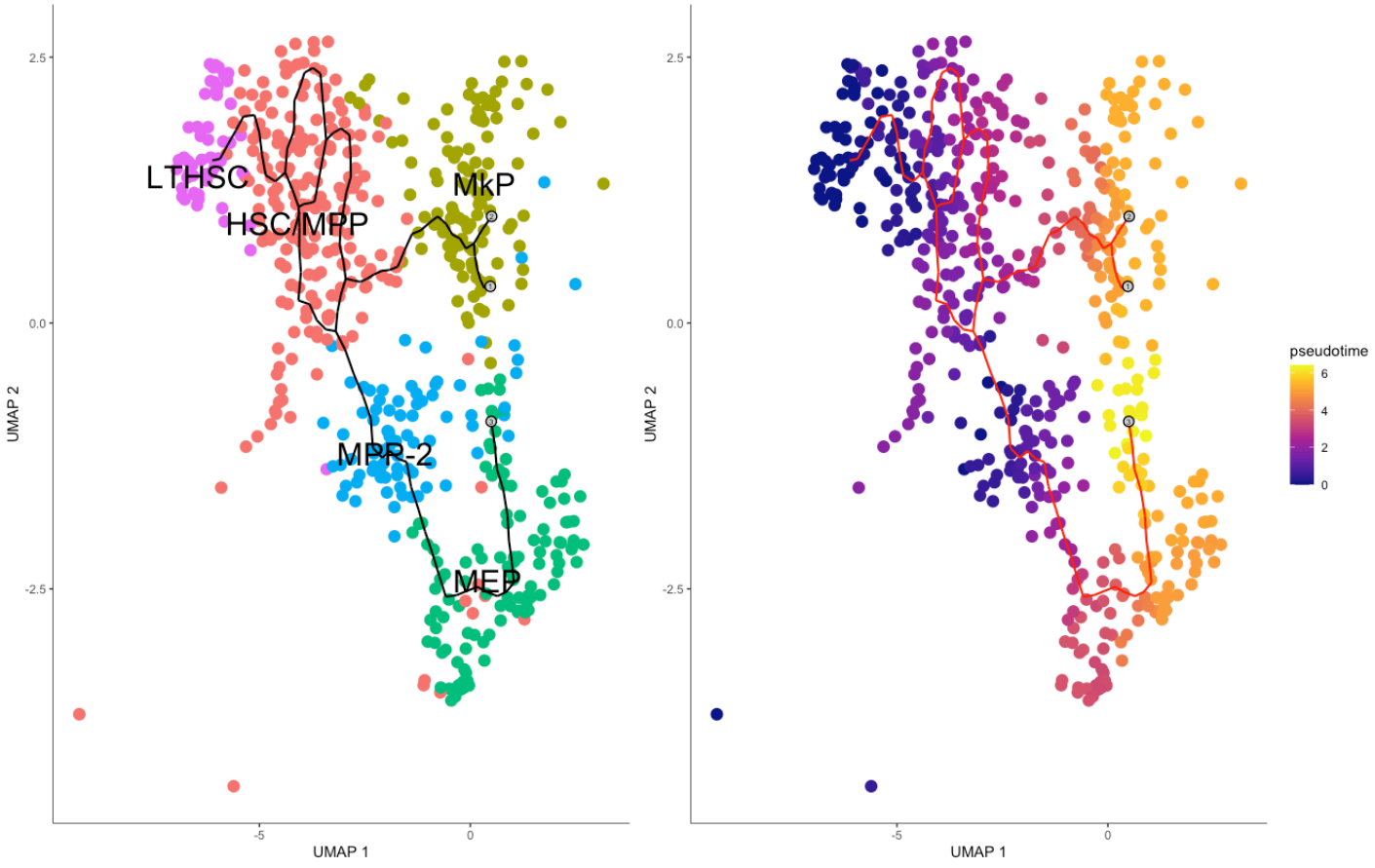


Figure 3.14. Subset excluding cells outside of the Mk trajectory to enable differential expression with pseudotime. Cell selection was performed using *Monocle3* choose_cells() excluding cells of the Ery lineage. This approach isolated the Mk lineage so analyses of differential gene expression along pseudotime reflected only differential expression within a single lineage.

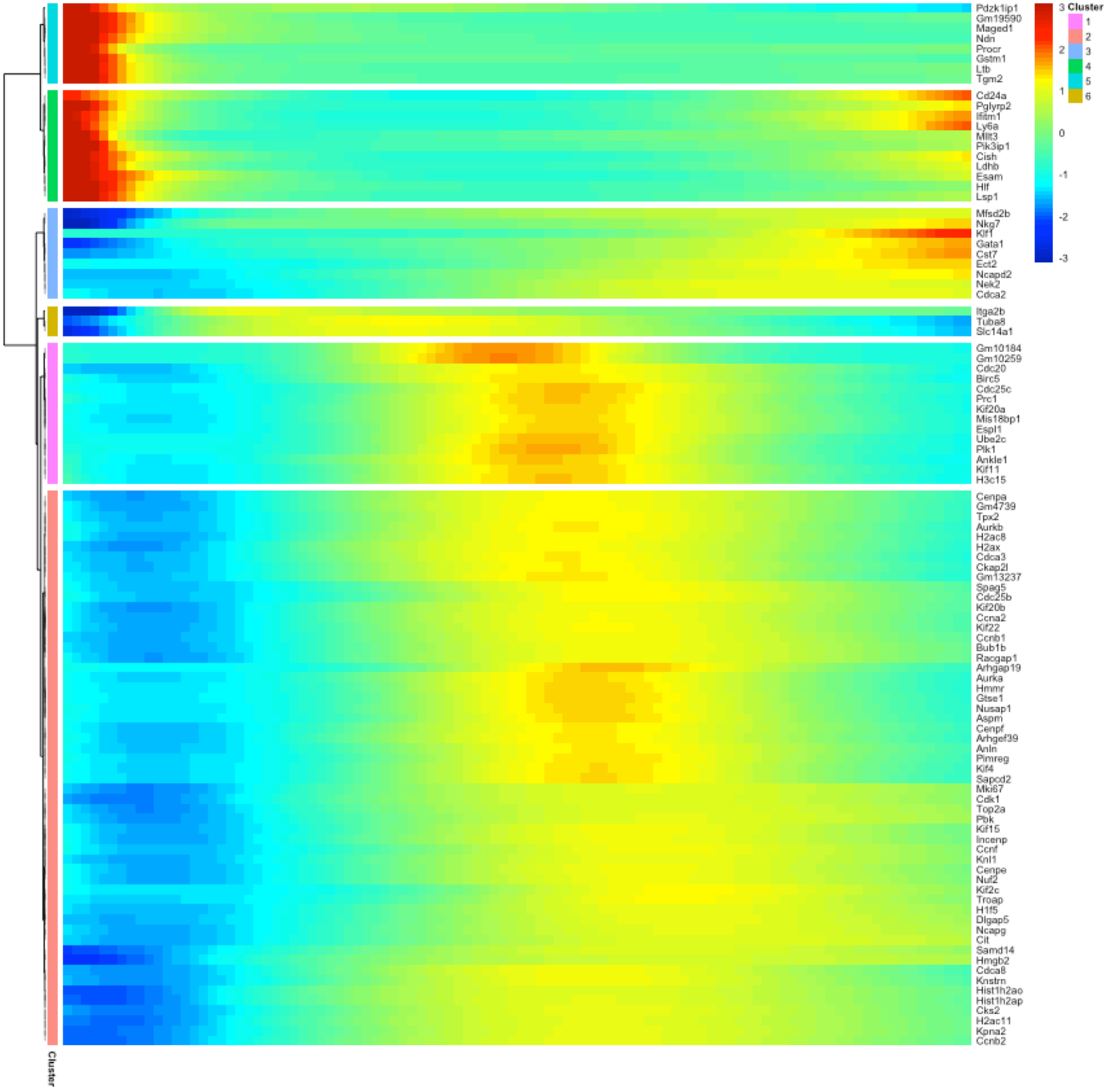


Figure 3.15. Heatmap of top 100 genes differentially expressed along Mk trajectory. Top genes were selected from those which were statistically significant (sig q < 0.05).

Differential expression genes (DEGs) as a function of the Mk pseudotime trajectory calculated genes' spatial autocorrelation, assessing the similarity or dissimilarity of neighbouring cells in terms of gene expression. Key markers of HSCs and committed-MkPs were among the most statistically significant genes with variable expression. A set of 830 genes exhibited a significant correlation with pseudotime, as indicated by adjusted p-values below 0.05. Inspection of the significant genes list, along with studying the distribution of their expression patterns across the cells, and existing gene annotations suggested the vast majority of genes broadly could qualitatively fit one of three main categories: 1. Genes with implications in HSC function; 2. Genes associated with cell cycle transition (particularly G2M/S cell cycle phase); and 3. Genes with implications in Mk function.

Plotting the top 100 most variable genes along pseudotime along the Mk trajectory shows a subset of these genes, and revealed in total 3 'major' gene clusters, and a total of 6 subclustered gene sets (Figure 3.15). Clusters represent grouped genes that exhibit similar expression dynamics, allowing for easier interpretation and analysis. Genes within the same cluster can often have similar functions or are co-regulated, hence this may be indicative of potential functional relationships and shared regulatory mechanisms across genes.

Clusters 4 and 5 are enriched for genes highly expressed in cells with low pseudotime values ie. LTHSCs and HSCs. This includes for example *Nectin (Ndn),* which encodes for a multifunctional protein that plays an important role in restricting excessive HSC proliferation during haematopoietic regeneration (Kubota *et al.*, 2009)**.** Another example is *Pdzk1ip1*; an important gene encoding a membrane-associated protein shown to modulate the levels of reactive oxygen species and an important determinant of HSC function. Moreover, previous work has shown *Pdzk1ip1* is within the expression domain of *Scl* with known implications in Mk maturation, sharing transcriptional enhancer elements (Pimanda *et al.*, 2007; Tijssen *et al.*, 2011).

Conversely, clusters 3 and 6 are enriched with genes that are lowly expressed at the beginning of the pseudotime trajectory. In particular, cluster 3 is instead composed of genes expressed in Mk-Ery progenitors such as *Gata1, Klf1 and Gata2* (Iturri *et al.,* 2021), while 6 grouped known MkP genes including *Itga2b,* and thrombin-receptors of the protease-activated receptor (PAR) family *F2r* and *F2rl2* (Figure 3.15) (Sun *et al.*, 2013) .

The remaining two subclustered sets of genes that correlate with pseudotime progression are highly composed of genes associated with cell proliferation (Figure 3.15). This includes canonical G2M/S phase transition genes such as *Ube2c* and *Birc5,* as well as *Ccnd1 and Ccne2;*

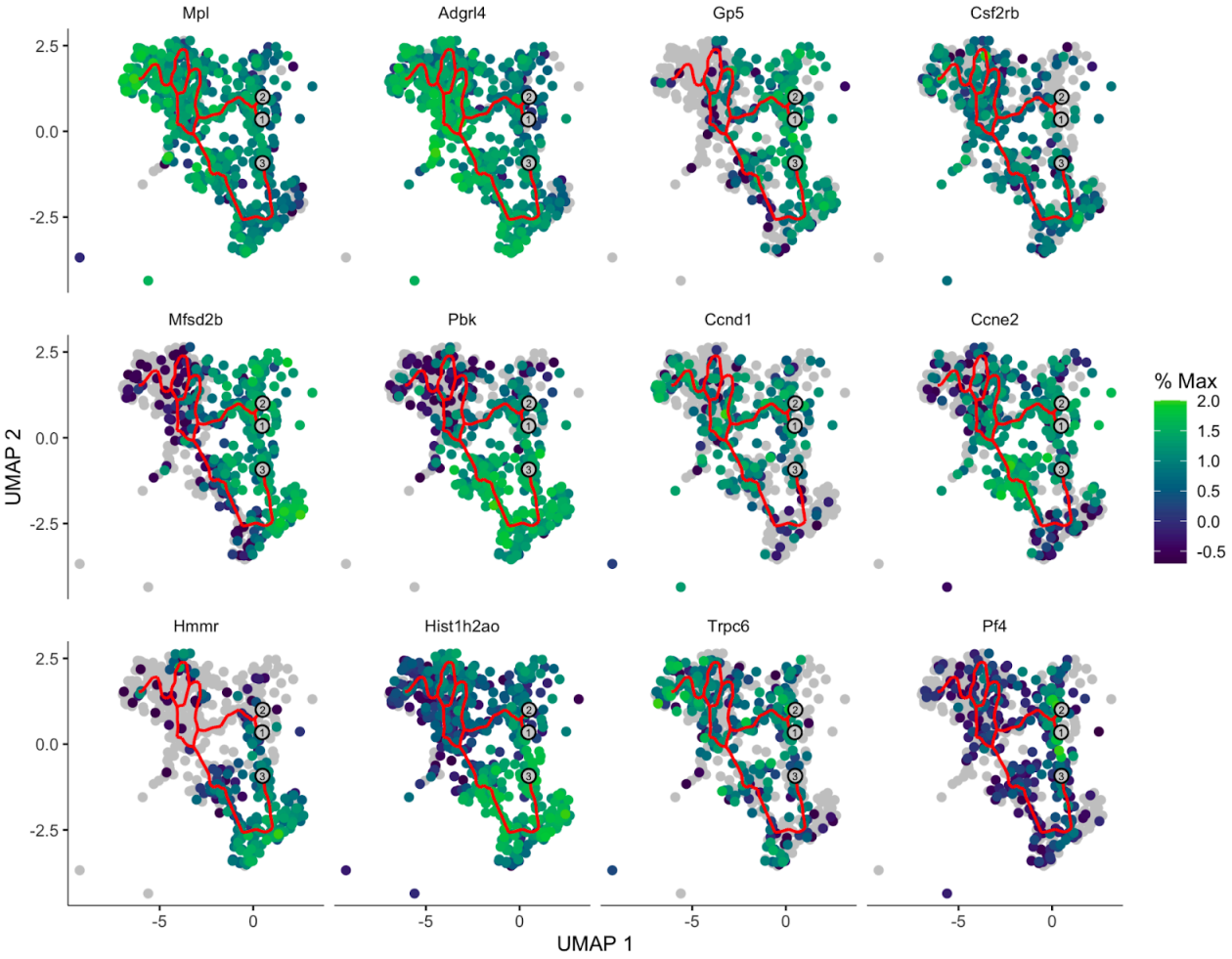


Figure 3.16. Dynamics of expression levels of genes that vary along pseudotime in Mk commitment**.**

which have additionally also been implicated in promoting endomitosis (repeated rounds of DNA synthesis without cell division) during megakaryopoiesis (Geng *et al.*, 2003; Muntean *et al.*, 2007). Mks are unique from all other blood cell types as committed MkPs proliferate to a relatively limited extent to give rise to colonies of Mks. Cells then undergo terminal differentiation where they bypass the latter stages of mitosis to increase their DNA content and size. Previous work has demonstrated TPO-induced Mk differentiation is in part linked to up-regulation of cyclin D1 (*Ccnd1)*, cyclin D2 (*Ccnd2)*, and cyclin D3 *(Ccnd3)*, where *Ccnd2* overexpression specifically was found to facilitate Mk differentiation even in the absence of TPO (Matsumura *et al.*, 2000). Similarly, E type cyclins (E1 and E2) are believed to drive cell entry into the S phase and are thus required for proliferation for many cell types. Mk endoreplication was found to be severely impaired in the absence of cyclin E (*Ccne),* where Cyclin E1 and E2 KO Mks exhibited significantly reduced ploidy profiles compared to WT Mks (Geng *et al.*, 2003; Eliades, Papadantonakis and Ravid, 2010). It is thought cyclins may

mediate their effects by promoting the expression of components of the pre-replication complex (Eliades, Papadantonakis and Ravid, 2010), and that cyclin expression is likely controlled by essential Mk-promoting TFs, such as GATA-1 however these mechanisms remain to be fully elucidated (Muntean *et al.,* 2007). This data shows an upregulation of cyclin genes with pseudotime, which is expressed from MPP-2s, Mk-MEPs to MkPs (Figure 3.16), and may suggest increased expression of cyclin genes contributes to the expansion of the Mk lineage, potentially acting as positive regulators of endomitosis.

Moreover, cell-cycle-associated-genes and cyclin expression across single cells is correlated with other important components involved with cell cycle regulation, spindle organisation, and chromosome segregation. This includes examples such as spindle assembly factor Hyaluronan Mediated Motility Receptor (*Hmmr)* (He *et al.,* 2020), protein kinases Aurora kinases A and B *(Aurka, Aurkb)* (Geddis and Kaushansky, 2004; Goldenson *et al.,* 2015), as well as histone gene family proteins *(Hist1h2ao and Hist1h2ap)* that are important for DNA packaging and nuclear organisation (Figure 3.16). Together the overall signature of this gene set is enriched for cell proliferation indicators, including both well established markers and co-expressed genes that have not previously been explored in the context of Mk differentiation.

To establish the expression of pseudotime-DEGs in the context of cell-type annotations, the expression levels of known and novel genes that exhibit dynamic expression during megakaryopoiesis was plotted across cells' *Seurat* annotations (Figure 3.17). This step confirmed the concordance between cell type annotations, *Monocle3* pseudotime state annotations and the available existing literature that have identified variable expression in certain genes during Mk differentiation. TPO receptor *Mpl* is most highly expressed in LTHSCs. It is co-expressed with *Trpc6* and *Ptk7* both of which are highly expressed exclusively in the earliest pseudotime states and have previously been implicated in megakaryocyte commitment. *Trpc6* is present on the platelet membrane and is thought to participate in calcium influx during platelet activation (Hassock *et al.*, 2002). Its' expression during the differentiation of Mks has been previously reported and has been proposed to play a role in the initiation and maintenance of TPO-induced calcium-mediated differentiation (Carter *et al.*, 2006; Ramanathan and Mannhalter, 2016). The *Ptk7* tyrosine kinase receptor on the other hand is a is a planar cell polarity receptor belonging to the Ig superfamily found largely on endothelial cells (Lhoumeau *et al.*, 2016), but has also been identified on the cell surface of human HSPCs and blast cells from patients with AML where it confers a poor prognosis to patients independently of other risk factors (Prebet *et al.*, 2010).

Another surface receptor associated with endothelial cells was also identified; *Adgrl4* is a G protein-coupled receptor that plays a role in angiogenesis and promotes tumour growth and metastasis (Schiöth and Fredriksson, 2005). Discovered in 2001, very little about the gene's function and its mechanism of activation has been elucidated. In 2019, *Adgrl4* silencing was found to regulate endothelial cell metabolism by suppressing the mitochondrial gene *SLC25A1*, as well as inducing cKit upregulation - where the authors suggest it may serve to maintain an equilibrium in endothelial metabolism and homeostasis (Favara *et al.*, 2019). *Adgrl4* has the potential to serve as a treatment target for multiple cancers, as it is frequently dysregulated in tumour-associated endothelial cells of patients with renal carcinoma, as well as other malignant and non-malignant diseases. Its aberrant expression correlates to tumour invasiveness (Kan *et al.*, 2018), tumour angiogenesis, as well as renal thrombotic microangiopathy amongst other known phenotypes (Niinivirta *et al.*, 2020).

Here, *Adgrl4* expression was revealed to be strongly correlated with Mk-lineage associated cell-types- from LTHSCs to MkPs (Figures 3.17 and Appendix Supplementary Figure 3.5). Literature connecting this gene to haematopoietic cell function / differentiation could not be identified. With important interactions between Mks and endothelial cells in other areas well documented, it is plausible to speculate this gene may represent another feature in common between the cell types that is yet to be explored. For instance, thrombotic microangiopathy (TMA) is a condition whereby endothelial injury and associated platelet activation contribute to microvascular thrombosis formation, tissue ischemia, and subsequent end-organ injury (Genest *et al.*, 2023). Most cases are caused by ADAMTS13 deficiency which results in accumulation of ultralarge vWF multimers, followed by widespread platelet aggregation and thrombosis (Noone *et al.*, 2016; Bettoni *et al.*, 2017). It can also be triggered by infections, autoimmune reactions, and other causes (Genest *et al.*, 2023). This is a pathological example of dysfunctional endothelial and thrombotic function that can lead to clinical features of microangiopathic hemolytic anaemia, thrombocytopenia, and ischemic end-organ injury (George and Nester, 2014). The identification of this gene along cells of the Mk lineage may represent an interesting avenue for further investigation.

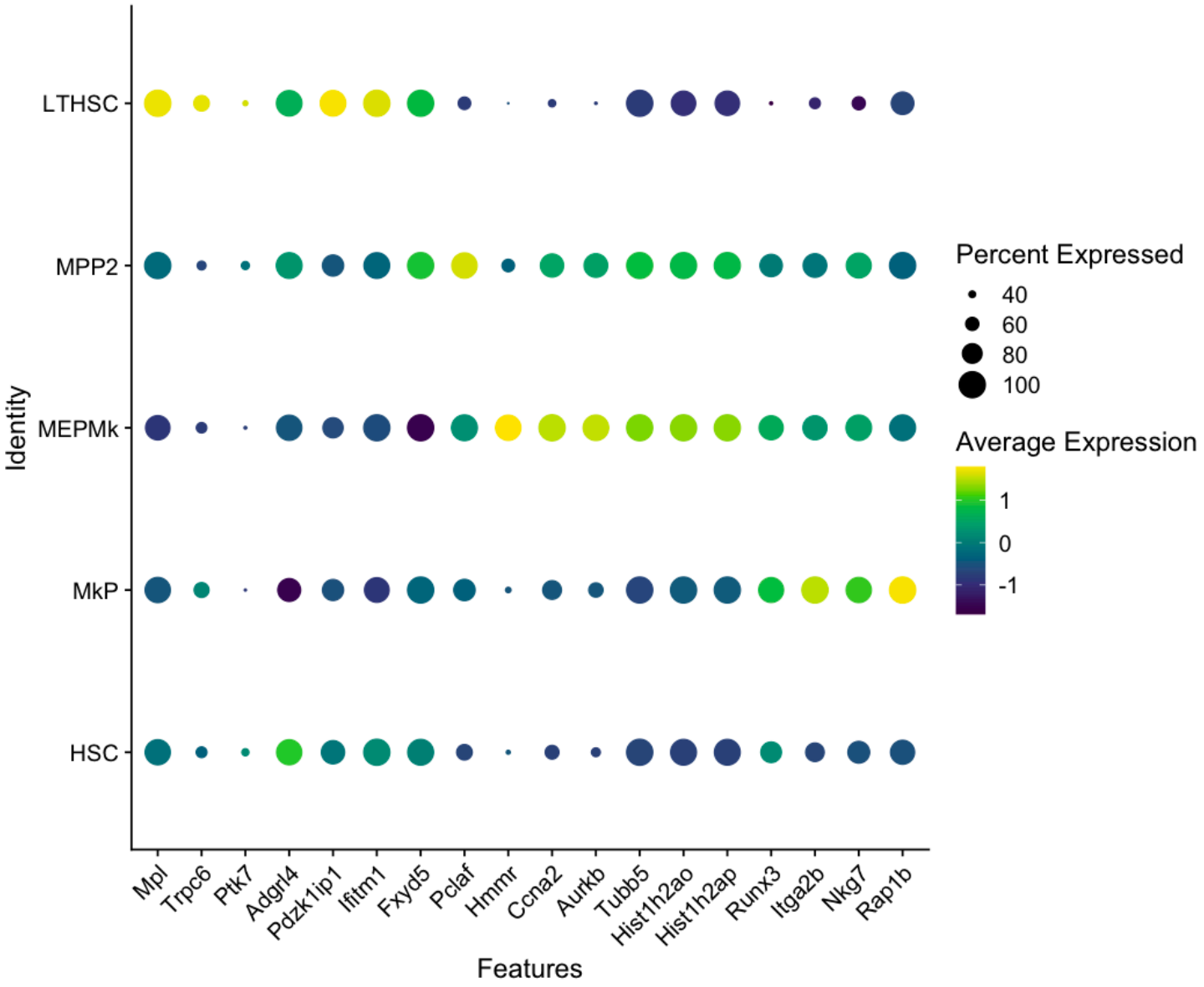


Figure 3.17. Average expression dotplot in a subset of genes identified as significantly enriched across distinct pseudotime states. Rows correspond to the annotated cell types

3.3.7 Pseudo-bulk differential expression analysis within clusters identifies significantly differentially expressed genes caused by platelet depletion

Single-cell specific tools for differential expression (DE) such as Seurat's *FindMarkers,* although useful for cell type annotation of clusters, often result in inflated p-values as each cell is treated as a sample. However, single cells within each sample are not independent of each other, and methods that ignore biological variation between biological replicates lead to biassed results and are prone to false discoveries. Failing to account for the intrinsic biological variation between biological replicates increases the probability of false discoveries in the presence of a real biological perturbation, leading to confounded results (see Appendix Supplementary Figure 3.4). Moreover, single-cell DE methods have a systematic tendency towards highly expressed genes, identifying highly expressed genes as DE even when their expressions were unchanged (Squair *et al.*, 2021).

To identify DE across cell types after platelet depletion, DE analyses were performed by aggregating single cell counts in pseudo-bulk replicates both by biological replicate (mouse ID) and cell type annotations (clusters). This was achieved by subsetting cells from the dataset by the cell type(s) of interest to perform the DE analysis within clusters, and using mouse ID metadata to split samples into their respective conditions and biological replicates from which gene counts were aggregated. DE was performed using the DESeq2 package (Love, Huber and Anders, 2014)**.** Functional analysis of DEGS between conditions was performed using gene set enrichment analysis (GSEA) with ReactomePA() and clusterProfiler() (Yu and He, 2016; Wu *et al.*, 2021). GSEA determines whether DEGS or pathways are overrepresented between conditions, using a background gene set of all expressed genes in the dataset.

DE in LTHSCs between treatment conditions revealed a total of 1348 DEGs, of which 186 were statistically significant (*adjusted P-values* <= 0.05) (Figure 3.18). To confirm pseudo-bulk DEGs could be seen at the single cell level, DEGS were also visualised in LTHSCs between the conditions, some of which are presented in Figure 3.18B.

DEGs and GSEA revealed that platelet-depleted LTHSCs had a marked increase in expression of genes associated with cell cycle progression, differentiation and DNA replication (Figure 3.18C-D). This includes examples such as: canonical proliferation markers like *Mki67* (Uxa *et al.,* 2021); kinetochore-associated proteins *(Knstrn)* that promote chromosome segregation during mitosis (Deng *et al.*, 2021)*;* DNA unwinding proteins e.g. *Helq (*Anand *et al.,* 2021) and epigenetic co-ordinators like *Uhrf1* which has previously been shown to be upregulated in proliferating cells and required for G1/S phase transition (Mousli *et al.*, 2003) (Figure 3.18B).

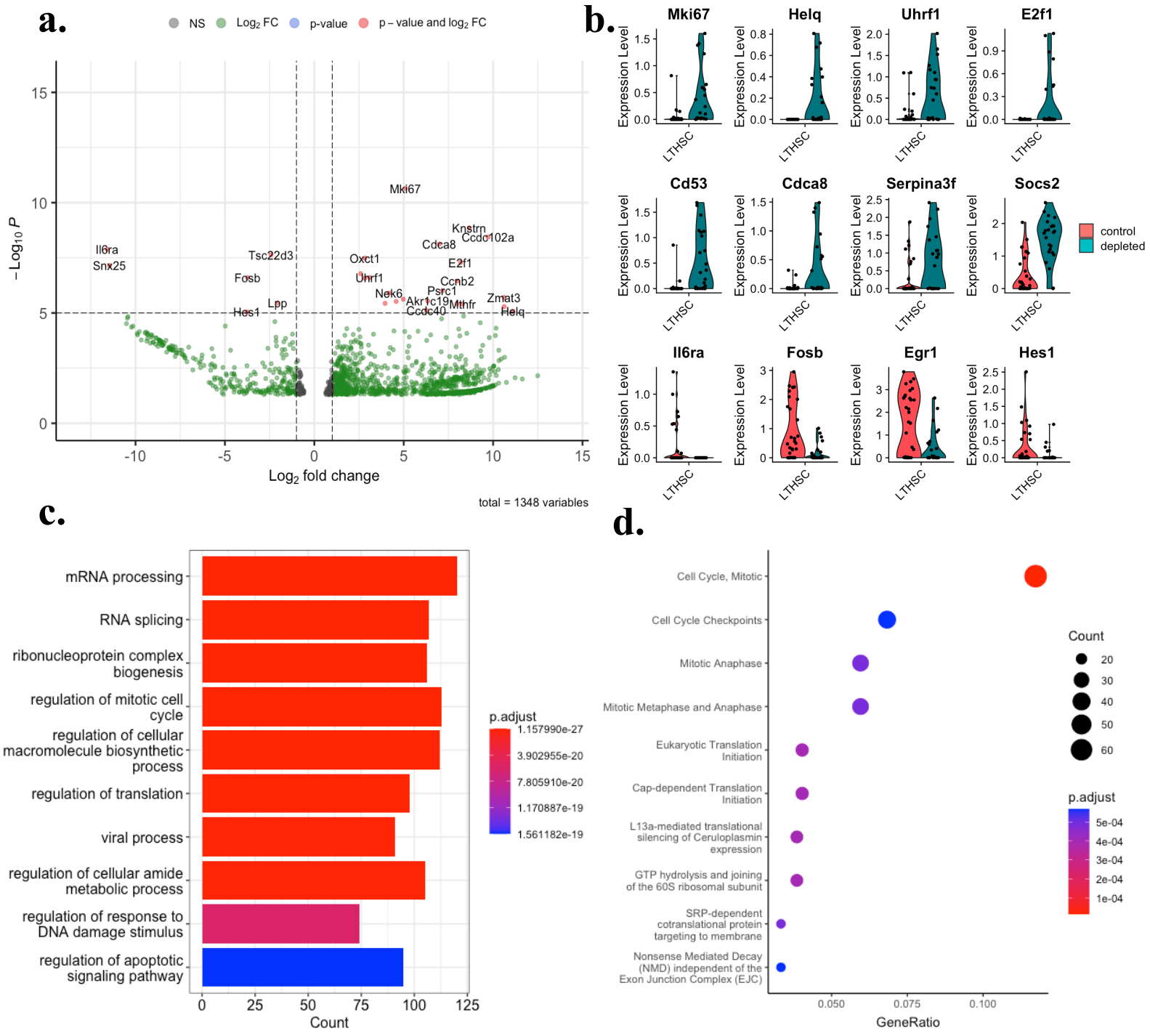


Figure 3.18. Differential expression analyses of LT-HSCs after platelet depletion. (a) Volcano plot of log2 fold change in expression of significantly DEGs in LT-HSCs post platelet depletion. (b) Violin plots of expression levels of a subset of DEGS in LT-HSCs across conditions. (c) Top 10 GO enrichment terms identified from significant DEGs. Bar length is equal to the number of genes identified corresponding to each term, and bars are coloured based on the adjusted p-value for each term (d) Top 10 enriched pathways identified from significant differentially expressed genes. Point size corresponds to the number of genes identified corresponding to each term, points are coloured based on the adjusted p-value for each term, and gene ratio corresponds to the degree of variance between control and post-platelet-depletion LT-HSCs.

Platelet-depletion also induced down regulation of expression of the genes encoding the TF EGR1, and *Fosb* which is part of the TF complex AP-1 together with *Jun* and other cofactors (Figure 3.18B). EGR1 can act as an activator or repressor of transcription depending on the regulation of distinct co-factors (Thiel and Cibelli, 2002). *Egr1* expression is enriched within the most primitive subset of LTHSCs under steady-state conditions, and is downregulated upon stimulation for cell division and migration (Min *et al.*, 2008). Similarly, previous data has shown prolonged expression of *Fosb* negatively controls cell cycle progression acting as a gatekeeper in cell cycle progression of primitive HSCs; whereby their activation in cell cycling and subsequent proliferation require coordination by early acting cytokines including SCF, IL-3, and IL-6 (Okada *et al.*, 1999). This existing data suggests platelet-depletion has resulted in downregulation of *Egr1* and *Fosb* in LTHSCs to promote cell division.

On the other hand, LTHSCs from platelet-depleted mice had significantly higher expression of *Socs2* (*p adj* = 0.007). This is a feedback inhibitor of JAK-STAT pathways expressed in primitive HSCs and can be upregulated in response to STAT5-inducing cytokines (Baker, Rane and Reddy, 2007; Kimura *et al.*, 2010). Previous work has implicated *Socs2* in the regulation of cell proliferative responses to myelopoietic stress (using myeloablation by 5-FU) (Vitali *et al.*, 2015). Specifically, this work revealed higher expansion of differentiated progenitors post myeloablation in *Socs2*−/− mice from LSK, CMP, GMP, and MEP amplification with a reduction in frequency of LTHSCs; suggesting *Socs2* expression acts as a regulatory mechanism at the HSC level. The authors proposed *Socs2* deficiency contributes to exhaustion of LTHSC stemness, whereby in WT conditions *Socs2* expression leads to negative regulation of STAT5 signalling; accounting for the increased myelopoietic response seen in *Socs2*−/− mice (Vitali *et al.*, 2015). Moreover, Vitalli *et al.* and others have shown *in vitro* TPO stimulation of BM Lin− cells is a potent inducer of *Socs2* expression (Bradley, Hawley and Bunting, 2002; Baker, Rane and Reddy, 2007; Vitali *et al.*, 2015) and induces both tyrosine phosphorylation and activation of STAT5 and STAT3 (Bacon *et al.*, 1995; Kato *et al.*, 2005). They showed TPO stimulation of LSKs from *Socs2*−/− mice exhibited a higher proliferative potential to WT LSKs - suggesting *Socs2* induction by TPO limits HSC proliferative response.

Here, platelet-depletion correlates with wide-spread *Socs2* upregulation (Figure 3.19A), suggesting its expression was induced in response to megakaryopoietic stress. Therefore, a possible explanation for this experiment may be that the targeted depletion of platelets led to an accumulation of TPO in mice, in turn inducing *Socs2* upregulation, which acted as a protective mechanism in LTHSCs by inhibiting JAK/STAT proliferation. Supporting this, Vitalli et al also showed that after multiple rounds of BM transplants (BMTs) from either *Socs2*-/- mouse BM or WT mouse BM led to an enhanced haematopoietic response in mice receiving Socs2-/- BM. However, this was only observed for the first 3 transplants and in turn resulted in the exhaustion

of LT-HSC repopulating potential by the 4th BMT and ultimately reduced survival (Vitali *et al.*, 2015). There was a decrease in the number of LTHSCs captured compared from platelet-depleted mice to control mice, with only 3.5% of platelet-depleted cells annotated as LTHSCs compared to 12.6% from control mice; while the inverse was the case for downstream precursors of the Mk lineage with higher proportions of cells from platelet depleted mice across both MPP2 and Mk-MEP cells. It is plausible that post antibody administration over time stem cell protective mechanisms were upregulated to safeguard the HSC compartment, which was induced to generate progeny as an emergency response and consequently depleted in numbers. It would be interesting to compare the proportions of cells within the Cd150+ HSPC compartment at time intervals, and correlate this to TPO cytokine levels in mice post platelet-depletion to assess the immediate and long-term effects across the BM cell populations.

Similarly, LTHSCs from mice who were platelet-depleted also exhibited higher expression of tetraspanin CD53 (Figure 3.18B). A recent publication has reported CD53 upregulation in HSCs in response to both inflammatory and proliferative stressors, revealing that the loss of CD53 is associated with a reduction in HSC function and prolonged cycling under haematopoietic stress (Greenberg *et al.*, 2023). CD53 was previously identified as a marker that segregates differentially in dividing human HSCs, localising to a more functionally primitive population (Beckmann *et al.*, 2007). Greenberg *et al.* are the first to suggest CD53 expression facilitates the return of cycling HSCs to quiescence. The RB-like, E2F and multi-vulval class B (DREAM) complex is a master transcriptional regulator that is known to repress cell cycling in response to stress (Sadasivam and DeCaprio, 2013). Greenberg *et al.* showed that CD53 promotes the activity of pocket proteins in response to HSC stress, facilitating DREAM complex binding and returning to quiescence (Greenberg *et al.*, 2023).

The DEG signature and GSEA (Figures 3.18C-D) show platelet-depletion induces LTHSC proliferation to produce the downstream progenitors necessary to restore platelet levels. Based on the literature, the expression of *Socs2* and CD53 in LTHSCs post platelet depletion suggests their upregulation may serve as a protective mechanism to help prevent HSC exhaustion during stress. By comparing the expression levels these DEGs to the *Fosb* and *Egr1* as well as a canonical cell-cycling marker shows there is variance within platelet-depleted LTHSCs of cells undergoing cell-cycling and cells in G1 phase (Figure 3.19B). Both downregulation of *Egr1* and *Fosb,* and upregulation of *Socs2* and *Cd53* with platelet-depletion is evident - but each of these signatures are largely found across different cells. This data suggests genes promoting both LTHSC differentiation and genes promoting stemness are within the DEG signature post platelet depletion, with some LTHSCs continuing to respond while others have either retained a quiescent phenotype, or reverted back to quiescence post exerting a stress response.

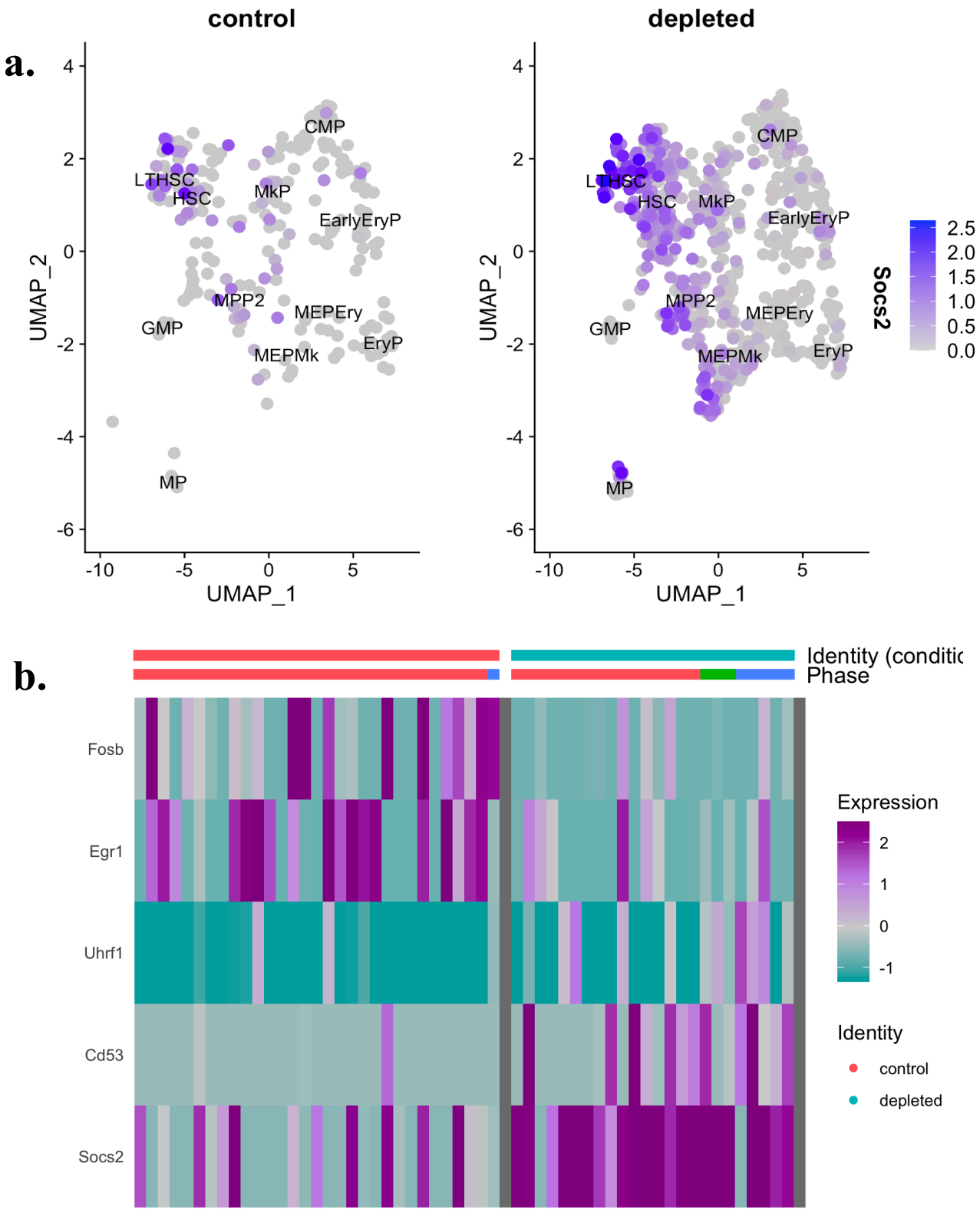


Figure 3.19. Signatures of LTHSC expression regulation post platelet depletion (a) Expression levels of *Socs2* on UMAP projections of single cells split based on treatment condition shows its upregulation across HSCs and cells of the Mk-lineage (b) Heatmap of expression in 5 genes across all LTHSCs from both treatment conditions shows LTHSCs from platelet depleted mice have variable expression in stem-cell protective (upregulated CD53, *Socs2*) vs cell cycle / proliferation-promoting genes (down regulated *Fosb and Egr1*, upregulated *Uhrf1*).

DEA between cells in the MPP2 cluster identified a total 2395 genes with DE with platelet-depletion, of which 395 were statistically significant (Figure 3.20A). Platelet depleted samples exhibited higher expression of G2M cell cycle phase genes and genes that encode for metabolic proteins. This includes *Ccnb1* – cell cycle progression gene (Grinenko *et al.*, 2018), *Golm1* - involved in protein biosynthesis in the rough endoplasmic reticulum and protein transportation through the Golgi apparatus (Q. Song *et al., 2021),* and *Rab40c* - involved in metabolism of GTP- and GDP-binding (Rossaint *et al.*, 2021). The top enriched GO terms associated with significant DEGs found multiple metabolic pathways were enriched, multiple of which were related to tRNA metabolism (Figure 3.20B). To study further the molecular pathways associated with the DEG signature, pathway enrichment was performed and interestingly identified a single enriched pathway (Figure 3.20C). Aminoacyl-tRNA synthetases (aaRSs) catalyse aminoacylation of tRNAs in the first step of protein synthesis in the cytoplasm. These charged tRNAs serve as adaptors during translation, bringing the correct amino acid to the ribosome during protein synthesis. In total, 13 DEGs were found to be related to aaRSs including *Aars2, Mars2* and *Sepsecs* which were indeed upregulated in platelet-depleted MPP2s (Figure 3.20D).

Previous work has demonstrated that aaRSs, in particular an activated form of tyrosyl-tRNA synthetase (YRS^{ACT}) is implicated to enhance megakaryopoiesis and platelet production both *in vitro* and *in vivo* (Kanaji *et al., 2018)*. Tyrosyl-tRNA synthetase is a type of aaRS responsible for attaching the amino acid tyrosine to its corresponding tRNA molecule, allowing for the accurate incorporation of tyrosine during protein synthesis. Kanaji *et al.* demonstrated that ex-translational activities of tyrosyl-tRNA synthetase YRS^{ACT} promote megakaryopoiesis in two ways: (i) inducing a distinct subset of Mks; and (ii) up-regulating secretion of 'monokines' (monocyte cytokines), including IL-6, that support Mk expansion and, ultimately, platelet production (Kanaji *et al.*, 2018). Moreover, the authors showed that YRS^{ACT} is secreted under stress and induces stress-stimulated signalling through toll-like receptor pathways as well as translocation to the nucleus where it contributed to trigger pathways for cell rescue. Importantly, the activity of YRS^{ACT} was independent of TPO, as evidenced by Mk expansion from iPS cell-derived HSCs from a patient deficient in TPO signalling (Kanaji *et al.*, 2018). This work suggests YRS^{ACT} may serve as an important mechanism in response to thrombocytopenia accelerating platelet-count recovery through a distinct and complementary mechanism to TPO stimulation.

More recently, they have also shown YRS^{ACT} mimics inflammatory stress in mice, inducing a distinct population Mks from Mk-biassed HSCs bypassing the MEP where in addition to promoting platelet production, platelets induce the release of pro-inflammatory cytokines from platelets and immune cells leading to an inflammatory response (Morodomi *et al.*, 2022).

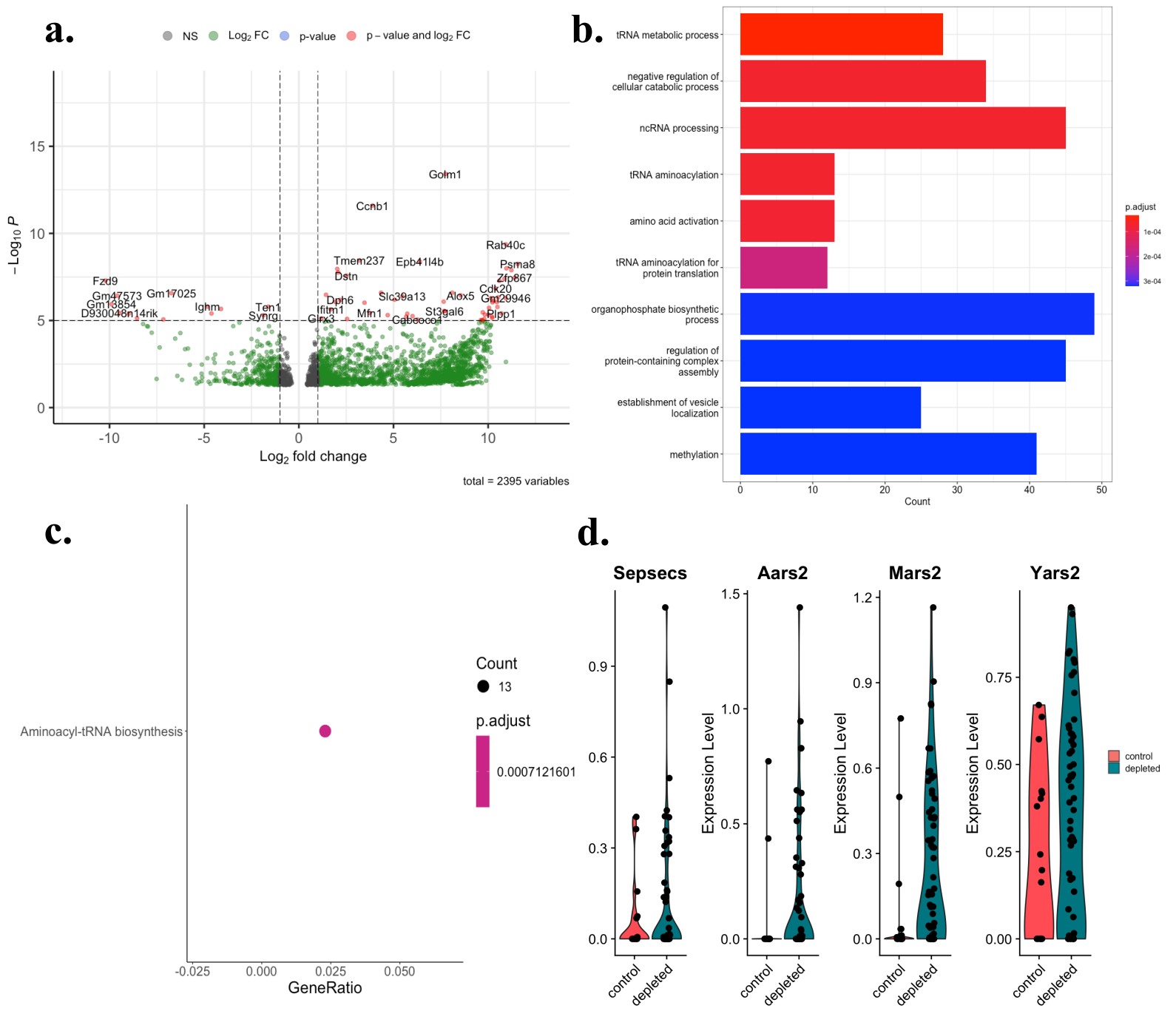


Figure 3.20. Differential expression analyses of MPP2 cells after platelet depletion (a) Volcano plot of log2 fold change in significant DEGs plotted in MPP-2 cluster post-platelet-depletion (b) Top 10 GO enrichment terms identified from significant DEGs. Bar length is equal to the number of genes identified corresponding to each term, and bars are coloured based on the adjusted p-value for each term (c) Enriched pathway identified from significant differentially expressed genes (d) Violin plots of expression levels between control and platelet-depleted samples across genes part of the Aminoacyl-tRNA biosynthesis pathway enrichment list.

Altogether their work implicates alternative functions of aaRSs contribute to platelet production under stress, suggesting they may represent possible pharmacologic modulator of inflammatory thrombopoiesis to replenish platelets and prevent haemorrhage (Morodomi *et al.*, 2022). Here, the robust enrichment of genes for aaRSs in MPP2 cells from platelet depleted samples corroborates this research (Figure 3.20B-D), and suggests a possible role for aaRSs in Mk generation to recover from acute platelet loss. Further research is needed to fully understand the roles of aaRS in the context of megakaryopoiesis but together these data indicate there may exist potential regulators of Mk differentiation within the aaRS family.

The Mk-MEP DEG signature was dominated by the upregulation of genes associated with cell division, DNA damage repair proteins, and cellular energy metabolism as indicated by the top GO enrichment terms from significant DEGs (Figure 3.21A-B). The *Haspin* gene, which encodes a protein kinase imperative for mitosis, is among the most significant upregulated genes in Mk-MEP cells of platelet depleted mice (Higgins, 2010). It is a member of the mitotic kinase family along with Aurora kinases, where specifically it phosphorylates H3T3 during cellular mitosis where it becomes crucial for recruitment of the chromosome passenger complex (CPC) (Dai, Sullivan and Higgins, 2006; Huang *et al.*, 2020). Previous work has demonstrated *Haspin* is an important regulator of cell cycle progression, where knock-down experiments showed multiple cell cycle defects along with suppressed cell proliferation by both induced cell-death or prolonged interphase progression (Wang *et al.*, 2021).

Moreover, upregulation of genes associated with DNA damage repair pathways was also observed across Mk-MEPs from platelet depleted samples (Figure 3.21C). This includes for example *Ercc4,* a gene which encodes a protein component of the nucleotide excision repair pathway tasked with repairing DNA damage caused by ultraviolet radiation, chemical agents, and other mutagens (H. Yu *et al.*, 2012). Working in coordination with other proteins it removes damaged sections of DNA and facilitates the synthesis and ligation of new DNA strands to restore the integrity of the genome. *Nthl1,* another gene upregulated in Mk-MEPs of platelet depleted mice (Figure 3.21C), encodes for a protein involved in DNA base-excision repair that is specifically responsible for repair of oxidative damage to DNA (Mjelle *et al.*, 2015).

In addition to DNA damage repair proteins, the upregulation of genes involved in RNA surveillance was also observed (Figure 3.21C). *Dis3l* encodes for an RNA exonuclease found in the cytoplasm which functions in RNA degradation and regulation (Brouze *et al.*, 2022). It is thought to participate in the surveillance and ‘quality-control’ of RNA molecules, whereby it helps to eliminate aberrant or defective RNAs, including those with incomplete or incorrect processing, modifications, or secondary structures (Houseley, LaCava and Tollervey, 2006). By removing faulty RNA species and regulating their abundance, it contributes to the overall

fidelity and functionality of cellular RNA and supporting cellular homeostasis. More recently it has been implicated in the coordinated control of cell proliferation, where mutated *Dis3l* led to decreased rates of proliferation (Towler *et al.*, 2020; Hojka-Osinska *et al.*, 2021)

The increased expression of cell proliferation, DNA damage repair and surveillance genes suggest platelet-depletion induced Mk-MEPs towards proliferation, likely to replenish Mks and restore platelet levels, while also upregulating safe-guarding mechanisms to cope with the increased levels of DNA damage that can occur during rapidly dividing and differentiating progenitors. Indeed, once driven into the cell cycle, the expressions of DNA damage repair genes in haematopoietic progenitor cells have previously been reported, allowing insults to be repaired to prevent cellular dysfunction and malignancy, particularly at the HSC level. It would be reasonable to conclude that post platelet depletion Mk-MEPs upregulate the expression of genes important to regulatory mechanisms at various levels to ensure the fidelity of transcriptomes as they differentiate into committed MkPs.

Altogether this data shows platelet-depletion induces changes in both the cellular composition and transcriptional signatures across multiple levels of megakaryopoiesis (Figure 3.22). DEA across individual cell-types along the Mk trajectory revealed cell-type specific signatures of stress in response to platelet-depletion, providing insights into how cells at different stages of Mk commitment regulate their transcriptomic repertoire to counteract thrombocytopenic conditions.

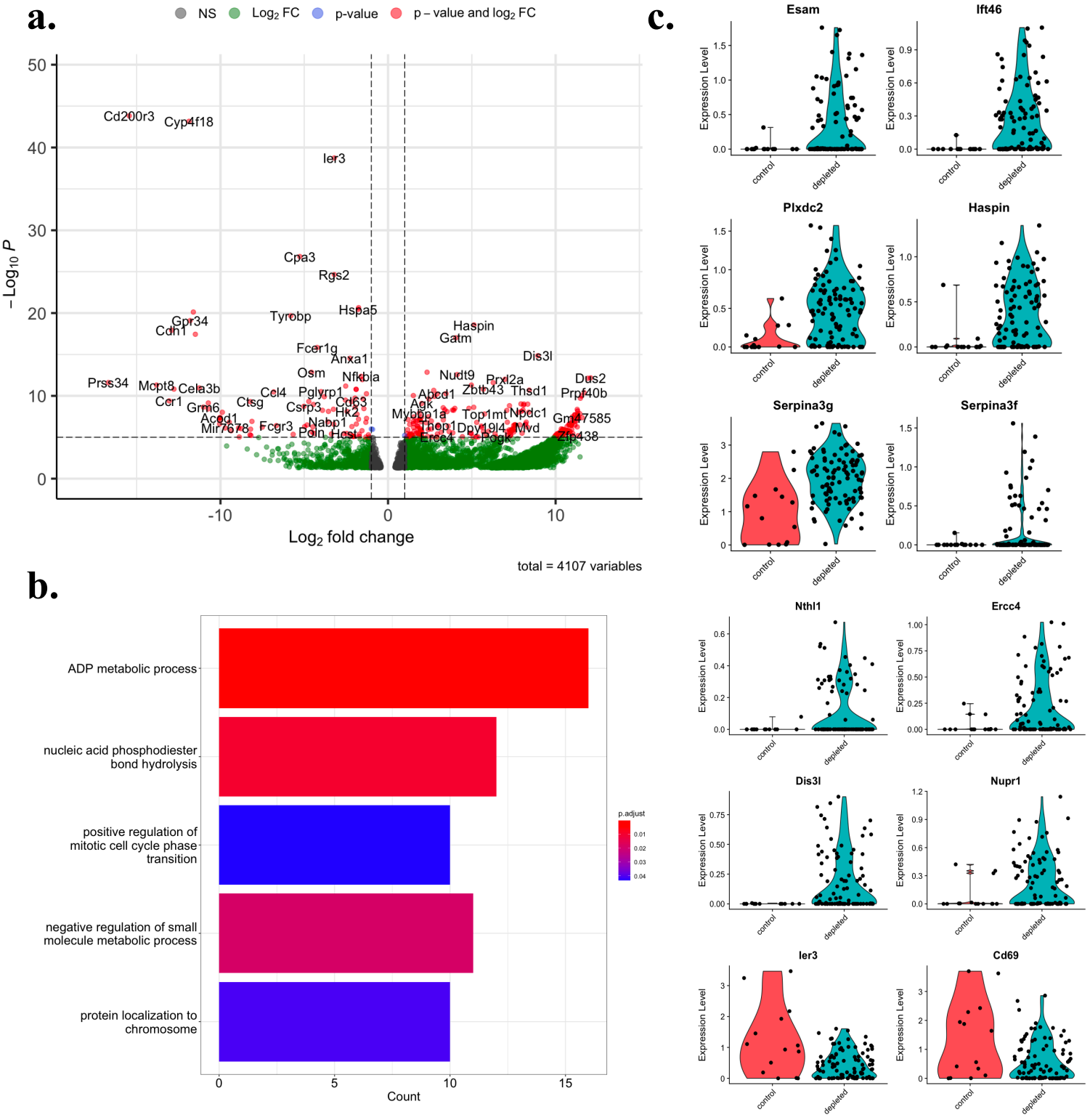


Figure 3.21. Differential expression analyses of Mk-MEP cells after platelet depletion. (a) Volcano plot of log2 fold change in significant DEGs plotted in Mk-MEP cluster post-platelet-depletion (b) Top 5 GO enrichment terms identified from significant DEGs. Bar length is equal to the number of genes identified corresponding to each term, and bars are coloured based on the adjusted p-value for each term (c) Violin plots of expression levels between control and platelet-depleted samples across Mk-MEP cells.

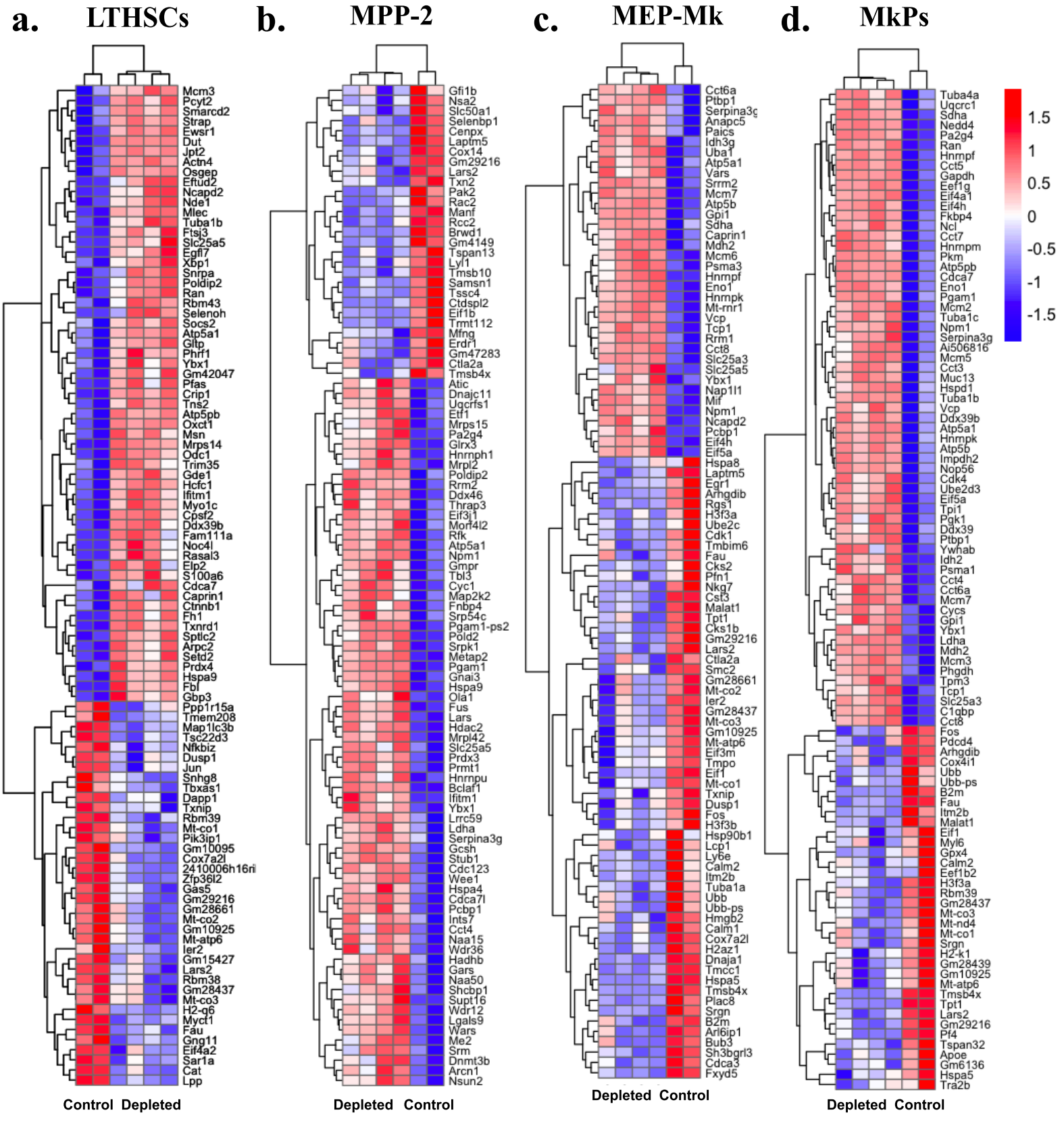


Figure 3.22. Heatmap of log-normalised expression levels in the top 100 DEGs after platelet depletion across cell types (a) LT-HSCs (b) MPP2 (c) Mk-MEPs and (d) MkPs.

3.4 Discussion

The primary purpose of this study was to capture the full differentiation trajectory from HSC towards MkP, and interrogate the response of this trajectory to platelet depletion using scRNA-seq. Mouse BM cells were isolated with FACS using a broad gate capturing LK Cd150+ cells, with an additional LT-HSC gate to ensure sufficient coverage of stem cells. Without using cell-type specific gates, the goal was to capture the full process including cell states that may exist outside of narrowly defined population gates, thus representing a complete continuum of HSC-MkP differentiation. The decision to employ this sorting strategy was based on data from Pronk *et al.*, that demonstrated *in vitro* high proportions of LSK Cd150+ cells generated exclusively Mk progeny, while the LSK Cd150- populations contained no stem cell activity and lacked Mk potential at a clonal level. Importantly, these findings were successfully recapitulated *in vivo,* with Cd150+ HSCs exhibiting robust platelet recovery after transplantation (Pronk *et al.*, 2007). Moreover, the MkPs isolated based on Cd150 expression were functionally equivalent to the CD9+ MkPs previously described (Nakorn, Miyamoto and Weissman, 2003). These results provided a strong rationale for using Cd150 expression to isolate cells along the Mk lineage, indicating Cd150 expression enables enrichment for cells of the Mk lineage at primitive and up to committed progenitor states. In total, this sorting strategy enabled the isolation of 1288 single cells from six mice treated with either platelet-depleting antibody or an isotype control, and provided a comprehensive dataset for further analysis and interrogation of transcriptional changes associated with Mk commitment under steady-state and stress conditions.

The hypothesis driving this study proposed that platelet depletion would trigger emergency megakaryopoiesis to rapidly replenish platelet levels, resulting in changes to the cellular composition and transcriptional profiles of cells. Specifically, an increased proportion of cells committed to the Mk lineage was anticipated, accompanied by activation of transcriptional programmes involved in driving megakaryopoiesis. By testing this hypothesis and carrying out bioinformatic analyses of the resulting data, the study sought to expand our understanding on genes governing different stages of Mk commitment, and also serve as a valuable resource enabling future research of the first steps in which HSCs commit to the Mk lineage.

Previous research into platelet recovery from induced thrombocytopenia has shown inflammatory signalling associated with depletion of platelet levels leads to the activation of an Mk maturation programme and expansion of a subpopulation of Mk-committed HSCs (Haas *et al.*, 2015). These findings were in concordance with the growing body of evidence for the existence of platelet-biassed HSCs, and that multiple differentiation routes can lead to Mk commitment, including direct HSC MkP differentiation (Sanjuan-Pla *et al.*, 2013; Yamamoto *et*

al., 2013; Shin *et al.*, 2014). Building upon this existing body of work, this study aimed to further elucidate this pathway, utilising antibody-induced thrombocytopenia as opposed to an infection-based model. While previous work focused on infection-induced thrombocytopenia, this study sought to specifically examine the impact of isolated platelet depletion on the transcriptome, a distinct strategy to elucidate molecular and cellular responses under stress. Moreover, this study employed Smart-seq2 to provide deep coverage of the transcriptional landscape of single cells, resulting in detection of an average ~7.8K genes per cell from an average sequencing depth of ~1.3M reads per cell. By leveraging this approach this set of experiments provided high resolution data on gene expression dynamics during Mk commitment during acute thrombocytopenia.

The findings of this study provide valuable insights into the dynamics of Mk lineage commitment and the transcriptional programmes involved in driving megakaryopoiesis under stress. Through unsupervised clustering, cells were grouped into 11 clusters based on their transcriptional profiles, which were cell-type annotated based on information from relevant studies in the existing literature (Pronk *et al.*, 2007; Haas *et al.*, 2015; Paul *et al.*, 2015; Pietras *et al.*, 2015; Psaila *et al.*, 2016; Miyawaki *et al.*, 2017; Dahlin *et al.*, 2018). Pseudotime analysis allowed cells to be ordered along a trajectory, revealing the differentiation continuum from LT-HSCs to Mk and Ery progenitors. The cell types captured were in alignment with those expected within the LK/LSK Cd150+ fraction and the reference literature, indicating a successful sorting strategy. Furthermore, no significant sample effects were observed, suggesting that the experimental conditions did not introduce any notable biases. A small proportion of cells captured (~3% of all cells) were found to express an immature myeloid cell signature, including markers of the GM lineage. However, the presence of these cells was deemed negligible, confirming the overall minor contamination of cells outside of the Mk lineage.

In addition to successfully capturing the expected cell populations within the LK/LSK Cd150+ fraction, this analysis revealed further heterogeneity within MEPs. Specifically, this data enabled MEPs to be subdivided into two distinct subpopulations, each characterised by differential expression levels of genes associated with the Mk lineage (Mk-MEP) or Ery lineage (Ery-MEP). This result aligns with previous studies that have reported differential expression patterns in key Mk and Ery genes across cells within the MEP compartment, rather than a homogenous cell type with equal propensity towards Mk and Ery fates. Importantly, these differential expression patterns observed have been linked to functional differences in lineage propensities between the subtypes (Psaila *et al.*, 2016). By identifying and characterising these MEP subpopulations based on their gene expression profiles, this study further supports the notion of functional heterogeneity and lineage-specific characteristics within the MEP

compartment (Psaila *et al.*, 2016; Lu, Krause, *et al.*, 2018). The analysis of cell population frequencies revealed platelet depletion induces the expansion of Mk-MEPs, with no change in the frequency of Ery-MEPs. This finding suggests that Mk-MEPs and Ery-MEPs possess distinct mechanisms of activation to elicit differentiation, enabling the selective expansion and differentiation of Mk-MEPs while retaining a stable Ery committed MEP population. The differential response of Mk-MEPs and Ery-MEPs to platelet depletion shows that the regulatory mechanisms governing their proliferation are distinct, with 3.5-fold more genes identified as significantly differentially expressed in Mk-MEPs, including upregulation of cell cycle and DNA damage repair genes in response to platelet depletion. Indeed, there is evidence to suggest that activation or repression of the cell cycle essentially serves as a rheostat to affect the Ery versus Mk specification of MEPs (Lu, Sanada, *et al.*, 2018), however further research of how the cell cycle affects MEP fate decisions are required to determine the underlying explanations of these observations.

To gain further insights into the distinct MEP subpopulations, it would be valuable to integrate the index-sort data in conjunction with scRNA-seq analysis to explore their surface marker expression profiles. This would help determine if the MEP subpopulations identified using scRNA-seq can be distinguished based on their surface marker profiles, while providing information of their phenotypic characteristics. In addition, integrating FACS analysis with scRNA-seq data could also improve the purity of future FACS sorting experiments, enabling more selective isolation of cells of interest. As previously stated, a low proportion of cells from the GM lineage were captured in this study. Correlating surface marker expression with single-cell annotations would reveal whether the captured GM cells are distinguishable by FACS analysis. For instance, if they are found to lie in proximity to the minimal threshold used for Cd150+ gating, this would indicate further restriction by raising the minimal threshold of Cd150 levels may enhance sort purity of future experiments. Therefore integration of index-sorting data will provide the complete phenotype of sorted cells, and may also be useful for further optimisation of sort gates leading to better enrichment of the cells of interest.

Among the results for this study, notable differences were observed in the population of LT-HSCs between platelet-depleted mice and the control group. Firstly, fewer LT-HSCs were captured from platelet-depleted mice, indicating a reduction in the LT-HSC pool in response to platelet depletion. Among the LT-HSCs that were captured, a higher proportion of stem cells from platelet-depleted mice were found in G2M and S phases of the cell cycle, suggesting increased proliferation and cell division compared to the mostly quiescent state observed in control LT-HSCs. Functional enrichment analysis of DEGs between the experimental conditions further supported these findings, revealing enrichment of multiple pathways related to increased cell proliferation. This supports existing evidence that Mk-associated signalling

affects the abundance of LT-HSCs and influences their cell cycle progression and proliferative behaviour (Hock *et al.*, 2004; Qian *et al.*, 2007; Yoshihara *et al.*, 2007; Nakamura-Ishizu *et al.*, 2014; Zhao *et al.*, 2014). Indeed, previous research has established a link between Mks and the regulation of HSC quiescence. Zhao *et al.* used Mk ablation to study HSC function *in vivo,* where they found Mks contribute to maintenance of HSC quiescence through TGF-β signalling under homeostasis, and switch upon stress to instead promote HSC expansion via fibroblast growth factor 1 (FGF1) signalling (Zhao *et al.*, 2014; Gong *et al.*, 2018). Bruns *et al.* instead implicated *Pf4* (CXCL4) production from Mks in the regulation of HSC quiescence (Bruns *et al.*, 2014). Multiple groups have established the role of TPO signalling in both Mk and HSC function (Kimura *et al.*, 1998; Qian *et al.*, 2007). However, this study specifically implicates platelet abundance in and of itself is capable of modulating multiple levels of megakaryopoiesis, independent of Mk levels, where platelet levels elicit a response that extends all the way back to the HSC compartment. The intravenous injection of anti-GPIbα was used to specifically target platelets for clearance, and has been shown previously not to affect megakaryopoiesis in the BM before 8 days post administration (Morodomi *et al.*, 2020). Therefore these results suggest that not only Mks, but also platelets can lead to changes in the HSC compartment to promote rapid expansion of stem cells when required.

Cytokine signalling is well established to play a crucial role in regulating HSC function, including in modulating HSC self-renewal, proliferation, differentiation, and mobilisation (Zhang and Lodish, 2008). It was suspected that if platelet depletion led to shifts in cells' transcriptomic signatures, altered cytokine signalling would likely also be observed. As part of the initial study design, plasma from peripheral blood samples taken from mice was going to be tested with a cytokine-array assay to complement the transcriptomic data. However, due to time constraints and limited access to the required imaging equipment, the cytokine assay could not be integrated into this study. Nevertheless, the cytokine array assay remains a crucial avenue for future work, as it holds the potential to provide valuable insights into the signalling pathways and thus molecular mechanisms involved in the regulation of megakaryopoiesis under thrombocytopenic conditions. By combining the transcriptomic analysis of this study with a cytokine array assay, it is anticipated a more comprehensive understanding of the changes associated in cytokine signals in response to platelet depletion can be achieved. For example, by comparing the cytokine profiles between the control and platelet-depleted groups, it would be possible to identify specific cytokines that are differentially expressed in response to platelet depletion, and to correlate this with differentially expressed transcriptomic signatures. This would include circulating levels of thrombopoietic and inflammatory cytokines, shedding light on both TPO-dependent and independent mechanisms involved in megakaryopoiesis under thrombocytopenic conditions. Therefore, incorporation of a cytokine array assay to these results

is part of the future outlook of this study, to create a more complete picture of the response to platelet depletion and ultimately a deeper understanding of regulatory mechanisms involved.

To further validate this study's results and provide phenotypic insights across cell populations in response to platelet depletion, integrating FACS analysis would also be advantageous. FACS analysis allows for the quantification of cell type abundance based on specific surface markers, providing specific measure of the cellular composition following platelet depletion. In this study, a smaller number of LT-HSCs and higher proportion of Mk-MEPs was observed in samples from platelet depleted animals. This finding, in conjunction with the transcriptomic analysis, indicates that platelet depletion induces the activation of HSCs into an active cycling state and selectively expands progenitors of the Mk lineage under conditions of scarce platelet abundance. To strengthen this conclusion and provide additional evidence, it would be valuable to quantify the abundance of LT-HSCs and Mk-MEPs in BM samples obtained from mice under steady-state conditions and post-platelet depletion using FACS analysis. This approach would allow direct comparison of the proportions in these cell populations in response to platelet depletion and validate the observed changes in cell composition. This complementary approach, combined with the transcriptomic analysis performed in this study, would provide a more comprehensive understanding of the phenotypic changes associated with platelet depletion and strengthen the conclusions drawn from this study.

Differential expression analysis along pseudotime provided insights into the dynamic expression of genes implicated in Mk function, including genes associated with Mk development and platelet function but crucially also those not previously implicated in Mk commitment. These findings contribute to our understanding of the molecular mechanisms governing Mk lineage commitment, and underscore the power of the approach to study the process of differentiation at the single cell level. Differential expression analysis across treatment conditions identified hundreds of genes with differential expression within cell types along the Mk trajectory. An important area for future work lies in the validation of these DEG signatures associated with platelet depletion identified across the observed cell types. Indeed, a substantial number of genes with significantly altered expression levels were identified across cell types following platelet depletion, encompassing both known genes implicated in Mk commitment and also novel candidate genes. To ensure the robustness and reliability of these findings, it is acknowledged that validation of DEGs through additional experimental techniques, such as employing quantitative polymerase chain reaction (qPCR) will be imperative to confirm any conclusions drawn from this data. The use of qPCR assays is widely implemented for such validation, providing accurate quantification of gene expression levels. This will involve the use of gene-specific primers and fluorescent probes to obtain quantification of individual target genes (in this case, designed to target DEG genes), providing

quantitative data of mRNA abundance in samples. The qPCR results will then be compared to the scRNA-seq data, evaluating the concordance between the two techniques either confirming the differential expression observed in the scRNA-seq analysis or identifying false positive results from this study. In this way, the validation of DEGs through qPCR will provide independent confirmation of their differential expression patterns, reinforcing the findings from the scRNA-seq analysis and will be performed as a future experiment.

Moreover, if this experiment were to be repeated the sample size of the control group would be increased to at least n = 3, to improve statistical power during DEA. Fewer cells in total were captured from mice in the isotype control treatment group, consequently resulting in relatively lower numbers of all cell types from control samples for downstream analyses. Moreover, it is acknowledged this limitation of sample size could impact the statistical power and generalisability of results. While acknowledging the limitation of sample size, single cell transcriptional profiling and bioinformatic analyses showed consistency across samples between treatment groups in terms of the cellular composition and transcriptional dynamics within the samples. Additionally, in an attempt to mitigate the small sample size of the control group, a pseudo-bulk DEA approach was employed to carry out statistical tests between conditions rather than single-cell DEAs. This approach involves aggregating the expression data from multiple cells within each condition and for each cell type to generate pseudo-bulk samples, allowing for statistical analyses to be representative of DEG signatures across cell types taking into account sample size. Although this approach addresses the low cell numbers issue to some extent, it is nonetheless an important consideration as it may overlook potential cell-to-cell heterogeneity within the samples. Therefore, it may be required to supplement this data with more replicates to increase sample size and ensure sufficient statistical power for single-cell analysis, which would provide a more comprehensive understanding of the cellular heterogeneity and dynamics in response to platelet depletion.

In conclusion, this chapter describes the analysis of 933 single cells, elucidating the process of murine megakaryopoiesis in steady-state and platelet depletion. Among the results this study shows differential gene expression along megakaryopoiesis, including identification of multiple trajectories towards MkP commitment, and revealed the dynamic expression of genes over pseudotime including novel signatures induced by stress. DEA across individual cell-types along the Mk trajectory revealed cell-type specific signatures of stress in response to platelet-depletion, providing insights into how cells at different stages of Mk commitment regulate their transcriptomic repertoire to counteract thrombocytopenic conditions. These findings provide valuable insights into the molecular mechanisms underlying emergency megakaryopoiesis and shed light on the dynamic changes occurring within the first steps of megakaryopoiesis.

Chapter 4

Exploring trajectories of megakaryopoiesis with age using scRNA-seq

Chapter disclosures:

Preprocessing of scRNA-seq libraries (genome alignment, sequencing quality-control, and gene count quantification) were performed by Anita Scoones using the *ScOmix* pipeline (unpublished) developed by Matthew Madgwick (see Materials and Methods 2.3.1 - 2.3.2).

4.1 Introduction

The process of ageing is accompanied by an overall loss of fitness and a dramatically increased prevalence of many diseases, including dementia, autoimmunity, and cancer. In the haematopoietic system, ageing is associated with profound changes within the bone marrow compartment that ultimately result in reduced adaptive immune system function, lower haematopoietic cellularity and increased incidence of haematological malignancies and anaemia (Geiger, de Haan and Carolina Florian, 2013). Unsurprisingly, ageing is also the major risk factor for several haematologic syndromes and malignancies, such as myelodysplastic syndromes (MDS) and acute myeloid leukaemia (AML) (Klepin, 2016).

Research into the ageing haematopoietic compartment is a large field in itself, with the overarching aim to explain the underlying mechanisms behind the phenotypic consequences of ageing to haematopoiesis. Long considered as one of the central mechanisms behind age-related haematopoietic defects is the accumulation of damage to cellular macromolecules, in particular, DNA damage (Kirkwood, 2005). Accumulation of DNA damage is a common feature of ageing in different tissues in many organisms (López-Otín *et al.*, 2013), and despite the many mechanisms in HSCs for genome protection from DNA alterations, specific mutations have been shown to be highly recurrent in the HSC compartment (Vas *et al.*, 2012; Beerman *et al.*, 2014; Flach *et al.*, 2014). Depending on the extent and nature of lesions on DNA, the repercussions may be cytotoxic or mutagenic leading to apoptosis or dysfunction of cells respectively. Apoptosis and dysfunction exceeding the rate of self-renewal would be expected to ultimately result in depletion of the stem cell compartment but with the HSC distinctive features of quiescence and attenuation of DNA repair pathways, DNA damage consequently is able to accumulate in HSCs with time (Beerman *et al.*, 2014). Moreover, the unique properties of stem cells potentiate the impact of damage because lesions can be both propagated through self-renewal and (horizontally) and conveyed into downstream progeny (vertically) meaning that damage starting at the stem cell level can have consequences across all levels of the haematopoietic system (Rossi, Seita, *et al.*, 2007). Whilst the important role of genetic defect accumulation in HSCs with age is evident, other factors have also been strongly implicated in HSC ageing including the interactions within the bone marrow microenvironment, changes in epigenetic regulation, and altered metabolism (Ho *et al.*, 2019).

The production of mature haematopoietic cell types has been shown to be altered with age in both mouse models and humans, with an increased myeloid output consistently observed at the expense of immune cell production (Beerman *et al.*, 2010; Pang *et al.*, 2011; Florian *et al.*, 2012). Research in support of HSC defects resulting in imbalanced haematopoiesis with age

has come from experiments assessing HSC developmental potential post-transplantation into preconditioned irradiated recipients; a model that has well-established long-term reconstitution of both lymphoid and myeloid lineages (Dorshkind *et al.*, 2020). These reports have collectively shown an increased frequency of HSCs with age but a generalised HSC functional decline, with key hallmarks including stem cell exhaustion where HSCs from aged individuals exhibit lower regenerative potential (diminished ability to self-renew) and a significant reduction in lymphogenesis capacity that correlates with the accumulation of cellular damage over time. This is coupled with an increased propensity towards myelopoiesis, with reports demonstrating increased cell cycling and production of myeloid cells (Morrison *et al.*, 1996; Sudo *et al.*, 2000; Liang, Van Zant and Szilvassy, 2005), ultimately suggesting myeloid biassed HSCs is an aged haematopoietic system phenotype. Moreover, even though the number of myeloid cells in aged individuals is higher, their quality is compromised (Signer *et al.*, 2007; Florian *et al.*, 2018). Old HSCs have been described as functionally inferior through both *in vitro* and *in vivo* assays, with delayed proliferation response in stromal co-cultures, a reduced efficiency for short-term bone-marrow homing, production of smaller clones of mature cells in transplanted recipients, and a reduced long-term *in vivo* self-renewal activity (Dykstra *et al.*, 2011). A culmination of these factors fueled research into the functional heterogeneity within the HSC compartment and led to the identification of subsets of HSCs that are lymphoid-biassed (Ly-HSCs), and myeloid-biassed (Mye-HSCs) distinguishable based on phenotypic differences - where Mye-HSCs which have been shown in mice to outnumber Ly-HSCs 6-fold with age (Montecino-Rodriguez *et al.*, 2019).

The identification of two functionally distinct HSC subtypes with differential lineage propensities, varying life spans, and cycling patterns revealed a further layer of complexity to understanding the ageing haematopoietic system. A number of groups used clonal composition assays of cells from the HSC compartment to show that clones with a balanced or Ly-biassed lineage output are depleted with domination by Mye-biassed clones(Muller-Sieburg *et al.*, 2004; Beerman *et al.*, 2010; Challen *et al.*, 2010). Single-cell transplantation of highly purified stem cells demonstrated that the clonal contribution to different lineages varies significantly and is maintained through serial passaging, showing these are stable phenotypes sustained *in vivo (Dykstra et al., 2007)*. Further insight into the heterogeneity of the HSC pool was enabled through the identification of cell surface markers that allowed prospective isolation of these subpopulations, firstly with the identification that CD150 expression within the LSK fraction could be used to sub fractionate Mye and Ly-biassed HSCs, where CD150low HSCs are lymphoid biassed, and Cd150+ cells are myeloid biassed (Beerman *et al.*, 2010; Challen *et al.*, 2010; Morita, Ema and Nakauchi, 2010). Gekas et al. also revealed a role for CD41 expression to identify myeloid-biassed HSCs and as a marker of specifically aged HSCs (Gekas and Graf 2013), classically known as a platelet marker required for platelet aggregation and clotting

(Shattil, Kashiwagi and Pampori, 1998). Technical breakthroughs including such marker identification improved purification strategies paving the way in assay refinement to directly interrogate the composition and regulation of the stem cell compartment. Moreover, the development of single-cell genomic technologies saw the shift in assessing HSC function based on lineage output or reconstitution capacity towards interrogation of the molecular mechanisms underlying HSC functionality at single-cell resolution, enabling the gene expression governing intrinsic changes in HSCs during ageing to be studied.

The study of lineage skewing of aged HSCs is supported by concording data showing an upregulation of myeloid-specific genes and a downregulation of lymphoid-specific genes (Rossi *et al.*, 2005; Chambers *et al.*, 2007; Dykstra *et al.*, 2011; Wahlestedt *et al.*, 2013). A seminal paper revealed a functionally distinct HSC subset that expresses high levels of *Vwf* that not only often exhibits myeloid bias, but instead a platelet-specific gene expression. Using a *Vwf-eGFP* reporter to study the distribution of expression of *Vwf* across HSCs and mature progenitors showed platelet-biassed HSCs exist within the phenotypically defined HSC fraction (LSK CD150+ CD48- CD34-), have long-term platelet-biassed of platelet/Mye-biassed reconstitution, are capable of self-renewal and are capable of generating *Vwf*- Ly-biassed HSCs (Sanjuan-Pla *et al.*, 2013). In the context of ageing, by measuring platelet output from single HSCs they established that myeloid-biassed HSCs also typically produce high levels of platelets in young individuals, and that a subset of HSCs exist with a distinct and stable platelet bias. They showed HSC ageing is accompanied by a coordinated upregulation of platelet-lineage gene expression, both in terms of the number of platelet-specific genes expressed per HSC and of their expression level (Grover *et al.*, 2016).

Data from Poscablo *et al.* showed that young and old MkPs have different gene expression programs, reflecting divergence in the molecular control of Mk differentiation during ageing. They showed that despite the reconstitution deficit of aged HSCs as consistent with pre-existing literature, *in vitro* experiments found that MkPs from aged mice in fact displayed greater proliferative potential and when transplanting young and old MkPs that old MkPs harboured a high capacity to engraft, expand, and reconstitute platelets. One would have expected that old MkPs would also display functional deficiencies in the same way limitations were consistently observed in old HSCs.

4.1.1 Aims

The aims of this chapter were to:

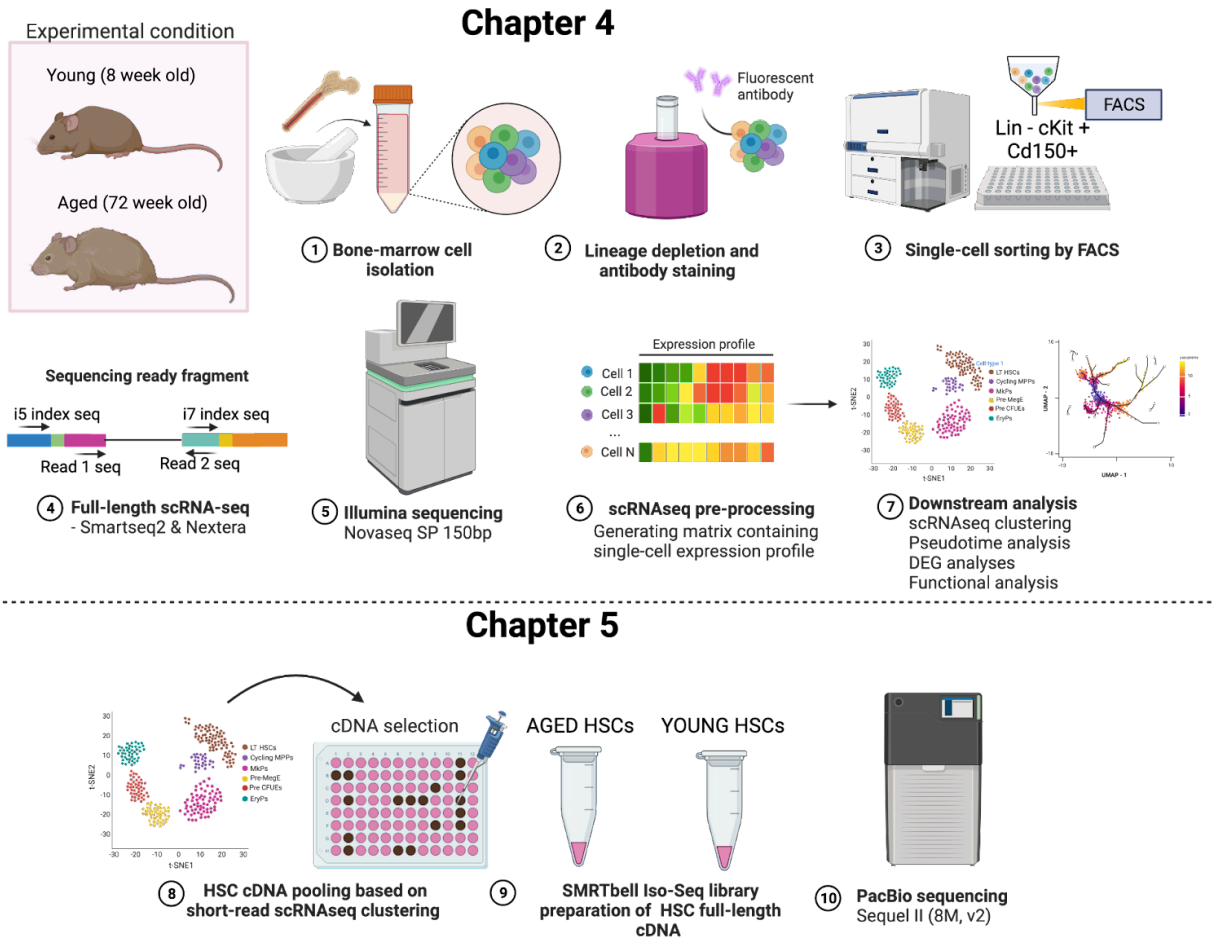
- Capture cells along the trajectory towards Mk commitment by FACS gating LK and LSK Cd150+ cells.
- Apply single-cell transcriptomics using Smart-seq2 to order cells along the continuum of differentiation between HSC and MkP.
- Interrogate the differentiation trajectory in both steady state and in response to ageing.

To investigate the changes along the trajectory towards Mk commitment with age, single-cell transcriptomics was employed to order cells along the differentiation continuum, and interrogate the differentiation trajectory in both steady state and in response to ageing.

Previous research indicates profound changes in the homeostatic control of haematopoiesis at multiple levels of differentiation during ageing, however, a thorough investigation of the LK Cd150+ compartment and targeted analysis of Mk lineage trajectories with age at single-cell resolution has not yet been produced. These aims were addressed by profiling HSPCs from mouse bone marrow using Smart-seq2, scRNA-seq profiles of cells were used to visualise and study the transcriptional heterogeneity of cells in the LK Cd150+ BM fraction. Pseudotime trajectory analysis was used to computationally order single cells along a differentiation trajectory. This chapter describes the analysis of transcriptional changes along Mk commitment and differential expression signatures associated with age.

4.2 Experimental approach

To investigate the trajectories of the Mk lineage with age, scRNA-seq experiments were performed using mice from two age groups (8 weeks and 72 weeks). A total of 6 mice were utilised, with 3 mice per age group. LK and LSK Cd150+ single-cells were sorted into 96-well plates (Methods 2.2.2 - 2.2.6.1) and processed for Smart-seq2, following previously described protocols (Picelli et al., 2014) (Methods 2.2.7.1-2.2.7.4). The cells from a total of 7 96-well plates were clustered and annotated into cell types based on their transcriptomic expression signatures and ordered using pseudotime analysis. Furthermore, differential expression and functional analyses were performed to identify signatures that are associated with pseudotime states and age.



13

Figure 4.1. Schematic workflow of the experimental approaches implemented for Chapters 4 and 5.

[13] Created with BioRender.com

4.3 Results

4.3.1 Isolation of megakaryocyte trajectory from young and aged bone-marrow samples for scRNA-seq

To study cells along the Mk trajectory in ageing, cells were isolated from mouse bone marrow and sorted by single-cell FACS coupled (Figure 4.2). In total, seven 96-well plates including LK Cd150+ and LSK Cd150+ single cells were sorted along with two positive controls (50 cells per well and two negative controls per plate (one positive and negative per condition) as indicated in Figure 4.2B. This was again performed as such to enable the detection of errors including contamination, reagent failures or technical problems. Cells from young and aged mice were sorted into each PCR plate, yielding a total of 644 cells for this set of experiments excluding control wells (Figure 4.2C). As before, combining both samples in each plate minimises the potential for batch effects or variation between samples due to technical factors either during sorting or downstream library preparation between each experimental condition. Plates were randomised and processed in two batches for Smart-seq2 single-cell RNAseq as previously described (Picelli 2014), and as before 0.2 ng/ µl of cDNA was used per sample to generate Illumina sequencing-ready libraries with NextEra. Sample quality was evaluated post-cDNA amplification product clean-up, and post-NextEra library preparation (Appendix Supplementary Figure 4.1). Plates were pooled at equimolar concentrations to ensure equal read coverage across libraries during sequencing to generate a minimum of 1M reads per cell.

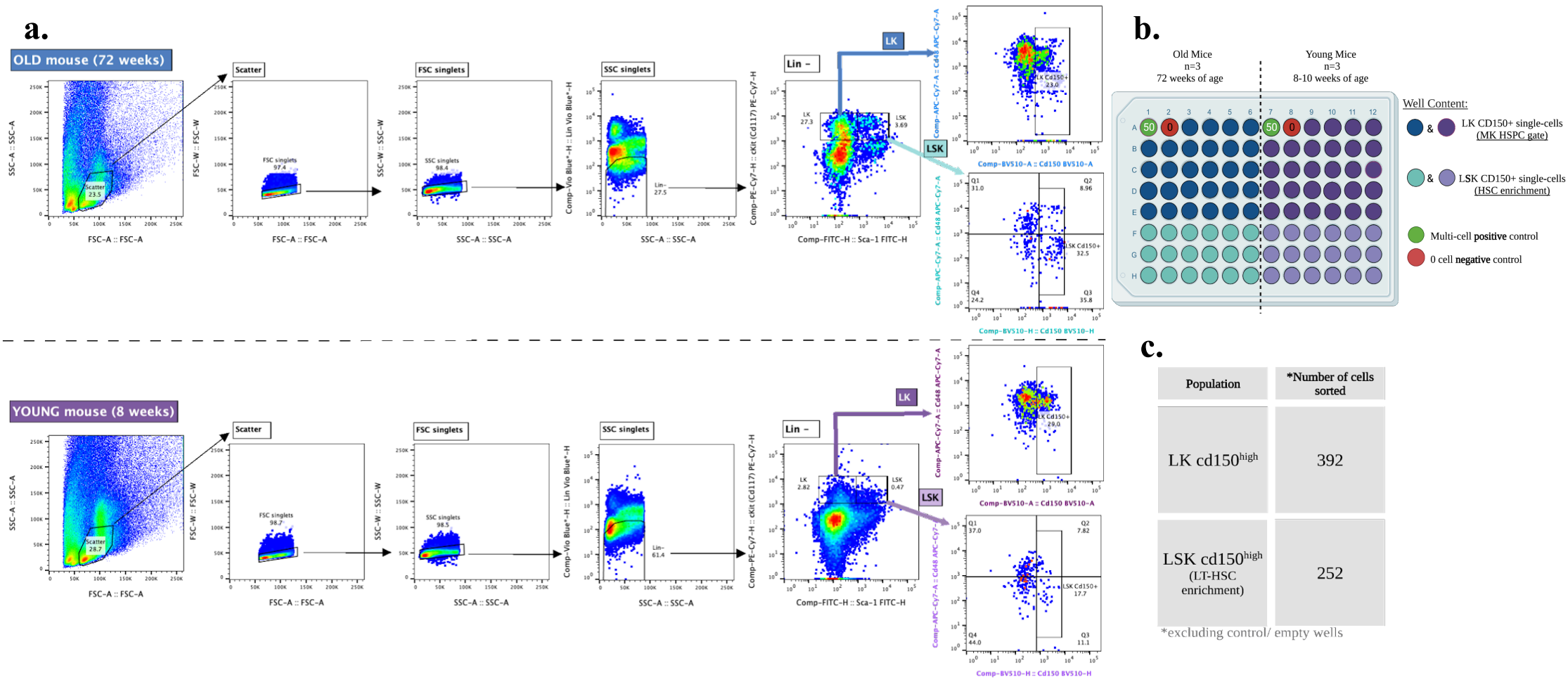


Figure 4.2. Cell isolation strategy of scRNA-seq data presented in Chapter 4. (a) FACS gating strategy for sorting LK and LSK Cd150+ cells. The same gating strategy was applied across all samples, shown is one representative example per experimental condition (top panel **aged,** bottom panel **young**). (b) Single-cell sort layout. Wells are coloured by FACS population, and the dotted line depicts the split across experimental conditions. Control well contents and locations are also shown. This layout was used across all plates in this chapter. (c) Summary of total cells isolated per FACS gate.

4.3.2 Quality control of scRNA-seq libraries and data integration for batch effect correction

After sequencing, scRNA-seq data preprocessing was performed as described (see Methods sections 2.3.1 - 2.3.2) and data quality was assessed per cell with MultiQC in the same way as described in Chapter 3. All data from every cell (including control samples) were retained as part of the dataset and parsed into *featureCounts* to generate one feature counts matrix based on mouse age, then split into two batches to generate 4 matrices (ie. young batch 1, old batch 1, young batch 2 and old batch 2). The decision to aggregate the data on both age and experiment, rather than age alone, was made as the cell sorting for these experiments was performed across two separate occasions, where on day 1 cell isolation was performed from old and young mice *30* and *31* respectively, whilst day 2 was the cell isolation from young mice *34* and *36* and old mice *35* and *37* (Table 4.1). As the primary objective for this set of experiments was to assess the differential gene expression signatures along the Mk lineage with age, it was important to account for batch effects between young and old samples to ensure that any observed differences in gene expression are genuinely due to age and not due to technical variability. Integrating scRNA-seq data from both experiments increased the sample size naturally improving the statistical power to detect age-related changes in gene expression. However, whilst the same methodology was applied, having involved two separate experiments adds the potential for experiment-specific technical artefacts which had to be accounted for during data integration, hence both batch and age variables were accounted for during data integration.

First, low-quality cells were excluded from the dataset prior to anchor identification for integration. The criteria for high-quality sample selection meant samples with over 50,000 total reads per cell, between 2,500-10,000 genes detected per cell and mitochondrial gene expression content below 15% were retained in the dataset. This left a combined total of 520 single-cell samples suitable for further analysis - excluding positive and negative control wells (Figures 4.3 and 4.4). After filtering for high-quality cells, each dataset was normalised (see Methods section 2.3.5) and the top 5000 variable features for each object calculated independently to identify integration features across the four objects. Once variable features per batch were calculated, a list of anchors between the batches were used to integrate samples back into a single dataset for further analysis. This approach uses the canonical correlation analysis (CCA) statistical technique that identifies shared sources of variation between multiple datasets by finding the canonical correlation vectors that maximise the correlation between the datasets (Stuart *et al.*, 2019). This cross-dataset alignment can then be used to generate a shared latent space that captures the biological variation across batches which can be used for downstream analyses. The goal of this was to improve the accuracy and reproducibility of downstream analyses, and facilitate the identification of biological insights in a way that takes into account the potential sources of batch effects, in this case, taking into account both mouse age and

experimental batch. The features identified as anchors using *SelectIntegrationFeatures* and *FindIntegrationAnchors* were provided as input to integrate the data using default parameters aside from *k.weight.* This is because the default value for *k.weight* in Seurat's *IntegrateData* function is *NULL*, meaning the optimal weighting for each dataset based on the number of cells and genes in each dataset is automatically calculated. However, after QC filtering the size of the datasets ie. the number of cells that remained in each of the 4 batches was not equal. In order to be able to combine the four objects, *k.weight* was set to 88, the highest value to enable integration which is equal to the number of cells in the object containing the least cells after QC.

Another common and unwanted source of variation that often plays an influential role in the downstream analysis is the strong cell-cycle gene expression signal inherent in scRNA-seq data. To minimise the influence of cell-cycle genes in downstream clustering and differential expression analyses, the integrated dataset was first scored into G2M, S and G1 phases based on the expression of pre-defined canonical cell-cycle-associated genes from the literature. Using these scores, the difference between the G2M and S phase scores was regressed prior to conducting downstream analyses.

Table 4.1. Summary of the total number of cells included per sequencing run, pooling strategy employed to sequence single-cell libraries. [14]

Sequencing Batch	Number of Plates	Young Mouse ID	Old Mouse ID	Illumina index set used for pooling	Total number of cells (excluding controls)
1	4	31	30	1, 2, 3 & 4	368
2	2	34	35	1, 2	184
	1	36	37	3	92

[14] Sequences for the four Illumina index sets of 96-well plates used in NextEra library preparation are listed in Supplementary Table 2.1.

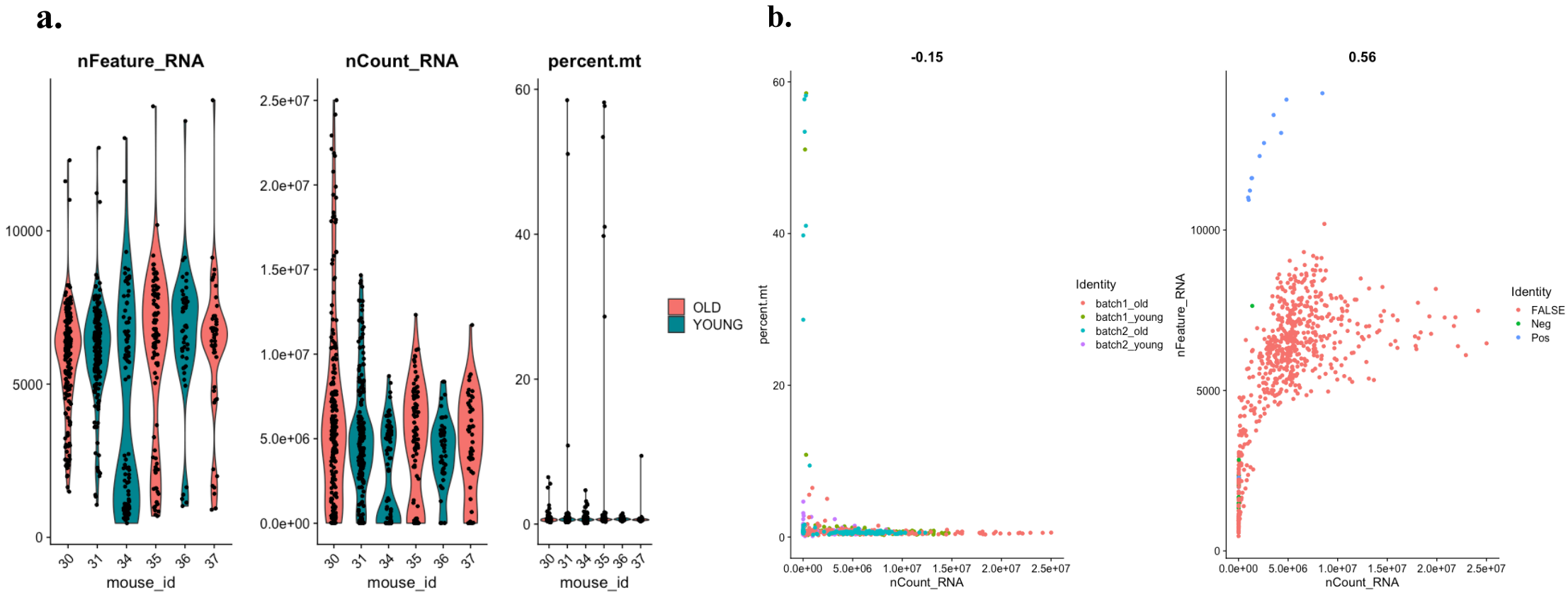


Figure 4.3. Single-cell sample quality control for selecting high-quality cells suitable for further analysis. (a) Violin plots showing: (1) number of genes (nFeature), (2) number of reads (nCount) and (3) percentage of reads attributed to mitochondrial genes (percent.mt) per single cell. Violins are coloured by mouse age (red = **old**, teal = **young**) and split by mouse ID. (b) (1) Correlation between the number of reads and the percentage of mitochondrial content per cell, coloured by batch ID and condition. (2) Correlation between the number of reads and the number of genes detected per cell, coloured by well type (**pos** = multi-cell well, **neg** = empty well, **false** = single-cell well).

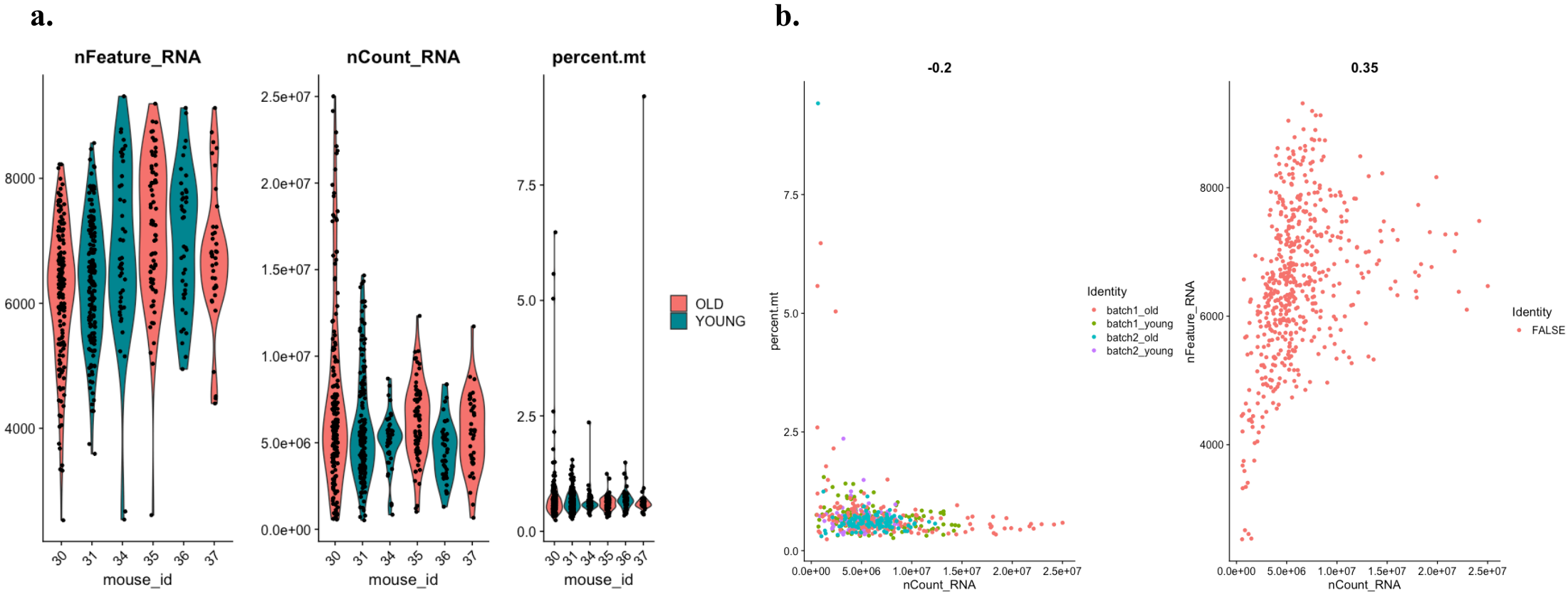


Figure 4.4. Single-cell sample quality control post-filtering based on quality metrics. (a) Violin plots showing: (1) number of genes (nFeature), (2) number of reads (nCount) and (3) percentage of reads attributed to mitochondrial genes (percent.mt) per single cell. Violins are coloured by mouse age (red = **old**, teal = **young**) and split by mouse ID. (b) (1) Correlation between the number of reads and the percentage of mitochondrial content per cell, coloured by batch ID and condition. (2) Correlation between the number of reads and the number of genes detected per cell, coloured by well type (**false** = single-cell well).

4.3.3. Principal component analysis of primary sources of variation across the dataset

PCA was performed on the preprocessed integrated data to identify the major sources of variability, also revealing the number of PCs that contained the most significant variability in the data and retained for downstream analysis. Of these, the gene found to display the highest standard variance across the dataset was Cd69, a transmembrane C-type lectin protein traditionally associated with early-leukocyte activation and expressed by activated T, B, and natural killer (NK) cells (Cebrián *et al.*, 1988; Testi *et al.*, 1990; Lauzurica *et al.*, 2000). The role of Cd69 is implicated during differentiation of progenitors, as a possible player in haematopoietic lineage commitment. Using integrated proteome, transcriptome, and DNA methylome analysis of HSCs and immature progenitors Cabezas-Wallscheid *et al.* found that Cd69 was among the most highly differentially expressed genes between HSCs and their downstream progeny (Cabezas-Wallscheid *et al.*, 2014). With that, heterogenous levels of Cd69 expression have also been reported within the HSC compartment, as an HSC-activation marker upon immune stimulation (Bujanover et al. 2018). More recently, heightened Cd69 expression was detected again in activated LSK CD48− 150+ stem and progenitor populations, most highly in HSCs, which was found to positively correlate with exit from quiescence (Thapa *et al.,* 2023). They also report Cd69 expression to be more responsive during the early phase of immune stimulation, and hypothesise that HSCs sense and can be activated at a low threshold, possibly in order to provide "training" through multiple immunological stimulations throughout life (Thapa *et al.,* 2023). Interestingly, Pang *et al.* identified increased Cd69 expression in myeloid-biassed HSCs with age (Pang *et al.,* 2011). Overall, further research is needed to elucidate the role of Cd69 expression in HSCs fully, but there exists some support for its role both in regulating haematopoietic differentiation in addition to serving as a marker of stem-cell activation.

Unsurprisingly, classic Mk markers *Pf4* and *Vwf* were among those most variable, as they are canonical markers for cells in the Mk lineage with known importance for Mk differentiation and identity (Figure 4.5A). Overall this initial insight, albeit preliminary, into variable genes detected across the dataset shows the heterogeneous expression of genes known to vary in expression across the haematopoietic stem and progenitor populations; particularly across Mk lineage commitment. It highlighted no indicators of low-quality data or batch effects (Figure 4.5C) among the results. To gain insights into genes likely driving the separation between different cell populations captured in the experiment, the contribution of genes in the first PCs was visualised using loading plots (Figure 4.5B). Here, genes strongly associated with a particular dimension are assigned a positive value, meaning that it is driving the separation whilst a negative value for a gene indicates that its expression is negatively correlated with a specific dimension ie. the gene is expressed at low levels in a particular population, and is therefore associated with the absence of that population along that dimension (Stuart *et al.*,

2018). This pulled out a high expression of markers for the expected haematopoietic cell populations, primarily HSCs, EryP and MkP, which were found to contribute to the same sources of variation in the data. For example, PC2 is negatively correlated with multiple genes associated with the Mk lineage (*Pf4, Mpl, Vwf, Pbx1,* and *Itga2b)* and has positive loadings for myelo/erythroid-associated genes *Car1, Trib2* and *Plac8* (Figure 4.5B) (Li *et al.*, 2013; Liang *et al.*, 2015; Upadhaya *et al.*, 2018).

Projecting cells into a reduced-dimensional space after PCA visualised the distribution of the cells captured in this dataset. By grouping cells based on both batch and experimental conditions (age), no discernable batch effect influencing the structure of the data was identified (Figure 4.5C). This PCA projection revealed the global structure of the data when condensed into its first principal components, but due to its limited capacity to visualise non-linear single-cell expression data was insufficient to reveal clusters. To enable cluster identification downstream, first, the number of components containing the highest sources of heterogeneity to include for downstream stages of analysis was determined by calculating the standard deviation of each PC. This was achieved using the Scree plot graphical method to identify significant PCs that capture meaningful biological variation. For this dataset, PC fifteen was determined as the cut-off for downstream analyses (Figure 4.5D).

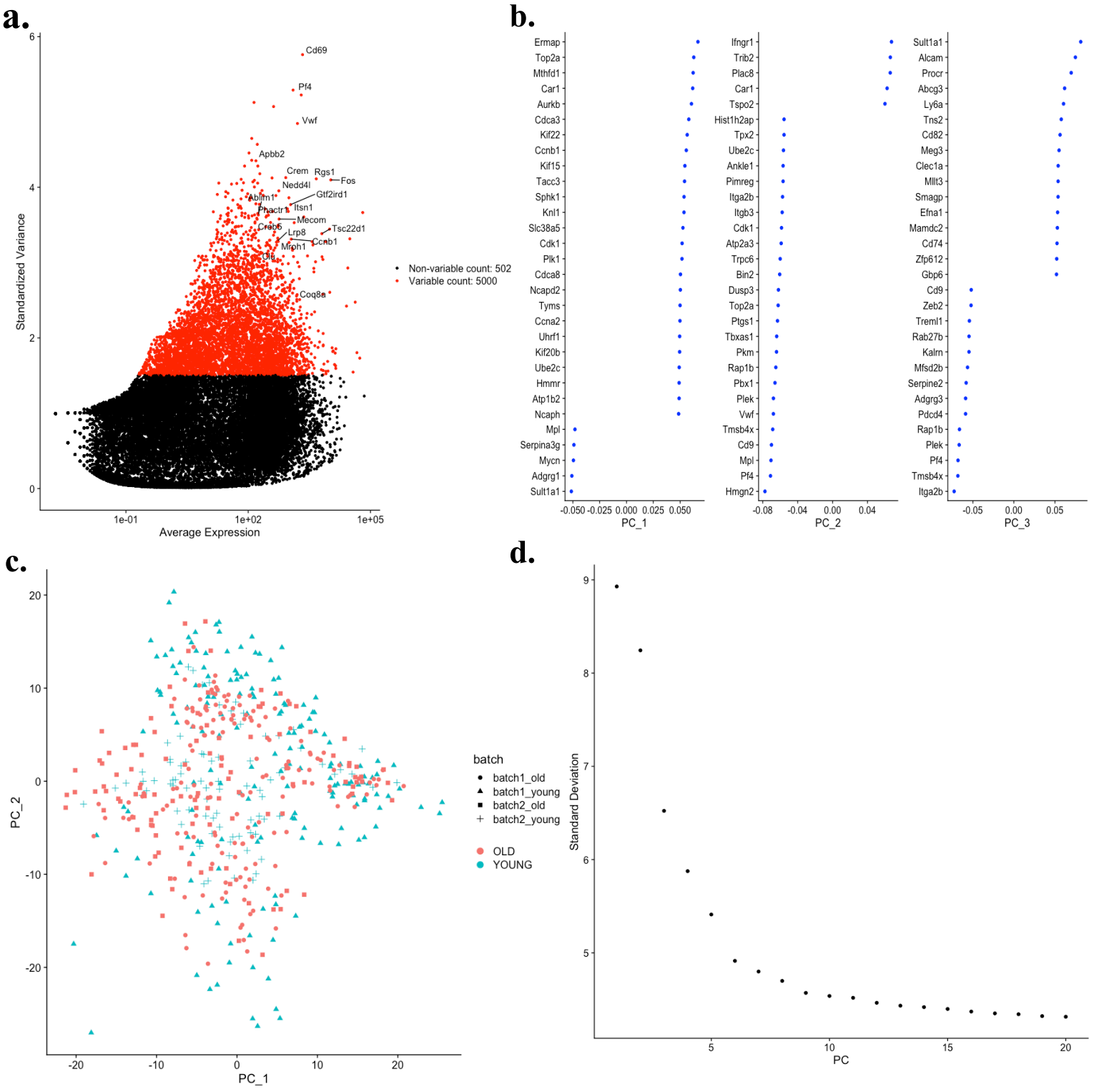


Figure 4.5. Principal component analysis of the dataset post-integration identifies the top principal components containing the most highly-variable genes and confirms no plate- or age-based batch effects. (a) The mean expression of each gene against its standard variance, with each dot representing a gene. Points in red indicate the top 5000 most highly variable genes used for downstream analysis, gene ID annotations shown for the top 20 (b) Gene loadings for the first three principal components showing the top genes contributing to each PC (c) PCA projection of all cells coloured by age where point shapes indicate batch in which samples were processed (d) Scree plot of standard deviation as a function of principal component number used to determine the number of components to include for downstream clustering.

4.3.4. Cell type annotation of clusters using marker expression signatures identifies cells captured in the ageing LK Cd150+ compartment

The integrated dataset containing all cells from both conditions that passed filtering was assigned into clusters using the *FindNeighbours()* and *FindClusters()* functions, respectively constructing a shared nearest-neighbour graph for the dataset based on the cell-cell similarities, followed by clustering cells using the first 15 PCs. Dimensionality reduction and unsupervised clustering were performed using UMAP and Louvain clustering with a resolution of 1.5 identifying 9 clusters across the dataset (Figure 4.6). Cells grouped into a cluster express a similar transcriptomics signature (set of genes) to one another relative to cells assigned into other clusters. Using these signatures in conjunction with known haematopoietic cell type-specific marker genes from relevant literature clusters were annotated by cell type (Paul *et al.*, 2015; Pellin *et al.*, 2019; Psaila *et al.*, 2020; Weinreb *et al.*, 2020; Roy *et al.*, 2021). Using the *FindAllMarkers()* function, sets of DEGs specific to particular clusters were identified and used to manually annotate clusters into cell types and compare expression signatures (Figure 4.7).

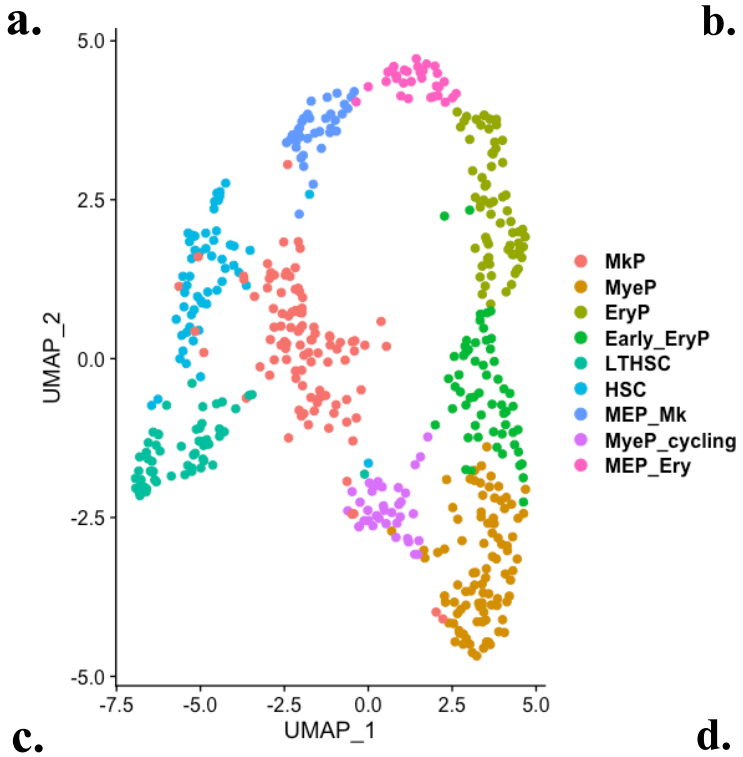
Cells in cluster 1 had a robust expression of canonical Mk-associated genes including Cd41 (*Itga2b*), *Pf4, Pbx1, Plek* and *Vwf.* This panel of genes is routinely used to identify phenotypic Mk-committed progenitors, thus with confidence cluster 1 was annotated as Mk progenitors (Supernat *et al.*, 2021).

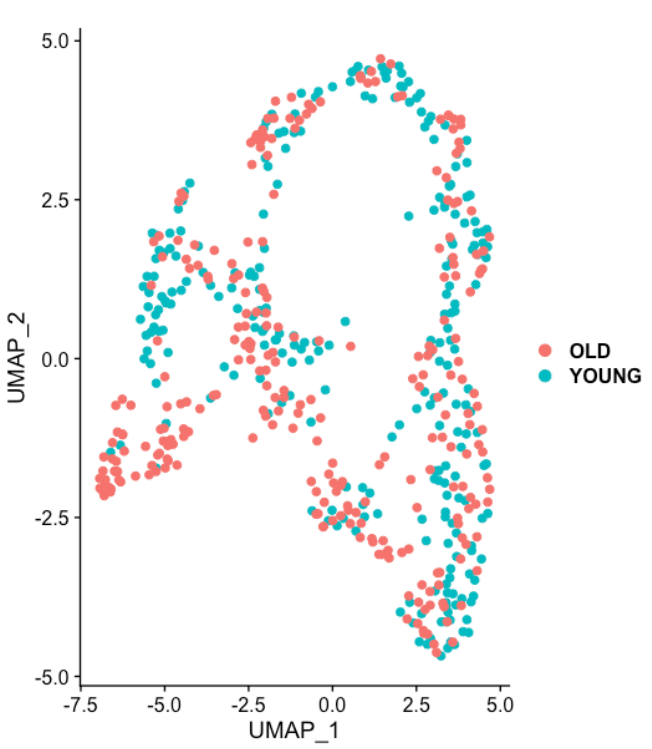
The signature of cells belonging to cluster 2 was more complex to assign to a specific cell type. One observation was that genes associated with the myeloid lineage - comprising neutrophils, monocytes, macrophages, and dendritic cells (DCs) - were found to be enriched in cluster 2. Some examples include *Plac8, Vim and Jun*, a transcription factor shown to be involved in activating myelomonocytic differentiation (Steidl *et al.*, 2006). *Irf1* was also identified as highly expressed in cluster 2, a member of the IFN regulatory factor (IRF) family that plays an important role in myeloid cell development and the maturation of the lymphoid lineage (R. Song *et al.*, 2021). Another gene of interest highly expressed in this population was *Egr1,* a zinc finger transcription factor known primarily for its role during cell growth, differentiation, and cellular depolarization (Sukhatme *et al.*, 1988). It is prominently involved in establishing early cell proliferative responses to extrinsic signals through transcriptional regulation of genes for growth and differentiation, but more recently has been associated with a shift in the differentiation behaviour of stem cell-enriched BM cells from granulocyte or erythroid lineages towards the macrophage lineage (Krishnaraju, Hoffman and Liebermann, 2001). The overexpression of *Egr1* in HSPCs was shown to result in failed BM engraftment upon transplantation in irradiated mice due to excessive differentiation bias towards the macrophage

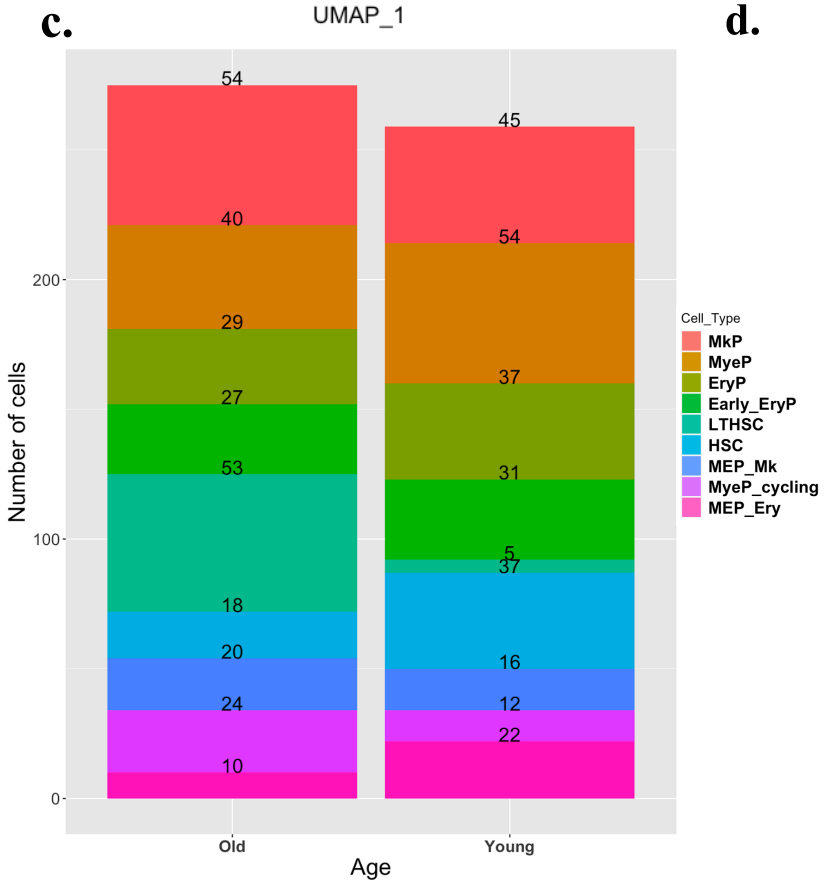
lineage - suggesting *Egr1* is a positive regulator for the myeloid fate. Asides from the increased expression in myeloid lineage-affiliated genes, it was also observed that markers typically affiliated with HSCs were also upregulated in cluster 2. This too includes *Egr1,* as not only has it been linked to myeloid fate but is also associated with HSC homeostatic regulation, specifically enhancing HSC quiescence and retention in the niche (Min *et al.*, 2008). Also highly expressed in cluster 2 was *Cd69*, known primarily as an early activation marker that is expressed in HSCs and lymphoid cells (Lauzurica *et al.*, 2000). *Cd69* upregulation is correlated with quiescence and low proliferative potential in HSCs, together with *Junb* and *Btg2,* which too were identified here as markers of cells in cluster 2 (Steidl *et al.*, 2006; Desterke, Bennaceur-Griscelli and Turhan, 2021). Taken together, the signature observed for cluster 2 is suggestive that this cluster may comprise a highly immature progenitor subset, with overlapping signature to HSCs, but the increased expression in myeloid lineage-associated genes thus was annotated as immature myeloid progenitors (MyeP).

Clusters 3 and 4 were assigned as EryPs and early EryPs respectively based on the high expression of canonical erythroid marker genes including the early specific marker of the erythroid differentiation *Car1,* erythroid membrane-associated protein (*Ermap),* erythropoietin receptor *Epor* and well-established erythroid TF *Klf1* (Su *et al.*, 2001; Dolznig *et al.*, 2006; Song *et al.*, 2012; Merryweather-Clarke *et al.*, 2016). Compared to one another, cells in cluster 3 expressed genes synonymously associated with committed erythroid progenitors, in particular genes associated with terminal stages commitment including blood antigen protein *Rhd, Sphk1* and *Gata1* (Kingsley *et al.*, 2013; Xiong *et al.*, 2014). Whilst cluster 4 exhibited higher expression of immature progenitor markers such as *Gata2,* known for being expressed in HSPCs before the commitment of progenitors (Suzuki *et al.*, 2013).

Cluster 7 markers show an overlapping expression signature with MkP cluster 1 including Mk-lineage associated genes *Itga2b, Vwf, Pf4* and *Mef2c* as well as a notable upregulation of G2M cycle phase genes which had a dominating effect during marker identification (Gekas *et al.*, 2009; Lambert *et al.*, 2009; Miyawaki *et al.*, 2017). In the same manner, Cluster 9 exhibits expression of Ery-associated markers *Ermap*, *Epor, Blvrb* and *Tfrc* (Cd71) (Dolznig *et al.*, 2006; Merryweather-Clarke *et al.*, 2016; Grzywa, Nowis and Golab, 2021). Comparing the markers in the earlier chapter with the signatures identified here clusters 7 and 9 were annotated sub-clustered MEP progenitors, with either higher Mk-lineage gene expression (Mk-MEP, 7) or Ery-lineage gene expression (Ery-MEP, 9) respectively. Exploring different resolution parameters during clustering, as before, showed these share an overlapping signature that can be seen in the heatmap of top marker genes per cluster (Figure 4.7).







d.

Cell type	% YOUNG	% OLD	Total cells per cluster
MkP	17.4	19.6	99
MyeP	20.8	14.6	94
EryP	14.3	10.5	66
Early EryP	12	9.8	58
LTHSC	1.9	19.3	58
HSC	14.3	6.6	55
MEP (Mk)	6.2	7.3	36
MyeP (cycling)	4.6	8.7	36
MEP (Ery)	8.5	3.6	32

Figure 4.6. Single-cell clustering. (a) UMAP projection of integrated dataset where each point represents a single cell, coloured by cell-type annotation (b) UMAP projection of single-cell clusters coloured based on mouse age (c) Stacked frequency plot showing the number of cells per cluster in each age group. Numbers above each stacked bar represent the number of cells found in each cluster (d) Percentage contribution of cells from each condition across clusters.

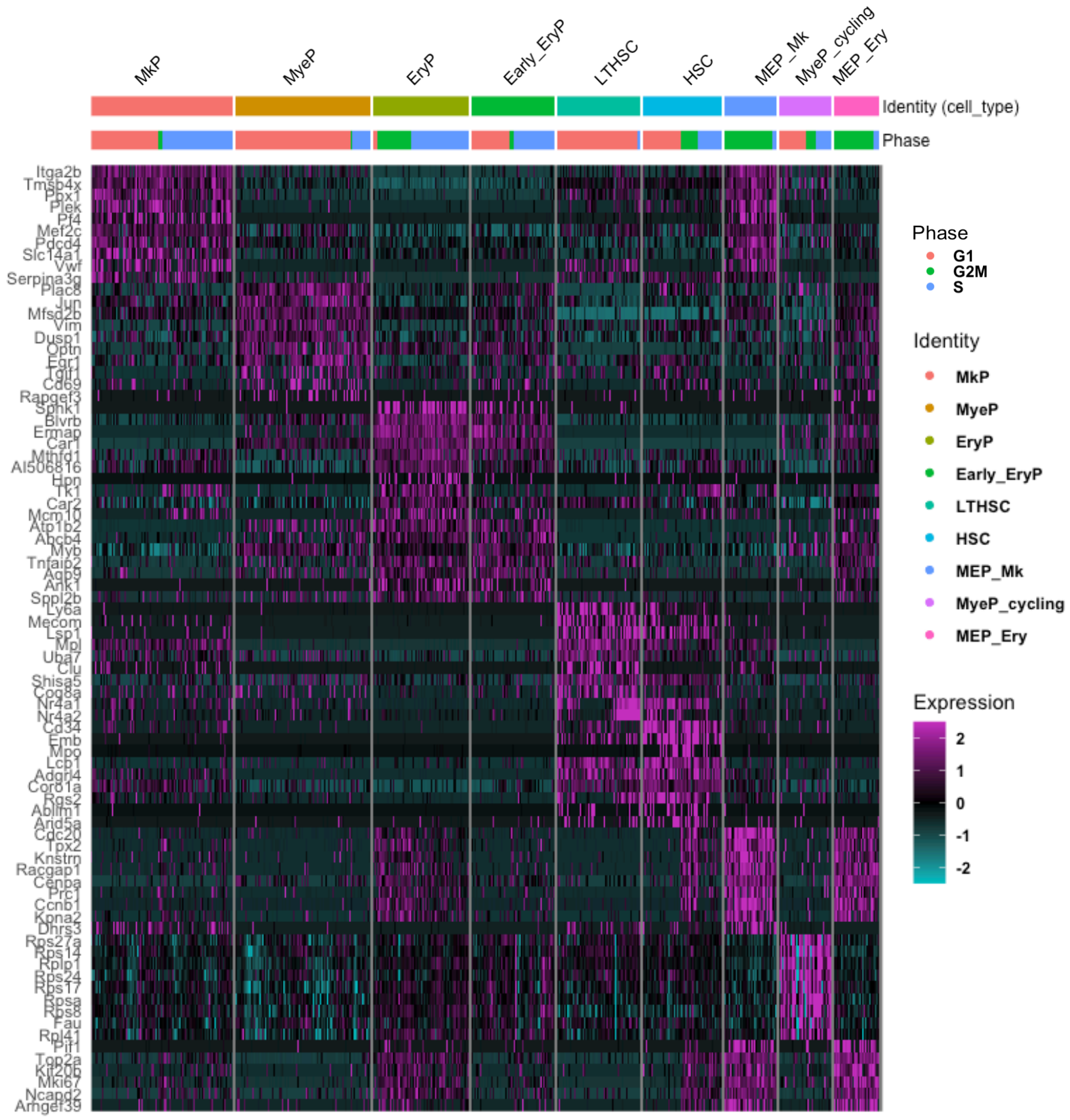


Figure 4.7. Heatmap showing the distribution of expression levels of the top 10 markers per cluster. Coloured bars on Y-axis indicate cluster cell-type annotation and cell-cycle phase classification.

Using canonical HSC-specific markers including *Mecom, Ly6a (*Sca-1), *Procr* and *Sult1a1* cluster 5 was annotated as LTHSCs, whilst cells in cluster 6 were assigned as HSCs that share a similar expression pattern with some exceptions that enable their distinction from one another (Osawa *et al.*, 1996; Balazs *et al.*, 2006; Qian *et al.*, 2007; Mohan Rao, Esmon and Pendurthi, 2014). This includes higher expression of *Cd69, Flt3,* and *Egr1* and conversely near absent levels of *Vwf* and *Clu* (Figure 4.8) (Peled *et al.*, 1999; Adolfsson *et al.*, 2001; Min *et al.*, 2008; Sanjuan-Pla *et al.*, 2013; Koide *et al.*, 2021).

The proportion of cells belonging to each cluster between the conditions was largely equivalent with a few notable distinctions. The number of LTHSCs from old samples vastly exceeded that of young samples; 19.3% of cells from aged mice were LTHSCs compared to only 1.9% from young samples (Figure 4.6D). The expansion of the HSC pool with age is well documented, with multiple reports demonstrating their increased frequency but reduced self-renewal and functional capacity (Rossi *et al.*, 2005; Chambers *et al.*, 2007; Yamamoto *et al.*, 2013; Yang and de Haan, 2021). Multiple intrinsic and extrinsic factors including changes to cell division kinetics, increased inflammation, and other age-associated changes to the HSC niche are among the hypotheses explaining this phenotype (Mirantes, Passegué and Pietras, 2014; Florian *et al.*, 2018; Pinho *et al.*, 2018). Specifically, it has been previously shown that Mk-biassed Cd150+ HSCs increase with age, with an altered expression signature and increased expression in key genes including *Slamf1* (Cd150), *Vwf* and *Mpl*; all of which were upregulated in cells from old vs young mice in this data (Figure 4.9) (Sanjuan-Pla *et al.*, 2013; Grover *et al.*, 2016). Mirroring this, 14.3% of cells from young mice contribute towards the HSC population compared with only 6.6% contribution from aged mice (Figure 4.6D). Notably, along with the expression of HSC markers, the expression of genes associated with the lymphoid lineage, such as *Flt3, Cd69, Dntt* and *Plac8* was also detected in this population (Figure 4.8). These differences in expression across the HSC populations and differential cluster contribution with age observed here are in agreement with the existing literature showing a heterogeneous HSC compartment, increased Mk lineage signature in ageing and that the expression of Mk-lineage genes correlates with signatures of the most primitive stem cells within the LSK compartment.

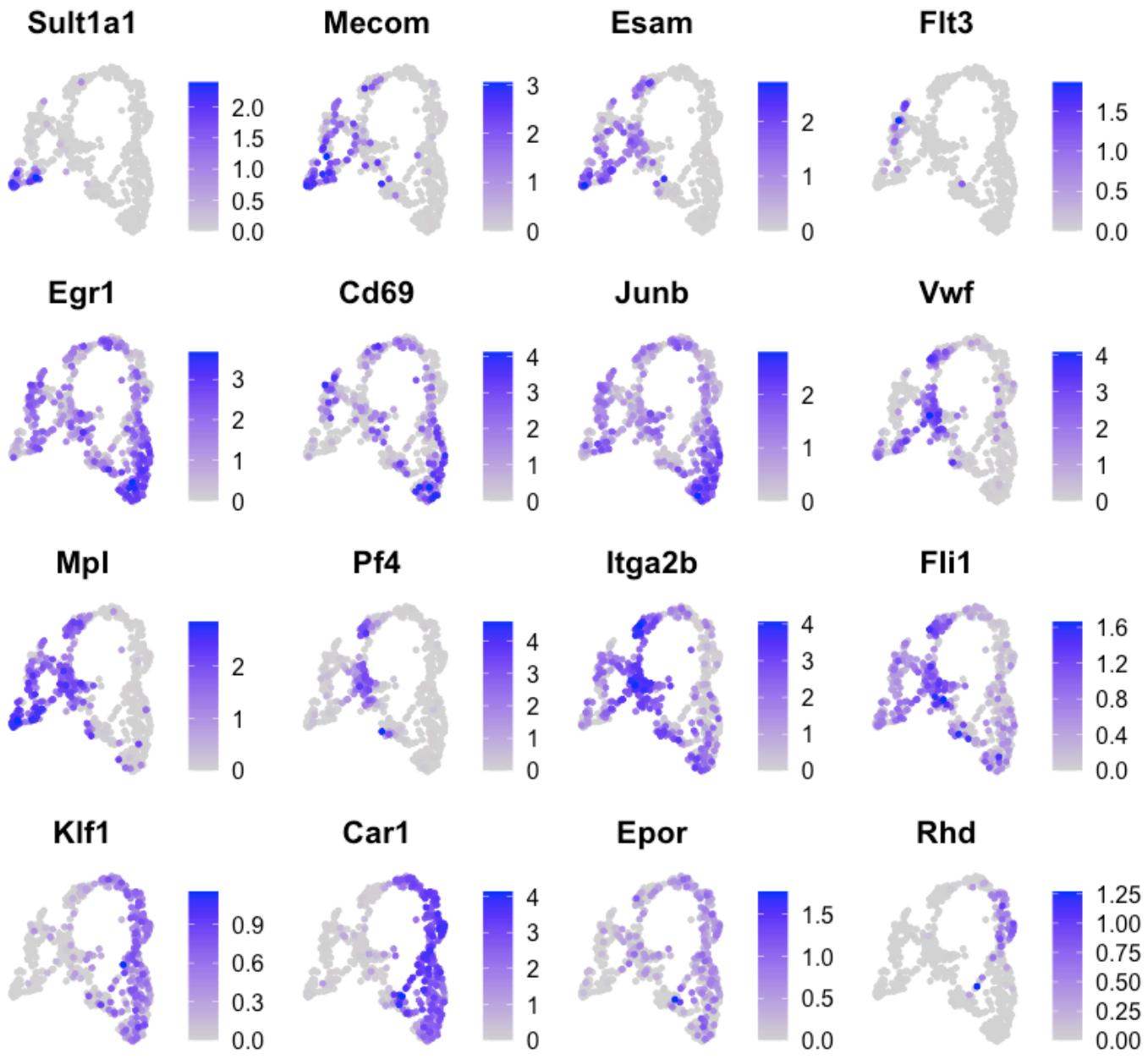


Figure 4.8. UMAP projection of single cells coloured by expression levels in a selected subset of canonical haematopoietic markers from existing literature used for manual cell-type annotation.

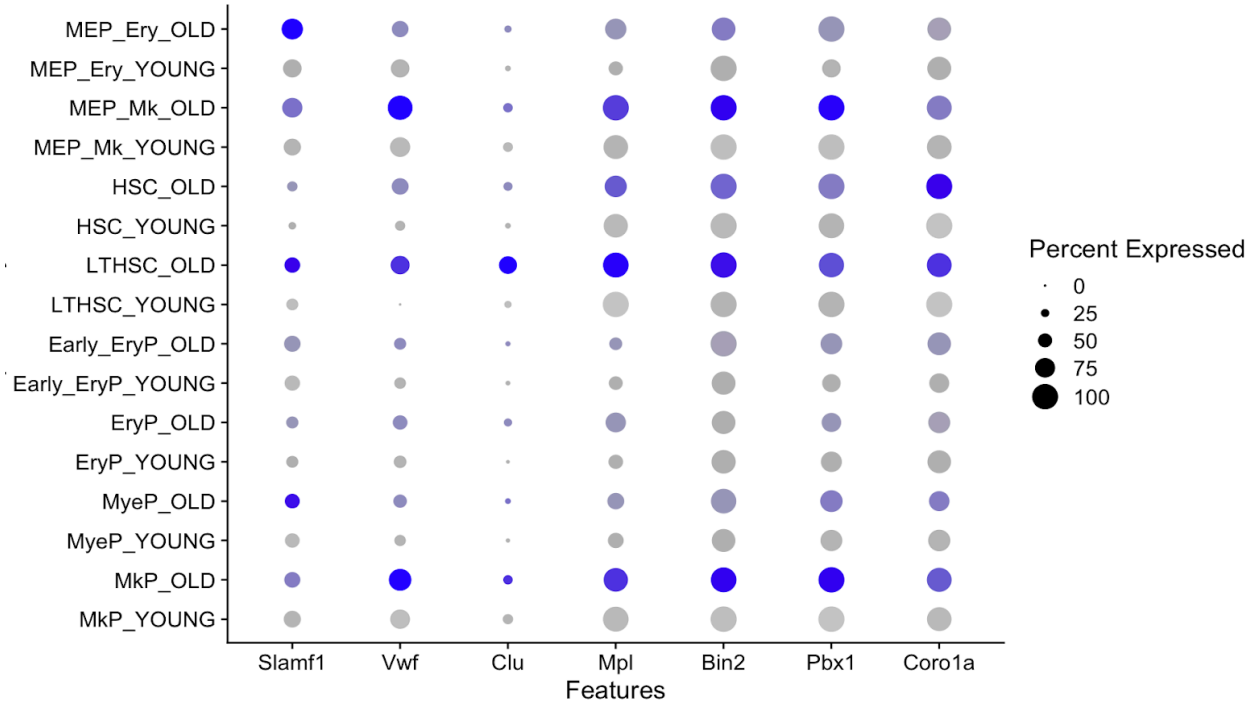


Figure 4.9. Expression of a subset of shared HSC and Mk markers across clusters separated by condition. Point size indicates the percentage of cells in the cluster that express the gene and opacity of colour indicates expression level.

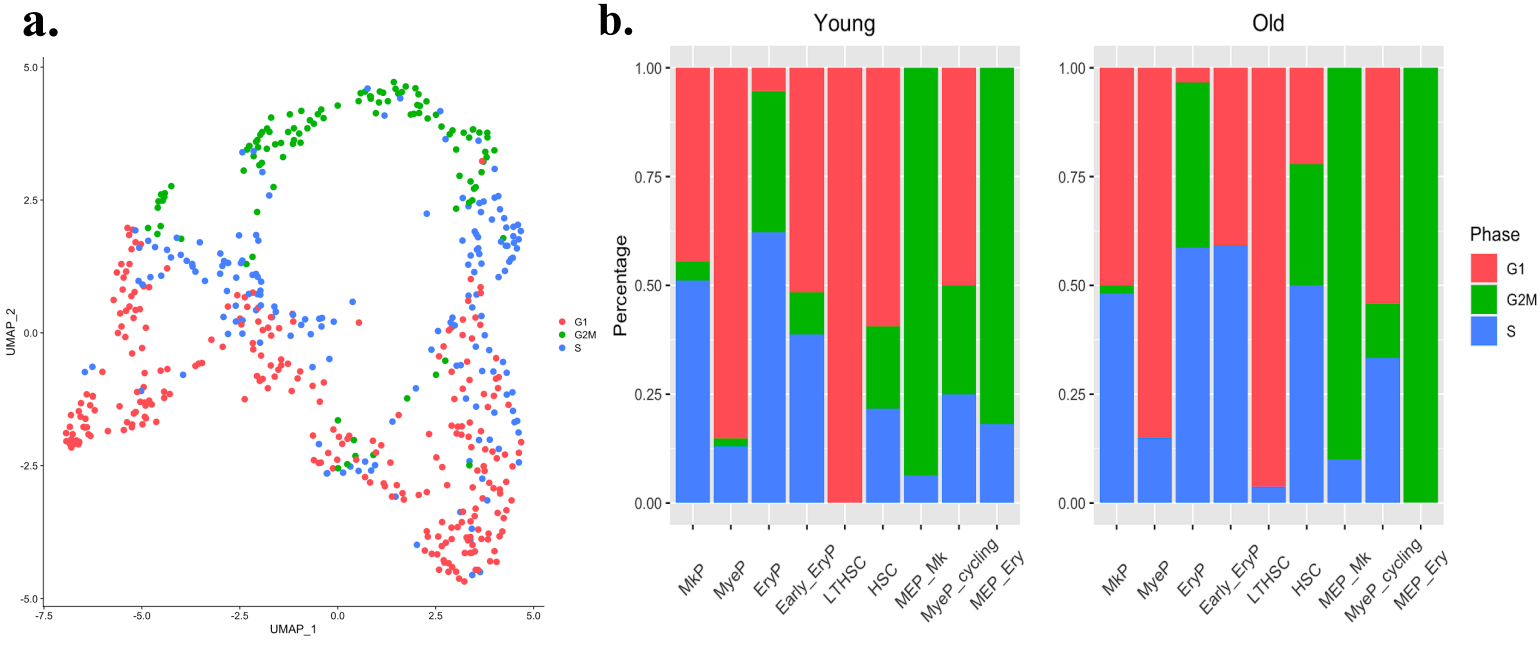


Figure 4.10. Cell cycle stage assignment across clusters. (a) UMAP projection of all single cells coloured by cell cycle stage (b) Distribution of cell cycle assignments shown as a percentage each cluster split by mouse age.

Cell cycle stage-specific patterns of expression were apparent throughout clustering and studying marker expression signatures. In particular, there was an evident abundance of G2M-associated genes identified as markers for both MEP subpopulations, such as high expression of *Spc24* and *Cdca3* (Figure 4.10A) (Xu *et al.*, 2022; Chen *et al.*, 2023). Cell cycle stage distribution across data from young and old samples was visualised to compare differential frequencies in cell cycle assignment between conditions (Figure 4.10B). LTHSCs and HSCs from aged mice had an increased proportion of cells undergoing cell division (in S and G2M phases) (Figure 4.10b). Previous work has demonstrated poor self-renewal of HSCs and a shift from asymmetric cell division (generating one stem and one differentiated progenitor) to symmetric cell division (generating either only stem cell clones or only differentiated progenitors) with age (Dykstra *et al.*, 2011; Florian *et al.*, 2018). Increased cycling in the aged HSC compartment correlates with deficiencies in DNA damage repair pathways and an accumulation of damage in the HSC compartment, which continues to serve as a strong explanation for the rapid deterioration of function in the HSCs and the heightened incidence of disease with age (Rossi, Bryder, *et al.*, 2007; Walter *et al.*, 2015). The underlying cause(s) of this behaviour remain to be fully elucidated, and whilst cell-cycle behaviour was not directly assayed the use of known cell-cycle associated markers provided supplementary support to the annotations cells were assigned, as well as additional insight into the state of cells at the time they were collected. This is of particular relevance as recent findings have implicated cell-cycle progression playing a direct role in the lineage fate determination of MEPs (Lu, Krause, *et al.*, 2018). The authors demonstrated that cell-cycle control influenced Mk vs Ery specification of MEPs, where MEPs with a smaller increase in cell cycle speed promoted Mk progenitor specification determined by the increased expression of Mk-specific genes, whilst a greater increase resulted in higher Ery-specific gene expression (Lu, Sanada, *et al.*, 2018). They classified MEPs as a single ‘transition state’ distinct from the CMP or MkP and EryP populations. The data presented here are in agreement with this but additionally, resolve the MEP state into subpopulations with distinct signatures. As predicted, both populations exhibit a highly active cell state across both age groups. Their profiles are retained with age with the exception of the MEP Ery population, which was found to contain only cells in G2M. Previous work has shown that the shortening of the S phase is essential for Ery differentiation (Hwang *et al.*, 2017), suggesting a prospective role for changes in cell cycle kinetics influencing MEP fate specification with age. Whilst this hypothesis is as of yet unsubstantiated, it poses an interesting avenue for further research.

Overall, the clustering results and analysis of cell-type specific marker expression and cell cycle states identified 9 populations of stem and progenitors of the Mk and Ery lineages that are in agreement with the existing literature and provided insights into the signatures and cell cycle stages of cells within the LSK Cd150+ compartment.

4.3.6 Pseudotime analysis orders differentiation trajectory from LT HSC towards Mk and Ery lineages.

A key objective of this chapter was to delineate the trajectory from stem cells toward the Mk lineage, demonstrating gene expression changes along different stages of Mk differentiation and how this is affected in ageing. After establishing the populations captured in the experiment through cell-type annotation in *Seurat,* the *Seurat* object was then converted into the necessary *cds* data format to perform pseudotime analysis with *Monocle3*. This ensured that only the same high-quality cells that passed QC and annotated were retained for trajectory inference and downstream analyses. Whilst the removal of samples was not required, other important preprocessing steps including log and size factor normalisation, scaling and dimensional reduction were performed before running trajectory reconstruction. These steps address depth differences across the dataset ensuring genes contribute equally to the analysis to account for possible differences in sequencing depth or other technical factors inherent in scRNA-seq data. Data preprocessing was performed using default recommended parameters, and the first fifty principal components for the data were calculated for further dimensionality reduction downstream. Batch correction was also performed with the *aling_cds()* function, using both experimental condition (mouse age, 'young' or 'old') and experiment (batch 1 or batch 2) as input groups for *Batchelor* alignment (Haghverdi *et al.*, 2018). Finally after preprocessing and batch correction and confirming the number of PCs containing the highest variance in the data (Figure 4.11A), data dimensionality was reduced with UMAP with *max_components()* set to 2 to facilitate trajectory inference downstream.

A necessary prerequisite for trajectory reconstruction with *Monocle3* is data clustering into partitions. Unsupervised *Leiden* clustering was performed with *k* and *resolution* parameters set to 5 and 1e-3 respectively, partitioning cells into 10 clusters. These assignments were consistent with clusters identified in *Seurat* and meant that annotations assigned to cells were sufficiently robust for labels to be transferred and used for ordering single cells along a pseudotime trajectory (Figure 4.11B).

To construct a trajectory the function *learn_graph()* was applied to the UMAP projection, creating a path of connecting points across clusters enabling semi-supervised pseudotime ordering. The only input parameter used for ordering cells was setting the root point of the trajectory as cells in the LT-HSC cluster. This assigned each cell with a numerical pseudotime value, which enabled cells to be ordered from low to high pseudotime states (Figure 4.11C-D).

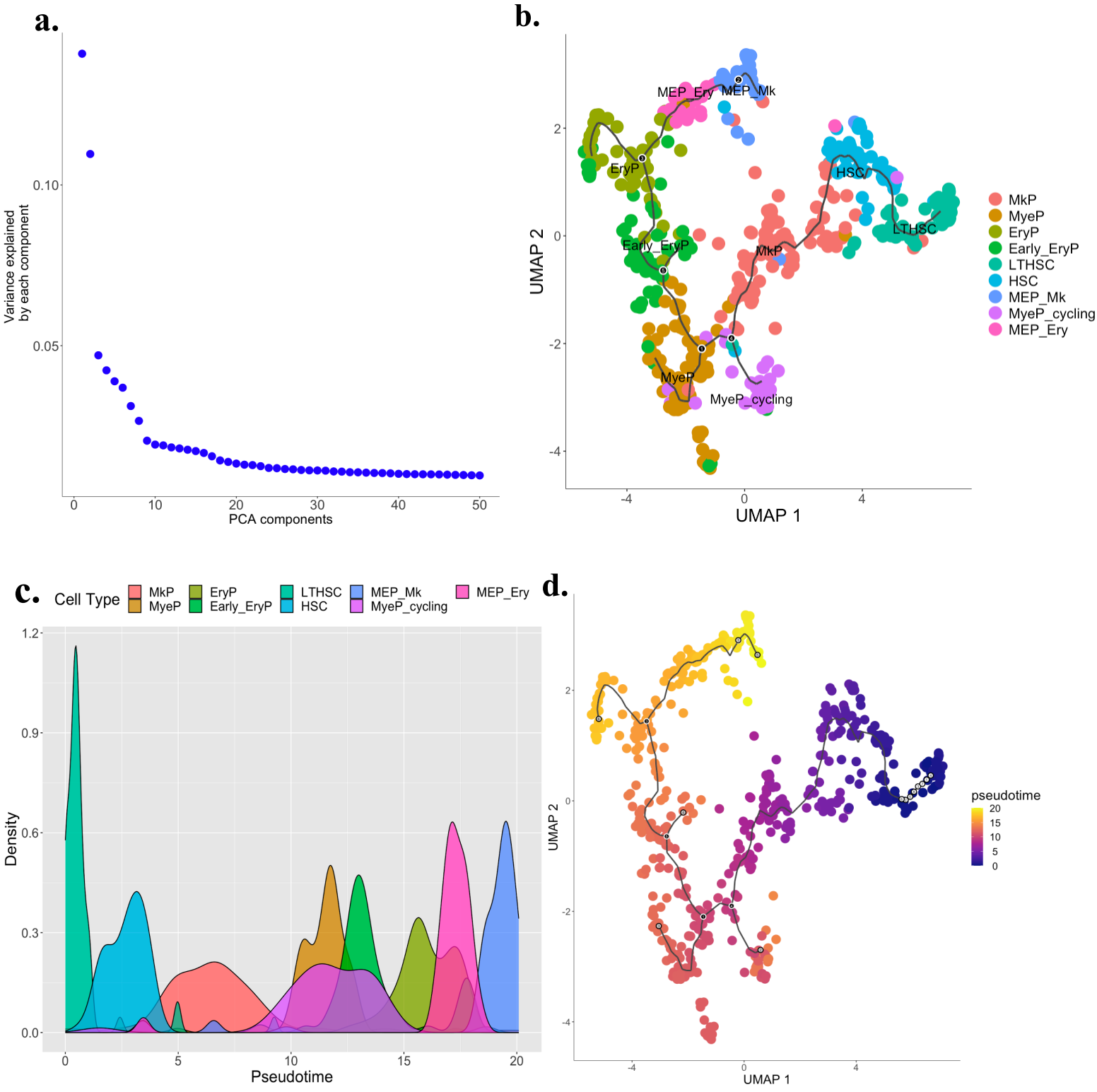


Figure 4.11. Pseudotime analysis of Mk/Ery differentiation. (a) Principal component analysis used to asses variance explained across principal components for pseudotime mapping (b) UMAP projection of single cell trajectory coloured by pseudotime (c) Density plot of the distribution pattern by all clusters across pseudotime (d) UMAP projection of single cell trajectory coloured by cell-type annotations performed in *Seurat*.

Pseudotime ordering generated a trajectory of single cells starting from LTHSCs, followed by HSCs and MkPs as the next earliest pseudotime states. This result was unsurprising as MkPs are known to share many common features with HSCs, and substantial evidence has shown expression of Mk-associated genes occurring in the earliest stages of HSC differentiation (Figure 4.11C) (Huang and Cantor, 2009; Sanjuan-Pla *et al.*, 2013).

The trajectory continues with the MyeP progenitor populations preceding early and late EryPs. This order supported the rationale from the cluster signatures; whereby the MyeP progenitor populations exhibit heightened expression in genes associated with immature stages of haematopoietic commitment, such as *Gata2,* whilst the EryP subpopulations exhibit lower *Gata2* and higher levels of *Gata1* and other markers associated with Ery lineage commitment (Suzuki *et al.*, 2013).

Finally, cells assigned the largest pseudotime values by *Monocle3* were both MEP populations. Whilst this result might dictate that MEPs exist later along the trajectory of lineage commitment from HSCs, it is important to take into account the influence of graph topology in the UMAP projection of cells used for pseudotime analysis. *Monocle3* utilises UMAP as the representation of the underlying structure of the data for clustering, which by definition places cells with similar expressions closer in proximity. The UMAP projection shows EryP and Ery MEPs in close proximity, followed by Mk MEPs which share a high degree of similarity in expression to Ery MEPs (Figure 4.7). However, tools for pseudotime analysis will utilise all shared expression patterns to calculate its projection irrespective of dominating signatures that may be of less biological significance or potential batch effects. Visual inspection of the UMAP shows the Mk MEPs proximal to MkPs, but the higher overlap with Ery MEPs, likely based on the higher proportion of cells expressing G2M-associated cell-cycle genes in Ery MEPs and EryPs than MkPs (Figure 4.10B). This meant that calculating the path connecting the data, on which pseudotime ordering relies, resulted in Mk MEPs being placed as the cell type furthest along the trajectory; which is not consistent with the expression of marker genes for this population (Figure 4.11D). This highlights a known limitation of graph-based pseudotime approaches, the importance of careful interpretation of pseudotime ordering results, and the requirement for validation through examining the expression of known marker genes and comparing results to independent datasets.

Cells were grouped into bins at regular intervals that spanned the pseudotime value range, assigning cells into pseudotime ‘states’ from 0-7 (Figure 4.12A). These states batched together cells within sets of pseudotime values ie. within close proximity to each other in the context of pseudotime, irrespective of cell type identity and enabling signatures across successive stages of pseudotime to be determined.

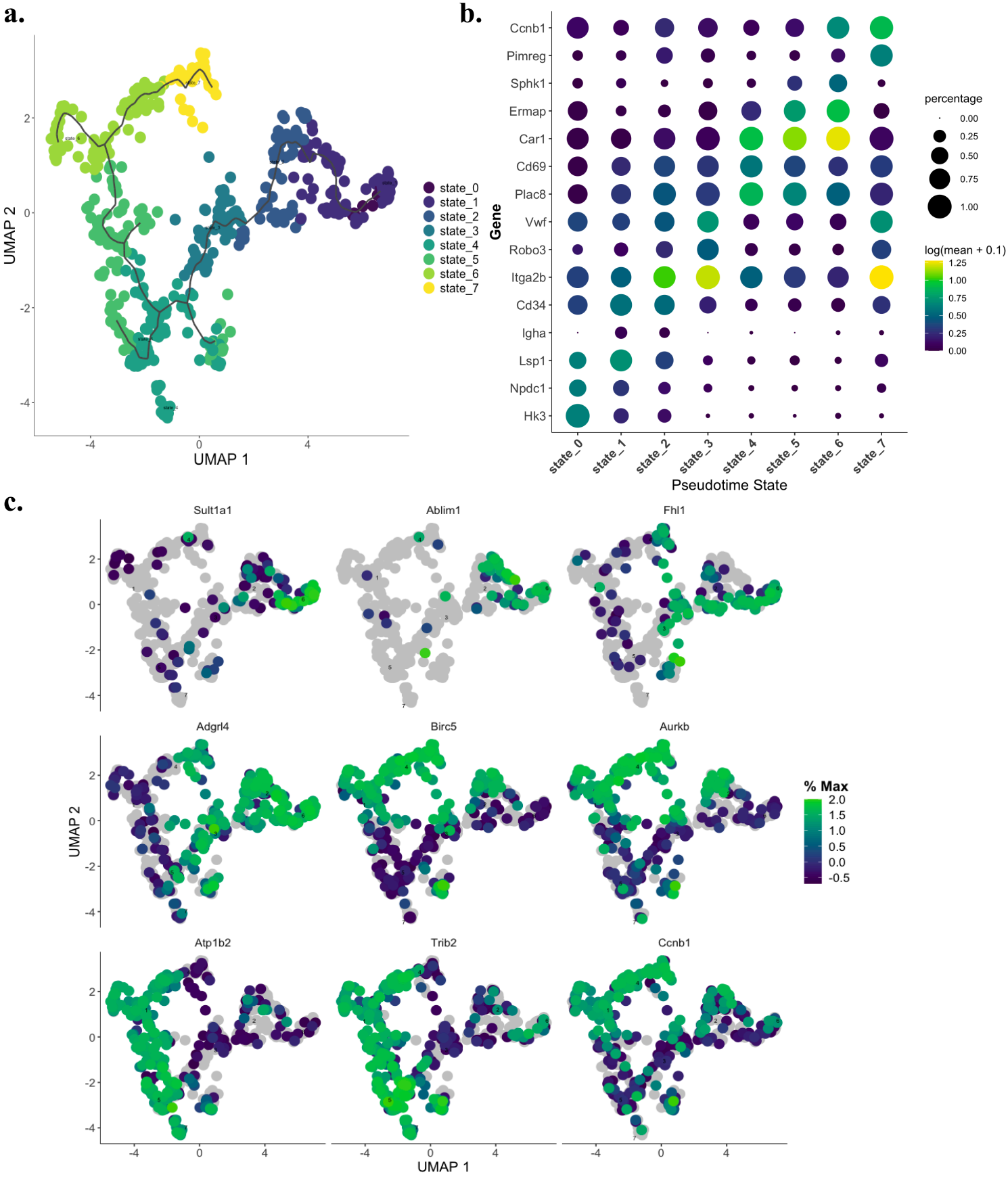


Figure 4.12. Pseudotime state analysis. (a) UMAP of pseudotime trajectory coloured by states (b) Gene markers with highest log fold-change and statistical significance across pseudotime states (c) Expression levels in genes which exhibit differential expression with pseudotime.

As one would expect, the more cells of the same cell-type (and therefore that have a greater set of genes in common) within a cell state, the more dominating the cell-type signature was on the signature for that state. This becomes clear when studying the markers across each pseudotime state with the highest log fold-change difference to other states. Visualisation of the top marker with the highest log fold-change, and the top marker with the highest adjuster p value per state exemplifies cell-type specific genes marking pseudotime states, for instance Mk marker *Itga2b* (Cd41) in state 2 and Ery marker *Car1* in states 5 and 6 having the highest fold-change difference in expression (Figure 4.12b). The gene with the highest statistical significance in marking cells at the most primitive pseudotime state (state 0) was *Hk3*. This gene encodes hexokinase 3, one of the four isoforms of the hexokinase family of enzymes that are important for catalysing the initial step of glycolysis (ATP-dependent phosphorylation of glucose to glucose-6-phosphate) (González *et al.*, 1964). It is primarily expressed in haematopoietic cells, and is notably distinct to its isozyme family members by lacking the N-terminal sequence necessary to bind to mitochondria but still exhibits prosurvival functions (Preller and Wilson, 1992; Wilson, 2003). Relatively little is known about the function of *Hk3*, but it has been identified as a target of the PU.1 transcriptional factor that is thought to be activated during neutrophil differentiation, potentially serving as a supportive mechanism for glycolysis under anaerobic conditions (Federzoni *et al.*, 2012). More recent work has revealed *Hk3* is significantly induced during myeloid differentiation, is upregulated during terminal differentiation in AML cell line models and upregulated during *in vitro* myeloid differentiation of HSPCs (Seiler *et al.*, 2022). Here, *Hk3* was found to be the most significant gene in identifying the earliest pseudotime state of the trajectory, while *Hk1* and *2* are expressed ubiquitously this isotype was exclusively expressed at the stem cell level. This, together with the existing literature that implicates a more lineage-specific role for *Hk3* compared to its isoform counterparts suggests a potential role of *Hk3* in serving a distinct metabolic purpose that is particularly important in stem cells and across specific haematopoietic lineages.

To calculate changes in gene expression along the full span of the trajectory at the pseudotime value level, the *graph_test()* function was performed across all cells. This identified 4,966 genes that vary significantly (adjusted $P<0.05$) in expression along the pseudotime trajectory, assigning each gene within this list a Moran's *I* value between -1 to 1 indicating the correlation in expression of each gene between neighbouring cells. Genes of known variable expression through differentiation were among the top genes with the highest Moran's test statistic, including HSC-specific markers, cell cycle-associated genes and lineage-restricted genes. Moreover, genes with implications in Mk differentiation and function were also among the top most significant such as *Tmsb4x* (Figure 4.12C). *Birc5* and *Aurkb,* two components of the central spindle that are upregulated during Mk differentiation with implications in cytokinesis and promoting polyploidization were also found to vary significantly across pseudotime (Y.

Zhang *et al.*, 2004; Wen *et al.*, 2012) (Figure 4.12C). The DEGs identified here make up the global picture of gene expression changes that were identified along the pseudotime trajectory. To compile this output and streamline the interpretation of these results, gene modules were calculated 'clustering' the list of genes co-expressed across cells into groups (Figure 4.13). This approach condenses the DEGs information into fewer variables, facilitating their interpretation based on the assumption that functionally related sets will be grouped. This enabled the modules to be analysed within variables of interest to provide a more biologically meaningful interpretation of the data such as the identification of cell type or condition-dependent patterns of expression. A total of 37 different gene modules were generated from the pseudotime DEGs.

Visualisation of the relationship between grouped genes across the different cell populations highlighted individual and sets of modules of interest (Figure 4.13A). Modules 10 and 15 were strongly correlated with both MEP subpopulations and poorly correlated with other cell types, indicating that the set of genes within these modules are highly expressed in most cells belonging to these populations (Figure 4.13B). Multiple cell types had more than 1 correlating module, for example, module numbers 9, 4 and 20 all were enriched on the LTHSC and HSCs populations, but 11 and 16 exclusively within LTHSCs - suggesting that genes that distinguish LTHSCs from HSCs are contained within these 2 groups. Of particular interest, modules 2 and 5 exhibited clear inversed enrichment patterns; where module 2 is high in LTHSCs, HSCs, MkPs and Mk-MEPs and a distinctly negative correlation with other cell types, whilst module 5 mirrors the inverse pattern showing enrichment exclusively across EryPs MyePs and Ery MEPs (Figure 4.13B). This reveals a clear distinction of the gene sets that were differentially expressed between the Mk and Ery lineages, enabling the interpretation of DEGs in the context of the lineage they were exhibiting variable expression along pseudotime (Figure 4.14).

Of the genes with variable expression in pseudotime identified, 245 and 218 genes were grouped in modules 2 and 5 respectively. Module 2 included multiple genes that have been previously demonstrated to be upregulated during Mk differentiation from HSC serving as a positive indication that module 2 is enriched for an Mk-associated DEG signature, with examples such as *Vwf, Rgs18, Fli1, Mpl, Selp, Pbx1, Bin2* and *Cavin2 (Debili et al., 1995; Klimchenko et al., 2009; Sengupta et al., 2013; Zhu et al., 2018; Walker et al., 2022)*. In the same way, module 5 was enriched for erythroid-associated genes including *Car1, Ermap, Klf1, Trib2,* and *Blvrb.* Pseudotime values assigned per cell were used to visualise the expression dynamics of individual genes, plotting the order of changes in their expression along pseudotime. By extracting the counts information for a given list of genes, it was possible to confirm known examples of genes from the literature that vary in expression across pseudotime, as well as the gene expression level in subsets of interest from DEG and module analysis revealing patterns of expression relating to different stages of commitment (4.15).

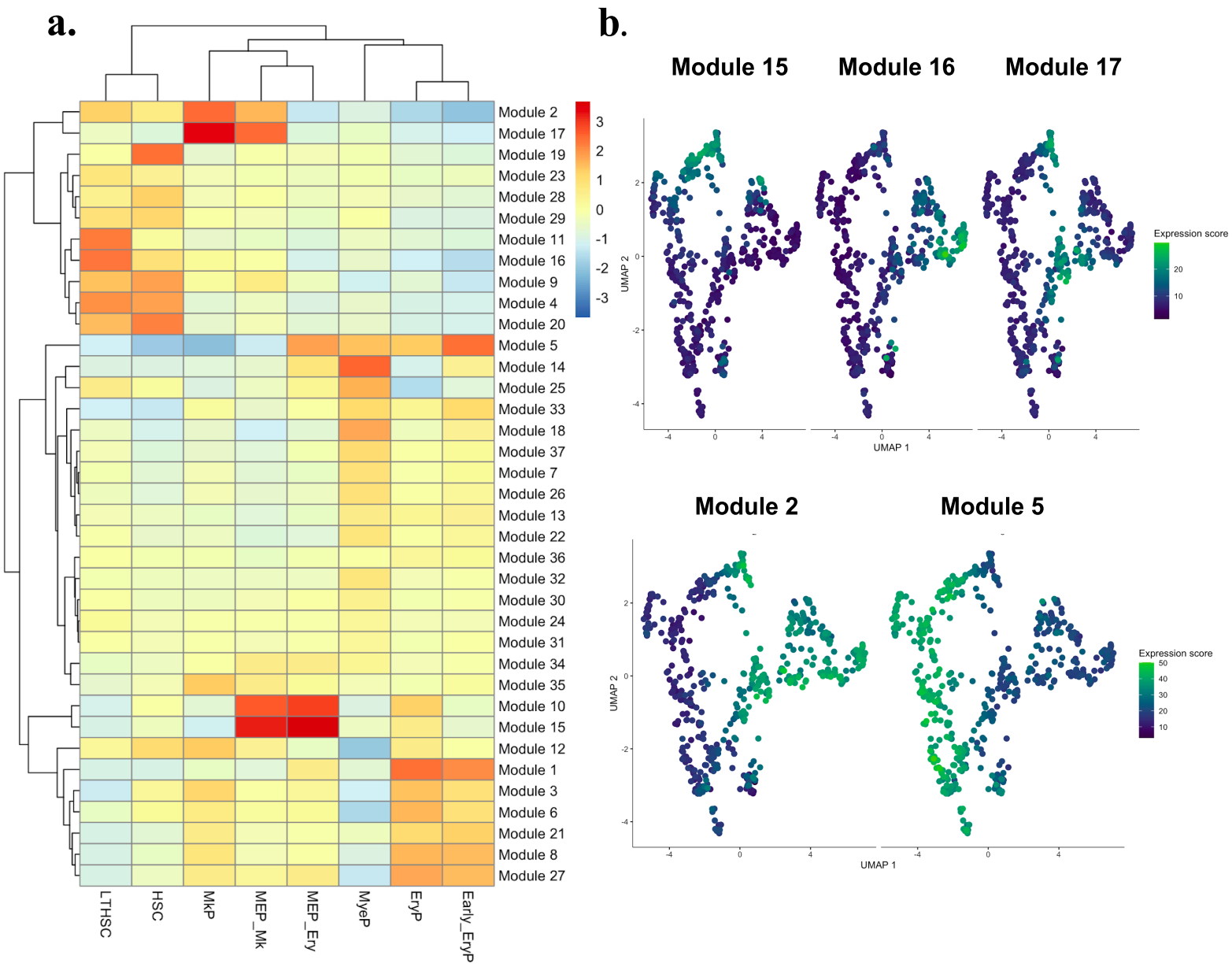


Figure 4.13. Gene modules across clusters capture co-expressed genes in cells at different pseudotime states. (a) Heatmap of cell populations and the enrichment score across 37 gene modules. (b) UMAP projection of all cells coloured by expression score for a subset of selected modules.

Aside from collating canonical lineage-associated markers, the module-based approach identified genes of interest with a less well-established connection with Mk differentiation towards the Mk and Ery lineages. *Inka1,* also known as *Pdcd7* (Programmed Cell Death 7), is a multifunctional gene that has been linked with diverse roles in various cellular processes including apoptosis, autophagy, and mitosis and is known to be expressed among HSPCs. It is clinically pertinent, with previous work demonstrating that *Inka1* overexpression stalls leukaemia stem cells in quiescence, resulting in the accumulation of cells in G0, and reduced output of differentiated progeny. This cell-cyle withdrawal allows leukaemia stem cells to retain disease-initiation properties whilst remaining at undetectable levels that have been linked to therapy failure and relapse in patients with AML (Shlush *et al.*, 2017; Kaufmann *et al.*, 2019). Whilst *Inka1* has been established as a stem-ness-promoting gene in HSCs, it has not been directly associated with Mk differentiation as this data indicates. *Inka1* is upregulated throughout all cells of the Mk lineage, at a sustained expression level from LTHSC to MkP, and has comparatively low expression in Ery cells.

Coro1a was also listed in module 2 and its expression level was found to be correlated with cells of the Mk lineage (Figure 4.15). This gene encodes Coronin-1A, part of the coronin family of F-actin– and Arp2/3-binding proteins, that is exclusively expressed in haematopoietic cells and serves as an auxiliary to cytoskeletal reorganisation processes that involve actin (de Hostos, 2008). Cytoskeletal remodelling is important for Mk maturation and platelet production, with many genes implicated for proplatelet formation encoding proteins involved in cytoskeletal dynamics such as Rho GTPases and their downstream targets (Rojnuckarin and Kaushansky, 2001; Ghalloussi, Dhenge and Bergmeier, 2019). Previous work using Coro1a-KO mice showed that this led to an inhibitory effect on steady-state F-actin formation (Föger *et al.*, 2006). More recently, research into the role of the TF serum response factor (*Srf*) in Mks found *Coro1a* and other cytoskeletal regulatory genes were downregulated in *Srf*-KO Mks. These Mks were found to exhibit abnormal maturation and function resulting in significant thrombocytopenia in mice (Halene *et al.*, 2010). Moreover, *Srf and Mrtfa* (a co-factor of *Srf)* overexpression *in vitro* enhanced megakaryopoiesis, exhibiting increased TF-binding at target sites during megakaryopoiesis and was associated with upregulated Mk-associated and cytoskeletal genes including *Coro1a (Rahman et al., 2018)*. Whilst it remains to be fully elucidated, these reports and the results presented here suggest *Coro1a* plays a role in megakaryopoiesis, likely facilitating proplatelet formation which is known to require actin-mediated forces for the bending/branching during the formation of processes (Italiano *et al.*, 1999).

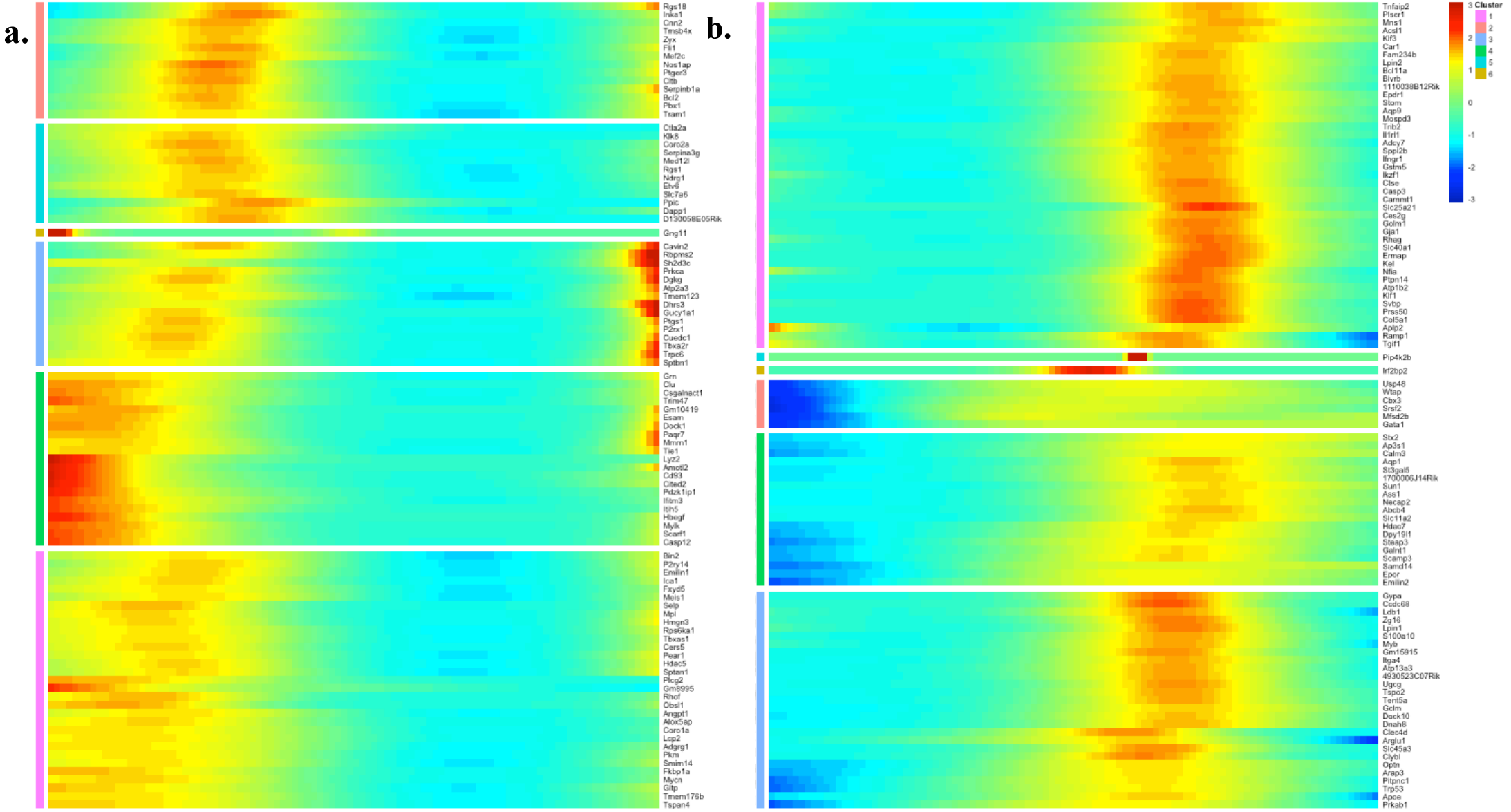


Figure 4.14. Heatmap of expression distribution in the top 100 DEGs of modules 2 (a) and 5 (b) X-axis represents pseudotime, heatmap colour indicates positive or negative expression and intensity indicates expression level. Correlating genes were clustered with default parameters (Y-axis).

Transmembrane proteins are an important class of molecules to investigate in the context of haematopoiesis due to their importance in processes including cell signalling, adhesion, and molecule transfer across cell membranes. In particular, transmembrane proteins that are expressed on the cell surface or have cytokine or growth factor extracellular binding domains mediate cell-cell interactions and have significant influence over haematopoietic regulatory processes and indeed lineage specification. Among many known Mk-affiliated genes in module 2, transmembrane-protein 176a (*Tmem176a)* and transmembrane-protein 176b (*Tmem176b)* were particularly well-correlated with cells along the Mk pseudotime trajectory. Both genes are expressed among all stages of Mk commitment, with high expression in LT-HSCs that is maintained through HSCs Mk-MEPS and MkPs. *Tmem176b* was found to be expressed at higher levels across pseudotime, whilst *Tmem176a* levels are highest amongst the immature cells (Figure 4.15). Despite the robust expression levels of both genes, specificity to clusters for the Mk lineage, and their co-expression with canonical Mk markers such as TPO receptor gene *Mpl,* literature connecting them to Mk commitment and/or function could not be found (Figure 4.16). The *Tmem176a* and *Tmem176b* proteins are considered related members of the MS4A family based on their shared topological characteristics, specifically their characteristic four membrane-spanning domains (Zuccolo *et al.*, 2010). The MS4A members have a broad expression profile across tissues, however, several are limited to cells in the haematopoietic system where they have known roles in immune cell functions. Notable examples include MS4A1 (CD20) in B cells, MS4A2 (FcεRIβ) in mast cells and basophils, and the involvement of MS4A3 in haematopoietic cell cycle regulation (Tedder *et al.*, 1988; Liang and Tedder, 2001; Donato *et al.*, 2002). Though their roles are diverse MS4A proteins comprise a family of ion channel/adaptor proteins facilitating intracellular protein–protein interactions (Eon Kuek *et al.*, 2016). The existing literature on *Tmem176a* and *(Wang et al., 2002) Tmem176b* (Lurton *et al.*, 1999) describe their expression in myeloid cells, specifically at the immature state of dendritic cells (Condamine *et al.*, 2010). There is growing evidence to suggest that MS4A proteins are associated with cell cycle control and differentiation, and most recently *Tmem176b* has specifically emerged as an immunoregulatory player (Hill *et al.*, 2022). This data suggests the expression of *Tmem176a* and *Tmem176b* is abundant across the Mk lineage, and prompts further investigation to verify the robustness of these findings and whether possible implications to the Mk function exist.

A final transmembrane protein-encoding gene also identified as upregulated in Mk differentiation was transmembrane protein 123 *(Tmem123).* This gene encodes for a highly glycosylated transmembrane protein that is mostly linked to mediating membrane dynamics and a form of cell death distinct from apoptosis known as oncosis, cell death induced by mechanical, chemical, and environmental factors (Zhang *et al.*, 1998). Oncosis is marked by cell and organelle swelling, membrane blebbing, and an increase in membrane permeability in

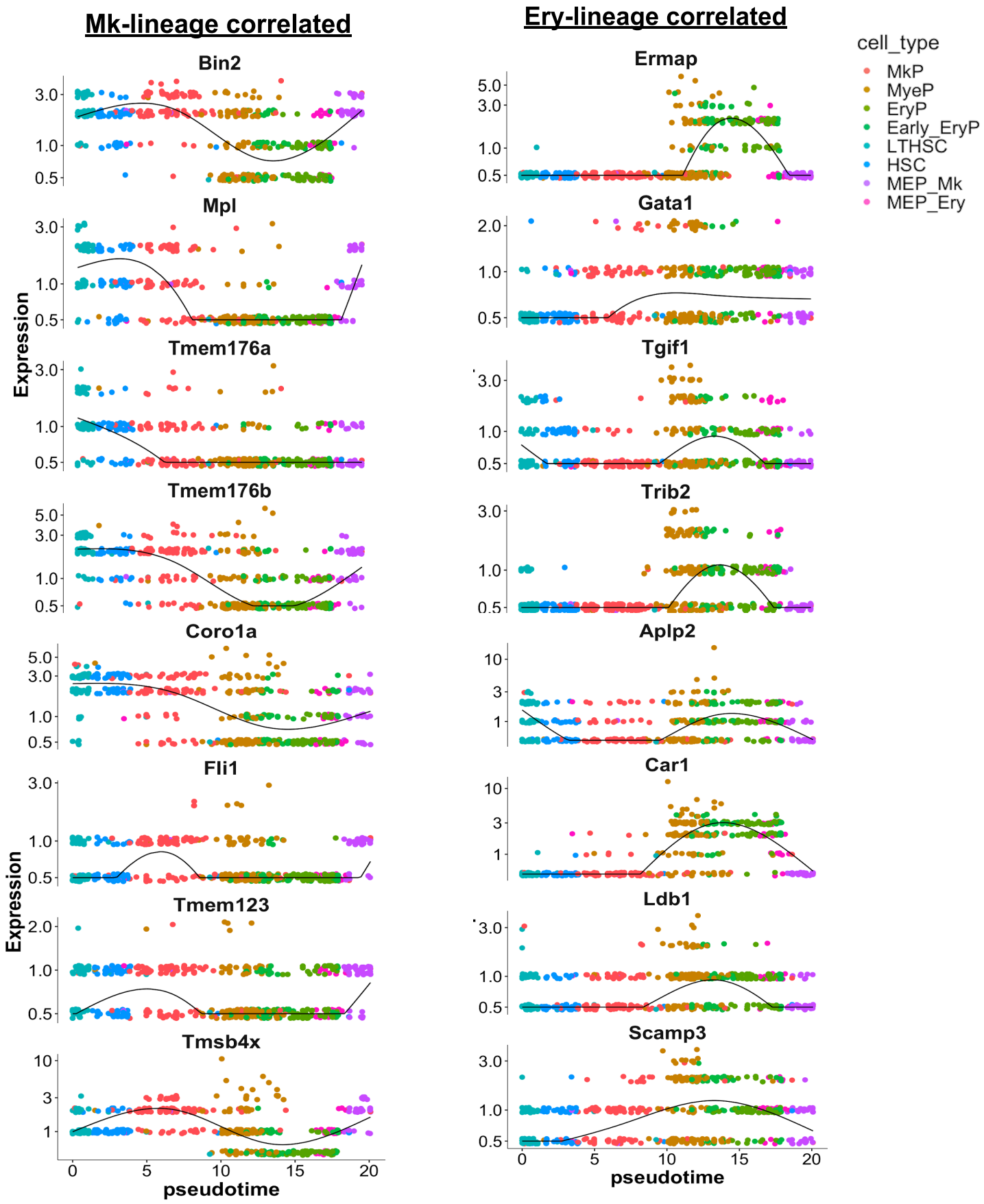


Figure 4.15. Expression dynamics of a subset of genes with differential expression along pseudotime. Plots show correlation with cells of the Mk (left) and Ery (right) lineages.

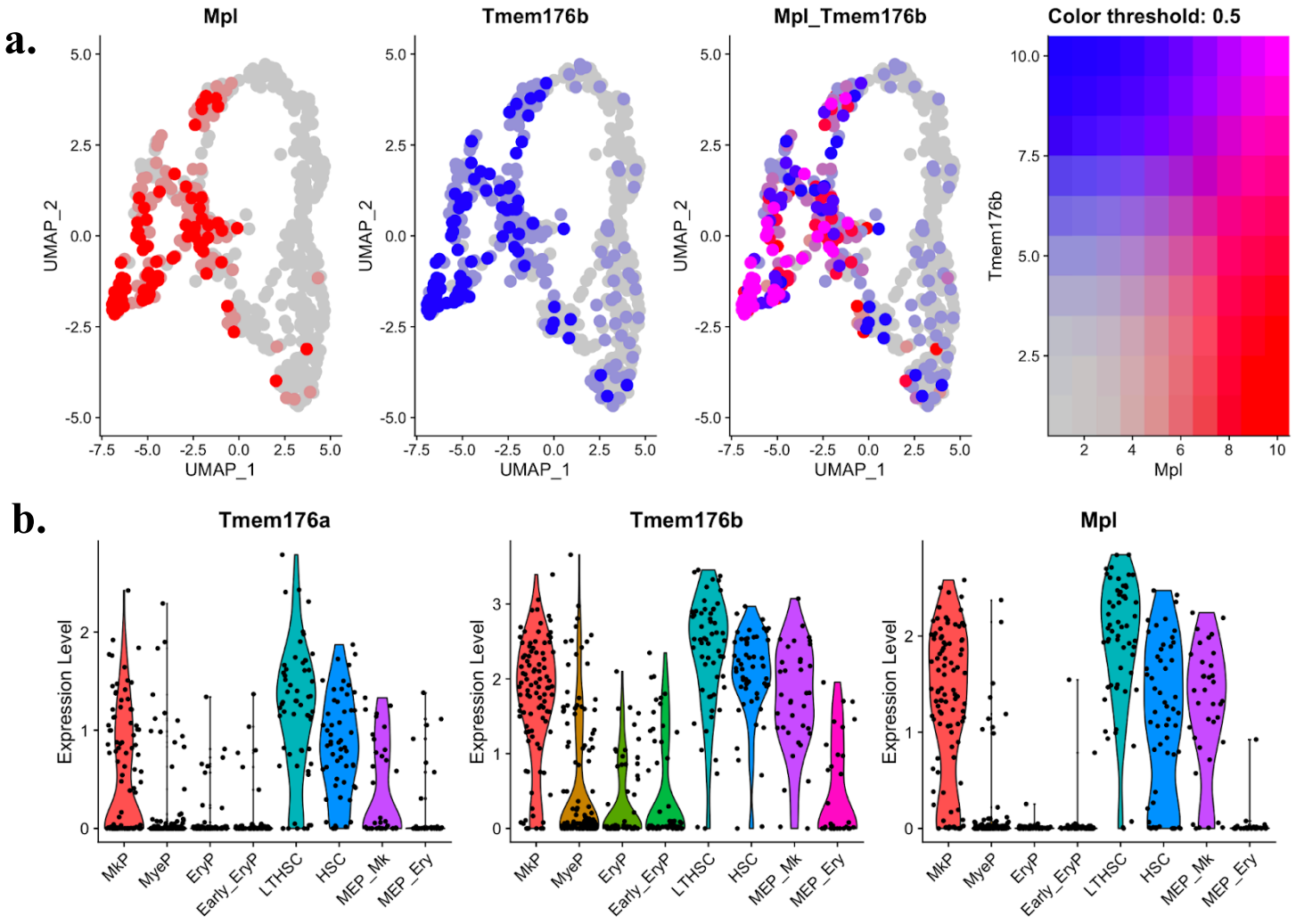


Figure 4.16. *Tmem176* gene expression is upregulated in cells of the Mk lineage and correlates with important canonical genes of Mk function. (a) UMAP co-expression projection of *Mpl* (red) and *Tmem176* (blue) across all cells. Cells are coloured by the expression level of each gene, where cells expressing both are projected in pink. (b) Violin expression plots of expression levels of *Tmem176* genes and *Mpl* across cell type annotated clusters.

contrast to apoptosis with cellular shrinkage and nuclear disruption (Majno and Joris, 1995). In humans, the protein Porimin (a member of the cell membrane-associated mucin family) shares the highest overall homology with *Tmem123* at 67% and is also thought to mediate cell death by oncosis (Ma *et al.*, 2001). Whilst it is primarily affiliated with oncosis across the literature, it is unclear whether *Tmem123/* Porimin may have other roles in cell function. Interestingly, a paralog of *Tmem123* is CD164, an important cell adhesion molecule that regulates the proliferation, adhesion and migration of haematopoietic progenitor cells (Zannettino *et al.*, 1998; Pellin *et al.*, 2019). This data identifies *Tmem123* expression across cells within the LK Cd150+ compartment, but shows it is most highly expressed along the Mk trajectory, with expression levels peaking in Mk-MEPs and MkPs (Figure 4.15). With other mucin-family members having reported roles in HSPCs, such as CD43-regulation of cell death in primitive progenitor cells (Bazil *et al.*, 1995), it would be interesting to explore a potential functions of *Tmem123* in Mk-commitment either in regulating cell death in Mks or other mucin-associated functions such as roles in membrane dynamics that facilitate Mk maturation and/or platelet formation.

Altogether this analysis describes the transcriptional changes from HSCs through to progenitors of the Mk and Ery lineages. The signatures of genes that were differentially expressed along pseudotime included genes that are known to vary in expression with increasing differentiation stages, as well as genes not previously associated with Mk commitment.

4.3.7. Differential signatures of Mk lineage in LK Cd150+ compartment with age

After computationally delineating states of differentiation in pseudospace and differential expression testing for genes that vary with pseudotime in the LK Cd150+ compartment, statistical differential expression analyses (DEA) were performed to determine ageing-specific transcriptional signatures. This was achieved by aggregating counts from single cells based on pseudobulk annotations, grouping data by cell type and experimental conditions (young vs old) whilst taking into account biological replicates (n = 3). This approach minimises the risk that dominating signatures ie. highly DEGs present in a small number of cells would lead to inflated p-values and false-positive DEG detection, and allows for better detection of cell-type specific differences with age. Reads from each mouse were pseudobulked, and genes with low expression levels (total read count per sample < 10) and likely to be noise or technical artefacts were excluded from the analysis. For additional robustness, only genes present above the minimum threshold in all three samples per condition were retained for downstream analysis to focus on genes that were observed across all technical replicates.

Differential expression was performed for each cell type using the *DESeq2* package, setting the design variable as mouse age (Love, Huber and Anders, 2014). This calculated differential expression from the normalised count data based on the negative binomial distribution model and statistical tests performed to determine significance. The lists of DEGs were filtered to include only genes that were statistically significant with age (adj P-values < 0.05), and the expression of the top DEGS specifically in cell types implicated in the Mk lineage were visualised at the sample level (Figure 4.17).

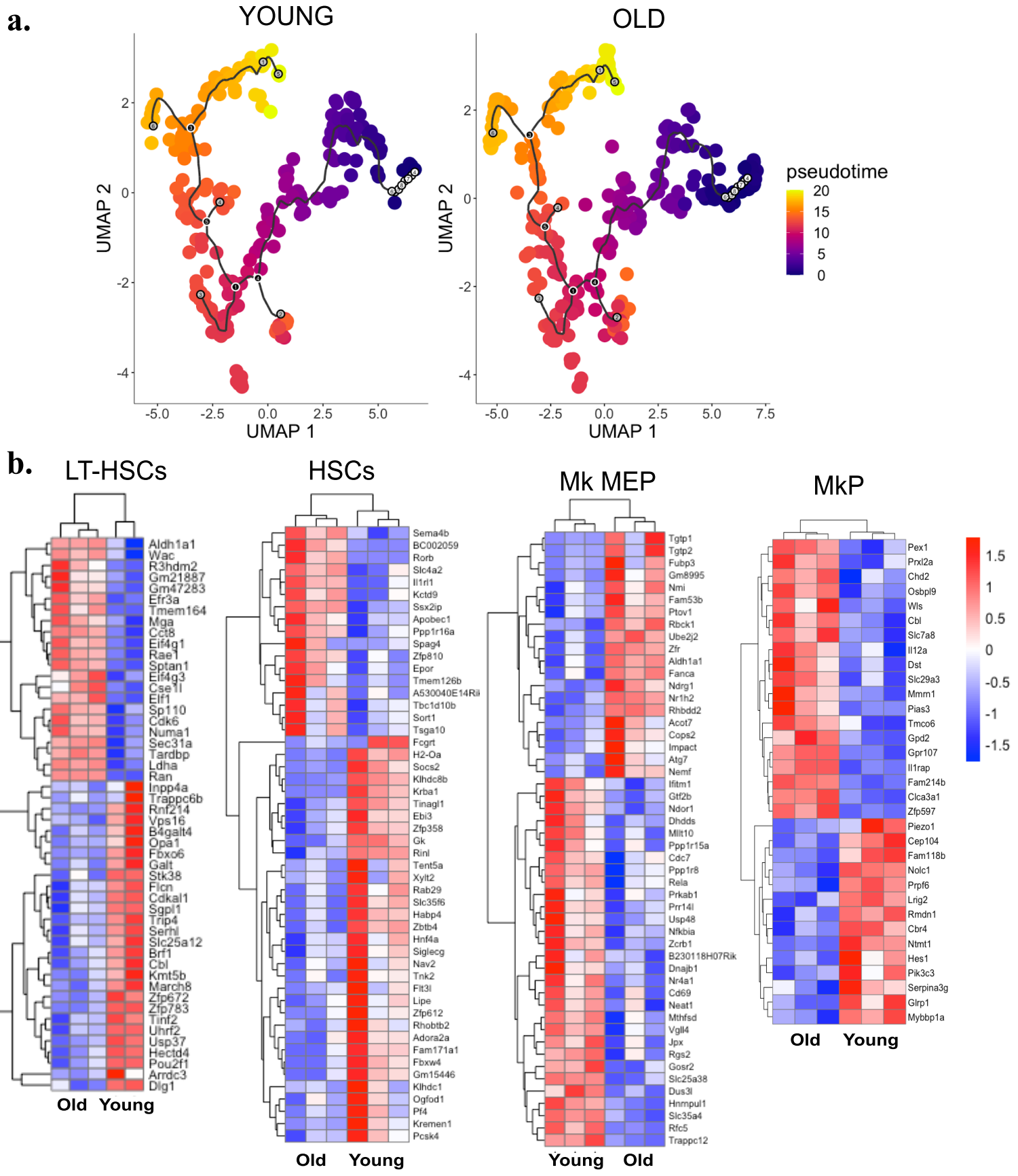


Figure 4.17. Differential expression across cells of the Mk lineage with age. (a) UMAP projection of pseudotime trajectory split by age (b) Normalised expression level of genes with DEGs with age in each cell population, split by biological replicate. Genes (rows) and biological samples (columns) have been clustered using agglomerative hierarchical clustering.

Pseudobulk DEA of MkPs identified 271 genes that exhibit significantly different expression levels with age, including DE among important regulators of Mk differentiation including *Selp*, *Vwf, Mapk7,* and *Arhgef6* (adj $p < 0.05$) (Figure 4.18 A-B). Aged MkPs were found to express significantly higher levels of *Plscr2,* a member of the phospholipid scramblase family of proteins that mediate calcium-dependent, non-specific movement of plasma membrane phospholipids. This gene has been previously implicated in early HSPCs with clonal output toward the Mk lineage *in vitro (Weinreb et al., 2020)*, and Sun *et al.* also showed it is upregulated in old HSCs (Sun *et al.*, 2014). There has been an increasing appreciation for the role of lipids in cell fate decisions during haematopoiesis over recent years (Bansal *et al.*, 2018; Pernes *et al.*, 2019), this result shows *Plscr2* is upregulated with age at multiple levels of the Mk lineage, including MkPs, Mk-MEPs and LTHSCs supporting the correlation of Mk clonal output *in vitro* and *Plscr2* expression. To further evaluate the expression signatures of young and old MkPs, functional analysis was performed using GSEA of DEGs. This analysis indicated that DEGs between young and old MkPs were highly enriched in categories related to regulation of platelet derived growth factor signalling and E-box binding, indicative that the transcriptional regulation mediated by E-box binding TFs may be altered during the ageing process.

DEA of Mk-MEPs identified 60 genes with significantly variable expression levels with age (Figure 4.18 C-D). This includes the RNA-binding protein *Rbpms2,* a gene that has previously been identified as a marker for progenitors with higher Mk clonal output, which was significantly upregulated in aged Mk-MEPs (Weinreb *et al.*, 2020). On the other hand, young Mk-MEPs had significantly higher *Nr4a1* gene expression. This is a member of the conserved subgroup of nuclear receptor TFs involved in the regulation of expression in a ligand-independent manner, with diverse roles in differentiation and function of distinct subsets of lymphoid and myeloid cells (Pearen and Muscat, 2010). Specifically, *Nr4a1* has been implicated in NF-κB signalling; a fundamental TF involved the expression of various genes involved in the inflammation and thrombotic processes (Mussbacher *et al.*, 2019). Moreover, *Nr4a1* has also been identified as a tumour suppressor in AML and pre-AML malignancies, including myelodysplastic/myeloproliferative disorders, whereby its abrogation provides a cell proliferation advantage in these disorders (Mullican *et al.*, 2007; Ramirez-Herrick *et al.*, 2011; Lin *et al.*, 2022). In the context of this study, high expression of *Nr4a1* was observed in LTHSCs, HSCs and Mk-MEPs and MkPs, with highest expression seen in young samples across all but one cell type where its expression is proportionally found across the two age groups (MkPs). A direct link between *Nr4a1* expression and a young cell phenotype as this data suggests has not previously been established, however with important roles of *Nr4a1* in regulating cell growth and preventing the development of haematological malignancies it would be interesting to explore its potential implications in this context.

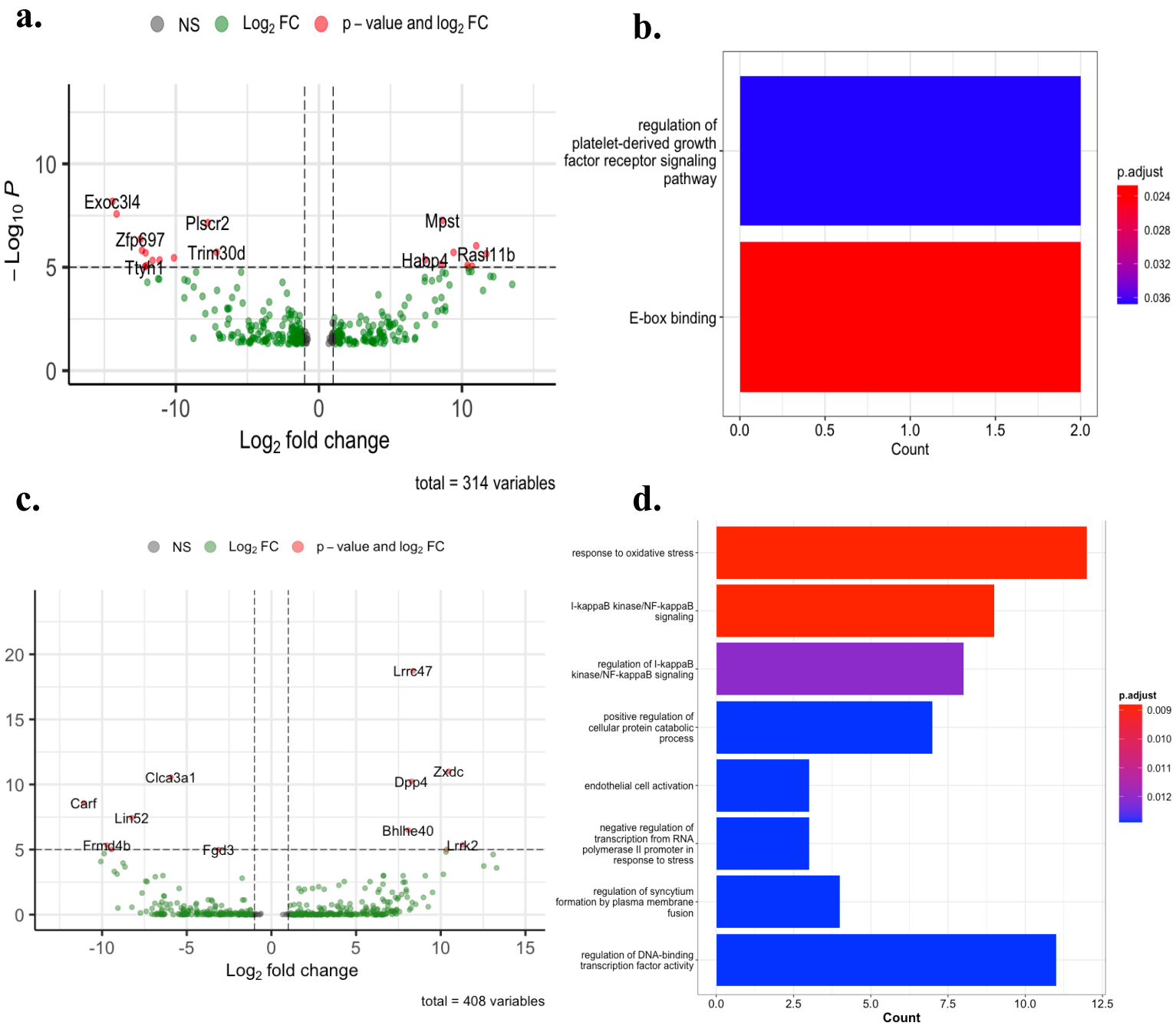


Figure 4.18. Differential expression across MkP and Mk-MEPs with age. (a) Volcano plot of top significant DEGs in MkPs with age. (b) GO GSEA result of enriched ontologies across significant MkP DEGs. (c) Volcano plot of top significant DEGs in Mk-MEPs with age. (d) GO GSEA result of enriched ontologies across significant Mk-MEP DEGs.

Aldehyde dehydrogenase (ALDH) activity is a known feature of HSCs and, as such, has been previously used as a marker to identify and purify HSCs (Storms *et al.*, 1999). In general, most ALDHs facilitate the oxidation of aldehydes to their corresponding carboxylic acids. *Aldh1a1,* the most abundant isoform found in HSCs, has a critical role in retinoid metabolism through its role in metabolising retinaldehyde to retinoic acid (Kavanagh *et al.*, 2008). DEA in MkP with age, and indeed most other cell types captured as part of this experiment, showed significant upregulation of *Aldh1a1* with age (Figure 4.19). However, not much is known about the underlying mechanism of *Aldh1a1* expression as an age-dependent haematopoietic phenotype. Poscablo *et al.* also found *Aldh1a1* upregulation in MkPs from aged mice (Poscablo *et al.*, 2021), however provide no indications as to biological significance and correlation with age. ALDH overexpression is strongly associated cancer cells with stem-like features, where they are thought to protect cancer cells by metabolising toxic aldehydes into less reactive and more soluble carboxylic acids (Jackson *et al.*, 2011; Xu *et al.*, 2015). Notably, ALDH expression has also been associated with functional roles in normal stem cells including promoting self-protection, expansion and differentiation (Ma and Allan, 2011). In the context of haematopoietic lineage expression, Rice *et al.* found *Aldh1a1* expression inhibited lymphopoiesis in favour of myelopoiesis suggesting its expression may be associated with HSC propensity towards myeloid differentiation (Rice *et al.*, 2008). The findings here show robust overexpression with age across almost all cell types captured, with the exception of Ery MEPs and EryP cells. However there is not sufficient data to establish a direct correlation between *Aldh1a1* expression with a distinct lineage.

Another gene revealed to exhibit age-associated expression was Calcium-activated chloride channel regulator 1 (*Clca3a1*), a gene encoding an accessory protein for calcium-activated chloride channels. *Clca3a1* expression was found to be significantly higher in aged LT-HSCs, MkPs, EryPs and Mk-MEPs compared to the same cell types from young mice (Figure 4.19). This finding is supported by previous work that has shown *Clca3a1* expression in LT-HSCs can be used as a marker to distinguish HSC sub-types in aged mice, where a low *Clca3a1* signal marks individual "young-like" HSCs within the pool of aged HSCs (Kim *et al.*, 2022). The authors found that the high expression of *Clca3a1* in HSCs (the aged phenotype) correlated with a myeloid-biassed Cd150+ LT-HSC expression profile, whereby the vast majority of genes that were upregulated in *Clca3a1* high cells were also upregulated in old vs. young LT-HSCs. Moreover, *in vivo* functional assays of these cells showed that recipients of *Clca3a1* high clones not only produced significantly more myeloid cells, but over time exhibited a defect in long-term repopulating activity (Kim *et al.*, 2022). These data are consistent with the upregulated expression of *Clca3a1* observed in LT-HSCs, MkPs, EryPs and Mk-MEPs in this analysis - all of which were isolated with a LK Cd150+ gate (as shown in Figure 4.19).

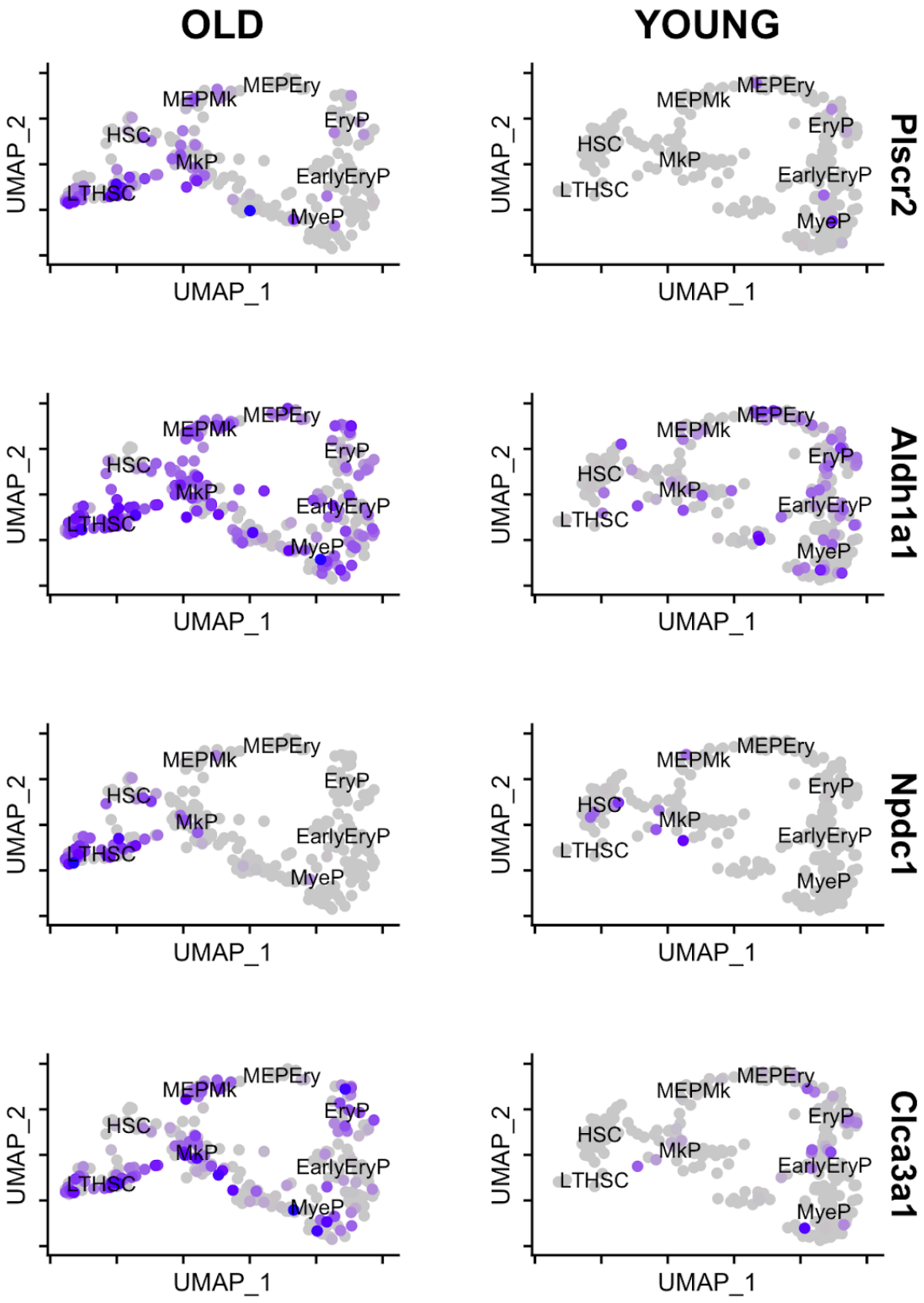


Figure 4.19. UMAP feature expression plots of genes significantly upregulated with age across multiple levels of Mk commitment.

Upregulation of the neural proliferation, differentiation and control 1 (*Npdc1*) gene was identified as part of the ageing signature in this study. This gene encodes for a neural factor involved in the control of neural cell proliferation and differentiation, where it is thought to down-regulate cell proliferation, however we know relatively little about how *Npdc1* functions in the context of haematopoiesis (Sansal *et al.*, 2000). Sansal *et al.* has shown that *Npdc1* interacts with some cell cycle proteins such as D-cyclins, cdk2 and most notably the TF E2F-1, in doing so reducing its binding to DNA and modulating its transcriptional activity. The E2F TF family is known to play key roles in the timely expression of genes required for cell cycle progression and proliferation. E2F1's role is extremely multifaceted, and has been implicated across diverse processes including both cell proliferation and antiproliferative processes such as apoptosis and senescence (Collins *et al.*, 1995; Ginsberg, 2002; Trikha *et al.*, 2011; Xu *et al.*, 2018). *Npdc1* has been identified as a significantly up-regulated gene associated in relapse incidences of AML, however has not previously been explored in the context of and ageing (Hackl *et al.*, 2015). The observed upregulation of *Npdc1* expression in LT- and ST-HSCs with age may therefore represent a noteworthy finding that warrants further investigation. To validate the robustness of *Npdc1* levels in HSCs with age, future experiments are necessary to help establish the consistency and reliability of this observed upregulation. The upregulation of *Npdc1* in ageing HSCs may raise questions about its potential role in age-related changes in haematopoiesis, including Mk commitment, and warrant further research to explore the significance of *Npdc1* in these contexts and understand its potential clinical relevance.

Gene *Fgd3* was found to be upregulated with age in Mk-MEPs. This gene encodes for the FYVE, RhoGEF and PH domain-containing protein 3, which is an activator or modulator of *Cdc42* signalling (Rho GTPase). Rho GTPases, including *Cdc42,* are the primary drivers in the dynamic cytoskeletal reorganisation process, leading to the development of filopodia and lamellipodia which increase platelet surface area upon their activation (Pleines *et al.*, 2010; Comer, 2021). *Fgd3* acts as a guanine nucleotide exchange factor fo*r Cdc42,* meaning it promotes the activation of *Cdc42 b*y facilitating the exchange of GDP (guanosine diphosphate) for GTP (guanosine triphosphate). This is necessary for its proper functioning in regulating the dynamic reorganisation of the cytoskeleton and cellular processes associated with it (Buchsbaum, 2007; Pleines *et al.*, 2010). In the context of Mks, higher expression of *Fgd3* with age would not be a phenomenon unique to *Fgd3*, as other important proteins, such as β-thromboglobulin and Pf4 (secreted from platelet α-granules), are also found at significantly higher levels in older individuals (Zahavi *et al.*, 1980; Bastyr, Kadrofske and Vinik, 1990).

Hyperactivity of platelet function is a known phenotype of ageing, that results in higher rates of both vascular and thrombotic disease with age. Several factors have been implicated to explain platelet hyperactivity with age, such as changes in the platelet-serotonin system, increased

oxidative stress, vascular prostaglandins alterations, and plasma membrane modifications with age (Le Blanc and Lordkipanidzé, 2019). Interestingly, multiple of these factors associated with platelet function hyperactivity were identified among the top enriched terms from functional GSEA of Mk-MEPs ageing DEGs, including significant enrichment for oxidative stress response, plasma membrane and aberrant signalling from the DEG signature (Figure 4.18D). This, in conjunction with the upregulation of *Fgd3* in aged Mk-MEPs, may indicate a possible role for *Fgd3 in* megakaryopoiesis with age, likely associated with its target *Cdc42* which has been established as essential for cytoplasmic Mk maturation and potentially also BM localisation (Pleines *et al.*, 2013; Dütting *et al.*, 2017; Pleines, Cherpokova and Bender, 2019; Heib *et al.*, 2021).

In summary, these DEAs identified hundreds of genes with significantly different levels of expression, providing an insight of the transcriptional landscape of megakaryopoiesis with age. Theses analyses confirmed several known signatures associated with ageing haematopoiesis, including upregulated expression of *Vwf, Nupr1, Sult1a1* and others (Figure 4.20), but crucially also revealed multiple novel genes which have not been previously studied in the context of ageing.

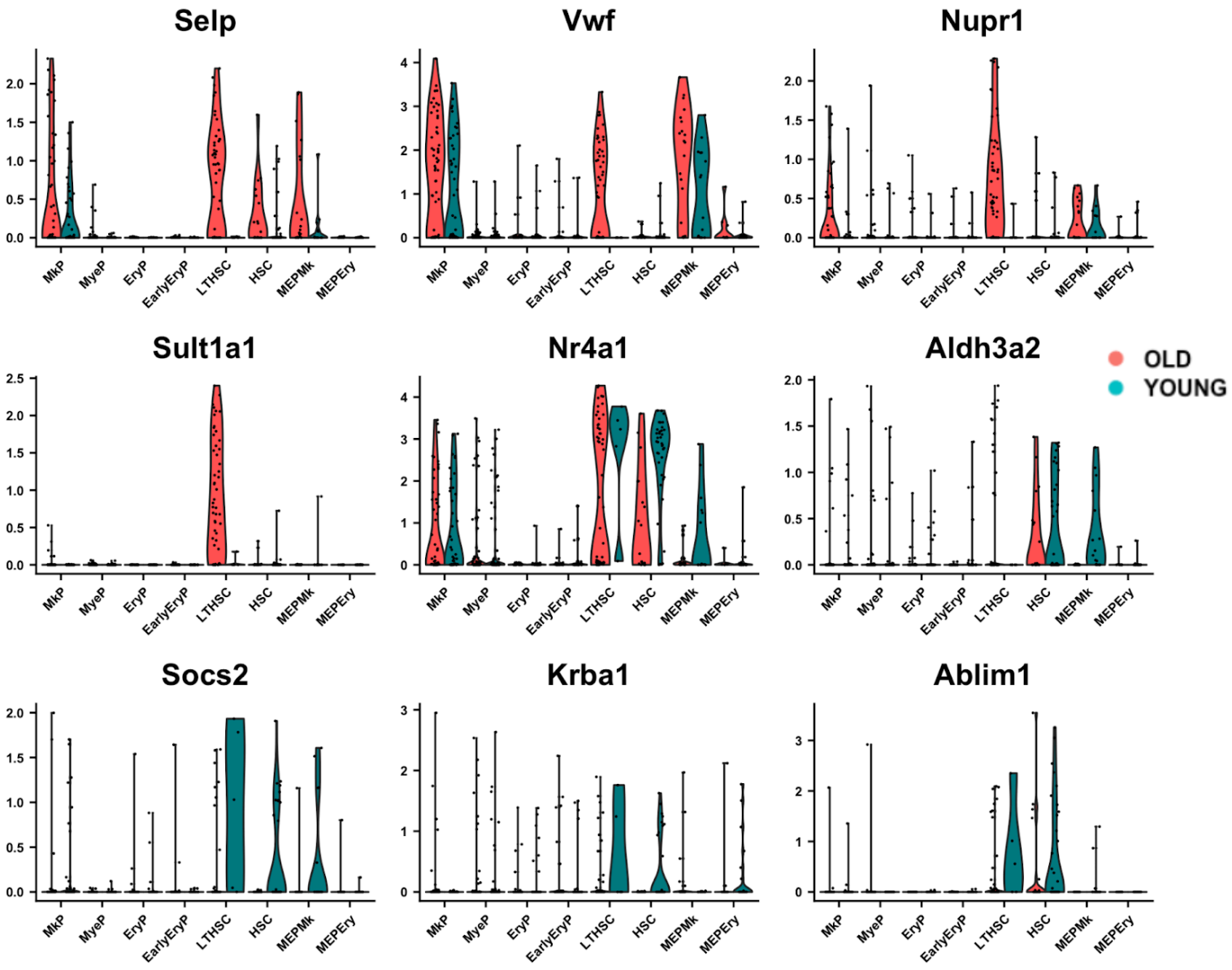


Figure 4.20. Violin expression plots of a subset of genes with differential expression signatures with age.

4.4 Discussion

There are significant changes in haematopoiesis with age, both in terms of the population frequencies within the BM and the functional capacity of cells. The consequences of ageing-associated aberrant gene expression and compositional changes within the haematopoietic compartment of the elderly contributes towards increased incidence in haematological diseases across the population, such as anaemia, arterial thrombosis, and myeloid and lymphoid malignancies (such as age-related clonal haematopoiesis, myelodysplastic syndromes and acute myeloid leukaemia) (Eisenstaedt, Penninx and Woodman, 2006; Steensma *et al.*, 2015; Zink *et al.*, 2017). This functional decline of the haematopoietic system is associated with several known phenomena some of which include diminished lymphoid potential, reduced regenerative capacity, heightened autoimmunity, and DNA damage accumulation (Sudo *et al.*, 2000; Rossi, Seita, *et al.*, 2007; Dykstra *et al.*, 2011; Beerman *et al.*, 2014; Flach *et al.*, 2014).

It is clear the challenges to the blood system with age result in multifaceted dysregulation of homeostatic regulation by committed progenitors. Previous research has shown the LSK Cd150+ fraction from old mice contain an increased proportion of myeloid-dominant HSCs and a lower output of mature blood cells per HSC (Cho, Sieburg and Muller-Sieburg, 2008; Weiskopf *et al.*, 2016). Moreover, an increase of Mk-primed HSCs have been consistently reported in the old HSC compartment (Sanjuan-Pla *et al.,* 2013; Yamamoto *et al.,* 2013; Grover *et al.,* 2016). This is also reflected by the heightened levels of expression in several Mk/platelet markers, such as CD150, CD41, CD61 and vWF in aged HSCs (Beerman *et al.*, 2010; Gekas and Graf, 2013; Sanjuan-Pla *et al.*, 2013; Mann *et al.*, 2018). The changes or aberrant expression in the Mk lineage genes with age contribute to a progressive increase in platelet responsiveness in both older men and women, which poses a significant problem in elderly populations with higher incidence of thrombotic disease and use of anti-platelet drugs (Jones, 2016; Zhang *et al.*, 2020). Age-related morbidity is associated with dysregulated differentiation from HSCs, including megakaryopoiesis, however a detailed insight of the transcriptional signatures of Mk lineage commitment with age has not yet been fully elucidated.

The primary objective of this study was to produce a detailed insight of the LK/LSK Cd150+ BM compartment with age using scRNA-seq. To achieve this, Smart-seq2 scRNA-seq was applied to FACS sorted LK and LSK Cd150+ cells from young (8 weeks) and aged mice (72 weeks), generating a total of 644 single cell libraries. This gating strategy builds on the work by Pronk *et al.* as described in detail in Chapter 3, and was employed in order to capture the full continuum of differentiation states from primitive HSCs to committed MkPs (Pronk *et al.*, 2007).

The single cell data were assessed based on quality metrics used to isolate only high quality samples for downstream analysis, leaving a total of 520 single cells. Data dimensionality reduction and Louvain-based clustering identified 9 clusters of single cells within the LSK Cd150+ compartment. Cell types were annotated based on thorough interrogation of highly expressed genes across each cluster, using canonical marker expression signatures of haematopoietic cell types from the existing literature as references (Pronk *et al.*, 2007; Haas *et al.*, 2015; Paul *et al.*, 2015; Pietras *et al.*, 2015; Miyawaki *et al.*, 2017; Dahlin *et al.*, 2018; Mincarelli *et al.*, 2023).

Semi-supervised pseudotime analysis was employed to establish a trajectory of cell differentiation within the LK/LSK Cd150+ compartment. The analysis utilised LT-HSCs as the starting point to construct the trajectory in pseudospace, revealing a clear continuum of differentiation from LT-HSCs to Mk and Ery progenitors. Transcriptional dynamics across cells were investigated, uncovering differentially expressed genes as a function of pseudotime. Notably, dynamic expression patterns were observed in genes previously known to vary during Mk commitment *(Mpl, Fli1, Gata1, Gata2, Vwf, Rgs18, Selp, Pbx1, Bin2* and *Cavin2)*, as well as in novel genes not previously associated with Mk commitment (Debili *et al.*, 1995; Klimchenko *et al.*, 2009; Sengupta *et al.*, 2013; Zhu *et al.*, 2018; Walker *et al.*, 2022). To facilitate the interpretation of genes exhibiting dynamic expression along the trajectory, gene co-expression modules were calculated to identify co-expressed genes at different pseudotime states. Among the 37 gene modules identified, modules 16, 15, and 17 displayed a strong positive correlation with LTHSCs, Mk-MEPs, and MkPs, respectively. Additionally, module 2 exhibited correlation with cells at various stages of the Mk lineage. These modules, in particular, revealed significant correlations of genes such as *Tmem176a & b, Adgrl4, Coro1a, Tmsb4x*, and others with Mk differentiation. These results provide valuable insights and potentially novel signatures underlying Mk commitment. Validation of these DEGs associated with Mk fate restriction based on this data will be critical to assess the robustness of these findings, and pave the way for further investigations into the regulatory networks governing Mk commitment.

There are well-established phenomena associated with haematopoietic ageing, including the expansion of the HSC compartment, myeloid skewing, increased lineage-biassed HSCs, and several age-related HSC niche changes (Dykstra *et al.*, 2011; Pang *et al.*, 2011; Grover *et al.*, 2016; Mead, 2021). Specifically, studies have shown that elderly mice exhibit a significant increase in the number of phenotypic myeloid-dominant HSCs compared to young adult mice. Previous research has demonstrated that myeloid-dominant HSCs are enriched in the CD150+ population of the LSK CD34^-flt3^- bone marrow compartment, and the size of this population increases dramatically with age (Beerman *et al.*, 2010; Dykstra *et al.*, 2011). Consistent with

the existing literature, the findings of this study align with these observations. Approximately 19% of cells sequenced from aged mice were identified as LT-HSCs, whereas only 1.9% of cells from young mice exhibited this phenotype. This stark difference further supports the notion of age-related changes in the composition of the HSC compartment. This is also reflected in the transcriptional signature of the LT-HSC compartment, which is dominated by genes that have been previously correlated with an aged HSC/ biassed signature including *Sult1a1, Vwf, Nupr1, Selp* and *Itgb3* (Grover *et al.*, 2016; Flohr Svendsen *et al.*, 2021; Mincarelli *et al.*, 2023). To perform cell-type specific DEA, counts from single cells were pseudo-bulked for each cell type based on age taking into account biological replicates. These analyses revealed significant cell-type specific signatures of ageing, further highlighting the impact of age on the molecular characteristics of different cell populations within the haematopoietic system. This includes *Plscr2, Aldh1a1, Npdc1 and Clca3a1* were among genes significantly upregulated with age across multiple levels of haematopoiesis.

To gain further insight into the functional implications of the age-associated DEGs, functional GSEA analyses were performed. Interestingly, the GSEA results revealed a correlation between the age-associated DEGs and factors previously implicated in platelet hyperactivity with age. Specifically, among these findings there was a significant enrichment for oxidative stress response, plasma membrane-related functions, and aberrant NF-κB signalling pathways (Liu *et al.*, 2017; Le Blanc and Lordkipanidzé, 2019). This result may indicate a link between the age-related DEGs identified in this dataset to with aberrant Mk/ platelet function in aged mice. Moreover, Svendsen *et al.* recently published a comprehensive, robust, and stable transcriptomic signature of HSC ageing, where they showed the HSC ageing signature is highly enriched for membrane-associated transcripts (Flohr Svendsen *et al.* 2021). This finding of enrichment in cell membrane–associated transcripts suggests that physiologically aged HSCs may communicate differently with their environment compared with their young counterparts, which is likely a significant factor influencing ageing associated phenotypes of the haematopoietic system.

The identification of age-associated differentially expressed genes (DEGs) that are linked to known clinical factors associated with age provides strong evidence for the functional implications and potential clinical relevance of these gene expression alterations. However, further investigation is necessary to validate the robustness of the observed expression levels across different cell types captured in this study. This validation will help confirm the reliability and consistency of the identified gene signatures in the ageing LK Cd150+ compartment.

To unravel the underlying mechanisms driving these gene expression changes and gain deeper insights into the ageing process, future experiments should focus on evaluating the *in vivo*

functional consequences of the novel genes associated with the ageing LK Cd150+ compartment. Specifically, functional assays that assess platelet production and function can be performed using manipulated expression levels of these genes specifically in LK Cd150+ cells. For example, overexpression assays can be conducted for genes that were identified as highly expressed in aged Mk-MEPs, and the effects on the production of functional platelets from the manipulated LK Cd150+ cells can be evaluated to determine whether they exacerbate a particular phenotype.

However, before designing and conducting functional assays, it is also important to consider other variables that may be of influence in these results. Ageing phenotypes exhibit high variability between individuals, and the observed signatures in this study may represent specific characteristics of the analysed samples. Therefore, validation of the identified signatures using a larger cohort of animals is essential to determine whether these observations are consistently present as ageing phenotypes across a significant number of samples. The sample size implemented in this study enabled the analysis of the Mk lineage with age, however despite capturing a consistent number of cells across conditions, due to the scarcity of LT-HSCs in young BM a very low number of LT-HSCs were captured in this study which poses a limitation in assessing DEG signatures of LT-HSCs with age. Consequently, further experiments should separately sort an enrichment of LT-HSCs to ensure sufficient coverage of future experiments enabling more robust statistical analyses of the primitive LSK Cd150+ young BM. This validation will help address limitations of this study, and enable the identification of transcriptomic signatures associated with ageing and determine whether findings are statistically significant at scale.

In conclusion, further investigation and validation of the observed gene expression alterations associated with ageing are necessary to confirm their reliability and establish their functional consequences. Future experiments should focus on conducting in vivo functional assays, particularly assessing platelet production and function, to gain a better understanding of the role of these gene signatures in the ageing LK Cd150+ compartment. Additionally, validation using a larger cohort of animals will help determine the stability and significance of these transcriptomic signatures as ageing phenotypes.

This study aimed to provide a detailed understanding of the LK/LSK Cd150+ bone marrow compartment with age using scRNA-seq. Through the analysis of high-quality single-cell data, different cell clusters within the LSK Cd150+ compartment were identified and annotated based on canonical marker expression signatures, capturing all levels of Mk commitment from primitive HSCs to committed MkPs. Pseudotime analyses revealed a continuum of differentiation from LT-HSCs to Mk and Ery progenitors, accompanied by dynamic expression

patterns of genes many of which were associated with Mk commitment. Co-expression module analysis further identified gene modules strongly correlated with different stages of Mk lineage commitment, suggesting potential novel signatures underlying differentiation. Age-related changes in the composition of the HSC compartment were also observed, with an increase in LT-HSCs in aged mice in line with a wealth of existing literature. Cell-type-specific DEA highlighted significant signatures of ageing at multiple levels of megakaryopoiesis. Functional GSEA revealed a correlation between the age-associated DEGs and factors implicated in platelet hyperactivity and aberrant NF-κB signalling pathways, suggesting a link between the identified gene expression alterations and dysfunctional Mk/platelet function in aged mice. Further investigation and validation are needed to confirm the reliability and functional consequences of the observed ageing signatures, including validation using a larger cohort of animals to determine the stability and significance of the identified transcriptomic signatures as ageing phenotypes. In conclusion, the data presented as part of this chapter sheds light on the age-related changes during megakaryopoiesis within the LK Cd150+ compartment. The identified gene expression alterations and associated functional consequences may have potential implications for understanding age-related haematological phenotypes. Future investigations and validations will deepen the understanding of the underlying mechanisms and regulatory networks governing Mk commitment with age, and provide insights into potential candidates for clinical applications.

Chapter 5:

Isoform profiling from single cell experiments using long-read sequencing

Disclosures:

Annotation of isoforms from pooled HSC libraries of young and aged mouse cDNA using SQANTI3[15] was performed by Anita Scoones with help from Dr David Wright and Sofia Kudasheva.

Results from experiments using an adaptation of the HIT scIso-Seq protocol[16] were generated as part of a collaborative project with Pacific Biosciences (PacBio). All experimental work implemented by Anita Scoones was performed under remote oversight by Dr Jason Underwood and Dr Jonas Korlach (CSO), and protocol adaptations are subject to a non-disclosure agreement between the Earlham Institute and PacBio. Data deconvolution was performed by Dr Roger Volden and Dr Elizabeth Tseng of PacBio.

MAS-seq experiments from mouse bone-marrow samples were performed by Anita Scoones, with laboratory support from Dr Eirini Lamprak. Data transfer and bioinformatics support were provided by Dr Elizabeth Tseng and Sam Holt of PacBio.

MAS-seq experiments from human PBMC samples were performed by Charlotte Utting, Ashleigh Lister and Lydia Pouncey. Computational data analysis and interpretation for this chapter were performed by Anita Scoones.

[15] (Tardaguila *et al.*, 2018)
[16] (Shi *et al.*, 2022)

5.1 Introduction

ScRNA-seq is a powerful tool in the field of genomic research, allowing scientists to study individual cells and their expression patterns with unprecedented resolution. This has revolutionised the way we can dissect the heterogeneity in cell types present within tissues, providing insights into complex biological processes underlying development and disease evolution.

The gold standard high-throughput methods for scRNA-seq rely on capturing the 3' end of RNA molecules to generate sequencing libraries. This is typically achieved through the use of oligo-dT primers that selectively bind to poly(A) tails on the 3' end of eukaryotic mRNAs, priming molecules for 3' cDNA synthesis. Two of the most widely used methods to achieve this are the Chromium microfluidics system developed and marketed by 10X Genomics, and the plate-based Smart-Seq2 protocol (Picelli *et al.*, 2014). While 3' end-based scRNA-seq methods have presented scientists with the opportunity to perform transcriptomic analyses at single-cell resolution, greatly advancing our understanding of cellular heterogeneity, they do have limitations.

One drawback of these approaches is that they rely on the fragmentation of cDNA into smaller inserts during library preparation. After full-length cDNA is generated from cellular mRNA the cDNA is cleaved along multiple sites into smaller inserts during library preparation. This is necessary to allow for the ligation of Illumina adapters, which enable sequencing of fragments on Illumina instruments, which historically have provided high sequencing depth and accuracy at a relatively lower cost as opposed to other technologies. With 10X Genomics, cDNA fragmentation is achieved using a combination of enzymatic digestion and heat. Similarly, Smart-Seq2 uses a transposase enzyme to cleave and insert transposons into cDNA molecules, resulting in fragments that are flanked by adapter sequences that enable Illumina sequencing. cDNA fragmentation is also important to mitigate PCR bias, which favours the 5' and 3' end of transcripts resulting in a lower coverage of the middle regions. By reducing the length of molecules, library complexity is increased ultimately leading to higher accuracy and specificity of the sequencing data.

However, as a consequence of this requirement, these technologies may miss spliced isoforms or non-polyadenylated RNAs. The structures of transcripts aren't preserved through library preparation so important information along the length of transcripts can often be lost. Additionally, the majority of short reads during sequencing fail to span successive splice sites, prohibiting the detection of alternatively spliced isoforms (Kanitz *et al.*, 2015). This means the

data produces an incomplete assessment of the full repertoire of transcript isoforms that underpin cell signatures and function.

Alternative splicing (AS), the post-transcriptional process in which exons of pre-mRNA can be selectively included or excluded during splicing to produce multiple mRNA transcripts from a single gene, plays a crucial role in regulating gene expression and proteome diversity. The production of multiple transcript variants from a single gene can alter protein structure and function, resulting in unique isoforms with distinct biological properties; often in a tissue- and development stage-specific way (Wang *et al.*, 2008; Barbosa-Morais *et al.*, 2012).

In haematopoiesis, AS has been shown to cause differential expression of isoforms that are important in regulating the differentiation and function of blood cells. For example, AS of three exons (4, 5, and 6) in the CD45 gene is known to produce multiple isoforms characterised by the differential inclusion of glycosylated segments of the CD45 cell-surface protein (Zikherman and Weiss, 2008). This alternative isoform expression has been shown to be regulated in a cell-lineage and state-specific fashion where the long isoform (CD45RABC) is almost exclusively found on B cells, whilst during differentiation of naive T cells (which express various larger isoforms) to memory T cells is accompanied by exon exclusion generating the RO short isoform, a canonical marker of T helper cells (Hermiston, Xu and Weiss, 2003). During erythropoiesis, AS of the *Bcl11a* gene results in the expression of different isoforms that regulate the switch from foetal to adult haemoglobin production (Uda *et al.*, 2008). This gene encodes a transcriptional repressor with essential functions during development, whereby *Bcl11a* haploinsufficiency causes Dias-Logan syndrome (a developmental disorder associated with the hereditary persistence of foetal haemoglobin) (Dias *et al.*, 2016).

Recent work has identified disease-causing variants in this gene that lead to the truncation of the BCL11A-XL protein through the absence of the C terminal components necessary for nuclear localisation signalling, rendering it inactive (Wessels *et al.*, 2021). In fact, splicing mutations are among the most recurrent genetic perturbations in haematological malignancies, common to all forms of myeloid malignancies including acute myeloid leukaemia (AML) and myeloid proliferative neoplasms (MPNs). Over 50% of patients with myelodysplastic syndromes (MDS), clonal blood disorders characterised by impaired haematopoiesis, carry 1 or more mutations affecting splicing factors with genes including *Sf3b1, U2af35* and *Zrsr2* (Graubert *et al*., 2011; Papaemmanuil *et al.*, 2011; Yoshida *et al.*, 2011; Cazzola, Della Porta and Malcovati, 2013; Wan and Wu, 2013; Genovese *et al.*, 2014; Haferlach et al., 2014; Desai *et al*., 2018). These are but a few known examples which demonstrate the important role of AS in regulating cellular function, highlighting how its dysregulation can also lead to disease.

However, while AS is of crucial importance for normal haematopoiesis and haematopoietic malignancies, the role it plays in haematopoietic lineage specification is still largely unknown.

Alternative methods are continuously emerging to address the limitations of current technologies with the overarching goal of providing a more comprehensive view of cellular expression whilst not compromising important practical factors of research, such as throughput and cost. The quantification of RNA isoforms has been a challenge within the field due to the requirement of both long enough reads that capture full-length isoforms and sufficient coverage depth to ensure sequence accuracy. This has been a particular challenge in single-cell experiments with the current platforms available for sequencing. An updated version of the Smart-Seq protocol by Picelli *et al.*, Smart-Seq3, has shown improvements in its ability to detect isoforms by enabling the reconstruction of single molecules through the integration of reads with matching 5' UMI (Picelli *et al.*, 2014; Hagemann-Jensen *et al.*, 2020). However, the majority of transcripts are often only partially reconstructed yielding only marginal gains in isoform quantification and therefore limited novel isoform discovery from single cells.

A solution for obtaining full-length transcripts is to transition from short- to long-read sequencing technologies. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) enable full-length RNA sequencing, capable of generating reads that are tens of kilobases in length without the computational requirement of transcript assembly but, conversely, suffer from comparatively lower read throughput and high costs. Fortunately the development of circularised consensus sequencing (CCS) by PacBio, whereby consensus reads from multiple sequencing 'passes' of individual library molecules are generated, has led to reduced error rates by improving the raw base calling accuracy in their long-read sequencing runs (Wenger *et al.*, 2019). The Phred scale is a widely-used measure of sequencing accuracy, expressed as a negative logarithmic scale that scores base-calling error probability. Higher Phred scores correspond to lower probabilities of base-calling errors, with a Phred score of Q30 indicating a base-calling accuracy of 99.9%, representing highly-accurate sequencing reads. PacBio sequencing generates ~Q30 quality reads (HiFi reads) after 10 circular passes, and is achieved from input libraries ranging between 15-20 kb. This poses a problem for sequencing single-cell transcripts which typically range 1-2 kb in length, as it results in excessive numbers of passes (50-60) and wasting sequencing potential by 'over-sequencing' molecules. This issue is particularly important when performing long-read sequencing from single-cell input, as it reduces throughput and limits the sequencing potential of the PacBio platform.

Two novel approaches were recently developed by independent groups to enable the generation of long-read single-cell libraries that are both effective in capturing full-length transcripts and more cost-effective for sequencing with long-read technologies. The first, *high-throughput and*

high-accuracy single-cell full-length isoform sequencing (HIT scIso-Seq) is a novel protocol that enables head-to-tail enzymatic concatenation of cDNA (Figure 5.1) (Shi *et al.*, 2022). This approach enables full-length cDNA generated using the 10X Genomics system to be concatenated using PCR-introduced palindrome-overhangs at both ends of cDNA inserts into single, longer molecules of multiple inserts. In addition, HIT scIso-Seq also incorporates an additional step for the removal of template switching oligo (TSO) artefacts, prior to the generation of concatenated library products. TSO artefacts represent another limitation in current scRNA-seq approaches, formed through priming errors during cDNA amplification producing cDNA products which lack the 3'-end barcode-containing sequence. These artefacts that lack UMI sequences mean that cDNA reads cannot be assigned to cells but are indistinguishable from 'desired' cDNA products and consequently are carried over during library generation and sequenced. Estimates suggest that cell-barcode-free TSO artefacts can constitute up to 50% of reads from libraries constructed with 10X Genomics (Lebrigand *et al.*, 2020).

A second method builds on the same idea of cDNA concatenation to generate longer reads, and also implements TSO-artefact depletion but is distinct in its concatenation strategy. Multiplexed Arrays Sequencing (MAS) of transcript isoforms (MAS-seq) uses barcoded adapters across parallel PCR reactions for programmable concatenation with a narrow length distribution (Figure 5.1) (Al'Khafaji *et al.*, 2021). This approach adds adapters of specific sequences to the 5' and 3' ends of cDNA across parallel reactions which, when pooled, are ligated; generating long fragments with multiple cDNA inserts. The authors of HIT scIso-Seq and MAS-seq reported 8- and 15-fold yield increase in total corrected read counts respectively, demonstrating the significant boosts in sequencing throughput that can be achieved with both methodologies.

scRNA-seq has revolutionised genomic research at the gene-level, but progress in the field of isoformed-resolved transcriptomics at single cell resolution has been hampered by the current technologies' high cost and low throughput. The development of methods such as these, which enhance long-read sequencing throughput from single cells, is crucial to enable the detection and hence study of isoform heterogeneity across cells. Early adoption of novel methods and testing across diverse experimental systems will be imperative to determine their efficacy, uncover potential applications, reveal limitations and drive further technological development.

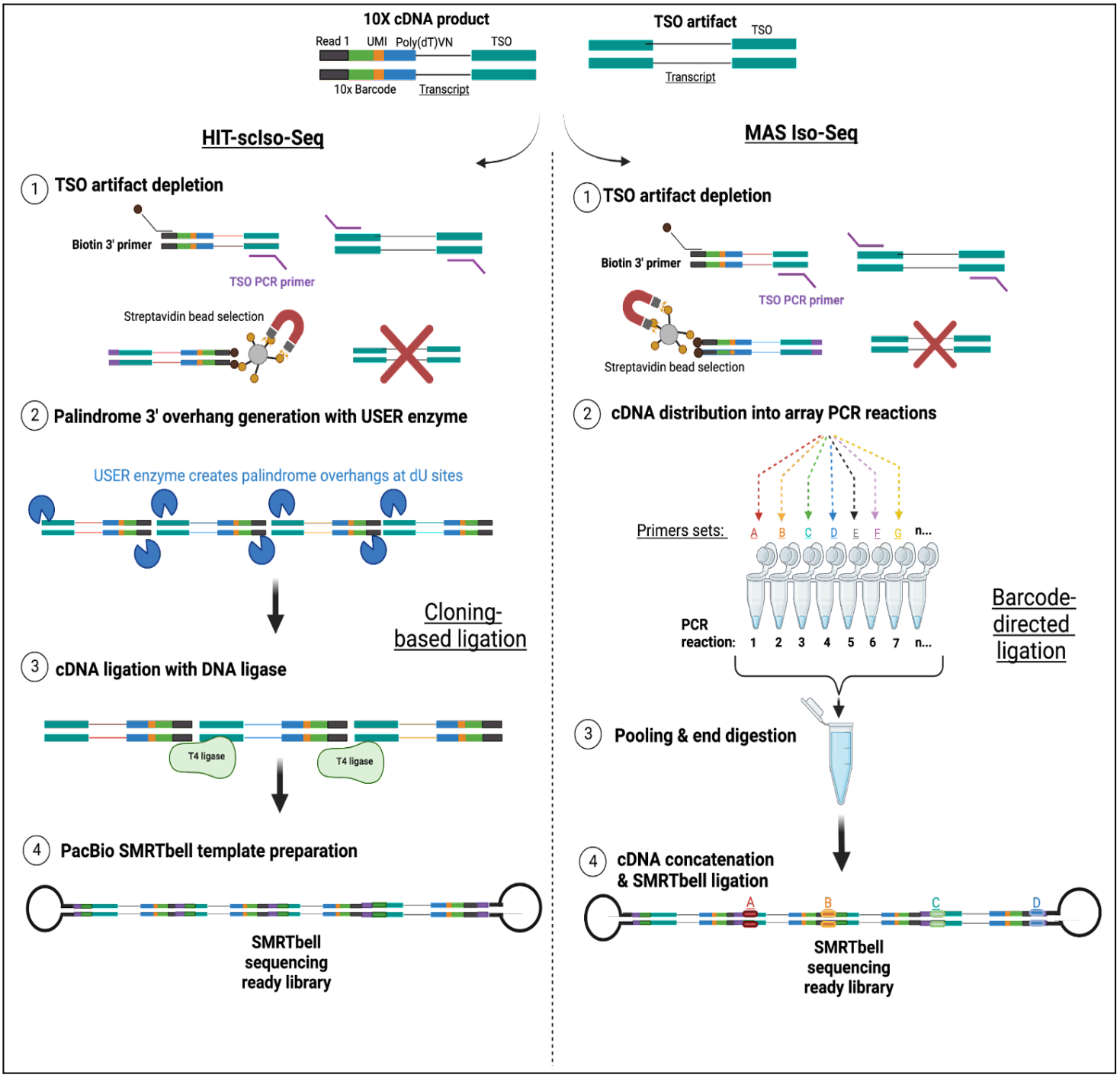


Figure 5.1. Side-by-side methodology workflows for HIT scIso-Seq (left) and MAS-seq (right) for concatenation of 10X Genomics scRNA-seq cDNA. HIT scIsoSeq uses specific biotinylated PCR primers that enable TSO artefact depletion and introduce deoxy uracil (dU) sites to the 3' ends of desired cDNA fragments. USER enzyme is used to generate palindrome overhangs and DNA ligase to concatenate complementary palindrome sequences thus concatenating cDNA through cloning-based ligation (Shi *et al.*, 2022). The first step of the MAS-seq workflow also applies TSO artefact depletion through streptavidin bead selection of desired fragments. After depletion, the TSO-depleted cDNA product is distributed across multiple tubes for parallel PCR using specific sets of primer pairs per reaction to append dU barcode adapters to both ends of cDNA fragments. After PCR, the arrayed reactions are pooled and undergo dU digestion and barcode-directed ligation generating concatenated cDNA arrays of programmable lengths (Al'Khafaji *et al.*, 2021).

5.1.1. Aims

The aims of this chapter were to assess approaches for long-read sequencing of cDNA in single-cell experiments. Two separate approaches were applied :

Part 1: Pooling strategy for cell type specific isoform sequencing

- Generate PacBio libraries from pooled cDNA from LK and LSK Cd150+ HSCs annotated from scRNA-seq data (Chapter 4).

Illumina single cell RNA sequencing provides highly accurate coverage to enable gene-level analyses from single cells. However, short-read libraries typically range between 150-300 bps which are often insufficient to resolve isoform expression. In the previous chapter, Illumina gene-level clustering was used to identify HSCs within the LK Cd150+ compartment. Drawing on this, it was hypothesised that long-read sequencing of the same HSCs would enable comprehensive isoform-resolved profiling of HSCs with age.

Part 2: Advanced methods for single-cell isoform sequencing

- Perform HITsc-IsoSeq PacBio library construction to generate cDNA concatemers from mouse BM single cells (10X Genomics)
- Perform MAS-seq PacBio library construction to generate cDNA concatemers from human PBMCs and FACS sorted mouse LK Cd150+ single cells (10X Genomics)

The hypothesis was that the utilisation of HITsc-IsoSeq and MAS-seq cDNA concatenation would increase the PacBio sequencing throughput for single-cell RNA generated using 10X Genomics, as excessive passes on individual molecules with fragments of < 3 kb are not advantageous for read accuracy. Additionally, it was predicted that the absence of fragmentation during library preparation, preserving full-length cDNA, would enhance isoform classification from single cells.

5.2 Experimental approach

Part 1

In order to investigate the expression of isoforms in the HSC compartment with age, PacBio libraries were generated of LK Cd150+ cells annotated as HSCs from scRNA-seq clustering (see Chapter 4 Results). To achieve this, IsoSeq libraries were constructed from intact full-length cDNA prepared using Smart-Seq2. The HSC cDNA samples were first pooled into two multi-cell suspensions based on the ages of the mice (young 8 - 10 weeks, or aged ~72 weeks) (Figure 4.1). The resulting SMRT bell libraries were subsequently sequenced using the PacBio Sequel II platform with long-read technology, as described in Methods 2.2.8.

Part 2

Recently, two novel methods have been published that aim to increase the PacBio sequencing throughput from single cells (Shi *et al.,* 2022; Al'Khafaji *et al.,* 2021). To investigate the effectiveness of these approaches, concatenated HITsc-IsoSeq and MAS-seq libraries were generated from single cell cDNA that was prepared using 10X Genomics from both mouse BM and human PBMCs. Furthermore, two MAS-seq libraries of LK Cd150+ FACS sorted cells were also created to test the compatibility of FACS cell-type enrichment for 10X Genomics. PacBio and Illumina sequencing were performed on all samples, and the obtained sequencing metrics for each approach were explored using Illumina data as a benchmark for the captured cell types (Figure 5.1).

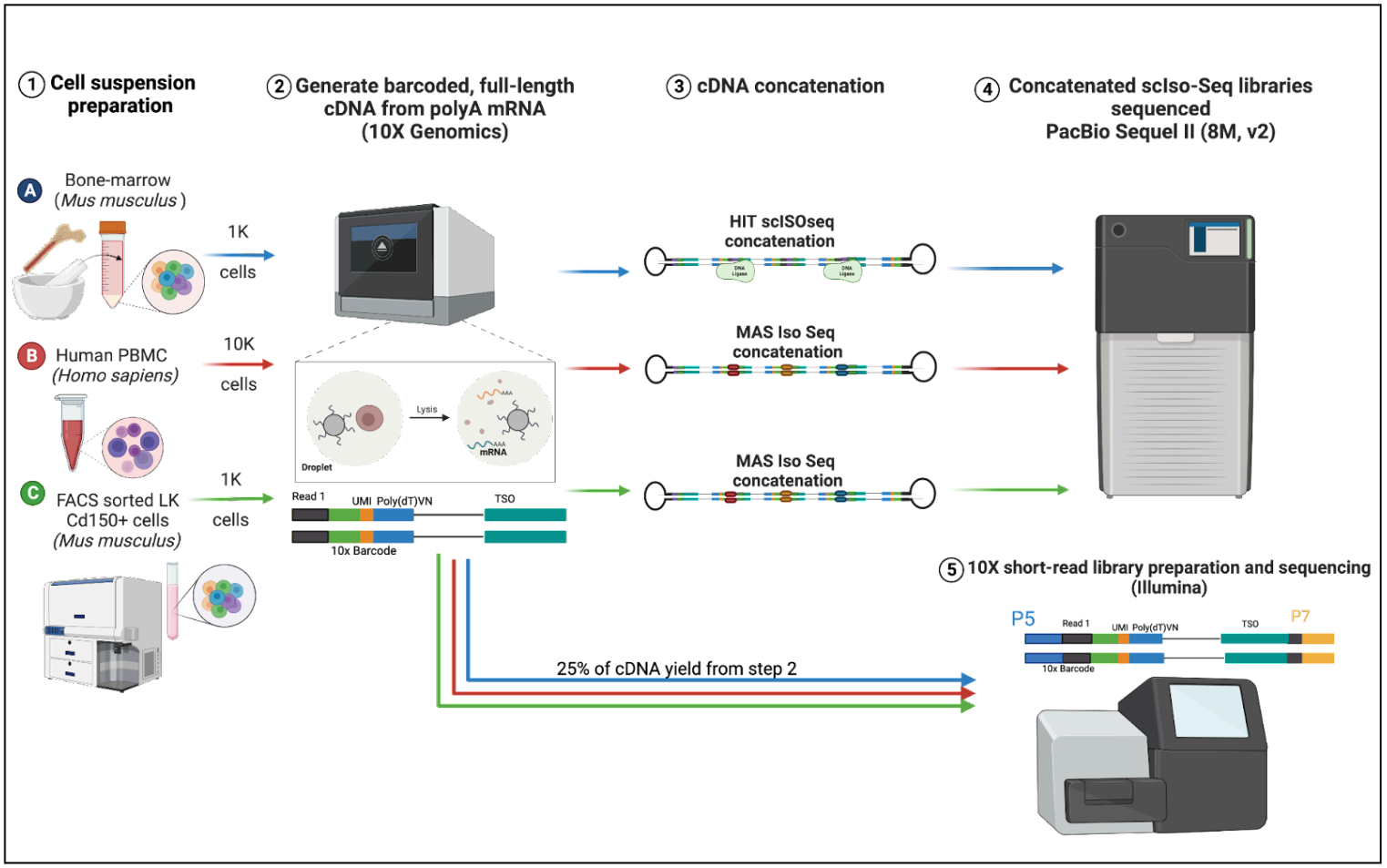


Figure 5.2. The summarised experimental approach implemented to obtain data presented in part 2 of Chapter 5[17].

[17] Created with BioRender.com

5.3 Results

Part 1:

5.3.1. Generation of long-read libraries from cDNA of purified HSCs

The scRNA-seq data generated in Chapter 4 using Illumina sequencing technology was used to annotate cell types based on their transcriptomic profiles. Cell IDs corresponding to HSCs, including the LT-HSC subset, were extracted from the *Seurat* object based on mouse age. Two lists containing metadata of the specific plate ID and well position across the 7 Smart-seq2 plates processed were generated, these consisted of 34 cells from young samples and 34 and 46 from aged. Using these lists up to 10 µl volumes were carefully collected from wells and pooled into two mini-bulk cDNA libraries, one for each experimental age group. After pooling, bead purification was performed to remove any degraded material, and the final total cDNA concentration was measured. The young sample yielded 273 ng of total mass, while the aged sample yielded 297 ng, both met the minimum input requirement of 160 ng for PacBio library construction.

SMRT bell libraries were constructed for each sample using the SMRTbell prep kit 3.0 (Materials and Methods 2.2.8.2 - 2.2.8.5). Post-library construction, QC was performed to assess library concentrations and size distributions were within the recommended range. The young IsoSeq library was 8.3 kb (sample 1), and the aged sample 6.2 kb (sample 2) (Appendix Supplementary Figure 5.1). Library concentrations were normalised and sequenced on 2 SMRT cells of the PacBio Sequel II (8 M, v 2) with 30 hr movies.

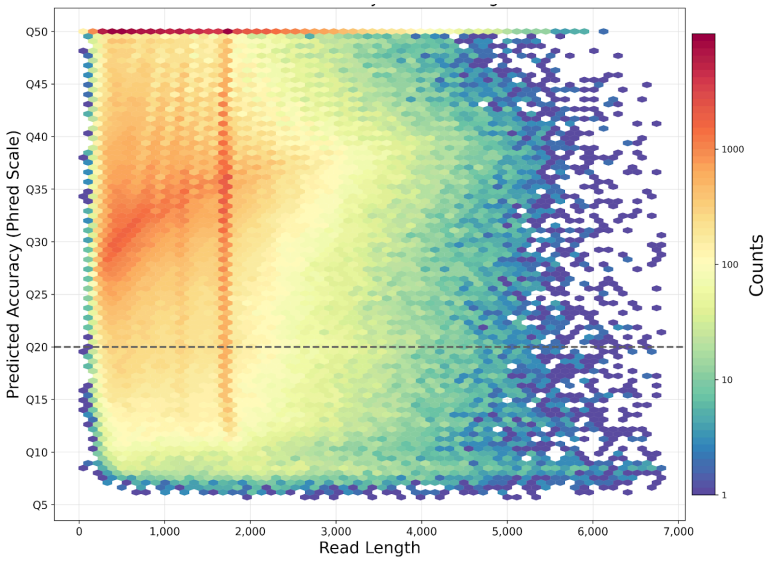
5.3.2. Sequencing metrics from HSC IsoSeq libraries

Before characterising isoform detection, the quality of SMRT cell runs was first evaluated based on a number of criteria; such as the number of reads generated in the run, average sequence length and polymerase read quality. This was achieved using PacBio's SMRT Link software, generating raw data reports for both libraries. SMRT Link was also used to extract High Fidelity (HiFi) reads from the raw sequencing output. These are characterised by high accuracy, low error rates and long read lengths. This makes them particularly useful for applications including transcript isoform discovery, alternative splicing analysis and gene expression quantification. The high accuracy of HiFi Iso-Seq reads is achieved by circular consensus sequencing (CCS), which involves multiple passes of the same DNA molecule to generate a high-confidence consensus sequence. In total, run 1 generated 1.4 million reads, of which ~687 K reads met HiFi criteria (Figure 5.3A). Figure 5.3B shows the correlation between predicted accuracy using the Phred scale metric and CCS read length, showing that most read

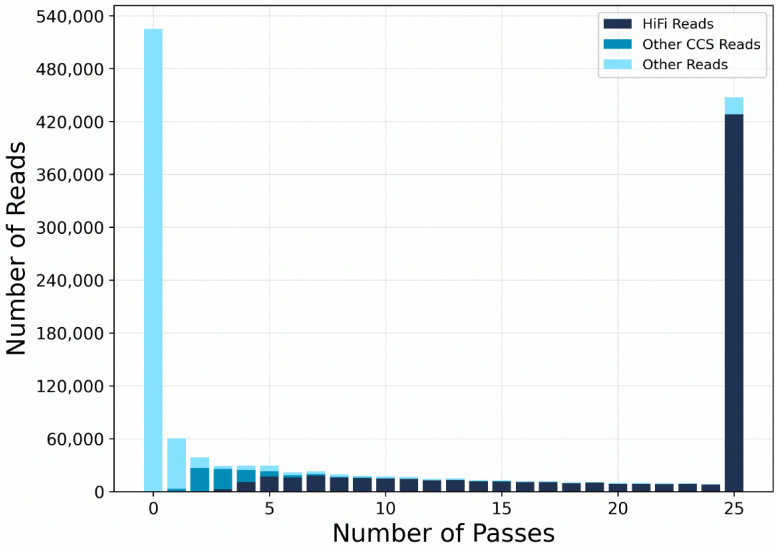
a.

Metric	Value
HiFi reads	686,809
HiFi yield (bp)	877,169,963
Mean HiFi read length (bp)	1,277
Median HiFi read quality (Q)	Q35
Mean HiFi number of passes	20
<Q 20 Reads	79,538

b.



c.



d.

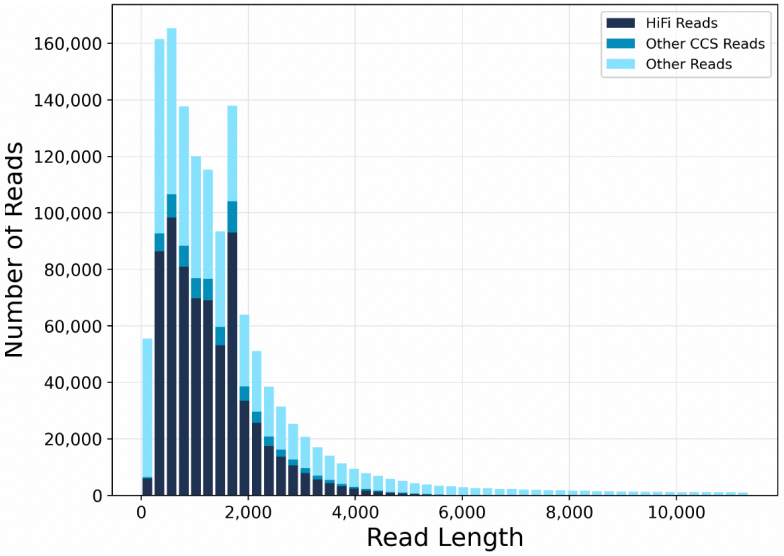


Figure 5.3. Sequencing statistics of PacBio IsoSeq library generated from pooled single-cell cDNA from young mice. (a) Summary of key HiFi quality metrics **(b)** Heat map of CCS Read lengths and predicted accuracies. The boundary between HiFi Reads and other CCS Reads is shown as a dashed line at QV 20. **(c)** Histogram distribution of HiFi Reads (QV >=20), other CCS Reads (three or more passes, but QV <20), and other reads, by number of passes **(d)** Histogram distribution of HiFi Reads (QV >=20), other CCS Reads (three or more passes, but QV <20), and other reads, by read length.

counts passed the Q20 threshold. The average HiFi read length was 1,277 bps, with a mean 20 passes per HiFi molecule (Figure 5.3C and D). Overall, these statistics suggest the production of high-quality reads, but that sequencing efficiency was suboptimal. The loading report revealed 17% productivity in P1, which represents the category of productive zero-mode waveguides (ZMW) with a high-quality sequencing region detected within the read. 81% ZMW productivity was categorised as P0 - representing a non-productive ZMW with no signal detected. The P0 metric provides an accurate estimate of sample loading on SMRT cells, where quantities exceeding the optimal inflection point of loading result in poorer sequencing performance. Collectively these sequence performance metrics suggest the library was overloaded, resulting in a poor overall yield, but give no indications suggesting library quality was compromised.

In comparison, a better sequencing performance from sample 2 (aged) is evident. 60% ZMW productivity was recorded in P1, with only 39% in non-producing P0. The number of HiFi reads reflects this, with ~3.6 fold more HiFi reads than sample 1 (Figure 5.4A). Despite the worse performance in sample 1, the HiFi quality across both libraries is consistent in terms of HiFi read average length, quality score and the number of passes; suggesting good library quality which was consistent between samples (Figure 5.4C and D).

a.

Metric	Value
HiFi reads	2,506,093
HiFi yield (bp)	3,501,432,015
Mean HiFi read length (bp)	1,397
Median HiFi read quality (Q)	Q35
Mean HiFi number of passes	20
<Q 20 Reads	309,056

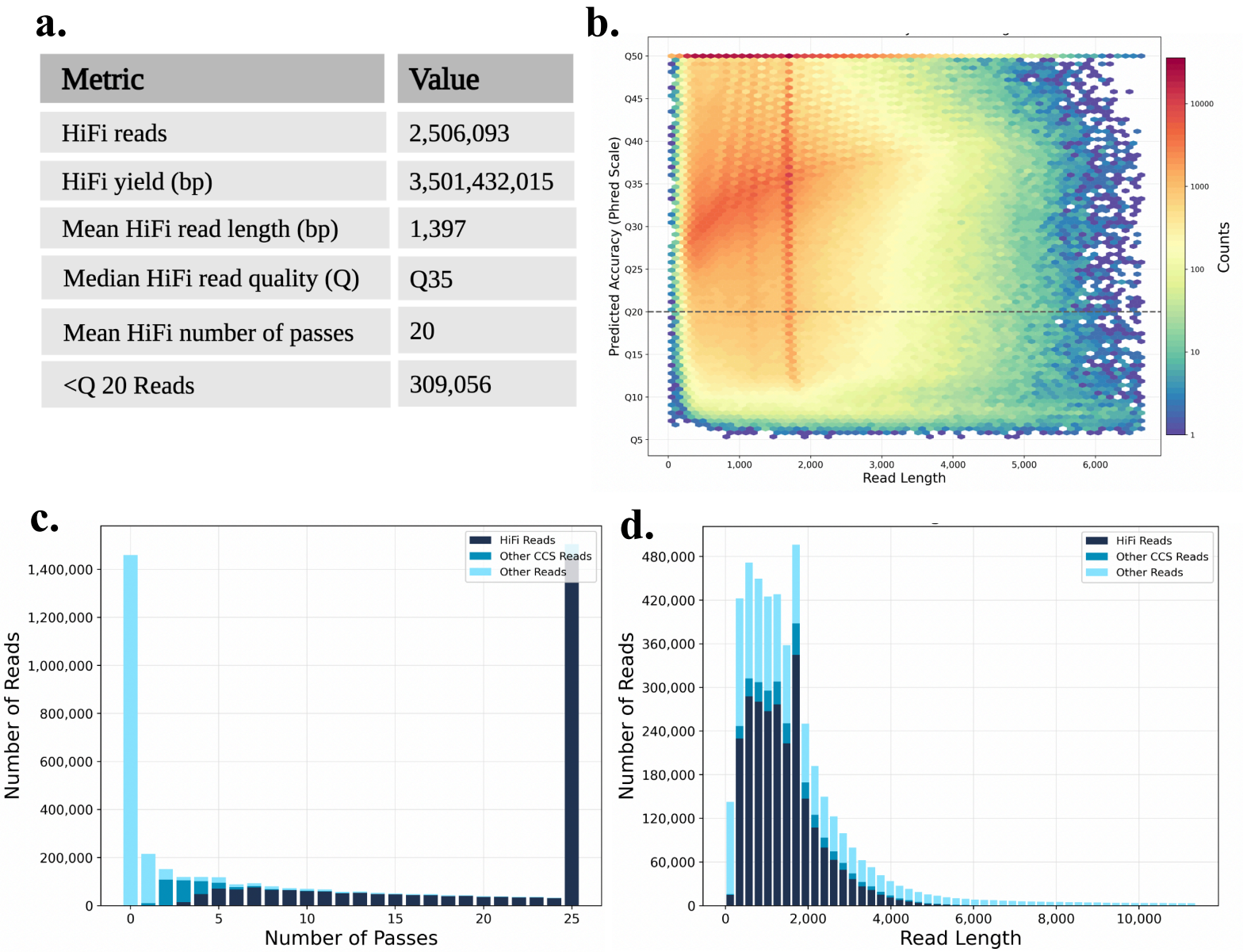


Figure 5.4. Sequencing statistics of PacBio IsoSeq library generated from pooled single-cell cDNA from old mice. (a) Summary of key HiFi quality metrics **(b)** Heat map of CCS Read lengths and predicted accuracies. The boundary between HiFi Reads and other CCS Reads is shown as a dashed line at QV 20. **(c)** Histogram distribution of HiFi Reads (Q value >=20), other CCS Reads (three or more passes, but QV <20), and other reads, by the number of passes **(d)** Histogram distribution of HiFi Reads (QV >=20), other CCS Reads (three or more passes, but QV <20), and other reads, by read length.

5.3.3. Quantification and annotation of isoforms based on structural categories

IsoSeq enables the production of full-length cDNA and thus quantification of isoforms. Smart-seq2 also produces full-length cDNA, and with high sensitivity and low input requirements is a popular approach for studying rare and heterogeneous cell types. Comparing the normalised coverage of IsoSeq and Smart-seq2 along transcripts shows that both approaches span the length of transcripts, but IsoSeq provides greater coverage of 5' and 3' ends (Figure 5.5).

To enable characterisation of the isoforms, transcripts first had to be mapped to the mouse genome and poor-quality reads removed. The *minimap2* sequence alignment program was used to perform spliced alignment of PacBio HiFi FASTQ reads to the mouse reference genome containing known splice junctions (Li, 2018; Frankish *et al.*, 2019). The aligned transcript sequences were then collapsed into unique isoforms using *cDNA cupcake*, clustering transcript isoforms that map to the same region into a representative consensus and removing redundant mapped reads.

QC and classification of isoforms were then achieved using *SQANTI3* (Tardaguila *et al.*, 2018). The collapsed mapped reads were first classified based on their mapping quality to the reference and overlap with known splice junctions. Isoforms were then classified into different categories based on their annotation status, splicing patterns, and genomic context, providing a comprehensive and detailed analysis of isoform diversity. This was performed for both samples individually and for a merged dataset of both libraries, as sample pooling is recommended to build a single transcriptome experiment on which to assign isoform identities.

A total of 5942 unique genes and 7456 unique isoforms were identified from the combined IsoSeq libraries. Transcript classification grouped isoforms into categories based on reference transcript categories and generated reports on data quality (Figure 5.6) (see also Materials and Methods section 2.3.15 and Figure 2.1). Transcripts matching all splice junctions of the reference genome are labelled as full splice match (FSM), while transcripts which do not contain all splice junctions of the reference but have matching consecutive junctions are designated incomplete splice match (ISM) transcripts. Novel transcripts are also annotated with SQANTI3, novel in catalogue (NIC) and novel not in catalogue (NNC). NIC isoforms contain novel combination(s) of splice junctions that have been previously annotated, or novel splice junctions from annotated donor and acceptor sites. While NNC are distinct by using novel donors and/or acceptors (Tilgner *et al.*, 2013; Tardaguila *et al.*, 2018). Further subtyping of isoforms is performed for those not matching the splice patterns of annotated references using existing annotations as a reference.

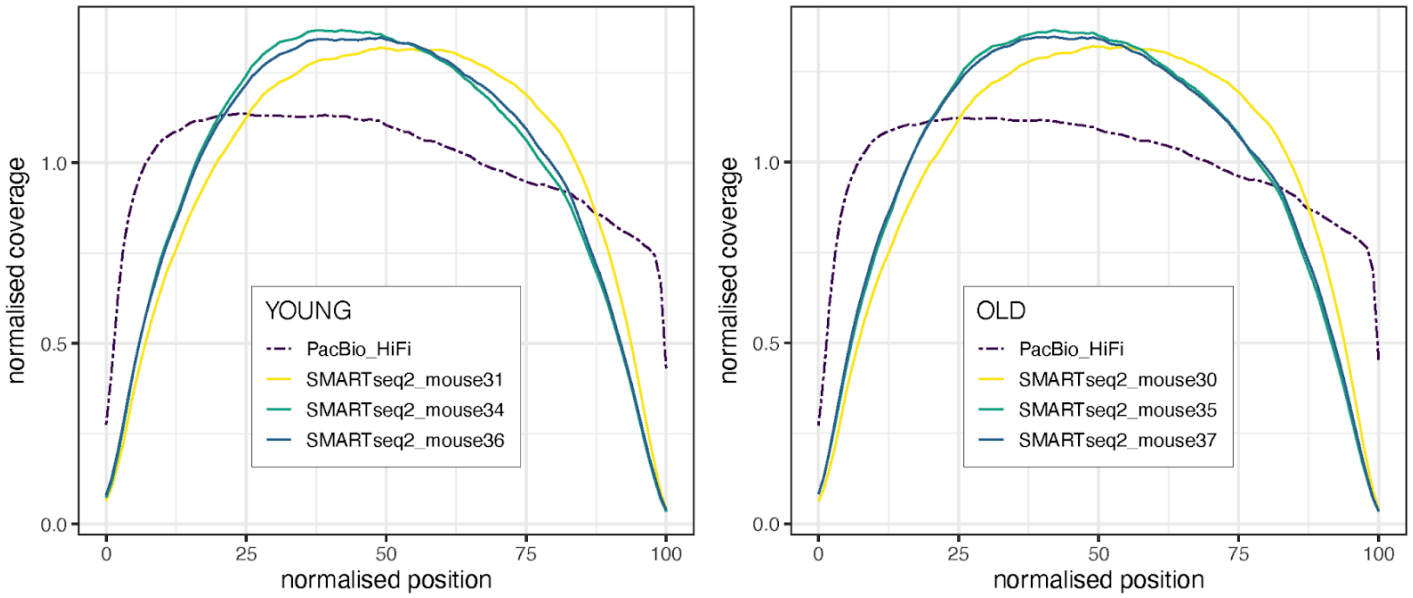


Figure 5.5. Normalised coverage across transcripts of young and old Smart-Seq2 Illumina and PacBio IsoSeq libraries from the same cells[18]. Data from Smart-seq2 samples were aggregated by batch and condition (mice 31 and 30 = yellow, mice 34 and 35 = green, mice 36 and 37 = blue).

[18] Figure generated by Dr David Wright

This includes transcripts in novel genes outside the boundaries of an annotated gene (Intergenic), those found entirely within boundaries of an annotated intron (Genic intron), and those with partial exon and intron overlap in a known gene (Genic genomic). In addition, polyA-containing transcripts that overlap the complementary strand of an annotated transcript as well as transcripts spanning two annotated loci are classed as antisense and fusion isoforms, respectively.

A significant proportion of the identified isoforms were classified as antisense, indicating that they were transcribed from the opposite DNA strand compared to the gene of interest (Figure 5.6B). These antisense isoforms can arise from various processes such as alternative transcription start or end sites, alternative splicing, or genomic rearrangements (Xu, Zhang and Zhang, 2018). SQANTI3 categorises isoforms as antisense based on their genomic orientation relative to annotated genes in the reference. However, this unusually high level of antisense isoforms suggests that experimental artefacts might be accumulating in these categories. It is important to consider the influence of reverse transcriptase-switching events. Reverse transcription (RT) switching during cDNA synthesis can introduce gaps that are erroneously interpreted as non-canonical splicing events (Houseley and Tollervey, 2010). RT is of course an essential component in Smart-seq2 cDNA generation, however this intrinsic property can lead to the generation of artificially deleted cDNA and result in false-positive detection of alternative transcripts (Tardaguila *et al.*, 2018). A higher percentage of RT-switching junctions may indicate a more complex and diverse library with a greater presence of novel transcript isoforms and AS events. However it could also indicate an increased occurrence of template-switching artefacts during cDNA synthesis, leading to inaccurate quantification of gene expression levels and incorrect annotation of transcript isoforms. RT switching events are associated with a direct repeat sequence between the upstream mRNA boundary of the noncanonical intron, and the adjacent intron region near the downstream exon boundary (Cocquet *et al.*, 2006). By exploiting this hallmark SQANTI3 is able to generate a prediction of the likelihood the observed event is a RT switching artefact.

Additionally, SQANTI3 assesses the possibility of intra-priming events, which involve the binding of the oligo(dT) primer during the first-strand cDNA synthesis to A-rich regions of the mRNA template. This off-target priming can lead to the formation of false cDNA molecules, as it occurs with intron-lariats or pre-messenger RNAs that still contain non-poly(A) tail adenine stretches (Nam *et al.*, 2002; Spies, Burge and Bartel, 2013). SQANTI3 addresses this by calculating the percentage of adenines (A) within a specific window downstream from the genetic coordinates corresponding to the 3' ends of transcripts to assign intra-priming events.

a.

Category	Isoforms, count
FSM	726
ISM	863
NIC	865
NNC	308
Genic Genomic	252
Antisense	2740
Fusion	20
Intergenic	1682
Genic Intron	0

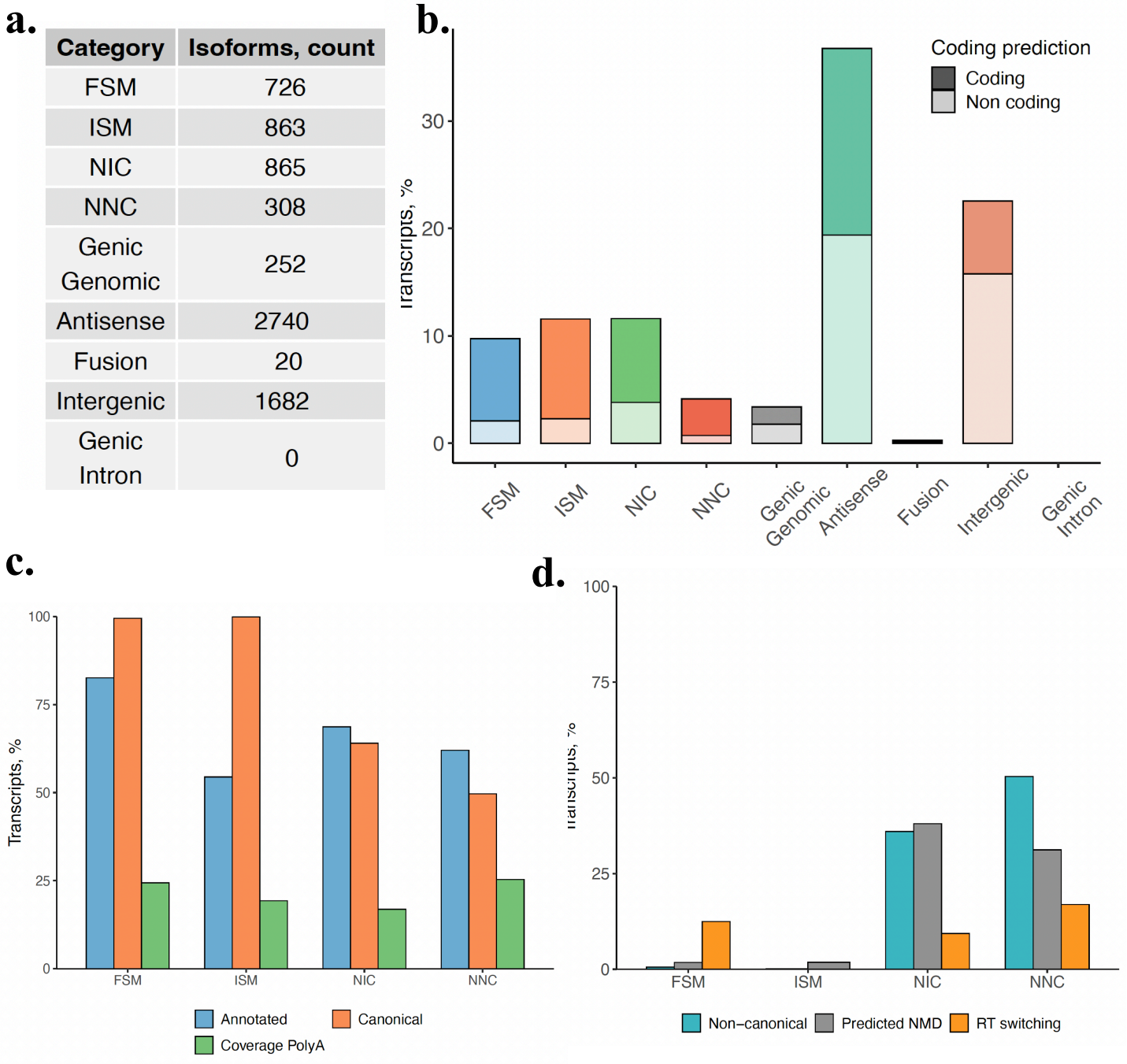


Figure 5.6. Isoform classification statistics from merged long-read data. (a) Isoform counts per structural categories post SQANTI *filter* artefact removal (b) Percentage distribution of isoforms across each structural category from both datasets combined (c) Good quality control attributes across structural categories including whether they are canonical, are supported by the annotation and poly(A) coverage (d) Poor quality control attributes across structural categories, including RT switching predicted nonsense-mediated decay and detection of non-canonical junctions.

RT switching has been previously shown to impact minor isoforms that coexist with a major isoform serving as the template for intra-molecular switching (Cocquet *et al.*, 2006). Examination of read tracks revealed the presence of many antisense isoforms as single exon reads in intergenic and intronic regions compared to the mouse reference annotation, potentially due to intra-priming resulting from RT switching (Tardaguila *et al.,* 2018). Moreover, the NNC transcript category was accordingly found to be enriched for the most predicted RT-switching events (Figure 5.6D). This suggests that QC parameters for excluding artefact isoforms need to be refined when utilising SQANTI3 for isoform annotation from Smart-seq2 cDNA data. These findings highlight the importance of considering RT switching and intra-priming events in the analysis and interpretation of isoform data, ensuring accurate annotation and quantification of transcript isoforms.

5.3.4. IsoSeq captures multiple isoforms of key genes for both HSC and Mk function

Analysis of the libraries revealed a notable disparity in the number of genes and transcript isoforms detected in each library. The aged HSC library, benefiting from higher sequencing efficiency, exhibited a significantly higher detection of unique genes (4723) and transcript isoforms (5737). In comparison the young library gave only 1642 unique genes and unique 1800 transcript isoforms. This represents an approximately 2.8-fold lower detection of genes and 3-fold lower detection of isoforms in the poorly sequenced library.

Considering the variation in sequencing coverage, the distribution of isoforms from the QC analysis of different structural annotations exhibited high consistency between both libraries. Despite differences in the total number of isoforms detected, the proportions of isoforms belonging to structural categories remained largely similar (Figures 5.7B and 5.7D). This suggests that the underlying transcript diversity was comparable between the two libraries, providing confidence in the identified isoform categories. However, since statistical analyses and comparisons of expression levels were not performed due to sequencing depth differences and sample size (n = 1), analysis from this point is exploratory in nature.

IGV was used to visualise individual read tracks of each library against mouse annotations provided by RefSeq. Most transcripts were within the 3 kb range across both libraries (Figure 5.7 B and D) and many canonical HSC markers were supported by multiple reads and were sequenced end-to-end in both libraries; examples include *Hlf, Mpl, Sult1a1,* and *Esam.* In particular, *Mpl* is known to be a critical gene for both supporting HSC function and promoting megakaryopoiesis. This gene encodes for the TPO receptor, the primary regulator of megakaryopoiesis (Kaushansky and Drachman, 2002). In HSCs, *Mpl* has been shown to support HSC quiescence and interactions with the osteoblastic niche, as well as metabolically prime HSCs towards megakaryopoiesis (Yoshihara *et al.*, 2007; Nakamura-Ishizu *et al.*, 2018). This gene is composed of 12 exons (Mignotte *et al.*, 1994), and is primarily expressed as two distinct alternate mRNA isoforms. The transmembrane variant Mpl-II is due to use of a cryptic splice acceptor in exon 4, resulting in an in-frame deletion of 60 amino acids. The second variant encodes a truncated soluble receptor, Mpl-tr, generated from AS of exon 8 directly to exon 11; eliminating the juxtamembrane extracellular part and the transmembrane domain (Skoda *et al.*, 1993; Coers, Ranft and Skoda, 2004). The deletion Mpl-tr removes the transmembrane domain, consequently Mpl-tr is expected to give rise to a secreted form which might antagonise Mpl signalling by sequestering TPO. Previous work has shown that Mpl-tr overexpression results in a decrease in Mpl protein abundance, exerting a dominant-negative effect on the proliferation and survival of HSCs (Coers, Ranft and Skoda, 2004).

To study the patterns of AS in *Mpl,* reads mapping to mouse *Mpl* locus were visualised and a graphical representation of the exon-exon junction reads was generated as a Sashimi plot (Figure 5.8). This shows the coverage of reads across AS sites, revealing insights into patterns of AS exhibited within the sample. Both *Mpl* isoforms were identified within HSC reads in both samples. Study of the aligned reads clearly shows exon 9 and 10 skipping (Figure 5.9). AS variants of cytokine receptors that result in protein isoforms with differential functional characteristics are important regulators of cytokine signalling (Heaney and Golde, 1998; Nakamura *et al.*, 1998; Ashman, 1999). Another example of a gene with multiple splice variants is *Racgap1*. It encodes for a protein belonging to the Rho family of small monomeric GTPases. Rac activity has been demonstrated to be important for such diverse functions as retention in the BM, including long-term engraftment of HSCs and HSC mobilisation (Jansen *et al.*, 2005; Cancelas and Williams, 2009). Moreover, this gene has been implicated as a regulator of microtubule stabilisation and dynamics, important regulatory features of Mk proplatelet formation (Pleines *et al.* 2013). It was identified as differentially expressed with age based on short-read scRNA-seq data, the combination of long-read sequencing provides full-length transcript coverage showing the AS of exon 2 between conditions.

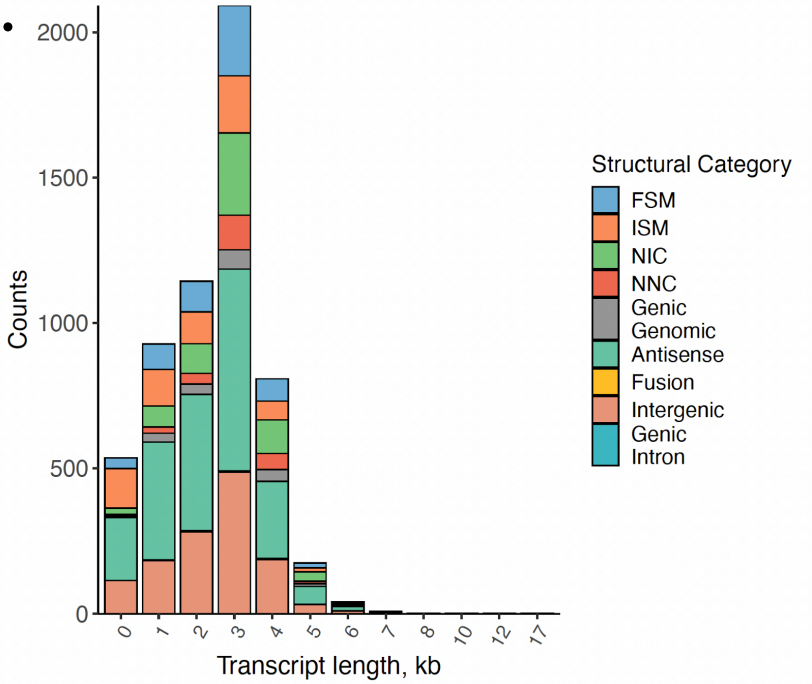
Chapter5: Part 1 - Results Summary:

In summary, this cell-type specific approach of long-read sequencing provided end-to-end coverage across RNA transcripts in HSCs, enabling the study of isoform-level expression heterogeneity. This technique allows the status of AS to be explored within a purified cell population through providing bulk-level sequencing coverage across the transcriptome HSCs that were characterised through both FACS analysis and their single-cell transcriptomic signature from short-read data.

a.

Category	Isoforms, count
FSM	568
ISM	647
NIC	631
NNC	252
Genic Genomic	196
Antisense	2127
Fusion	15
Intergenic	1301
Genic Intron	0

b.



c.

Category	Isoforms, count
FSM	195
ISM	219
NIC	238
NNC	58
Genic Genomic	60
Antisense	639
Fusion	5
Intergenic	386
Genic Intron	0

d.

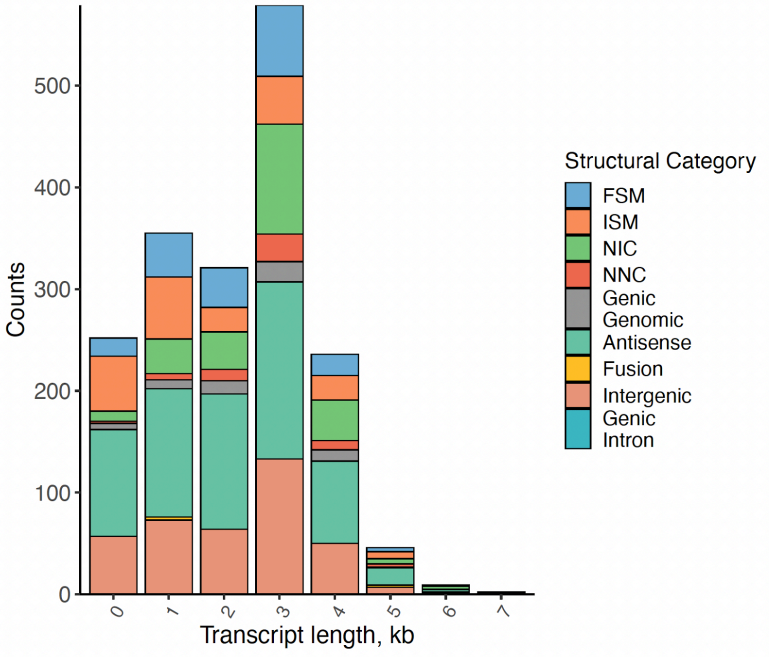


Figure 5.7. Isoform classification statistics between young and aged IsoSeq libraries. (a) Transcript classifications of aged IsoSeq library (b) Aged IsoSeq Structural Categories by Transcript Length (c) Transcript classifications of young IsoSeq library (d) Young IsoSeq Structural Categories by Transcript Length.

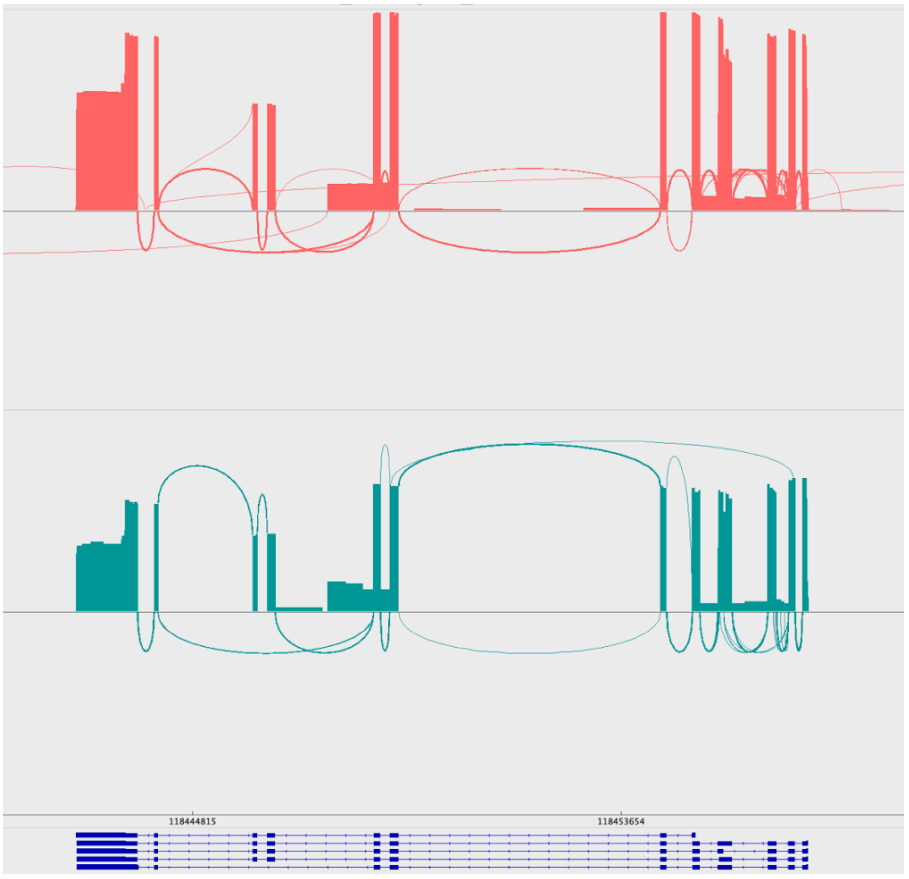


Figure 5.8. Sashimi plot of splice junction coverage of *Mpl* in IsoSeq libraries. Refseq annotations of known Mpl isoforms showing exon positions are shown in blue. The height or width of each rectangle represents the number of reads supporting a particular exon or splice junction. A**ged (top, red) and young (bottom, teal) HSCs.**

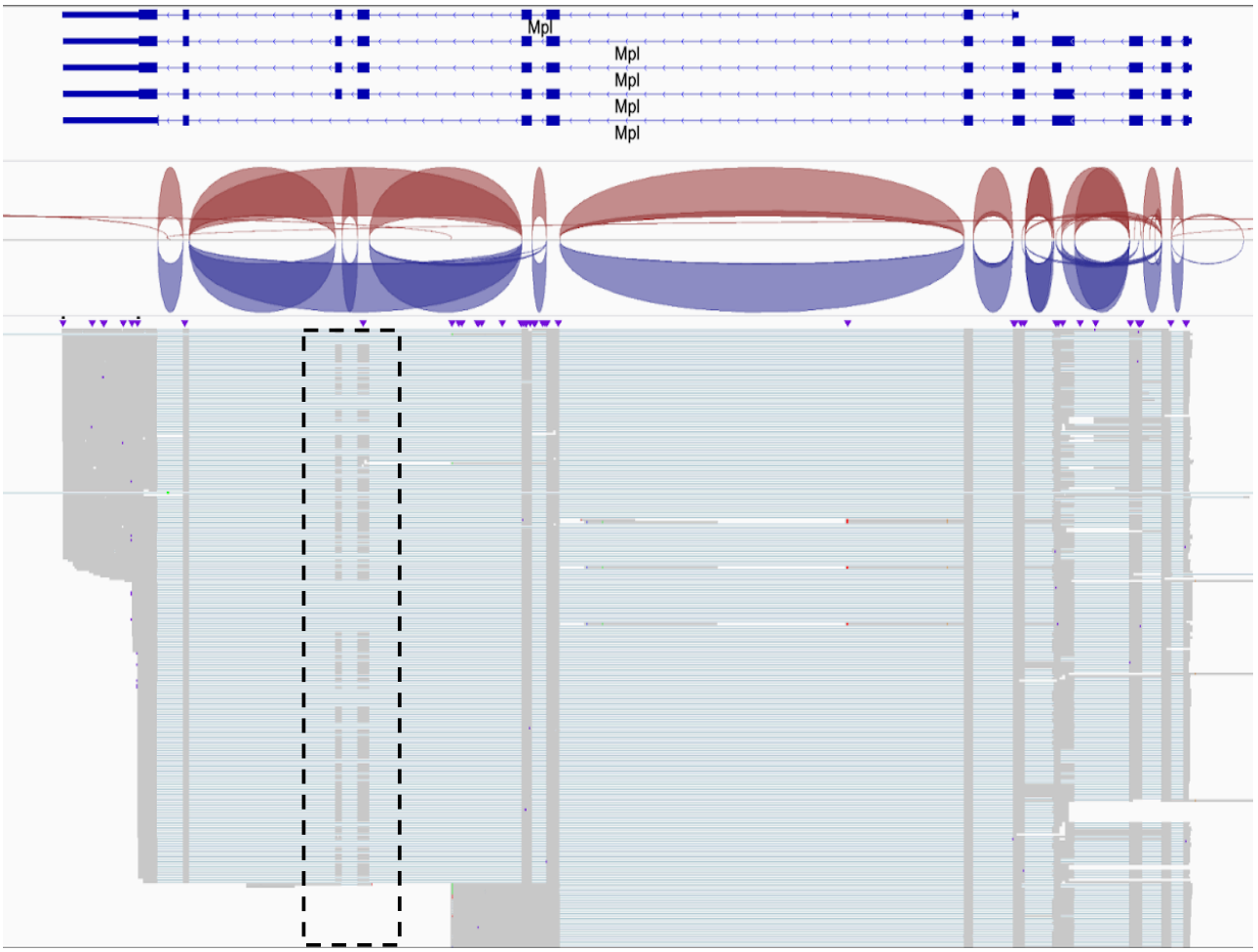


Figure 5.9. Alignment tracks of HiFi reads mapping to *Mpl* from aged samples. RefSeq reference annotations of known splice isoforms of *Mpl* are shown in blue (top). Splice junction coverage is shown coloured by read strand. Notably, these tracks show AS of exons 9 and 10 (highlighted) which are known to produce a truncated isoform of Mpl.

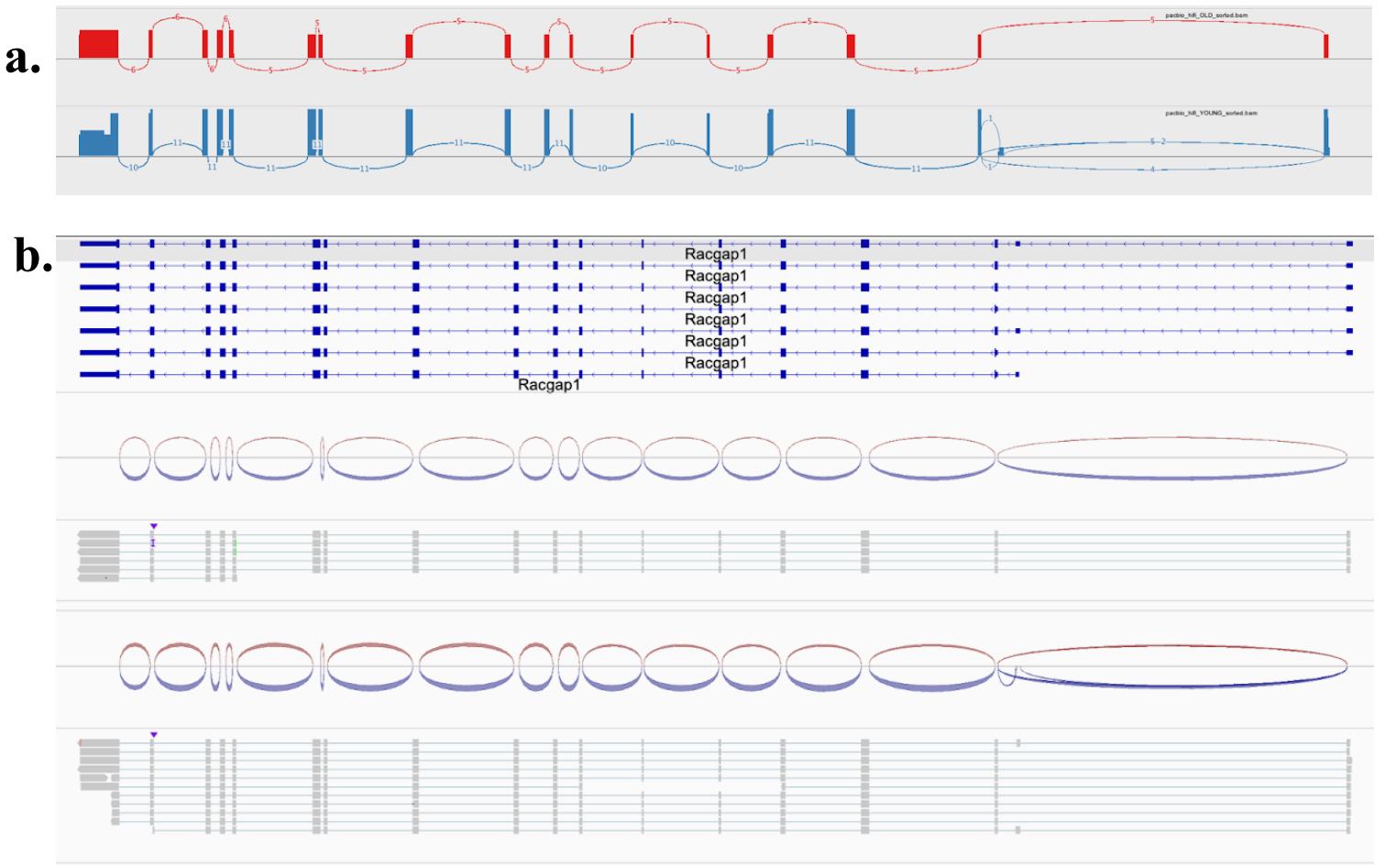


Figure 5.10. Splice junction coverage from IsoSeq libraries and HiFi read alignment for *Racgap1* from aged and young samples. (a) Sashimi plot of splice junction coverage of *Racgap1* in IsoSeq libraries from aged (top, red) and young (bottom, blue) HSCs. (b) Read alignment tracks of all HiFi reads mapping to the *Racgap1* from aged and young samples. RefSeq reference annotations of known splice isoforms of Racgap1 are shown in blue. The top tracks are from the aged library, bottom tracks are from the young library. Splice junction coverage is shown for each track, coloured by read strand. Notably, these tracks show the inclusion of Exon 2 in only the young library (bottom).

Part 2:

5.3.5. Cell isolation and library yields from 10X single-cell cDNA concatenated libraries

5.3.5.1. HIT scIso-Seq

To obtain a single-cell suspension of whole mouse bone-marrow cells, bones were dissected and prepared as described (see Materials and Methods 2.2.2 - 2.2.3). After processing bone-marrow samples and determining a cell viability of 92% the single-cell suspension was diluted in DPBS medium to the optimal concentration of 600 cells/uL and loaded onto the 10X Genomics Chromium NextGem Chip (L) as per the manufacturer's instructions for GEM generation. The sample was then processed with the low-throughput (LT) 3' v3.1 10X Genomics chemistry to obtain a full-length barcoded cDNA library. This yielded a total mass of 102 ng of cDNA with an average size distribution of 1882 bp (see Appendix Supplementary Figure 5.3). 21 ng of this cDNA was used for Illumina library construction, resulting in a final library of 144 ng. Sequencing was performed on 1 lane of the MiSeq v3 flow cell with 28-10-10-90 configuration.

27 ng of the total cDNA generated was used as input for cDNA concatenation with an adapted version of the HIT scIso-Seq protocol (Shi *et al.*, 2022). Briefly, this approach uses Uracil-Specific Excision Reagent (USER) enzyme to generate palindrome overhangs at specific deoxy-Uracil sites introduced to both 5' and 3' ends of inserts through PCR. DNA ligase is then used to concatenate complementary palindrome sequences resulting in the concatenation of cDNA through cloning-based ligation.

Prior to the concatenation of cDNA, the sample was depleted of TSO-artefacts through streptavidin magnetic bead separation and PCR amplified for primer ligation, enabling USER enzyme to create necessary overhangs for concatenation on cDNA 5' and 3' ends. Concatenation yielded 285 ng of SMRTbell library, with an average size distribution of 10.4 kb (Figure 5.11).

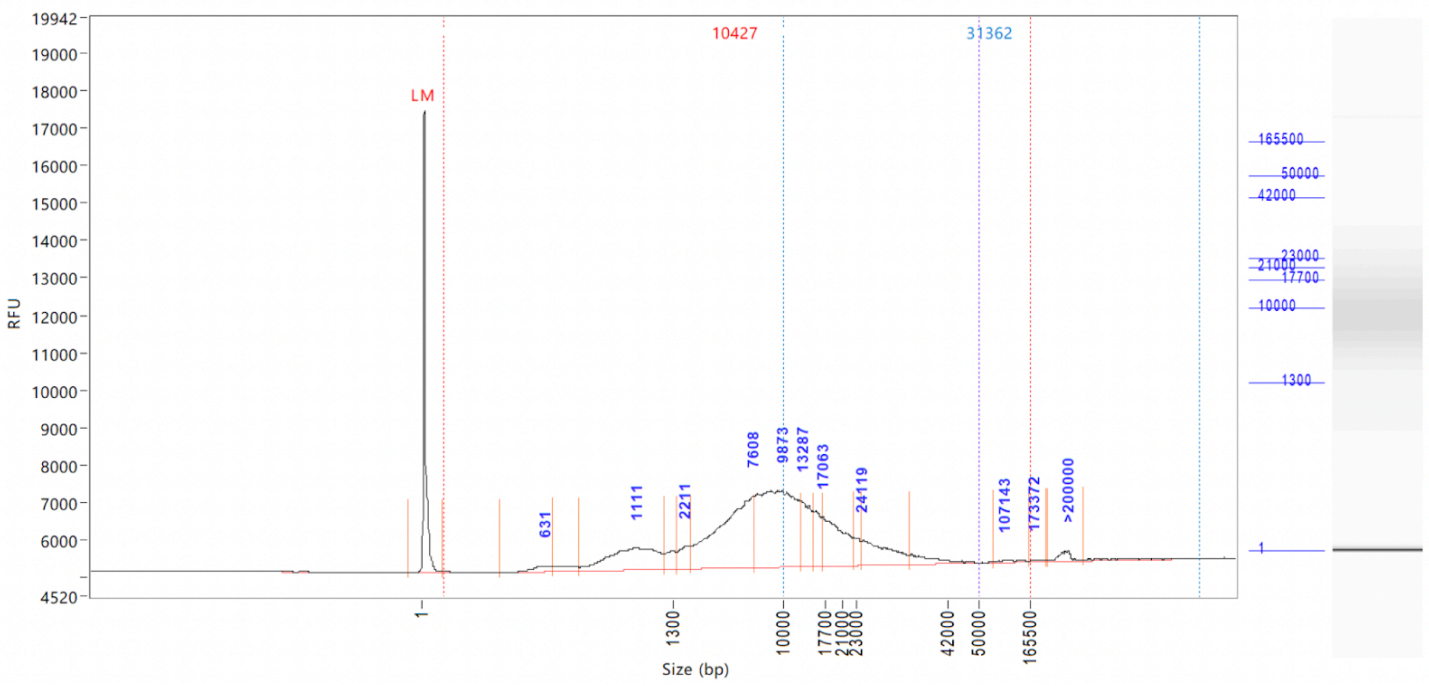


Figure 5.11. Library quality of single-cell library post HIT sc-IsoSeq concatenation. Femto Pulse size distribution trace of concatenated SMRTbell library generated from barcoded mouse whole bone-marrow cDNA.

5.3.5.2. MAS-seq of PBMCs

The viability and concentration of two human PBMC aliquots were determined using the Countess 2 automated cell counter (see Materials and Methods section 2.2.1: Human PBMCs). Cells were highly viable, with approximately 97% viability measured across both samples. The single-cell suspensions were diluted in DPBS medium to the optimal concentration of 1100 cells/μL and loaded onto the 10X Genomics Chromium NextGem Chip (M) as per the manufacturer's instructions with a target recovery of 8000 cells. Samples were processed with the Chromium Next GEM Single Cell 3' High-Throughput (HT) v3.1 kit to obtain two full-length barcoded cDNA libraries. The resulting cDNA products were measured on a Bioanalyzer with an average fragment length of 1.2 kb and total yields of 89.9ng and 257.4ng (see Appendix Supplementary Figure 5.4). 25% of the cDNA generated from each sample was used for Illumina library construction, providing 60 ng and 123 ng respectively, and sequenced on 1 SP lane of the Illumina NovaSeq.

To first generate biotinylated DNA fragments enabling TSO-artefact depletion a total of 15ng from each sample was PCR-amplified with MAS capture primers that selectively bind to desired cDNA inserts. After 5 cycles of amplification and bead-purification a total of 372 ng and 365 ng cDNA PCR products were available to be carried into TSO-artefact depletion. TSO artefact removal was performed on both libraries leaving 56 ng and 45 ng per library. This difference in yield means ~86% of the cDNA library was discarded during TSO artefact removal, double the predicted ~40% estimates based on previous work. However, bead purification is known to result in some loss during clean-up steps, so as much as >80% of libraries consisted of artefact inserts.

The TSO-depleted samples were then arrayed across 16 PCR reactions per sample, and amplified with pre-mixed MAS array primer sets for a total of 9 cycles based on the cDNA input concentrations. This anneals the compatible sequences at 5' and 3' ends of cDNA between reactions that enable cDNA segment concatenation into linear arrays. After PCR reactions for each sample were pooled, combining each array into a single reaction. After bead-purification this gave a total of 13.2 ng and 13.5 ng of each cDNA sample which were used as input for downstream concatenation by MAS ligase. The final concatenated inserts were used to generate SMRTbell libraries of 716 ng and 842 ng and 11.3 kbp and 11.4 kbp respectively (Figure 5.12).

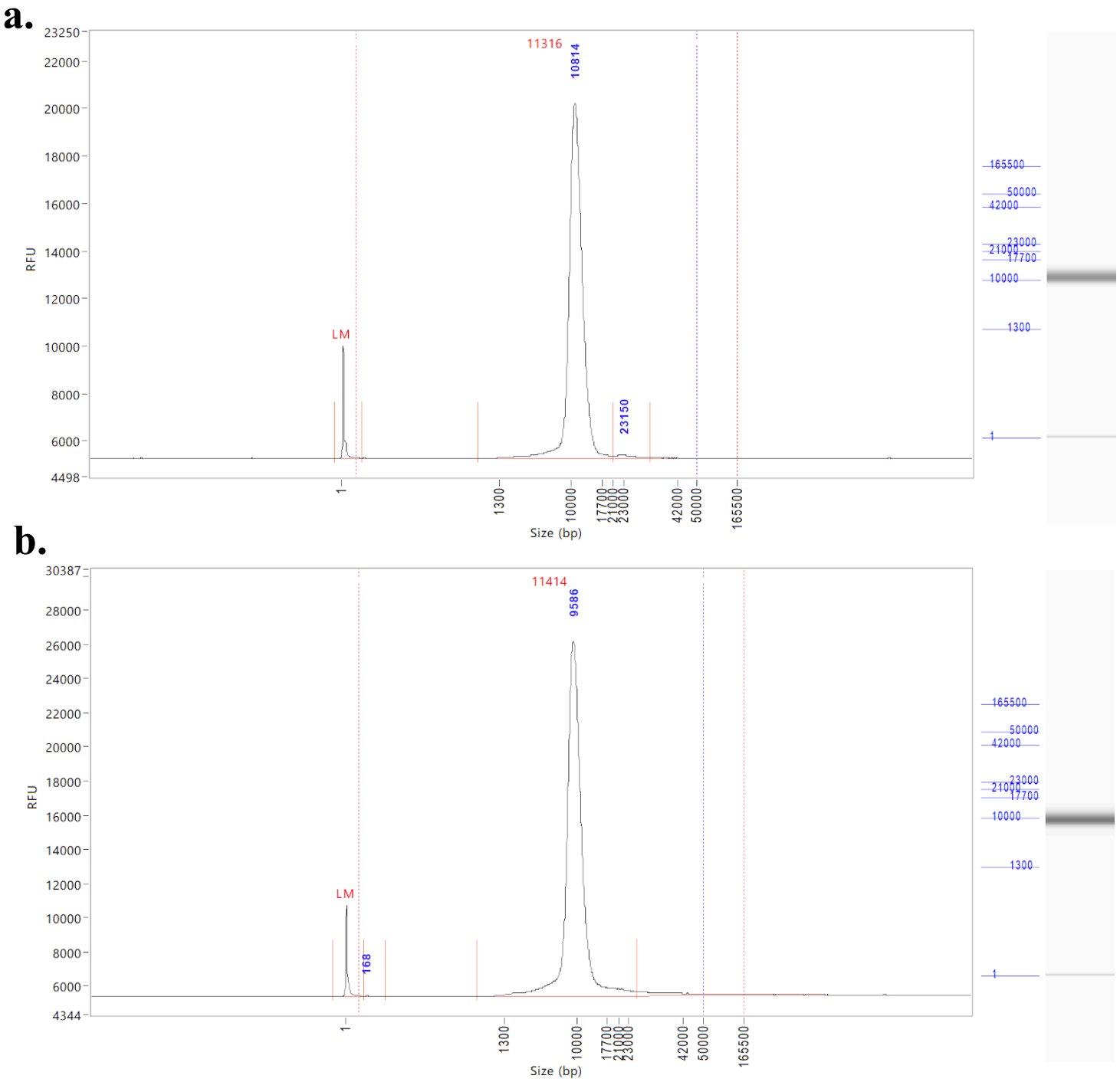


Figure 5.12. Library quality of single-cell library post MAS-seq concatenation of PBMC cDNA. Femto Pulse size distribution traces of concatenated SMRTbell libraries generated from (a) PBMC Library 1 (b) PBMC Library 2.

5.3.5.3. MAS-seq of FACS sorted mouse LK Cd150+ cells

10X Genomics offers an efficient high-throughput solution for scRNA-seq across various cell types. However, a limitation of this method is cell loss during sample loading and GEM generation. The capture rate of cells with the 10X HT3' v3.1 kit is 65%, requiring a minimum of 3000 cells per reaction, which can be challenging for rare cell types. For instance, it is estimated that mouse bone marrow comprises only 0.01% HSCs, and approximately 5000 can be isolated from an individual mouse (Challen *et al.*, 2009). FACS enrichment prior to loading is a popular approach to improve capture rates for rare cell types, whilst also selecting for live cells and removing debris. By enriching the population of interest, sensitivity and specificity can be increased. To test the compatibility of combining FACS population enrichment for 10X Genomics cDNA generation, cells were sorted into two wells of a 96-well plate containing 7 µl of FACS sorting media (see Buffers & Solutions 2.1.2: 3). Cells were gated using forward- and side-scatter for LKCd150 and LSK Cd150+ cells (Figure 5.13). A total of 3,600 and 800 LK Cd150+ and LSK Cd150 were sorted into wells yielding approximately 4,400 cells per sample at an approximate concentration of 600 cells per µl, which is the recommended optimal concentration to capture approximately 1000 cells. Cells were immediately loaded onto the 10X Chromium Next GEM Chip (L) and run on the Chromium Controller with the L programme.

cDNA generation yielded 90 ng and 145 ng of total mass for samples 1 and 2 respectively, with mean cDNA fragment lengths of 771 bp and 720 bp (Appendix Supplementary Figure 5.5). 25% of cDNA from each sample was used for Illumina library construction, yielding 60 nM and 65 nM respectively, and were sequenced on 1 lane of MiSeq v3 flow cell with 28-10-10-90 configuration.

TSO-artefact depletion and MAS-seq preparation of cDNA were performed as previously described using the recommended input of 15 ng per sample. The final MAS-seq library concentrations generated were 214 and 106 ng, and the mean library fragment lengths were 9271 bp and 7134 bp (Figure 5.14). Libraries were sequenced on 2 SMRT cells on the PacBio Sequel IIe (8M, v2) with 30hr movies.

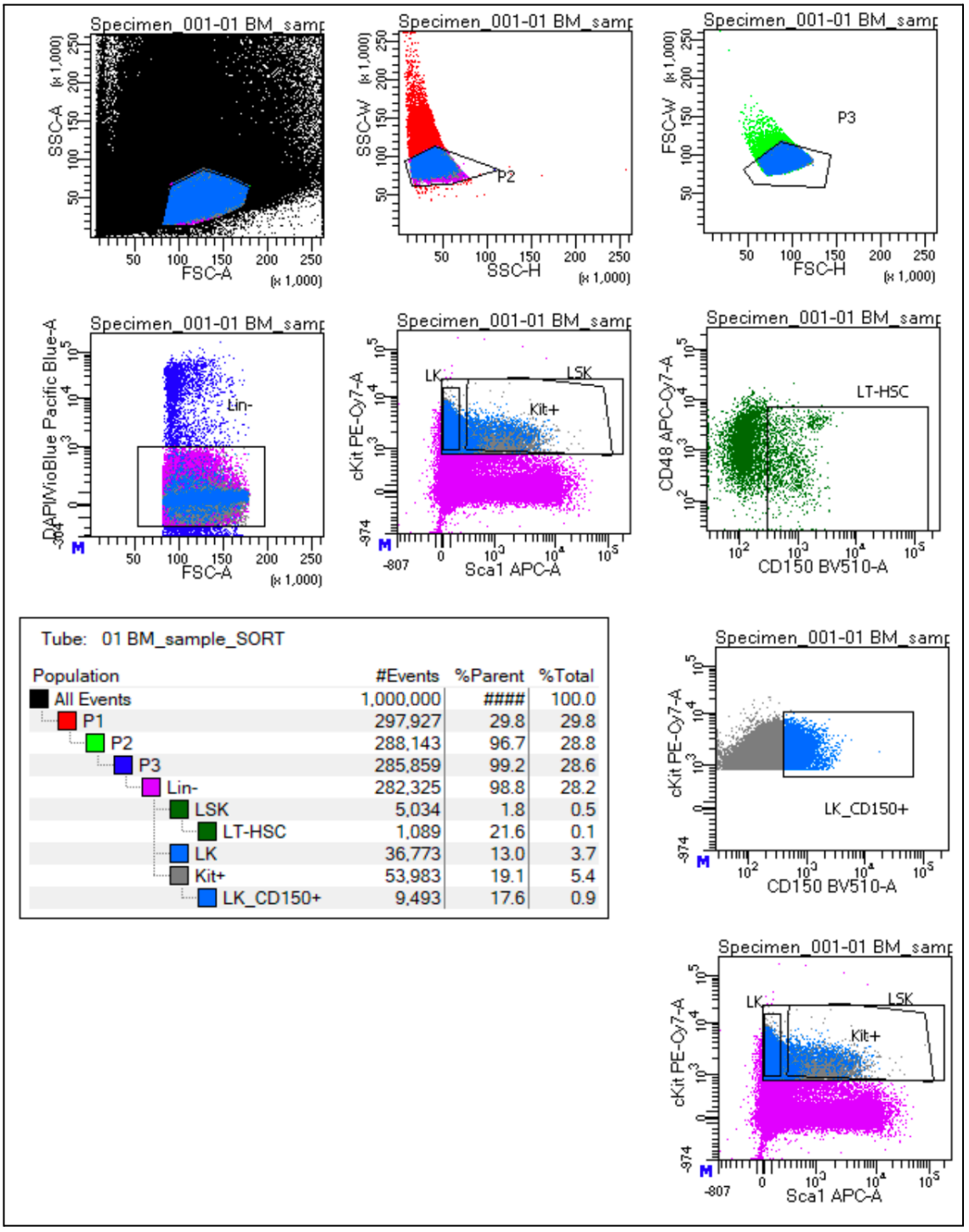
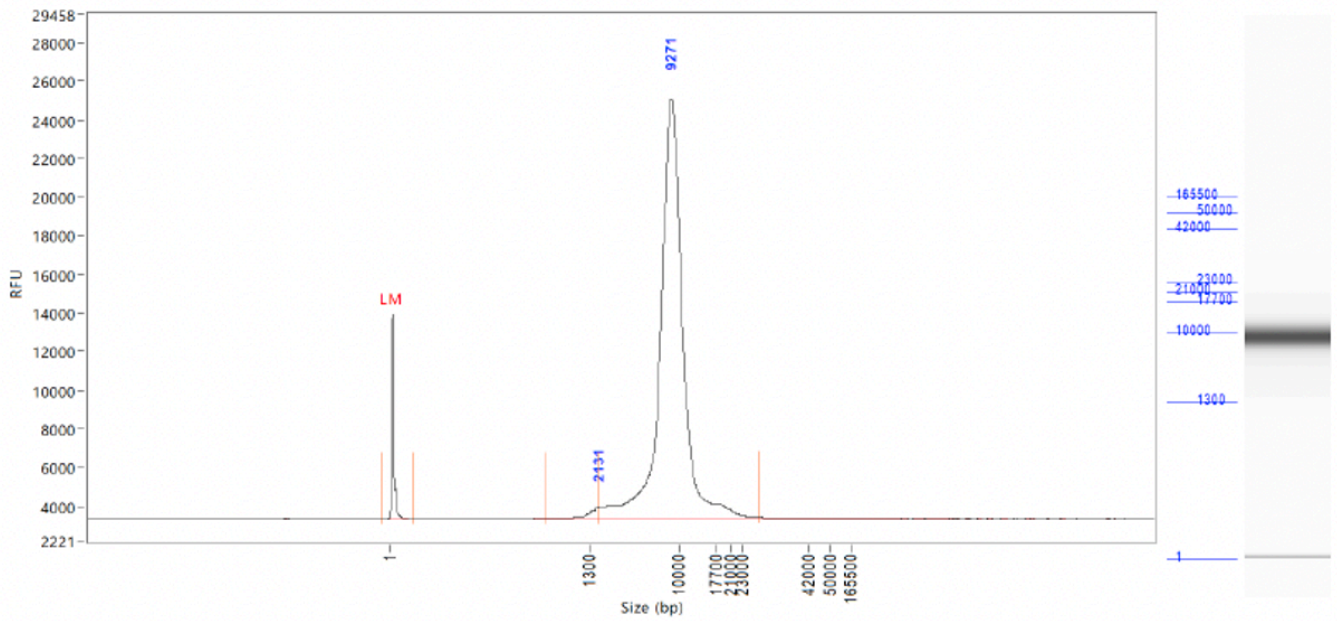


Figure 5.13. Gating strategy for FACS sorting mouse LK Cd150+ and LSK Cd150+ cells for 10X Genomics cDNA generation. Gate P1 set uses forward scatter (FSC) and side scatter (SSC) signals to exclude debris from the sample based on size and granularity. FSC singlet and SSC singlet gates (P2 and P3) are used to exclude doublets or aggregates of cells. Lin- isolates haematopoietic progenitors which are negative for lineage-specific markers. cKit (Cd117) and Sca-1 (Ly6a) are used to distinguish progenitors from immature progenitors and HSCs. Cd150 expression is used to enrich for HSCs and cells of the Mk lineage.

a.



b.

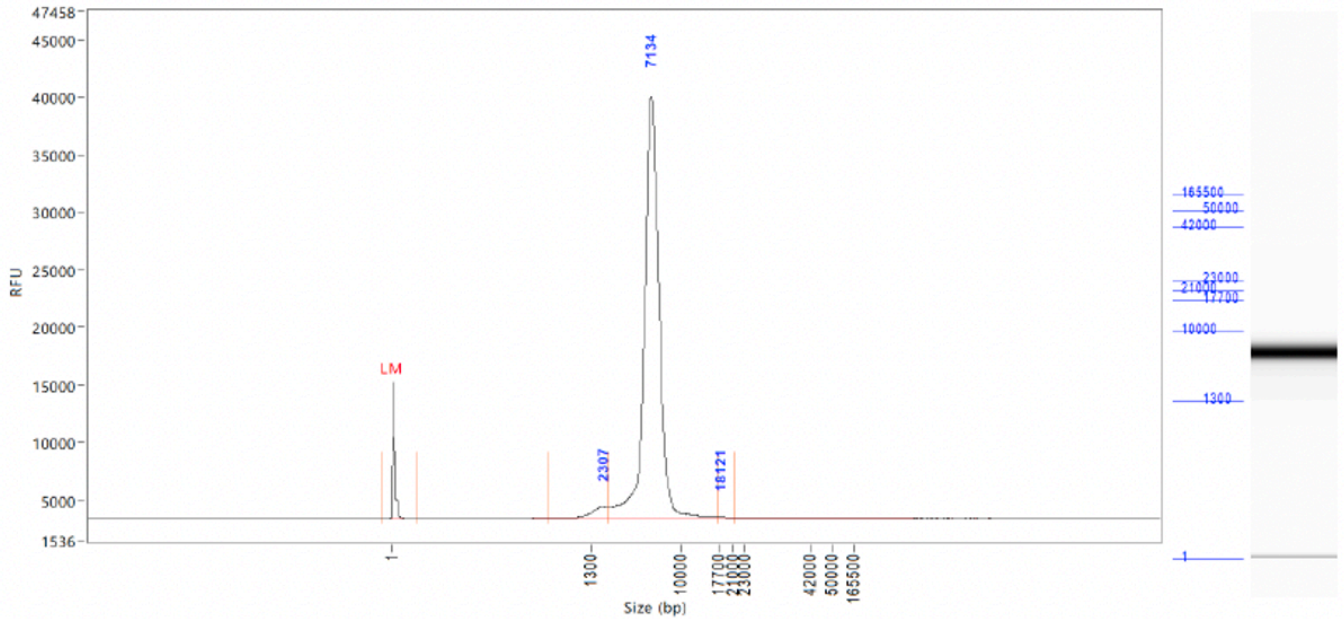


Figure 5.14. Library quality of cDNA libraries generated through MAS-seq concatenation of FACS sorted LK Cd150+ single-cells: Femto Pulse size distribution trace of final concatenated SMRTbell libraries generated from single-cells (a) Sample 1 (b) Sample 2.

5.3.6. scRNA-seq data processing of Illumina 10X Genomics data

Generating short-read single-cell libraries can be a useful complementary approach to long-read sequencing, particularly when testing new sequencing technologies from the same single cells; providing validation, complementarity, efficiency, and benchmarking advantages. For this reason Illumina libraries were generated for every experiment to provide validation and complementary information. The 10X genomics scRNA-seq data were first pre-processed using CellRanger, performing read alignment to the mouse/human genome, cell barcode demultiplexing, gene count quantification and preliminary QC.

5.3.6.1 Mouse whole bone-marrow Illumina sequencing

This experiment yielded a total of 1040 single cells, with sequencing coverage capturing an average of 1,528 genes from ~21K reads per cell. The optimal range in the number of cells captured using the LT 10X Genomics chemistry is >100 and <1000. Inspection of the distribution of UMIs and genes across the cells captured provided initial insights into the quality of the data and the number of cells that were captured during the experiment (Figure 5.15A). Most cells have a relatively high number of UMIs and genes, indicating that they were successfully captured and sequenced. To ensure any multiplets and low-quality cells captured were excluded from downstream analysis, only cells expressing between 200-4K genes with at least 1K reads and under 15% mitochondrial content were retained (Figure 5.15B and C). This left 825 cells available for downstream analysis.

The data was analysed using *Seurat* following the same workflow described previously (see Materials & Methods 2.3.5 - 2.3.8). The top genes with the highest standard variance across the data largely consisted of canonical markers of BM cell types (Figure 5.14D). No cell-type enrichment was performed during the preparation of mouse BM so it is unsurprising that genes with the highest standard variance are markers among the most abundant cells in BM. Dimensional reduction was performed using PCA with the first 10 PCs, Louvain clustering grouped cells in 9 clusters (Figure 5.14E). Clusters were then annotated using the *TabulaMurisData* package which uses pre-labelled reference scRNA-seq data from the Tabula Muris Consortium (Tabula Muris Consortium *et al.*, 2018). Using the "Marrow" dataset from droplet experiments as a reference in combination with cluster markers identified with the *Seurat FindMarkers()* function all clusters were assigned to unique cell types (Figure 5.14F).

The results from the analysis of this Illumina dataset represent a foundational reference of the cell types captured in this experiment and their viability to use as a benchmark to the PacBio sequencing results.

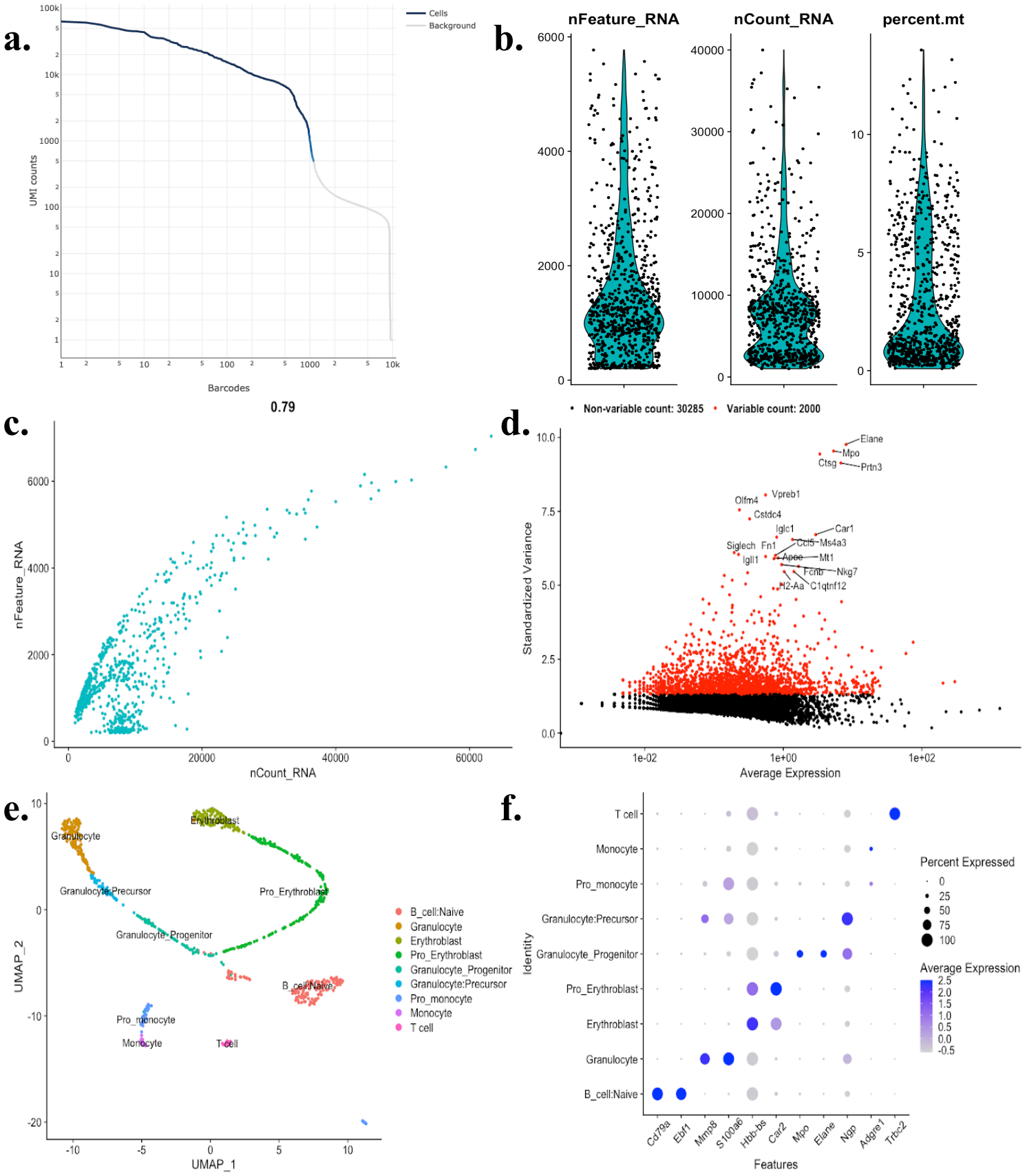


Figure 5.15. 10X Genomics scRNA-seq analysis of whole mouse BM from Illumina sequencing. (a) Distribution of the number of UMIs and the number of cells (barcodes) sequenced (b) Violin plots of the distribution of genes, reads and percentage of mitochondrial

content (c) Correlation between the number of reads (X-axis) and genes captured (Y-axis) per cell from Illumina sequencing (d) Standard variance of genes across the dataset, with top 2000 genes in red and top 20 genes annotated (e) UMAP projection of single cell clusters annotated by cell-type (f) Dot plot of marker expression across cell clusters (rows), with point size indicating the percentage of cells expressing genes and colour opacity showing expression level.

5.3.6.2. Human PBMC 10X Genomics Illumina

The PBMC Illumina libraries were sequenced on 1 SP lane of the NovaSeq generating on average 227K total reads. Read alignment was performed in CellRanger, where ~95% were mapped confidently to the human genome. CellRanger statistics calculated an estimated total of 4,875 single cells, with a median 1,530 genes from a mean 46,961 reads per cell. On the other hand, sample 2 captured 7,384 cells, 1,562 genes and a median 32,174 reads per cell. Data QC was performed to select high-quality cells by removing cells with >15% of reads mapping to mitochondrial genes. In addition, the lower and upper bounds for the number of genes per cell were calculated based on the 0.01 (lower boundary) and 0.99 (higher boundary) percentiles of the distribution of all genes in each experiment. Cells were only retained if they expressed more genes than the lower boundary and less than the upper boundary. This left 4,506 and 7,061 cells for downstream analysis from samples 1 and 2 accordingly. Removing the poor-quality samples increased the mean number of genes per cell to 1638 and 1841 respectively and improved the correlation between read depth and genes detected due to higher consistency across the retained cells (Figure 5.16A and B).

The two datasets were normalised and then integrated using *IntegrateData()*. Data integration was accomplished by identifying cross-dataset anchors in the top 2000 most variable features across both datasets. This was critical to enable the datasets to be merged, as sample identity was identified as a dominating batch effect of downstream analysis and clustering (Figure 5.16C and D). After data integration, scaling and PCA were conducted to determine the number of components. Dimensionality reduction was used to reduce the dataset into the first 20 PCs. Cells were grouped into clusters as previously described and annotated by cell type using the *SingleR()* package (Aran *et al.*, 2019). This matched the expression profile of each individual cell to the corresponding cell-type using the *HumanPrimaryCellAtlasData()* reference dataset (Regev *et al.*, 2017). The function outputs a matrix of scores indicating the similarity of each cell to every cell-type in the reference dataset. Then the cell-type with the best similarity score based on its average expression profile is assigned to each cell. Overall, this method assigned 17 cell types to the dataset, consisting of primarily T cells, B cells, natural killer (NK) cells and monocytes (Figure 5.16E); found at highly consistent proportions across the two samples (Figure 5.16F).

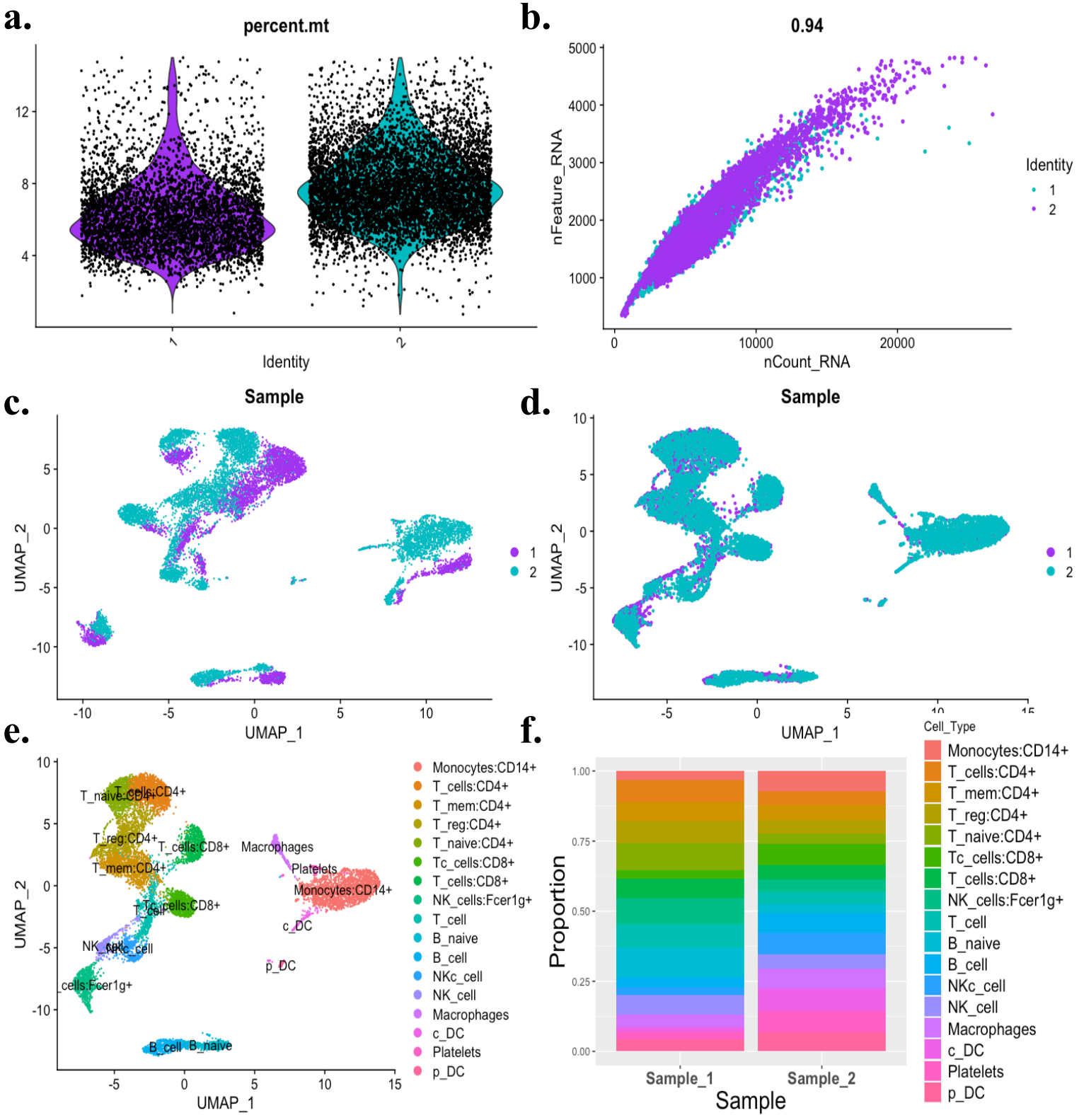


Figure 5.16. 10X Genomics PBMC analysis from Illumina sequencing. (a) Violin plots of the distribution of the percentage of mitochondrial content in cells post-QC separated by sample (b) Correlation between the number of reads and genes captured per cell coloured by sample ID (c) UMAP of single cells coloured by sample prior to dataset integration (d) UMAP of single cells coloured by sample post dataset integration (e) UMAP projection of single cell clusters annotated by cell-type (f) Stacked bar plot of the proportion of cells across cluster populations separated by sample ID.

5.3.6.3. Mouse FACS sorted LK Cd150+ 10X Genomics Illumina

The libraries generated with the 10X Genomics 3' LT chemistry from FACS sorted LK Cd150+ mouse BM cells were sequenced on 1 lane of the MiSeq v3 flow cell. The total number of read pairs that were assigned to each library in demultiplexing was 3.6M and 17M and 80% mapped with high confidence to the mouse genome. However, data pre-processing and QC in CellRanger revealed both samples had low fractions of valid-barcodes, which are confidently-mapped-to-transcriptome reads with cell-associated barcodes (55.8% and 65.7%). The estimated number of cells captured in each experiment was unexpectedly low for both libraries, capturing 23 and 50 cells in each sample. This was surprising, given the concentration of cDNA obtained from each library was not unusually low (90 ng and 145 ng), and the final yield and size distributions for both libraries were comfortably within the optimal range (see Appendix Supplementary Figure 5.5). FACS reports from the experiment confirmed ~4,400 cells were sorted for each sample, with no indications of poor quality cDNA or issues with library preparation; it suggests a high number of cells may have been lost in the transfer from wells of the 96-well plate to the chip during loading.

To assess whether the cells recovered were viable for downstream analysis, the datasets for each sample were processed in *Seurat* as previously described. Notably, of the 73 captured in total only 3 cells were excluded during QC filtering based on the percentage of mitochondrial content exceeding 15%, with 70 cells exhibiting a viable signature (Figure 5.17A). On average 3430 genes were detected per cell, from ~18K counts per cell and a strong correlation between the number of features and counts for both samples (Figure 5.17B and D). Inspection of the genes with the highest variance across the data revealed expression of genes canonical to cell types within the LK Cd150+ compartment, including Mk-marker *Pf4,* Ery-marker *Car1* (Figure 5.17C). This confirms that despite the low cell recovery of the experiment some viable LK Cd150+ cells were captured.

Due to the size of the dataset the ability to interpret downstream analyses was limited. Small datasets are more susceptible to noise and prone to overfitting which obscures the underlying patterns in the data and can result in the identification of artificial clusters. Dimensionality reduction was performed using UMAP to project cells into a latent space, from which two major clusters were distinguished. These two clusters were largely driven by a difference in sequencing coverage whilst cell-type specific signatures can be seen across cells in clusters 1-3 (Figure 5.17E and F).

These results revealed a low recovery of cells, suggesting that further protocol development is required to refine any future 10X Genomics experiments from FACS-sorted cell populations.

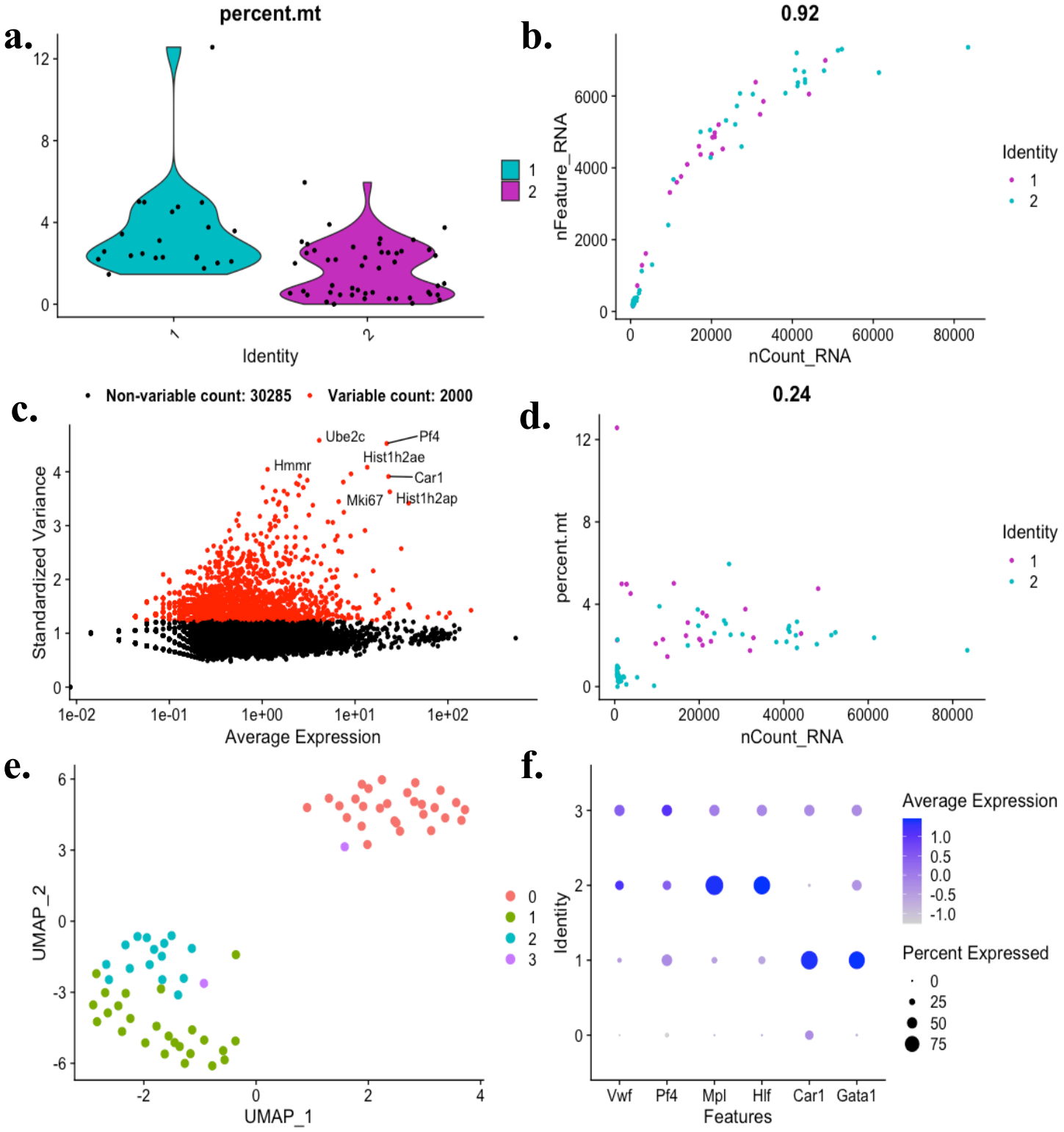


Figure 5.17. 10X Genomics short-read analysis of FACS sorted LK Cd150+ single-cells. (a) Violin plots of the distribution of the percentage of mitochondrial content in cells post-QC separated by sample (b) Correlation between the number of reads and genes captured per cell coloured by sample (c) Standard variance of genes across the dataset, the 2000 genes with highest standard variance (red) (d) Correlation between the number of reads and mitochondrial content per cell coloured by sample (e) UMAP projection of single-cells coloured by cluster identity (f) Dot plot of Mk and myelo-erythroid gene expression across cell clusters (rows), where point size corresponds to the percentage of cells with expression detected across genes (columns), and colour opacity signifies expression level.

Table 5.1. Experiment summary of all 10X Genomics Illumina scRNA-seq libraries.

Metric	HIT-scIsoSeq	MAS IsoSeq PBMC	MAS IsoSeq LK cd150
Number of cells captured	1,040	12,259	73
Total number of reads	22,511,462	232,596,896*	13,772,386
Fraction of reads in cells	90.1%	97.9%*	65.7%
Number of cells after QC	825	11,567	70
Mean counts per cell after QC	8,529	5,884**	18,566
Mean features per cell after QC	1,528	1,775**	3,383

* Average across libraries
** After data integration

5.3.7. Exploring single cell PacBio sequencing data generated from two different cDNA concatenation approaches

2.3.7.1 HITsc-IsoSeq PacBio sequencing of mouse BM single cells

A single cell IsoSeq library generated using an adapted version of the HIT scIsoSeq protocol and was sequenced with PacBio on one SMRT cell of the Sequel II instrument (8M v2) with a 30hr movie and 2-hour pre-extension. The SMRT Link Circular Consensus Sequencing (CCS) pipeline was used to obtain highly accurate consensus reads from multiple passes of DNA molecules, generating CCS reads. Briefly, raw sequencing data is processed using PacBio's base caller which assigns a quality score to each base of the sequencing read. A consensus read is generated by circularising the subreads and using multiple passes to generate a consensus sequence, followed by filtering of CCS reads based on quality metrics such as read length, number of passes, and quality score. CCS Reads with a quality value equal to or greater than 20 are classified as HiFi Reads, with a raw accuracy of up to 99.99% after error correction and polishing. This library yielded a total 189,339 HiFi reads, with a median read quality (Q) of Q34 at an average of 5.5 kb read length from ~19 passes per molecule on average (Figure 5.16A and B).

The read-length distribution observed is largely consistent with the size distribution recorded for the library prior to sequencing (Figure 5.11), and most reads had a predicted accuracy well above the Q threshold for HiFi quality (Figure 5.18A). While the reads obtained were of high quality, the library overall displayed signs of low loading efficiency, with only 20% of polymerase-template complexes active during sequencing. This indicates that fewer polymerases successfully bound to the template resulting in lower sequencing output, reflected by the lower number of reads obtained than anticipated (target yield was 3M molecules). The mean polymerase read length was also shorter at 32.6 kb, indicating the polymerase enzyme was not able to read through the entire template during sequencing. This could be due to several factors, such as low template concentration or technical issues during sequencing.

Altogether the sequencing metrics indicated the HIT sc-IsoSeq library was underloaded, which caused low sequencing efficiency. However, sufficient levels of high-quality data were generated to allow demultiplexing of 10X barcodes, enabling the assignment of reads to single cells and further downstream analysis. Segmentation of concatemers into cDNA reads was performed using *Longbow*; a demultiplexing command-line tool as part of the IsoSeq analysis suite. *Longbow* employs a generative modelling approach to accurately annotate adapter and transcript boundaries and in turn segments annotated reads for use with other downstream tools. This revealed the distribution of concatemers per molecule, which varied from 1 (ie. non-concatenated) to >10 (Figure 5.18C). The average concatenation achieved was 4.7-fold.

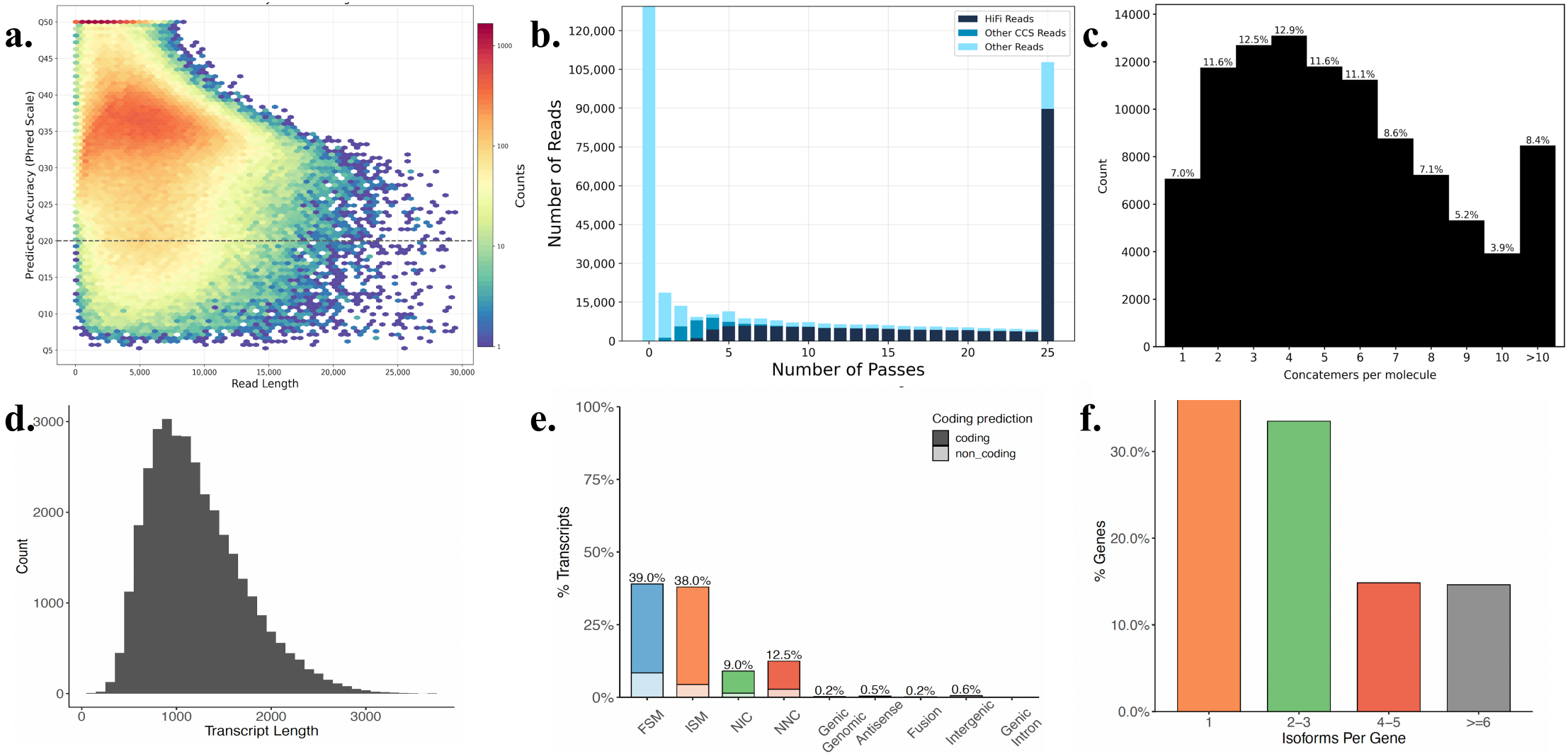


Figure 5.18. HIT sc-IsoSeq long-read sequencing of concatenated cDNA from mouse bone marrow. (a) Heat map of CCS Read lengths and predicted accuracy from PacBio sequencing (b) Distribution of HiFi Reads and other CCS Reads by the number of passes (c) Histogram the counts distribution across concatemers per molecule (d) Distribution of transcript lengths after read segmentation (e) Percentage of transcripts across SQANTI3 isoform structural categories (f) Number of isoforms detected per gene.

Primers and polyA tails were used to orient the read into 5' – 3' orientation and then trimmed from the final transcript reads. This was followed by extraction of single-cell barcode and UMI information. Reads were then mapped to the mouse genome and classified against the *GENCODE* transcript annotation (Frankish *et al.*, 2020). *SQANTI3* was applied following the previously described method and an isoform-level single-cell matrix was generated which served as the output for downstream analysis in *Seurat*.

After deconcatenation, the average transcript length was approximately 1.2 kb (Figure 5.18D). In the HITsc IsoSeq data, a total of 10,664 unique genes and 33,963 unique isoforms were identified by *SQANTI3*. These isoforms were classified into various structural categories, as illustrated in Figure 5.16E. The majority of genes (over 35%) were found to have a single isoform. Notable proportions of genes with multiple isoforms were also detected across the single cell dataset, indicating additional levels of transcript diversity and highlighting the heterogeneity in isoform expression across genes (Figure 5.18F).

Among the isoforms, the largest structural category observed in this dataset consisted of FSMs, with over 13,000 transcripts matching all of their reference sequences in the mouse annotation. This shows a significant number of isoforms that align well with known references were captured. Moreover, the analysis revealed a comprehensive coverage of splice junctions, with approximately 75,000 junctions detected. Of these junctions 8% were classified as novel, indicating the presence of previously unreported splice junctions within the dataset.

Seurat analysis of the isoform expression matrix was performed following the standard analytical workflow to see how this data translated at the single cell level. 825 single cells were retained after QC, filtering out cells which had mitochondrial content of >10% and a read count of < 100. The average number of isoforms detected per cell was 200, with a strong correlation between the number of isoforms and reads detected across cells (Figure 5.19A). Multiple isoforms of the same genes were identified within the first 4 components of PCA, indicating that isoform diversity contributes towards driving cell variability (Figure 5.19B). Single cell Louvain clustering was performed identifying 6 populations of cells based on isoform-level signatures. Using the *TabulaMuris()* reference dataset of mouse BM data as performed in the Illumina single cell data analysis, cell cluster annotation was achieved and found to be consistent with the short-read data (Figure 5.19C).

Altogether these results demonstrate HITsc IsoSeq is a viable strategy for long-read sequencing from single cells and produces isoform resolved single-cell data. Higher throughput sequencing from single cells was achieved with an effective concatenation factor of 4.7, with concatemer molecules of variable lengths being formed.

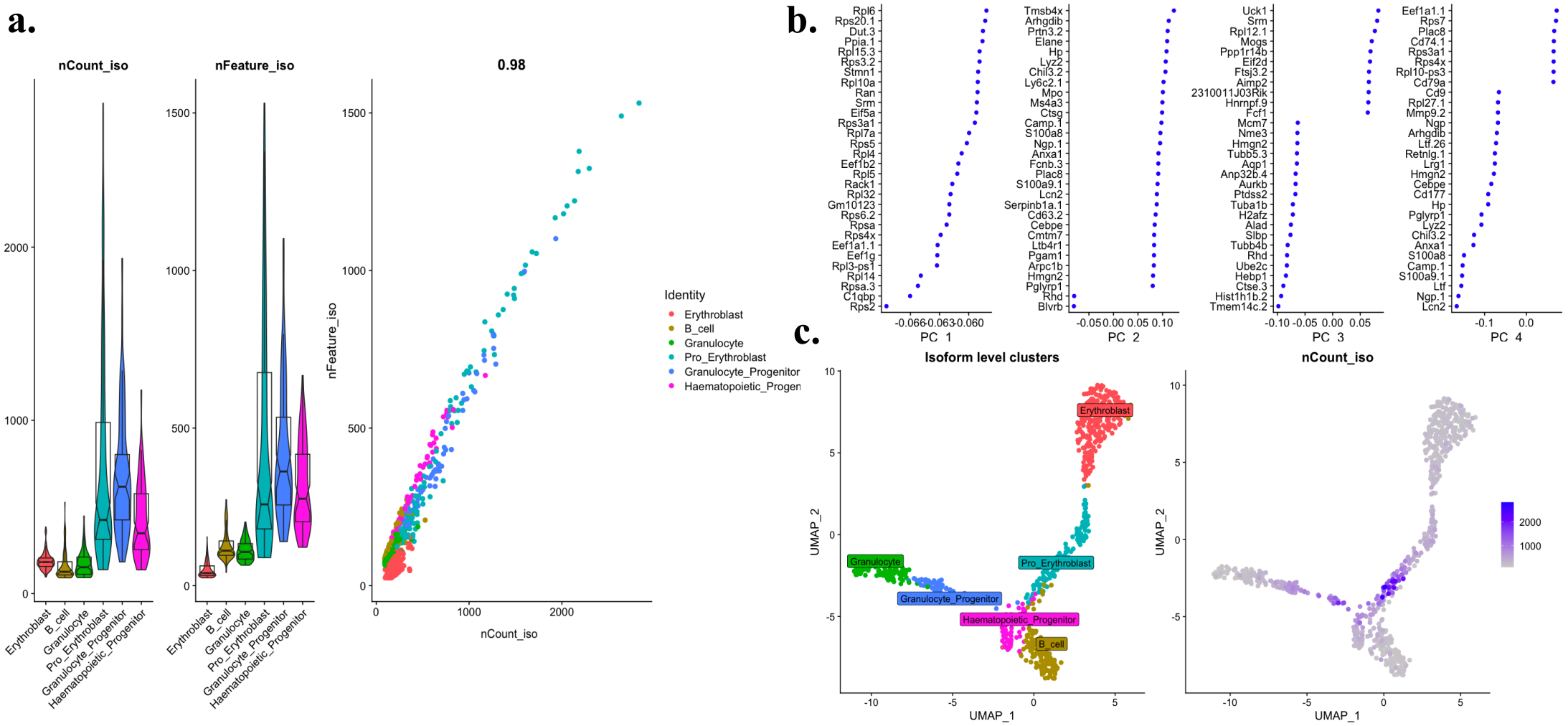


Figure 5.19. HITsc-IsoSeq analysis of mouse BM single cells. (a) Violin plots of distribution of single cell isoform counts and isoform-level features across cell clusters, and the correlation between isoform features and isoform counts (b) PC loadings of top 4 PCs from single cell data show multiple isoforms of genes contribute to first PCs driving the variability among the cells (c) UMAP projections of isoform-resolved single cell RNAseq data coloured by cell type and distribution of count levels in single cells.

5.3.7.2. MAS-seq from human PBMCs

To assess the quality of sequencing data from single-cell MAS-seq PBMC libraries, sequencing data were processed using the SMRT Link software to obtain initial statistics of the raw sequencing results.

The sequencing efficiency of the two libraries differed greatly. Library 1 generated 1.9 million HiFi reads with a mean length of 116 kb, while library 2 only generated 227,300 HiFi reads with a mean length of 82 kb, resulting in an almost 9-fold difference in the HiFi read yield. Although both libraries had an equal number of productive zero-mode waveguides (ZMW) which contained the DNA polymerase enzyme and template DNA during sequencing, only 9% of ZMW in library 2 detected high-quality reads compared to 50% in library 1. This indicates that there may be underlying differences in library quality or characteristics that were not apparent during library quality control prior to sequencing. This initial QC of PBMC MAS-seq libraries revealed a noticeable discrepancy in sequencing efficiency despite being prepared and sequenced under the same conditions. Illumina sequencing data generated from the cDNA used for library preparation did not indicate any difference in cell viability between the two samples. Based on this and the difference in high-quality read detection in ZMWs, the difference in sequencing efficiency is more likely attributable to differences in library loading.

In terms of HiFi read length, sample 1 had an average of 12.5 kb while sample 2 had an average of 11.5 kb, both libraries had a median accuracy prediction score of 34 (Figures 5.18A and 5.19A). The difference in read length can be attributed to the number of passes across reads, with library 2 having on average 18 passes compared to 16 in library 1 (Figures 5.20C and 5.21C).

Each multiplexed array was split into individual sequence segments (S reads) corresponding to their original cDNA fragments. The segmentation of a total of 1.9 million and 227,000 reads from samples 1 and 2, yielded 29.3 and 3.4 million S reads, of ~750 bp average read lengths. Concatenation of MAS-seq libraries was highly consistent, with between 82% - 85% of reads with full MAS arrays and a concatenation factor of 15 concatemers per molecule (Figures 5.20B, D and 5.21B, D), demonstrating the effectiveness of targeted array formation.

Demultiplexing cell barcodes from the PacBio S reads estimated 4773 and 7277 cells captured in each experiment, consistent with the number of cells estimated from Illumina libraries. Of all detected reads, 95% were assigned to single cells, giving a total of 5495 reads per cell in PBMC sample 1 and only 434 in sample 2. This disparity between the two libraries is also evident by the median number of UMIs transcripts per cell, or the number of distinct RNA molecules captured and sequenced from each cell, 3788 and 324 for library 1 and 2 respectively.

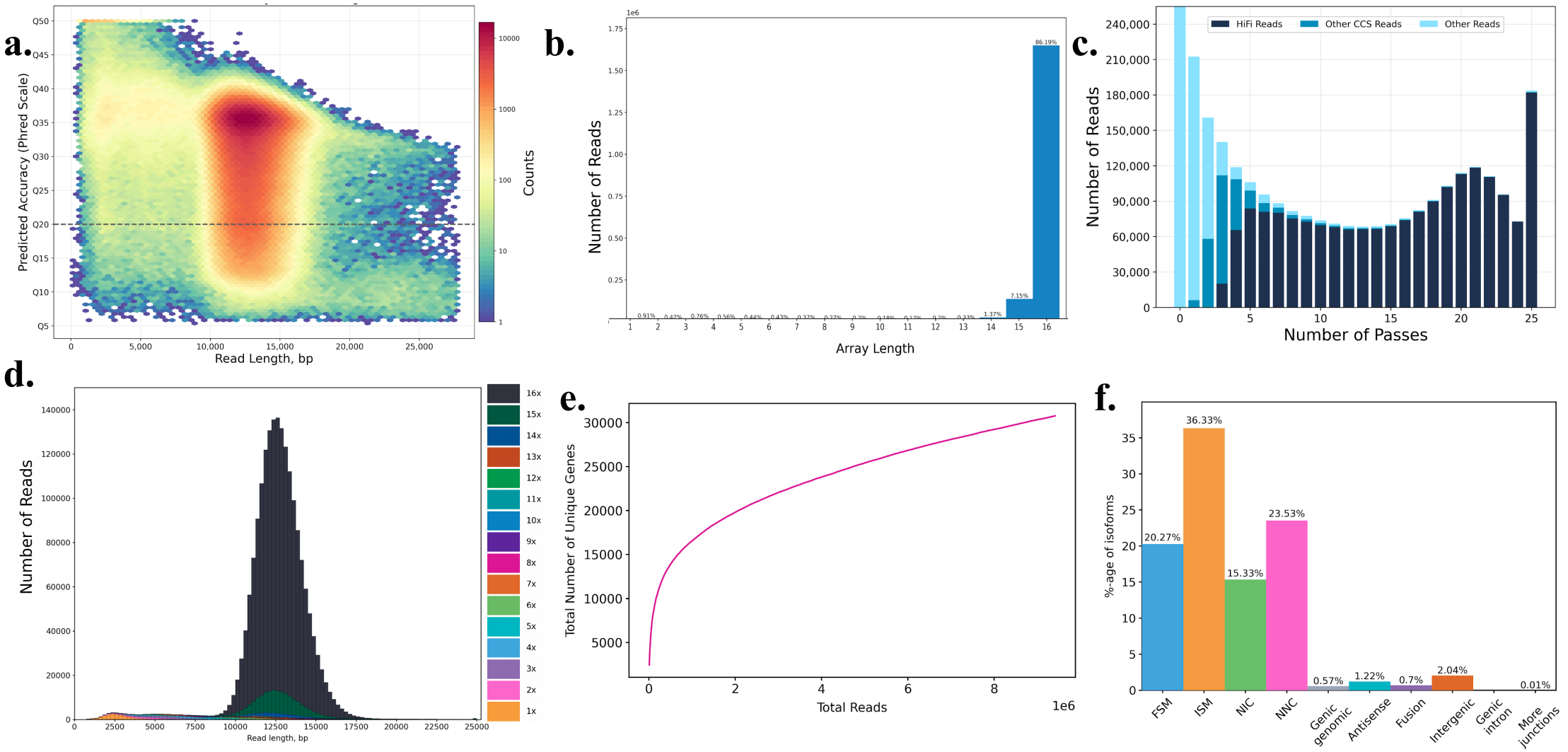


Figure 5.20. MAS-seq long-read sequencing of concatenated cDNA from PBMC sample 1. (a) Heat map of CCS Read lengths vs predicted accuracies (Q scores) (b) Number of reads across MAS array lengths (c) Distribution of HiFi Reads and other CCS Reads by number of passes (d) Number of reads against concatemer read lengths (e) Gene saturation plot: Total number of unique genes detected across reads sequenced (f) Percentage of transcripts across SQANTI3 isoform structural categories.

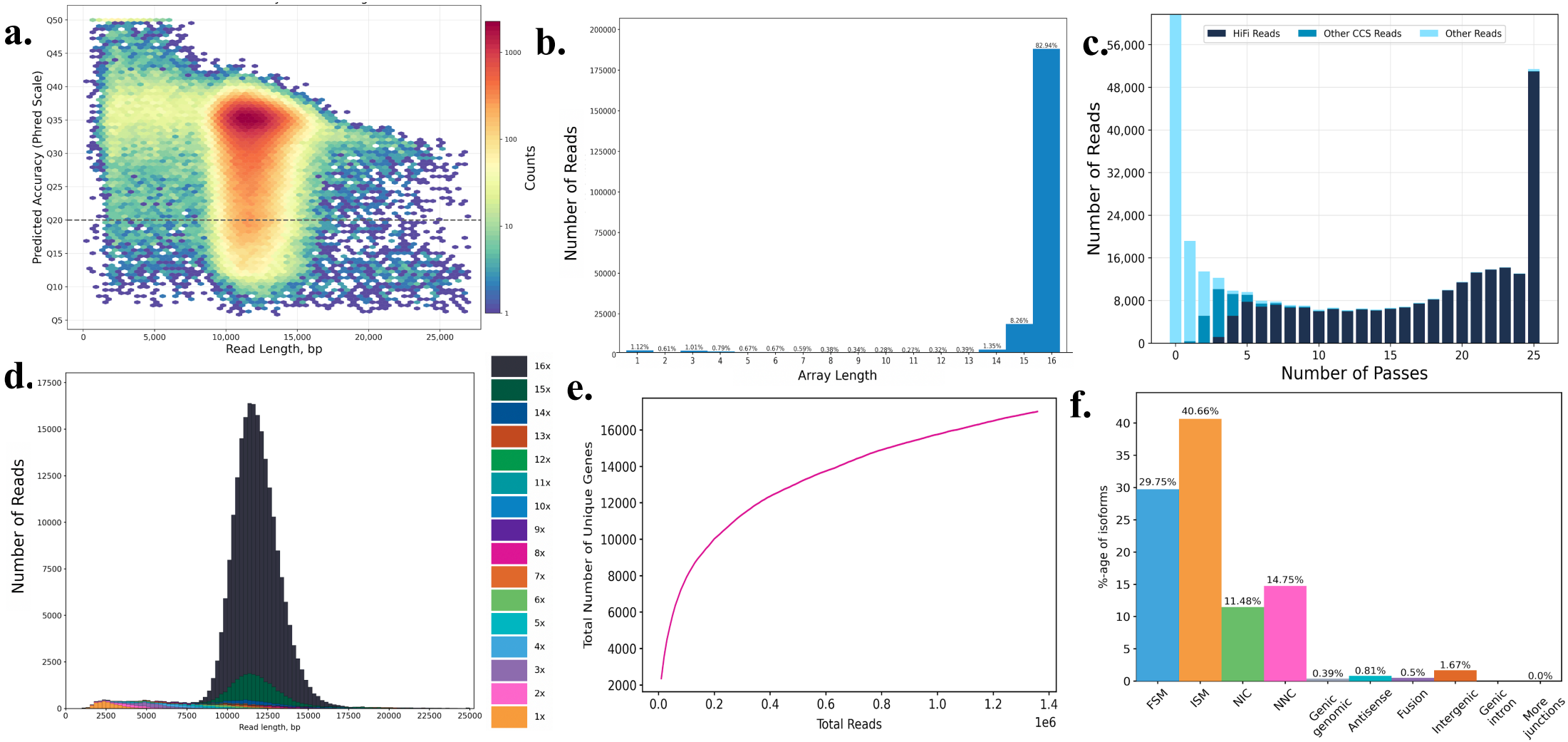


Figure 5.21. MAS-seq long-read sequencing of concatenated cDNA from PBMC sample 2. (a) Heat map of CCS Read lengths vs predicted accuracies (Q scores) (b) Number of reads across MAS array lengths (c) Distribution of HiFi Reads and other CCS Reads by number of passes (d) Number of reads against concatemer read lengths (e) Gene saturation plot: Total number of unique genes detected across reads sequenced (f) Percentage of transcripts across SQANTI3 isoform structural categories.

Gene saturation plots were used to visualise the relationship between sequencing depth and the number of unique genes detected for both libraries. Where the plot shows a plateau it indicates adding further reads gives diminishing returns with regard to capturing additional unique genes. The total number of unique genes was 30,750 in sample 1 and 17,014 in sample 2 (Figures 5.20E and 5.21E). The difference was greater at the transcript level, where the total number of known unique transcripts detected in sample 1 was 331,684 and only 80,095 in sample 2.

SQANTI3 transcript analysis was performed as previously described for each library, producing isoform annotations of the transcripts captured within the data (Tardaguila *et al.*, 2018). ISMs were the most prevalent structural category isoform identified for both PBMC libraries followed by FSM isoforms (Figures 5.20F and 5.21F); showing that a high proportion of detected isoforms were well supported by existing annotations.

To perform downstream expression analyses across single cells, isoform-level and gene-level expression matrices for each PBMC dataset were processed using *Seurat.* This created single objects for each PBMC library but was split into two assays with gene-level data as the "RNA" assay and isoform-level expression data stored within the "Iso" assay. Single cell QC was performed for each assay individually, excluding cells which did not express the number of genes within the lower threshold (10th percentile of the number of features expressed) and upper threshold (99th percentile of the number of features expressed) calculated for each object. This retained a total of 4242 cells in PBMC sample 1 and 6528 cells in sample 2 for downstream analysis. The average number of counts across each of the gene-level datasets were 1099 and 102 counts per cell, the average counts per cell at the isoform-level were 1017 and 114 for PBMC 1 and 2 respectively (Figures 5.22A and 5.23A). This drastic difference seen in the number of counts per cell was equally observed in feature detection, and was expected based on the sequencing data QC. Integration of the two datasets was attempted, however even after the integration of the dataset, the sample identity was a major batch effect, dominating the analysis and cluster identification (Appendix Supplementary Figure 5.8). For this reason, each PBMC library was analysed independently.

Following the standard *Seurat* workflow previously described, the data were normalised, scaled, and reduced into fewer dimensions using the top most variable features for each dataset. Single cell clustering at the gene- and isoform-levels was performed for each dataset by setting the appropriate assay as input for clustering (Figures 5.22B and 5.23B). Isoform-level clustering recapitulated clusters found at the gene level for both datasets.

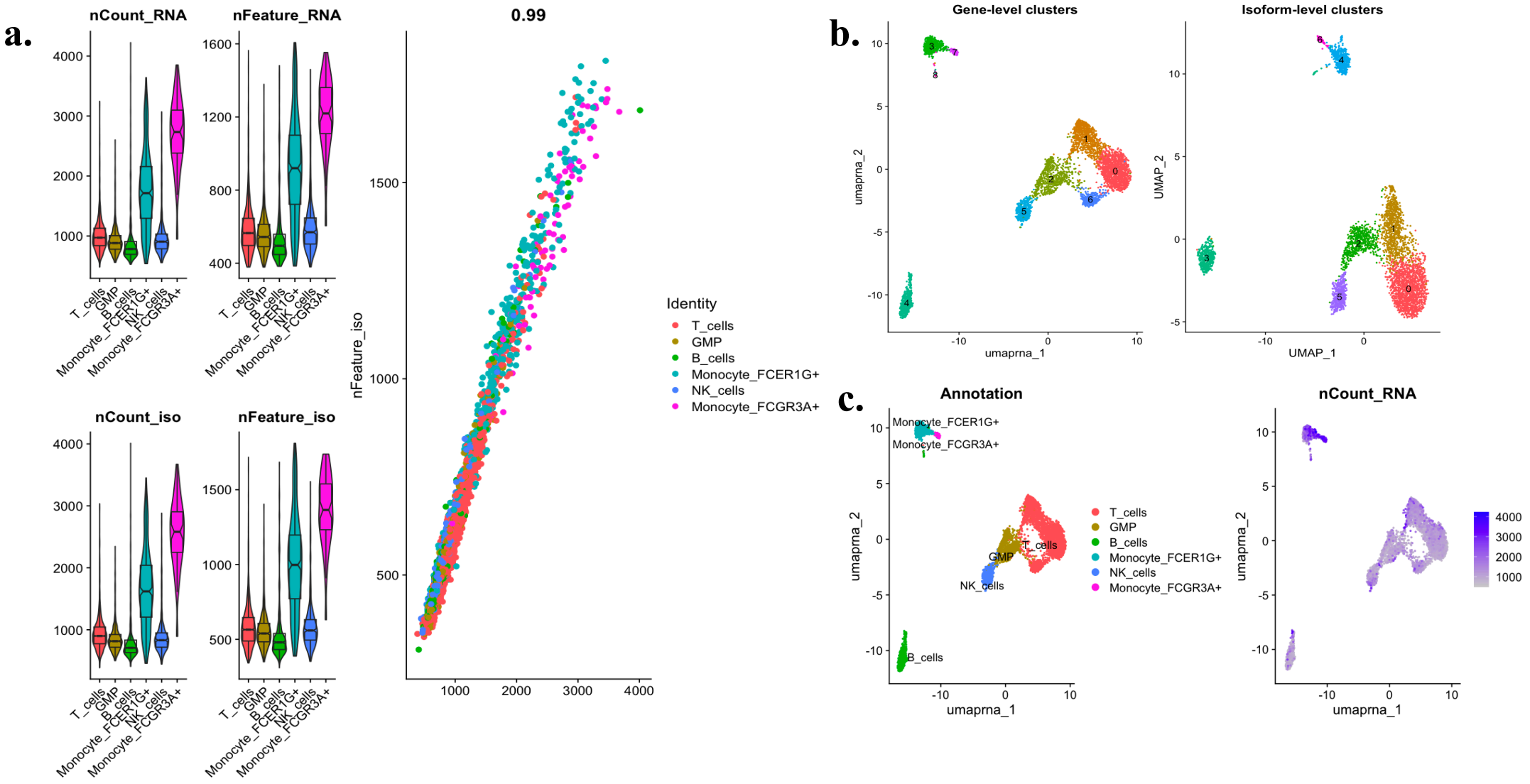


Figure 5.22. MAS-seq analysis of PBMC sample 1. (a) Violin plots of the distribution of single-cell gene and isoform counts and features across cell clusters, and the correlation between isoform features and isoform counts (b) UMAP projections of gene- and isoform-level clusters coloured by cluster identity (c) UMAP projection annotated by cell type annotations and distribution of count levels in single cells.

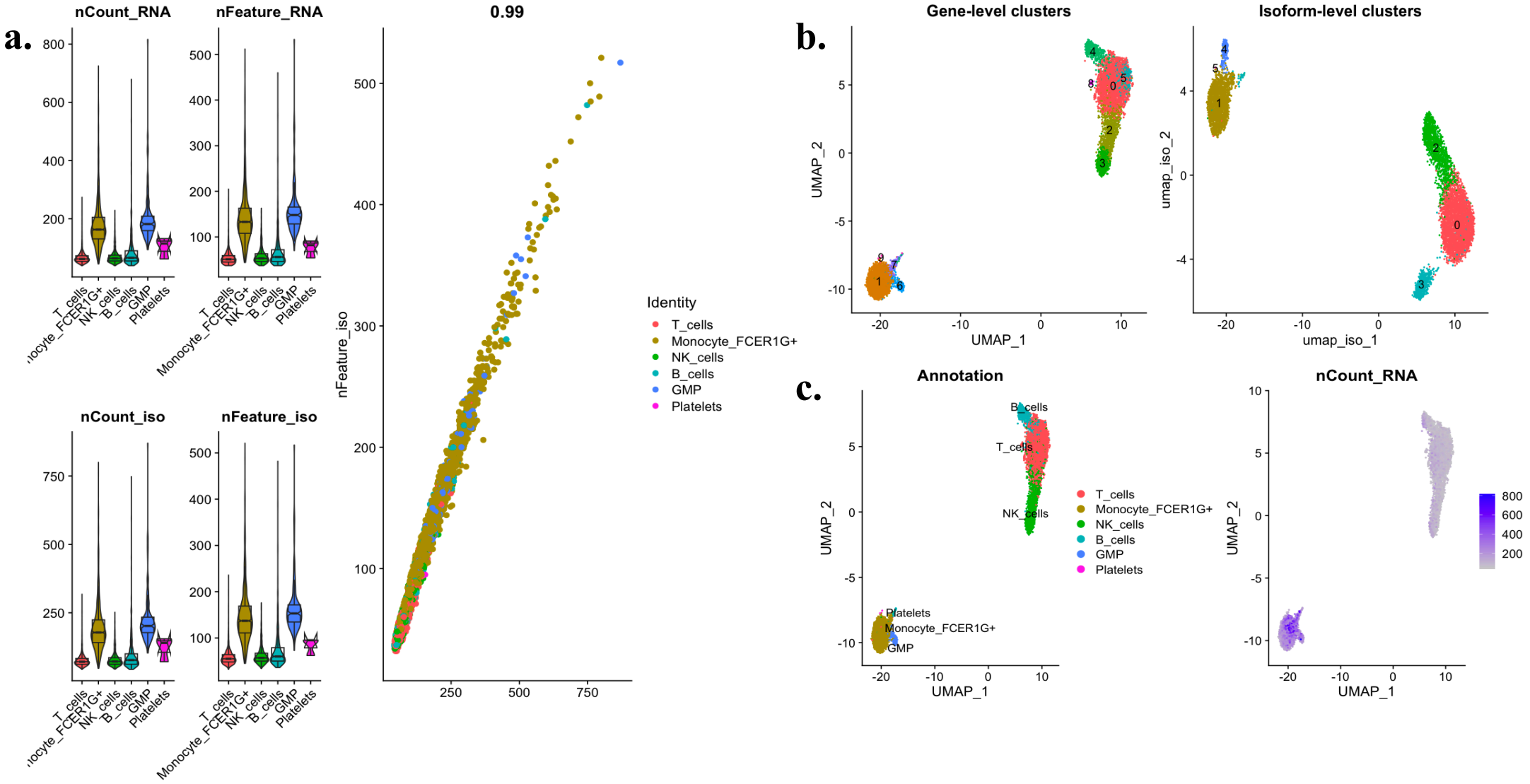


Figure 5.23. **MAS-seq analysis of PBMC sample 2.** (a) Violin plots of the distribution of single cell gene and isoform counts and features across cell clusters, and the correlation between isoform features and isoform counts (b) UMAP projections of gene- and isoform-level clusters coloured by cluster identity (c) UMAP projection annotated by cell type annotations and distribution of count levels in single cells.

Clusters were annotated into cell types using the same approach as implemented for short-read data analysis, using the *HumanCellAtlas()* BM reference dataset to estimate cell type ontologies. This identified a total of 6 cell-types in each PBMC dataset, and corroborated the populations identified in the short-read data of the same libraries (Figures 5.22C and 5.23C).

Notably, single cell clustering was highly influenced by the distribution of counts expressed by cells, with cells containing the highest counts being grouped into single populations. This batch effect was apparent across both samples, but was particularly evident in PBMC sample 2, underscoring the significant impact sequencing coverage has on single cell cluster identification.

Overall, these data show the output of MAS-seq library construction and sequencing of human PBMCs, enabling gene- and isoform-level analysis of single cells.

5.3.7.3. MAS-seq from FACS sorted LK Cd150+ cells

Sequencing of PacBio MAS-seq libraries and Illumina 10X libraries from FACS sorted LK and LSK Cd150+ cells was performed in parallel. PacBio libraries generated an average total of 5.8 million polymerase reads in each sample, yielding between 1.9-2 million HiFi reads from an average of 15 HiFi passes. This output is consistent with metrics observed from PBMC library 1, containing ~4773 cells while these libraries were expected to contain ~1K cells (Figures 5.24A-B and 5.25A-B). The polymerase reads translated to a total 28,865,911 and 32,071,873 S-reads across the two samples, which were on average 617 bp and 386 bp in length respectively. This difference in S read length corresponds to shorter concatemer units in sample 2 compared to sample 1, which corroborates the overall shorter MAS array observed prior to (Figure 5.14) and post sequencing (Figures 5.24A and 5.25A).

As observed for both PBMC libraries, MAS-seq generated concatenated libraries with a narrow distribution in the number of concatemers per molecule, with ~80% and 92% of reads forming full-length MAS arrays. This is equal to a concatenation factor of 14.5 and 15.8 for libraries 1 and 2 respectively, calculated based on the mean array size (Figures 5.24C and 5.25C). A higher number of partial arrays (< x16 concatenation) were generated in sample 1, with ~ 5% of reads attributed to 15-fold arrays and approximately 12% of all reads in arrays of <= 10 segments per molecule (Figure 5.24D). The source behind this higher proportion of incomplete arrays is at present unknown, but factors including high input cDNA concentration and large size of fragments are known factors that contribute to lower concatenation efficiency. With this said, these data demonstrate efficient cDNA concatenation from FACS-sorted LK and LSK Cd150+ single cells.

Despite the lack of indications from PacBio sequencing data QC, the recovery of cells was unexpectedly low for both samples - with only 17 and 22 cells estimated per library after demultiplexing barcodes (Appendix Supplementary Figure 5.7). This meant that only 51% - 57% of total reads were found in cells and was an unexpected finding given FACS sort reports and all preceding QC steps to determine library quality (both post cDNA generation and post MAS-seq library construction) gave no indication of poor cell recovery from the 10X Genomics experiment. As libraries were sequenced under the expectation they comprised cDNA from approximately 1K cells per sample, this meant that these cells were sequenced at a significantly high sequencing depth, on average obtaining 602,482 and 429,834 reads per single cell across samples 1 and 2. These reads equated to a median of 51,588 and 52,723 UMIs being detected per cell, capturing on average >11K genes per cell (Figures 5.24E and 5.25E).

The QC and annotations of transcripts based on *SQANTI3* categories were in agreement with the demultiplexing statistics. After de-duplication and the exclusion of transcripts classified as potential artefacts, a total of 11,230 and 10,785 unique genes (equating to 38,069 and 35,669 total unique transcripts) were identified across the two samples. In terms of the isoform distribution, on average 83% of isoforms detected between the two samples were classified as FSM and ISM, meaning most of the transcripts detected are well supported by existing references (Figures 5.24F and 5.25F).

Across both libraries, 3401 and 2830 novel non-canonical (NNC) isoforms were detected, these are isoforms which deviate from the standard canonical annotation of a gene. These isoforms may represent previously unannotated transcripts or rare events that are not well-characterised in canonical gene annotations, representing interesting targets for future investigations.

Single cell resolution analyses were performed in *Seurat* as previously described. The same viable single cells that were analysed from Illumina data were retained for downstream analysis of PacBio single cell data using cell barcodes to identify the correct cells, and included a total of 70 cells across both samples (Figure 5.26). Visualisation of the number of counts and features detected in each sample both at the gene- and isoform-level showed an average of 29,000 counts and 5753 genes captured across single cells (Figure 5.26A and B). PCA and dimensionality reduction based on the most variable genes (top 2000) across all cells revealed gene-level data from single cells exhibited the greatest similarity and clustered together when plotted on the first 2 PCs. The distance between data points in PCA reflects their similarity or dissimilarity in terms of expression signatures. This result shows that isoform-level expression was more variable between the two samples than gene-level expression (Figure 5.26). The small size of the dataset meant that limited tertiary analyses could be performed, however it was possible to perform exploratory analysis of genes known to be expressed within cells of the

LK Cd150+ compartment. For example, isoform-level data revealed three isoforms of *Mpl* were expressed across single cells (Figure 5.26D).

Chapter 5: Part 2 - Results Summary:

In summary these results show that a low-cell yield was obtained from both experiments, suggesting future work that aims to combine FACS cell type enrichment prior to 10X Genomics sample loading will require protocol optimisations. The cells captured however were in the most part viable LK Cd150+ cells, with only 3 cells removed during QC. With adjustments to sample suspension loading, this data shows MAS-seq from FACS sorted cell populations is a successful strategy to obtain gene and isoform level expression from single cells.

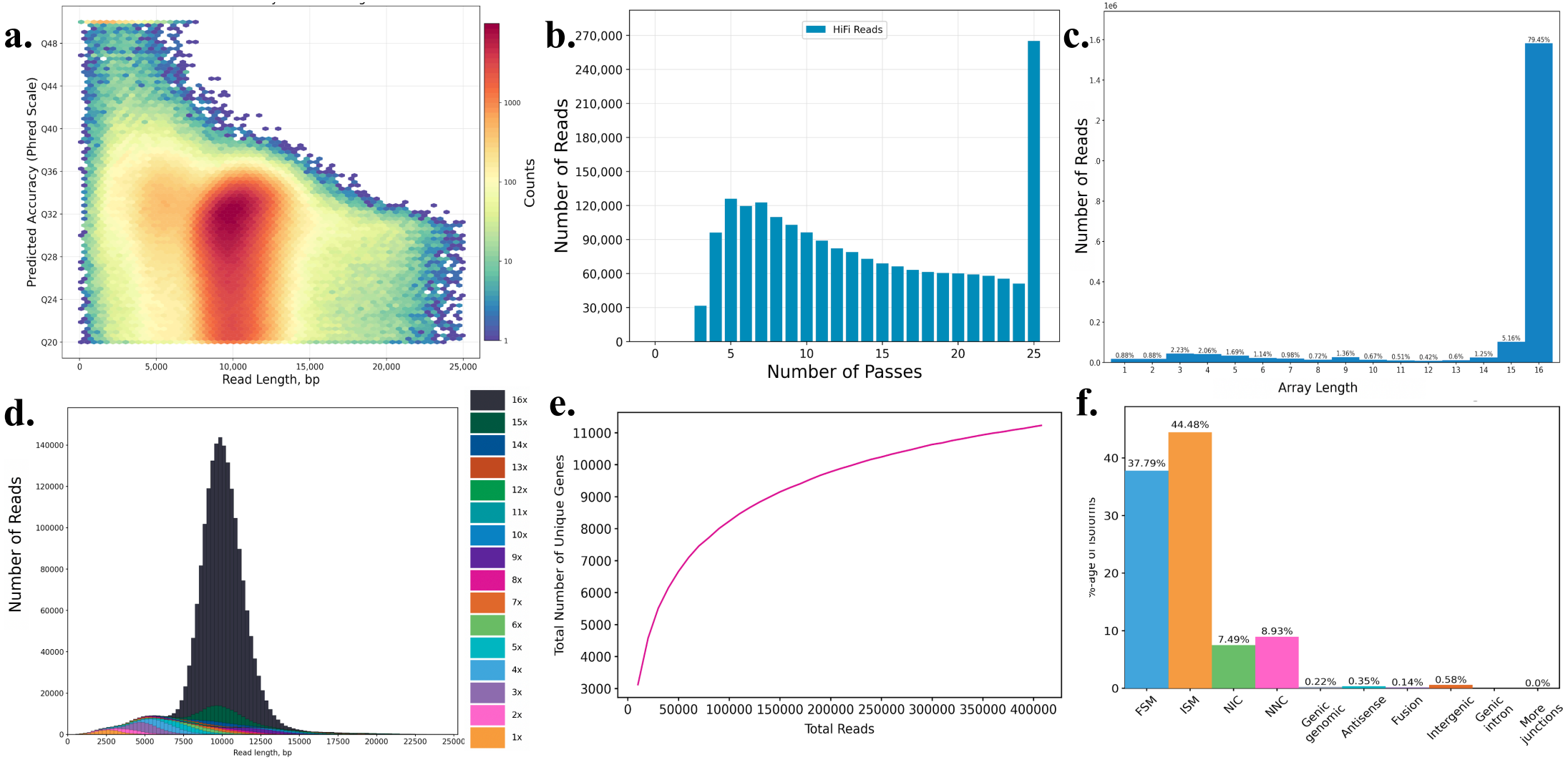


Figure 5.24. MAS-seq long-read sequencing of concatenated cDNA from FACS sorted LK Cd150+ sample 1. (a) Heat map of CCS Read lengths vs predicted accuracies (Q scores) (b) Distribution of HiFi reads by number of passes (c) Number of reads across MAS array lengths(d) Number of reads against concatemer read lengths (e) Gene saturation plot: Total number of unique genes detected across reads sequenced (f) Percentage of transcripts across *SQANTI3* isoform structural categories.

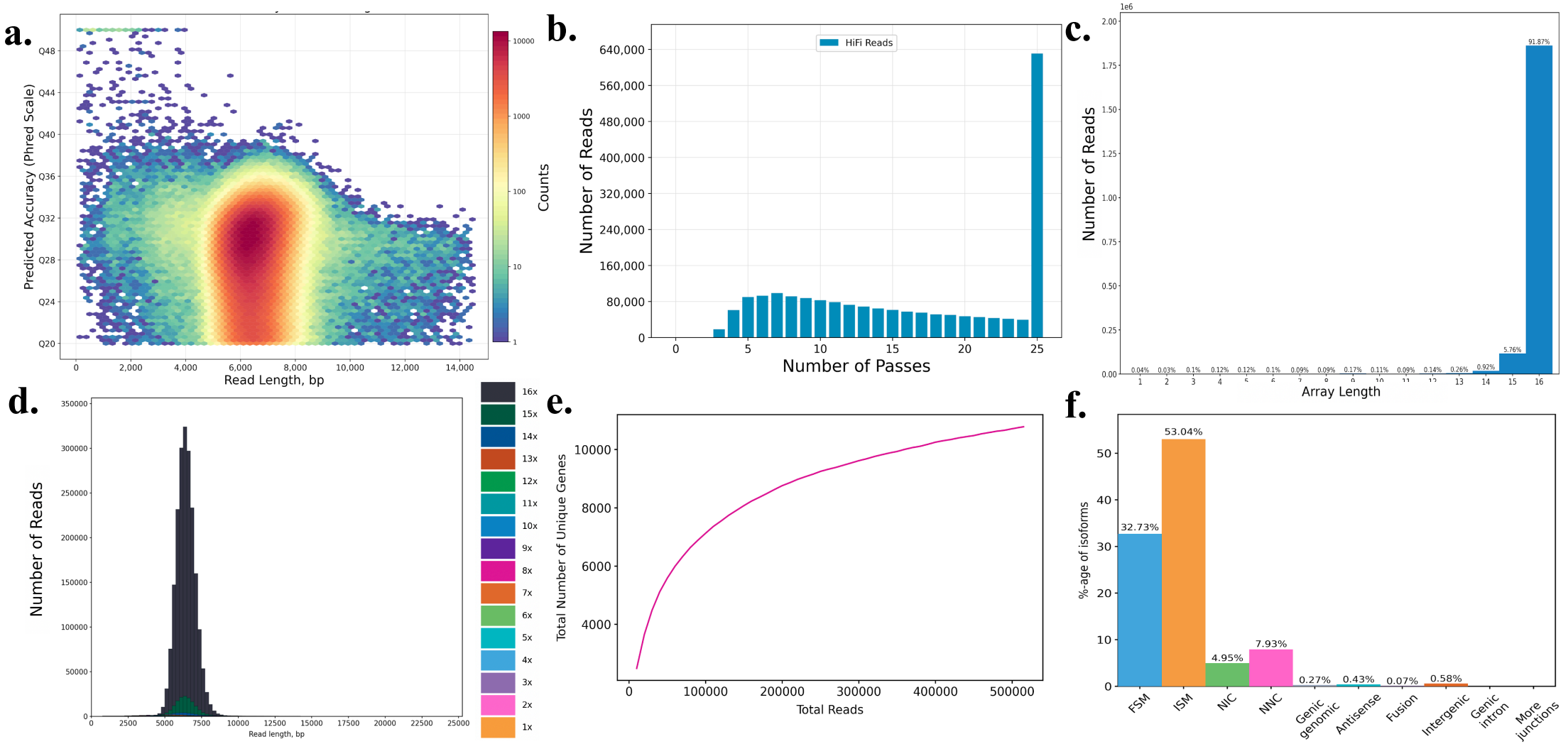


Figure 5.25. MAS-seq long-read sequencing of concatenated cDNA from FACS sorted LK Cd150+ sample 2. (a) Heat map of CCS Read lengths vs predicted accuracies (Q scores) (b) Distribution of HiFi reads by number of passes (c) Number of reads across MAS array lengths (d) Number of reads against concatemer read lengths (e) Gene saturation plot: Total number of unique genes detected across reads sequenced (f) Percentage of transcripts across *SQANTI3* isoform structural categories.

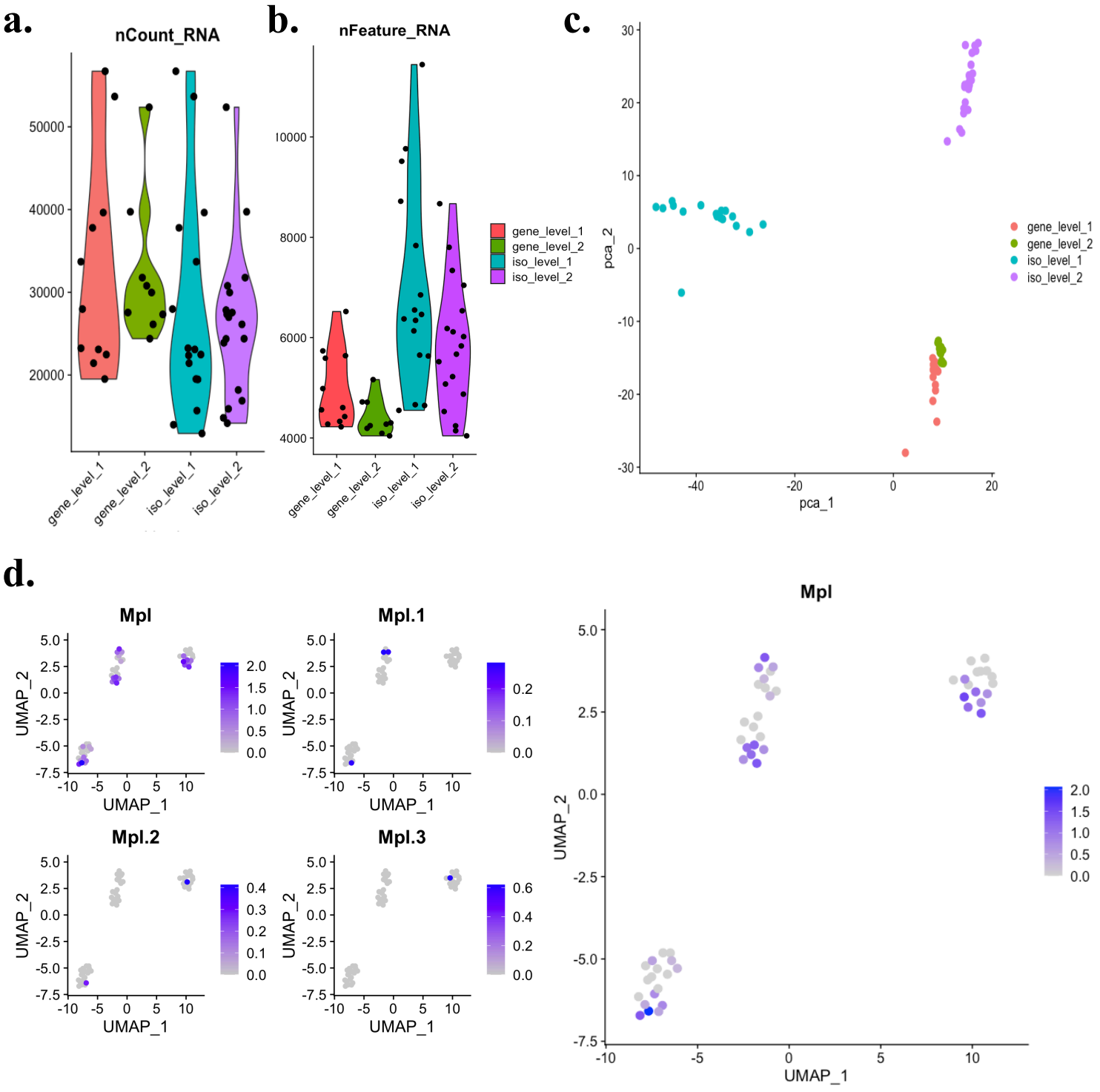


Figure 5.26. Gene and isoform expression analysis from the same FACS sorted LK Cd150+ single cell libraries sequenced with PacBio. (a) Violin plots showing the distribution of total counts across each modality and dataset. (b) Violin plots showing the distribution of total features across each modality and dataset**.** (c) PCA of all datasets labelled by their modality. (d) UMAP projection of the datasets labelled by the normalised expression levels of isoforms of Mpl.

Table 5.2. Experiment summary of all PacBio libraries made from concatenated 10X cDNA.

Metric	HIT-scIsoSeq	MAS IsoSeq PBMC		MAS IsoSeq LK cd150	
Number of polymerase reads	1,635,859	4,056,237	747,483	5,806,174	5,880,204
Number of HiFi reads	189,339	1,921,796	227,330	2,028,642	1,996,733
Mean HiFi read length (bp)	5,501	12,468	11,531	6,432	9,791
Number of cells detected	1040	4,875	7,384	22	17
Proportion of reads in cells	90%	95%	96%	57%	51%
Median UMIs per cell	6,456	4,993	5,062	52,723	51,588

5.4 Discussion

RNA-seq that utilises short- next generation sequencing (NGS) or long-read sequencing is a powerful tool to enable interrogation of the transcriptional diversity among cells. Historically, NGS has been more reliable in expression quantification but limited in its ability to study AS events due to generating reads with poor connectivity across splice junctions (Trapnell *et al.*, 2010). The advancements in long-read sequencing methods such as PacBio and Oxford Nanopore provide the unique ability to sequence full-length isoforms, and in particular, PacBio IsoSeq can now produce reads with over 99.99% accuracy. This removes the need for reconstructing possible transcript isoforms from fragmented short-reads and can improve our understanding of alternatively spliced isoforms of complex genes (Steijger *et al.*, 2013; Chen *et al.*, 2019). With the advances over the past 10 years that enable accurate sequencing of long-range exon connectivity we are now beginning to learn the full extent of alternative isoform expression (Sharon *et al.*, 2013; De Paoli-Iseppi, Gleeson and Clark, 2021). The expression of different RNA isoforms has been shown to drive cellular differentiation and regulate cell function, while aberrant splicing contributes to various diseases and oncogenic progression (Trapnell *et al.*, 2010; Graubert *et al.*, 2011; Haferlach *et al.*, 2014).

This chapter describes several approaches that were implemented to study isoform expression from single cells. One of these strategies was the generation of long-read libraries from cDNA derived from HSCs. Instead of focusing on single cells, a 'mini-bulk' approach was implemented, where each library was composed of 40-50 cells. Although this approach did not provide data at single-cell resolution, it offers a distinct advantage in capturing isoform diversity from a defined cell population. One major advantage of this approach was the higher loading concentration of libraries, which resulted in a greater yield of data. This increased data output provides greater coverage compared to what is typically expected from low-input 'bulk' samples with the crucial factor being that the sample purity was known prior to sequencing. This meant that prior to analysis, it was known that the results were specific to the HSC population, whereby long-read sequencing of HSCs was used to augment the information already obtained from short-read data.

Previous work has demonstrated intricate hematopoietic gene expression and AS programmes in HSCs that are associated with differentiation towards specific blood cell lineages (Chen *et al.*, 2014; Edwards *et al.*, 2016; Goldstein *et al.*, 2017). This includes the expression of transcripts unique to specific progenitor populations and AS events in important regulatory genes that were associated with the gain or loss of functional domains such as exon skipping and premature stop codon introduction (Chen *et al.*, 2014). For example, the isoform nuclear

factor IB (*Nf1b)* was revealed as important for Mk differentiation and was identified to be transcribed from a previously-unannotated transcription start site. This isoform lacks the DNA dimerisation domain that is necessary for binding the *Nfic* (its partner protein) illustrating how AS can result in functional consequences within the haematopoietic system (Chen *et al.*, 2014).

Moreover, a dynamic intron retention programme was identified in the murine Mk and Ery lineages involving hundreds of introns and genes with higher loss of intron retention in Ery cells and Mks compared with MEPs. The authors showed despite the common origin of Ery and Mk cells and some overlap of their transcriptomic signatures, the Mk intron retention programme was entirely distinct from that of the Ery lineage suggesting a lineage-specific regulation of intron retention (Edwards *et al.*, 2016). This finding adds a layer of complexity to our understanding of Mk and Ery differentiation and suggests that the intron retention process may be specific to each lineage, and might serve as a regulatory mechanism. For instance, the higher preservation of intron retention in Mk cells may have functional ramifications in the Mk lineage, potentially contributing to the regulation of genes involved in Mk function. This example of distinct intron retention programmes between Mk and Ery cells illustrate the need for further investigations to unravel the cell lineage-specific AS mechanisms.

The approach to study HSC isoform expression presented in this chapter was performed as a proof-of-concept experiment to determine whether the combination of single-cell transcriptome clustering and cell surface-marker expression (provided by FACS) enables the enrichment of a highly purified sample to interrogate the HSC AS landscape. Results confirmed the successful generation and sequencing of IsoSeq libraries, capturing a total of 5942 unique genes and 7456 unique isoforms which were annotated into structural categories along with AS analysis across key genes involved in HSC function. Hence this serves as a powerful strategy to obtain a comprehensive view of isoform expression patterns from a single population, enabling a more in-depth analysis of isoform heterogeneity at a cell type specific level.

The sample size in this experiment posed a limitation of this study. Due to technical constraints and the availability of cells, single cells from multiple young and aged mice were pooled to generate a single library per condition. The nature of this experiment means sample size is inherently restricted to the number of cells captured in the experiments presented in Chapter 4, and insufficient HSCs were captured to generate multiple replicates per condition. Having now established that the selection of Smart-seq2 full-length cDNA products can be successfully used to generate long-read libraries, to address this limitation and enhance the statistical power of future experiments, a future study could be designed to generate sample-level pools for long-read sequencing. In this way, instead of pooling cells from different samples into a single library, cells belonging to a specific population (ie. HSCs) would be pooled together based on

biological replicates (mice). This would involve generating multiple libraries across different experimental conditions, thus providing greater statistical power for conducting differential isoform usage analyses across different conditions. With a larger sample size, it would be possible to more reliably identify and quantify isoform expression differences between groups, thereby strengthening the robustness and validity of the findings. Furthermore, sample-level pools would also enable the exploration of inter-individual variability and the identification of potential age-related differences within each condition; therefore enabling interrogation of the impact of ageing on isoform expression and its potential implications for HSC function.

In summary, the first part of Chapter 5 presents a proof-of-concept strategy utilising long-read sequencing to study isoform expression in HSCs with age. The approach involved the targeted sequencing of cDNA from highly purified HSCs, which were selected based on their transcriptomic signatures identified through short-read scRNA-seq data clustering. By manually pooling the cDNA from HSCs and preparing long-read libraries, the study enabled the investigation of isoform expression in HSCs. The utilisation of long-read sequencing allowed for end-to-end full-length coverage of isoforms, providing valuable insights into the isoform expression of important genes for HSC and Mk function. The findings to date from this approach lay the foundation for further investigations into isoform diversity in HSCs. While the limited sample size and absence of statistical analyses present limitations, the initial results demonstrate this is a viable strategy to study isoform diversity in defined cell populations, and set the stage for further investigations. Future analysis of this data has the potential to uncover novel isoforms that may have been overlooked in traditional bulk long-read sequencing approaches. Moving forward, further exploration of this dataset will be used to set the stage for future investigations into isoform diversity in HSCs which could provide valuable insights into the mechanisms underlying HSC function.

The utility of scRNA-seq using long-read technologies has been constrained due to throughput limitations. Two key factors contributing to these limitations are the presence of undesired sequencing TSO-artefacts, and the library construction of short cDNA inserts which are not optimal with the long-read sequencing capacity (Wenger *et al.*, 2019; Lebrigand *et al.*, 2020). However recent advancements in the field have introduced two protocols that address these challenges, opening up new opportunities for accelerating the field of long-read single-cell transcriptomics (Shi *et al.,* 2022; Al'Khafaji *et al.,* 2021).

The first published protocol, HITscIso-Seq, revolutionised the field by enabling the concatenation of cDNA from single cells for PacBio sequencing. This approach incorporates a step to deplete TSO-artefacts using a PCR biotin-assisted capture procedure, followed by USER cloning-based cDNA concatenation (Shi *et al.,* 2022). This strategy effectively tackles

the issue of TSO-artefacts, significantly improving the quality of long-read sequencing data obtained from single cells.

The second strategy, developed by Al'Khafaji *et al.* from the Broad Institute, also addresses the challenge of TSO-artefacts but employs a different approach for concatenation. This method involves dU digestion followed by barcode-directed ligation of cDNAs to generate a long cDNA array (Al'Khafaji *et al.*, 2021). The development of this technique into a commercialised kit by PacBio, released in December of 2022, has provided a high-throughput solution for single-cell isoform sequencing. Fortunately, early access to this kit was granted by PacBio, facilitating the experiments presented in this thesis.

The primary objective of this chapter was to evaluate the effectiveness of these two approaches to enable gene- and isoform-resolved sequencing at single cell resolution. Concatenated HITsc-IsoSeq and MAS-seq libraries were generated from single cell cDNA that was prepared using 10X Genomics from both mouse BM and human PBMCs. Additionally, MAS-seq libraries of LK Cd150+ FACS sorted cells were also created to test the compatibility of FACS cell-type enrichment for 10X Genomics long-read sequencing.

The results of both approaches, HIT scIsoSeq and MAS-seq, provided valuable insights into their performance and the information obtained from long-read libraries using each approach. In the case of the HIT scIsoSeq 10X Genomics run, a total of 1,040 single cells were sequenced using short-read sequencing, resulting in an average of 1,528 genes from ~21K reads detected per cell. The Illumina data was processed following the standard workflow of *Seurat* scRNA-seq data analysis and enabled single cells to be grouped into 9 clusters. Cluster annotation was achieved using the *TabulaMurisConsortium()* as a reference; this data served as a ground truth for the cell-types captured in the experiment.

The concatenated product from 10X Genomics cDNA inserts generated a SMRTbell library of 285ng with an average size of 10.4 kb. Sequencing with PacBio yielded 189,339 HiFi reads, with an average read length of 5.5 kb. However, it is worth noting that the library displayed signs of low loading efficiency, with only 20% of polymerase-template complexes active during sequencing. Gene- and isoform-level analyses identified a total of 10,664 unique genes and 33,963 unique isoforms, which translated to an average of ~200 isoforms detected at the single cell level. Single cell clustering based on isoform expression was found to recapitulate the short-read gene-level clusters, and cell type annotations using the same reference were in concordance between the two datasets. These results demonstrate the successful generation of long-read libraries and the detection of isoform diversity using the HIT scIsoSeq approach.

In contrast, two PBMC samples were prepared into MAS-seq libraries. Illumina sequencing of 10X Genomics libraries revealed sample 1 consisted of 4,875 single cells, while sample 2 comprised 7,384 cells. Illumina sequencing yielded a median of 1,530 and 1,562 genes and mean ~42K and ~32K reads per cell for each library respectively. This data was processed following the standard workflow of *Seurat* scRNA-seq data analysis and after integration of the two runs single cells were grouped into 17 clusters. Cluster annotation was achieved using the *HumanPrimaryCellAtlasData()* reference dataset and again served as a ground truth for the cell-types sequenced from the PBMC samples.

The final MAS-seq concatenated inserts from each sample generated SMRTbell libraries of 716 ng and 842 ng and 11.3 kbp and 11.4 kbp insert length respectively. PacBio sequencing revealed an almost 9-fold difference in the HiFi read yield between the samples (1.9 million and 227,300 HiFi reads), and meant that integration of the two runs was not feasible therefore PBMC PacBio data were processed independently. In terms of HiFi read length, sample 1 of the MAS-seq approach exhibited an average length of 12.5 kb, while sample 2 had an average length of 11.5 kb. These values demonstrate a higher concordance with their average fragment lengths prior to sequencing compared to the HIT scIsoSeq approach, which had an average HiFi read length of 5.5 kb despite an average fragment length of 10.4 kb prior to sequencing. These values indicate a higher concordance with the average fragment lengths of the cDNA by MAS-seq. The closer alignment between the HiFi read lengths and the average fragment lengths in MAS-seq samples suggests that sequencing was effective in preserving the integrity of the cDNA array molecules. In comparison, the discrepancy observed with the HIT scIsoSeq approach could be due to a number of factors including poor polymerase sequencing efficiency that was shown as part of the analysis.

SQANTI3 analysis of PBMC libraries was performed to classify the isoforms identified in each sample based on structural categories. In total, 38,069 and 35,669 total unique transcripts were captured in the two samples with MAS-seq. Further downstream analyses of MAS-seq data was performed in *Seurat*, to analyse the data at single cell resolution. Each PBMC library was processed as two assays split based on isoform-level or gene-level expression data. This analysis revealed at the isoform-level 1017 and 114 isoforms per cell on average for each sample. This result clearly illustrates the large discrepancy between the coverage of the two libraries highlighting the impact of sequencing depth on capturing a comprehensive representation of isoform expression. Analysis of both samples independently showed isoform-level clustering recapitulated clusters found at the gene level. However, it is worth noting that sample 2 exhibited poorer cluster separation compared to sample 1, which can be attributed to the relatively shallow coverage in sample 2.

One of the main distinctions between the results of the two methods can be attributed to their different concatenation approaches. HIT scIsoSeq concatenation relies on stochastic enzymatic end-to-end ligation of cDNA inserts, whereas MAS-seq utilises a programmable strategy with the incorporation of specific 15 bp barcoded complementary adapter pairs at 3' and 5' ends of cDNA to generate the final array length. The contrasting concatenation methods employed by HIT scIsoSeq and MAS-seq contribute to the observed differences in the consistency and distribution of fragment lengths. MAS-seq ensures highly concatenation, with a controlled number of concatemers per molecule resulting in uniform distribution of read and fragment lengths. Indeed, 82% - 85% of reads from MAS-seq libraries were full arrays with a concatenation factor of 15 concatemers per molecule. On the other hand, HIT scIsoSeq generated a wider distribution of fragment lengths with varying numbers of concatemers per molecule, which varied from 1 (ie. non-concatenated) to >10 cDNA inserts per library resulting in an average 4.7-fold concatenation achieved. The stochastic enzymatic ligation process introduces more variability in the concatenation step, leading to a broader range of fragment lengths within the library.

The programmable nature of MAS-seq's concatenation approach provides greater control and reproducibility in generating libraries with consistent fragment lengths, enhancing also the reliability of isoform-level data analysis. The variable fragment lengths in HIT scIsoSeq introduces additional challenges in downstream data analysis including more complex processes for read segmentation, and read alignment. Ultimately, this variability may result in reduced sensitivity and accuracy in downstream analyses in comparison to libraries of highly uniform fragment lengths that facilitate accurate alignment and comparison of reads.

In order to compare the isoform sensitivity of each experiment, the number of isoforms detected per cell was used as a metric. The ratio of isoforms per cell was calculated by dividing the total number of unique isoforms detected in each experiment by the number of cells. The HIT scIsoSeq experiment exhibited a ~4-fold higher number of isoforms per cell (32.65 isoforms/cell) compared to MAS-seq (7.81 isoforms/cell for Sample 1 and 4.83 isoforms/cell for Sample 2). While this suggests that HIT scIsoSeq may be more effective in capturing a greater diversity of isoforms per cell, indicating higher isoform sensitivity, it's important to consider several factors. These include differences in sample isoform diversity (mouse BM vs. human PBMCs), sequencing depth, and other experimental variables that can influence the results. When comparing the two PBMC libraries, which had the same sample input but a significant difference in sequencing coverage, Sample 1 exhibited a 1.6-fold higher sensitivity in isoform detection per cell. This underscores the influential role of sequencing coverage on isoform detection.

This chapter also details a supplementary experiment involving the FACS sorted LK and LSK Cd150+ single cells from mouse BM for subsequent loading on the 10X Genomics platform. The objective of this experiment was to assess the feasibility of using FACS enrichment for the target cell population in generating MAS-seq libraries. However, the results revealed suboptimal cell loading, resulting in the capture of only 73 cells across 2 10X Genomics LT 3' scRNA-seq runs. Analysis of the cells after sequencing on both Illumina and PacBio platforms showed that they exhibited high expression of genes associated with the Mk lineage. This finding showed that although the number of captured cells was limited, the transcriptional profile of these cells indicated their viability as LSK Cd150+ cells. Several factors could explain the low cell recovery, with the most likely one being the loss of material during the transfer of samples from the wells of the 96-well plate, where they were initially sorted, to the 10X Genomics chip. Notably, assessments of sample quality at various stages did not indicate any issues with cell recovery.

Despite the challenges faced in cell recovery, MAS-seq libraries were successfully generated for both samples, resulting in exceptionally high coverage of single cells. While this experiment will be repeated in the future using an optimised cell isolation protocol, further analysis of this dataset is expected to reveal insightful information on isoform expression across these cells. Future iterations of this experiment will incorporate improvements to enhance cell isolation efficiency and for comprehensive analysis of isoform expression patterns in the LK Cd150+ compartment.

In conclusion, this chapter has provided a comprehensive overview of approaches for long-read sequencing of cDNA in single-cell experiments, with a focus on novel advanced methods for single-cell isoform sequencing. Through the testing of two recently published protocols to increase PacBio throughput from single cells, these results demonstrated that both methods are capable of achieving high-throughput gene- and isoform-level single-cell RNA sequencing. The programmable approach of MAS-seq has emerged as a particularly attractive method, offering the advantage of control over concatemer generation. This programmability and flexibility make MAS-seq a valuable tool that is likely to be easily adopted by the research community. The results presented in this chapter provide a solid foundation for further exploration of isoform diversity and alternative splicing events at the single-cell level. By comparing the performance of different sequencing methods and analysing the relationship between the number of isoforms detected and various experimental parameters, valuable insights have been gained into the strengths and limitations of each approach.

6. General Discussion &

Concluding Remarks

The purpose of this thesis was to interrogate differentiation trajectories of Mk commitment in the LK Cd150+ BM compartment with isoform-resolved single-cell transcriptomics. Specifically, scRNA-seq with Smart-seq2 was employed to study the transcriptomic signatures among stem and Mk progenitors under steady-state haematopoiesis and stress; including normal ageing and platelet depletion. The scRNA-seq datasets generated were subject to bioinformatic analyses to identify the cell types captured in experiments, interrogate patterns of gene expression across single cells and perform differential expression analyses across cell types and experimental conditions. Furthermore, semi-supervised pseudo-temporal ordering of the single-cell data was implemented to delineate trajectories of differentiation from stem cells toward Mk and Ery progenitors.

Additionally, this thesis explores the technological advances in the field of isoform resolved single cell transcriptomics. Three strategies were implemented to enable long-read sequencing and isoform profiling from haematopoietic cells. The first involved pooling of highly pure HSC cDNA samples from young and aged mice to produce a detailed overview of isoform expression heterogeneity among HSCs in the context of ageing. The latter two approaches involved the implementation of two novel methods for advanced long-read sequencing from single cells.

It is important to acknowledge that by focussing on the transcriptional programmes of Mk commitment, this inherently sidelines other potential important contributors of Mk differentiation. Haematopoietic lineage commitment, as demonstrated by a large and diverse body of previous research, is influenced by a myriad of factors beyond transcriptomic signatures alone. Such factors include differences in epigenetic status, cell cycling kinetics, BM microenvironment characteristics, HSC niche elements, cell metabolism, and the intrinsic transcriptional 'noise' amongst cells (Passegué *et al.*, 2005; Arias and Hayward, 2006; Hayashi *et al.*, 2008; Losick and Desplan, 2008; Raj and van Oudenaarden, 2008; Shahrezaei and Swain, 2008; Roundtree and He, 2016; Ho *et al.*, 2019). In the ensuing discussion, other important factors involved in Mk commitment emerging in the literature will be explored that are not addressed by work presented in this thesis.

First, a noteworthy recent study by Meng *et al.* used transcriptome and chromatin profiling of clonal HSC populations from single-cell transplantations to study epigenetic and transcriptional programmes in haematopoietic fate restriction (Meng et al. 2023). Their scRNA-seq data from platelet-biassed HSCs (determined based on platelet restricted lineage output of Vwf+ BM HSCs [Lin–c-Kit+Sca-1+CD150+CD48–Gata1-eGFP–]) includes complementary insights to the findings presented in this thesis. For example, their research reinforces the observations regarding the heightened expression of genes associated with HSC quiescence, stemness, and

platelet-lineage-specificity in platelet-biassed HSCs. Moreover, they also revealed an enrichment of gene signatures related to inflammation, including elevated NF-κB signaling in the context of Mk fate restriction, which is in agreement with the identification of age-associated DE in factors implicated in platelet hyperactivity and NF-κB signalling (Meng *et al.,* 2023). However, while our scRNA-seq data shares many points of convergence with their research, Meng *et al.* also emphasise the significance of epigenetic priming in fate restriction. They conclude that it is epigenetic rather than transcriptional lineage priming that most accurately predicts the differential lineage output from platelet-biassed vs. multilineage HSCs, based on ATAC-seq data revealing increased chromatin accessibility to both platelet-lineage-specific upstream regulatory elements and promoter regions in platelet restricted HSCs (Meng *et al.,* 2023). Their work, among various previous publications, highlight an intricate interplay between epigenetic and transcriptional regulation in governing haematopoietic lineage commitment (Heuston *et al.* 2018; Rodrigues *et al.* 2020; Adelman *et al.* 2017; Zhao *et al.* 2023; Meng *et al.* 2023). Therefore, it is essential to emphasise that while this thesis contributes to the understanding of transcriptional programs of Mk fate restriction, it acknowledges that data overwhelmingly indicates transcriptional regulation is not the sole mechanism at play.

Another emerging area of research is the role of uptake, functionalisation, and metabolism of fatty acids in Mk function. Lipidomics has revealed lipids fulfill distinct yet crucial roles in cell function, including energy provision, signaling, and perhaps most notably for Mks, membrane architecture. Indeed, the importance of a lipid-rich cell membrane composition in facilitating the extensive membrane remodeling is key for Mk functionality, generating the lipid-rich demarcation membrane system and Mk polarisation toward the protrusion of proplatelets into the sinusoids of the bone marrow (Geue et al. 2019; Eckly et al. 2014; Kelly et al. 2020). Recent investigations have also shed light on the involvement of lipids in the Mk maturation process itself. Utilising mass spectrometry to elucidate a quantitative lipidomics map of Mk differentiation, Jonckheere *et al.* lipid uptake increases significantly during Mk maturation, both *in vitro* and *in vivo*, as evidenced by the heightened expression of fatty acid receptors such as CD36 (de Jonckheere et al. 2023). Furthermore, *de novo* lipogenesis has emerged as a potential regulatory mechanism on megakaryopoiesis, particularly in the late stages of Mk maturation (Barrachina et al. 2023). Liquid chromatography tandem-mass spectometry was used to identify lipidome alterations throughout Mk differentiation stages, and also revealed platelet production can be manipulated based on the availability of exogenous lipid availability including dietary polyunsaturated fats (Kelly et al. 2020; Barrachina et al. 2023). Our understanding of the role of lipid biosynthesis and utilisation in Mk differentiation is in its infancy, but these reports suggest metabolic shifts and changes to membrane transporters play an important role in Mk differentiation. Further work in this field poses attractive opportunities

for clinical strategies as simple as dietary intervention to regulate megakaryopoiesis and manipulate platelet production.

In the context of discussing limitations of this work, it is imperative to recognise that an inherent limitation of scRNA-seq is its provision of a mere snapshot of a cell's transcriptome at the time of sampling. This limitation necessitates the reliance on computational algorithms to infer cell state transitions, which means that the inferred trajectories of cellular differentiation are based on extrapolations from a single time point, potentially overlooking transient or dynamic states that occur between sampling intervals. Consequently, there's a risk of oversimplifying the complexity of such processes, as the actual transitions between cellular states may involve intricate temporal dynamics that are not fully captured in a static snapshot. Despite these challenges, computational algorithms used in scRNA-seq data analyses have been continuously improving, with the development of sophisticated methods for data normalisation, dimensionality reduction, and trajectory inference (Haghverdi and Ludwig, 2023). And so although assumptions are made about cell state transitions, these assumptions are often based on well-established biological knowledge and are supported by experimental validation.

Moreover, while scRNA-seq provides valuable insights into gene expression at the single-cell level, it is suceptible to both technical as well as biological variability. Indeed, data from scRNA-seq experiments can be affected by technical noise and variability, which may can introduce bias and compromise the accuracy of downstream analyses (Hicks *et al.*, 2018). Hence, it was imperative to employ strategies to mitigate technical variability and ensure the reliability of the data obtained for analysis and interpretation throughout this thesis. Various strategies were deployed for this purpose, such as inclusion of library preparation and sequencing controls, stringent data QC measures, batch correction during data integration, normalisation of sequencing coverage and robust statistical analyses tailored to account for the inherent noise in single-cell data. The implementation of these methods demonstrates a conscientious effort to minimise technical variability effects and underscores the robustness of the data obtained for downstream analysis and interpretation across this thesis.

Lastly, it's also important to note that disparities in mRNA abundance among certain genes may not necessarily translate to discernible functional consequences. Meaning although mRNA often serves as a proxy for protein expression, the traditional method for defining cell types, the relationship is not strictly one-to-one (Vogel and Marcotte, 2012; Edfors *et al.*, 2016). Nevertheless, the vast majority of genes exhibit a strong correlation between RNA and protein expression, even at the single-cell level (Darmanis *et al.*, 2016). And so while this simplification overlooks crucial structural elements of cells and external influences, it still provides profound insights into cell state and identity. Therefore despite these inherent

limitations, scRNA-seq remains an invaluable tool for investigating cellular differentiation and uncovering the pivotal transcriptional regulators that govern cell fate decisions.

6.1. Delineating megakaryopoiesis using scRNA-seq under steady-state and stress

The primary focus of this body of work was to gain a deeper understanding of the transcriptional signatures activated during the commitment pathways of the Mk lineage. To achieve this, a consistent single-cell sorting strategy was implemented across experimental chapters 3 and 4, that builds upon the foundational research conducted by Pronk *et al.* on the LK/LSK Cd150+ compartment (Pronk *et al.*, 2007). By adhering to a single sorting strategy across these chapters, isolating the same cellular compartment across experiments, the separate datasets presented corroborate to reveal the specific cell types involved in Mk differentiation, alongside Ery cells and a small proportion of myeloid committed cells in the context of both normal ageing haematopoietic stress.

To delineate the changes occurring during Mk lineage commitment, rigorous bioinformatic analyses were conducted using pseudotime trajectory construction. This powerful tool enabled the exploration of transcriptional changes along different states of Mk commitment, providing insights into the hierarchical nature of these changes. Pseudotime trajectory construction enabled the dynamic expression patterns of genes from LT-HSCs to committed progenitor states during both steady-state megakaryopoiesis and under stress conditions, including acute thrombocytopenia and ageing. Through differential expression analyses, genes associated with specific cell states along the pseudotime trajectories were statistically assessed, resulting in comprehensive collections of genes exhibiting significant variable expression as a function of pseudotime. Furthermore, in addition to pseudotime trajectory analysis DEA were performed throughout this thesis to identify DEGs across experimental conditions, such as platelet depletion treatment or mouse age. Notably, these analyses were conducted at the pseudo-bulk level to minimise the likelihood of false-positive detections caused by inflated p-values, which can be a concern with many single-cell DEA approaches. This approach successfully identified hundreds of genes associated with age and/or acute thrombocytopenia at the cell-type level, including previously unexplored genes that have not been implicated in these contexts before.

It should be underscored that conducting independent validation experiments testing the identified DEGs described in Chapters 3 and 4 will be necessary to fortify the robustness of these results. For instance, qPCR assays are a robust way to obtain precise quantification of gene expression levels in a targeted manner, and are often used to provide validation of the scRNA-seq results. By measuring the abundance of select genes identified as differentially expressed from pseudotime analysis in cells at different stages of Mk commitment, it will be

possible to validate whether their expression levels corroborate the levels seen in the RNA-seq datasets presented in this thesis, and whether their expression are associated with specific Mk commitment stages. In the same way, qPCR assays of the same cell populations but different sample conditions, namely age and presence of haematopoietic stress, will be essential before concluding the correlation of transcriptomic signatures observed here are biologically significant. Another set of relevant validation strategies worth considering includes gene knockdown or overexpression techniques like RNA interference and CRISPR/Cas9 gene editing. These approaches that enable modulation of expression of genes, would help reveal functional roles of the signatures identified here, and the effect of specific genes of interest identified in this work on cellular phenotype. Such validation experiments will support the observations of the identified DEGs and enhance the confidence in their biological significance. Additionally, these experiments may uncover novel candidates that are specifically associated with Mk fate restriction, further expanding our understanding of Mk lineage commitment.

6.2 Implementation of long-read sequencing approaches for studying isoform expression at single cell resolution

Recent advancements in the field of isoform-profiling at the single-cell level have yielded significant progress in the past few years. The growing recognition of the role of isoform heterogeneity and the functional implications of alternative splicing in haematopoiesis underscore the need for methodologies that facilitate isoform-resolved single-cell transcriptomics at high-throughput (Chen *et al.*, 2014; Yap and Makeyev, 2016; Song *et al.*, 2017). Specifically, a crucial question that remains unanswered is the impact of alternatively spliced isoforms during haematopoietic cell fate decisions, such as in the context of lineage commitment towards the Mk fate. While lineage-biassed HSCs that are primed for platelet-specific gene expression have been identified, the precise mechanisms governing the lineage bias towards the Mk fate are yet to be fully understood (Sanjuan-Pla *et al.*, 2013).

In this thesis, a comprehensive analysis of the LK Cd150+ compartment has contributed towards addressing this question, elucidating gene-level transcriptional signatures associated with Mk differentiation. However, as the field of single-cell isoform sequencing continues to evolve, it becomes imperative to transition from a gene-centric analysis of lineage restriction to exploring isoform-resolved heterogeneity among differentiating cells.

The integration of isoform-level resolution into the investigation of cell fate decisions will not only unravel the complex interplay between alternative splicing events and commitment but also likely provide a more comprehensive understanding of the underlying regulatory mechanisms. By delving into isoform diversity and dynamics during Mk lineage commitment, the objective is to potentially uncover novel regulatory networks and signalling pathways that influence the fate determination of haematopoietic progenitor cells. Moreover, this approach may reveal specific isoforms that act as key drivers of the commitment process, steering cells towards particular lineages.

To achieve this objective, this thesis also focused on implementing long-read sequencing approaches that capture full-length isoforms from single cells. Chapter 5 outlines multiple strategies employed to enable isoform sequencing. The first strategy involved a targeted pooling approach, where cDNA from highly purified HSCs was manually pooled based on their transcriptomic signatures identified through short-read scRA-seq data clustering. By preparing long-read libraries from these pooled cDNA samples, the study successfully investigated isoform expression in HSCs. The utilisation of long-read sequencing technology facilitated

end-to-end full-length coverage of isoforms, providing valuable insights into the isoform expression patterns of genes crucial for HSC and Mk functionality in the context of age.
Additionally, Chapter 5 explored two novel advanced methods for single-cell isoform sequencing, which involved concatenating multiple cDNA inserts from single-cells to maximise long-read sequencing throughput. These methods, known as HITsc-IsoSeq and MAS-seq, were applied to haematopoietic cells with the goal of increasing PacBio throughput from the 10X Genomics scRNA-seq platform. The hypothesis underlying these experiments was that the utilisation of cDNA concatenation would enhance sequencing throughput, enabling both gene- and isoform-level analysis at the single-cell resolution. By adopting these innovative techniques, this set of experiments aimed to establish the effectiveness of both approaches in enabling the study of isoform diversity and evaluate potential applications of these protocols in haematopoiesis.

The analysis of short- and long-read sequencing data derived from experiments utilising both concatenation strategies revealed the distinct characteristics of each method, shedding light on their individual strengths and limitations. Importantly, these findings highlighted the sensitivity of the approaches in enabling isoform sequencing analysis of single cells, providing valuable insights into the potential of these techniques to advance our understanding of gene- and isoform-level scRNA-seq in the context of haematopoiesis.

Further analysis of the collected data will include extracting the lists of isoforms identified using SQANTI3, with a specific focus on the identification of novel isoforms (NNC) from the MAS-seq data. By filtering and studying the sequence alignments of this subset of isoforms, the goal would be to assess potentially novel alternative variants of genes. Priority will be given to investigating NNC isoforms found in genes known to play crucial roles in haematopoiesis. In this way identifying if any interesting novel signatures, with potential functional consequences, have been identified in the data collected thus far. This may provide valuable insights into the landscape of isoform diversity within the MAS-seq datasets, shedding light on potential novel mechanisms and functions underlying cellular processes that were captured.

By successfully demonstrating the high-throughput gene- and isoform-level scRNA-seq capabilities of both HITsc-IsoSeq and MAS-seq, this research contributes to the advancement of the field. These novel methodologies pave the way for in-depth investigations into the complex isoform diversity, offering a more comprehensive understanding of the regulatory mechanisms governing complex processes such as haematopoietic cell fate decisions. The ability to capture and analyse isoform-level information at single-cell resolution will enable researchers to begin to unravel the functional consequences of alternative splicing and its impact on lineage commitment in haematopoiesis. Indeed, by advancing our technical

capabilities, these methodologies contribute to the ongoing progress in the field of single cell long-read sequencing, opening new avenues for investigating isoform diversity, unravelling the complexity of alternative splicing, and furthering our understanding of the molecular mechanisms driving haematopoietic cell fate determination.

6.3. Future perspectives

Single-cell transcriptomics has predominantly focused on gene-level expression to investigate the signatures and functional states of biological systems. These gene expression measurements capture the collective expression of multiple isoforms originating from individual genes (Wen, Mead and Thongjuea, 2020). This tendency can be attributed, at least in part, to the advantages of short-read sequencing, which has historically dominated genomics due to its high accuracy and relatively low sequencing costs. However, recent advancements in long-read sequencing technologies, especially in the past few years, have significantly progressed.

The rapid development of long-read sequencing technologies, such as ONT and PacBio instruments, has facilitated the generation of highly accurate full-length reads that cover the entire length of transcripts. While the significance of alternative splicing in driving protein diversity has long been acknowledged, the interest in studying isoform diversity at the single-cell level has grown substantially with the increased accessibility of profiling isoforms using long-read sequencing (Pan *et al.*, 2008; Wang *et al.*, 2008; Trapnell *et al.*, 2010; Chen *et al.*, 2019). Furthermore, the integration of high-throughput approaches for scRNA-seq with long-read sequencing has made the utilisation of these technologies more affordable and worthwhile for single-cell analyses.

This thesis implemented and evaluated the two newest strategies, HIT scIsoSeq and MAS-seq, to enhance the throughput of long-read sequencing in profiling haematopoietic cells at both the gene and isoform level (Al'Khafaji *et al.*, 2021; Shi *et al.*, 2022). By successfully demonstrating the effectiveness of both approaches in generating concatenated libraries from 10X Genomics cDNA inserts, it is evident that they have tremendous potential in enabling alternative splicing analysis at the single-cell resolution. Notably, the MAS-seq concatenation strategy offers the advantage of consistency through its programmability, ensuring stable and uniform concatenation. This feature not only guarantees expected fold-concatenation in experiments but also facilitates downstream read segmentation and analyses.

This thesis has contributed towards the understanding of the gene-level signatures of megakaryopoiesis, however several important questions remain unanswered. A burgeoning area

of research has revealed the critical impact of alternatively spliced isoforms on cell phenotypes, including cell fate decisions. Biassed haematopoiesis towards the Mk lineage has been investigated in various contexts, such as the effects of BM localization, cell-cell interactions within the BM micro-environment, and clonal haematopoiesis (Sanjuan-Pla *et al.*, 2013; Haas *et al.*, 2015; Frisch *et al.*, 2019; Psaila *et al.*, 2020; Estevez *et al.*, 2021). However, the potential role of alternative spliced isoforms as a mechanism contributing to lineage bias remains unexplored.

As a proof-of-concept experiment, FACS-sorted LK Cd150+ cells were isolated for single-cell RNA sequencing (scRNA-seq) using the 10X Genomics LT 3' scRNA-seq chemistry. However, the results revealed a low cell recovery rate after Illumina and PacBio sequencing, suggesting a loss of sample during the experimental library preparation process. Despite this issue, 10X Genomics successfully generated cDNA, which was then used to generate MAS-seq libraries. This outcome suggests that with improvements to the cell loading strategy for 10X Genomics, this approach could be a viable method for performing isoform-resolved single-cell transcriptomics of cells along the Mk lineage.

To explore cell-type-specific signatures of alternative splicing in the Mk lineage, a future experiment will be conducted using a combination of the 10X Genomics HT 3' chemistry, which has since become the recommended chemistry to combine with PacBio MAS-seq. This experiment will focus on the LK Cd150+ population, which, thus far in this thesis, has only been examined at the gene level.

An important consideration for the success of this approach will be to ensure adequate and as uniform coverage across single cells as possible. In scRNA-seq experiments, there is a well-known trade-off between the number of cells captured and the sequencing coverage obtained per cell. In this experiment, prioritising sequencing coverage over the number of cells captured will maximise the sensitivity for potentially detecting rare isoforms. The PBMC data presented as part of Chapter 5 could be useful to help determine the optimal cell loading concentration in order to obtain as many isoforms per cell whilst ensuring sufficient cells are loaded for successful cDNA generation.

Moreover, to provide a "ground-truth" for the cell types captured in the experiment, 10X Genomics cDNA from LSK Cd150+ cells will also be sequenced using an Illumina platform. This sequencing data will establish the cell types present in the experiment to supplement isoform-level analyses generated from PacBio MAS-seq libraries.generated from the sample cDNA samples. This approach will enable the study of isoform heterogeneity across cell types

within the LSK Cd150+ compartment. Analysing differential isoform usage between cell types may reveal enriched expression of specific isoforms in different cell types.

An additional aspect that could be considered is the integration of protein-level expression alongside isoform-level expression, which would provide an additional layer of information for exploring the relationship between isoform expression and protein expression. CITE-seq, for example, is a method that enables the detection of protein expression at single-cell resolution (Stoeckius *et al.*, 2017). This approach allows for the simultaneous measurement of gene expression and the proteins expressed on the cell surface within individual cells. By incorporating this additional modality, it may be possible to address questions regarding whether isoform heterogeneity translates to the protein level. This integrated approach, covering multiple levels of expression, holds potential for exploring the relationships between isoforms and protein expression. However, it is important to note that with each additional layer of complexity, there are increased risks in experimental design and execution. Therefore, careful considerations regarding the feasibility of introducing an additional modality must be taken into account.

Despite the challenges faced in the first attempted experiment, the combined use of 10X Genomics and MAS-seq shows promise in enabling isoform-resolved single-cell transcriptomics of cells along the Mk lineage. Future experiments utilising the 10X Genomics HT 3' chemistry, along with PacBio MAS-seq, will provide valuable insights into cell-type-specific signatures of alternative splicing. By combining these strategies and considering the mentioned considerations, it is anticipated that this future work will contribute to a deeper understanding of alternative splicing and its impact across diverse settings, including megakaryopoiesis.

In conclusion, this body of work represents a significant contribution to our understanding of transcriptional dynamics involved in Mk lineage commitment. Through the utilisation of robust single cell experimental methodologies and bioinformatic analyses, this research provides valuable insights into key genes and processes governing different states of Mk differentiation, thereby delineating a high-resolution transcriptomic roadmap of Mk differentiation signatures. This work uncovered several novel potential candidates associated with Mk cells in the context of age and acute thrombocytopenia. These findings serve as a valuable resource for future investigations aiming to address important unanswered questions in the field. By shedding light on the intricate molecular profiles of Mk cells, this research enhances our understanding of the complex mechanisms governing megakaryopoiesis. It also seeks to enable continued exploration to unravel the molecular mechanisms underlying Mk lineage-biassed differentiation and its potential implications in disease.

Bibliography

Abdulhay, N.J. *et al.* (2019) 'Impaired human hematopoiesis due to a cryptic intronic GATA1 splicing mutation', *The Journal of experimental medicine*, 216(5), pp. 1050–1060. Available at: https://doi.org/10.1084/jem.20181625.

Adelman, E. *et al.* (2017) 'Integrative epigenetic and single-cell RNA-seq profiling of human hematopoietic stem cells reveals epigenetic reprogramming of enhancer and regulatory elements during normal aging', *Blood*, 130(Suppl_1), pp. 770–770. Available at: https://doi.org/10.1182/blood.v130.suppl_1.770.770.

Adolfsson, J. *et al.* (2001) 'Upregulation of Flt3 expression within the bone marrow Lin- Sca1+ c-kit+ stem cell compartment is accompanied by loss of self-renewal capacity', *Immunity*, 15(4), pp. 659–669. Available at: https://www.sciencedirect.com/science/article/pii/S1074761301002205.

Adolfsson, J. *et al.* (2005) 'Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment', *Cell*, 121(2), pp. 295–306. Available at: https://doi.org/10.1016/j.cell.2005.02.013.

Akashi, K. *et al.* (2000) 'A clonogenic common myeloid progenitor that gives rise to all myeloid lineages', *Nature*, 404, p. 193. Available at: https://doi.org/10.1038/35004599.

Alexander, W.S. *et al.* (1996) 'Deficiencies in progenitor cells of multiple hematopoietic lineages and defective megakaryocytopoiesis in mice lacking the thrombopoietic receptor c-Mpl', *Blood*, 87(6), pp. 2162–2170. Available at: https://www.ncbi.nlm.nih.gov/pubmed/8630375.

Ali, M.A.E. *et al.* (2017) 'Functional dissection of hematopoietic stem cell populations with a stemness-monitoring system based on NS-GFP transgene expression', *Scientific reports*, 7(1), p. 11442. Available at: https://doi.org/10.1038/s41598-017-11909-3.

Al'Khafaji, A.M. *et al.* (2021) 'High-throughput RNA isoform sequencing using programmable cDNA concatenation', *bioRxiv*. Available at: https://doi.org/10.1101/2021.10.01.462818.

Amezquita, R.A. *et al.* (2020) 'Orchestrating single-cell analysis with Bioconductor', *Nature methods*, 17(2), pp. 137–145. Available at: https://doi.org/10.1038/s41592-019-0654-x.

Anand, R. *et al.* (2021) 'HELQ is a dual-function DSB repair enzyme modulated by RPA and RAD51', *Nature*, 601(7892), pp. 268–273. Available at: https://doi.org/10.1038/s41586-021-04261-0.

Andrews, S. (no date) *FastQC: A quality control analysis tool for high throughput sequencing data*. Github. Available at: https://github.com/s-andrews/FastQC (Accessed: 24 February 2023).

Aran, D. *et al.* (2019) 'Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage', *Nature immunology*, 20(2), pp. 163–172. Available at: https://doi.org/10.1038/s41590-018-0276-y.

Arias, A.M. and Hayward, P. (2006) 'Filtering transcriptional noise during development: concepts and mechanisms', *Nature reviews. Genetics*, 7(1), pp. 34–44. Available at: https://doi.org/10.1038/nrg1750.

Arinobu, Y. *et al.* (2007) 'Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages', *Cell stem cell*, 1(4), pp. 416–427. Available at: https://doi.org/10.1016/j.stem.2007.07.004.

Ashman, L.K. (1999) ‘The biology of stem cell factor and its receptor C-kit’, *The international journal of biochemistry & cell biology*, 31(10), pp. 1037–1051. Available at: https://doi.org/10.1016/s1357-2725(99)00076-x.

Bacon, C.M. *et al.* (1995) ‘Thrombopoietin (TPO) induces tyrosine phosphorylation and activation of STAT5 and STAT3’, *FEBS letters*, 370(1-2), pp. 63–68. Available at: https://doi.org/10.1016/0014-5793(95)00796-c.

Bahr, C. *et al.* (2018) ‘Author Correction: A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies’, *Nature*, 558(7711), p. E4. Available at: https://doi.org/10.1038/s41586-018-0113-3.

Baker, S.J., Rane, S.G. and Reddy, E.P. (2007) ‘Hematopoietic cytokine receptor signaling’, *Oncogene*, 26(47), pp. 6724–6737. Available at: https://doi.org/10.1038/sj.onc.1210757.

Balazs, A.B. *et al.* (2006) ‘Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow’, *Blood*, 107(6), pp. 2317–2321. Available at: https://doi.org/10.1182/blood-2005-06-2249.

Balenci, L. *et al.* (2013) ‘Bone morphogenetic proteins and secreted frizzled related protein 2 maintain the quiescence of adult mammalian retinal stem cells’, *Stem cells*, 31(10), pp. 2218–2230. Available at: https://doi.org/10.1002/stem.1470.

Bansal, P. *et al.* (2018) ‘Current Updates on Role of Lipids in Hematopoiesis’, *Infectious disorders drug targets*, 18(3), pp. 192–198. Available at: https://doi.org/10.2174/1871526518666180405155015.

Barbosa-Morais, N.L. *et al.* (2012) ‘The evolutionary landscape of alternative splicing in vertebrate species’, *Science*, 338(6114), pp. 1587–1593. Available at: https://doi.org/10.1126/science.1230612.

Barrachina, Maria N et al. (2023) “Efficient megakaryopoiesis and platelet production require phospholipid remodeling and PUFA uptake through CD36.” *bioRxiv : the preprint server for biology*. doi:10.1101/2023.02.12.527706.

Bastyr, E.J., 3rd, Kadrofske, M.M. and Vinik, A.I. (1990) ‘Platelet activity and phosphoinositide turnover increase with advancing age’, *The American journal of medicine*, 88(6), pp. 601–606. Available at: https://doi.org/10.1016/0002-9343(90)90525-i.

Bazil, V. *et al.* (1995) ‘Apoptosis of human hematopoietic progenitor cells induced by crosslinking of surface CD43, the major sialoglycoprotein of leukocytes’, *Blood*, 86(2), pp. 502–511. Available at: https://doi.org/10.1182/blood.V86.2.502.bloodjournal862502.

Becht, E. *et al.* (2018) ‘Dimensionality reduction for visualizing single-cell data using UMAP’, *Nature biotechnology*, 37(1), pp. 38–44. Available at: https://doi.org/10.1038/nbt.4314.

Becker, A.J., McCULLOCH, E.A. and Till, J.E. (1963) ‘Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells’, *Nature*, 197, pp. 452–454. Available at: https://doi.org/10.1038/197452a0.

Beckmann, J. *et al.* (2007) ‘Asymmetric cell division within the human hematopoietic stem and progenitor cell compartment: identification of asymmetrically segregating proteins’, *Blood*, 109(12), pp. 5494–5501. Available at: https://doi.org/10.1182/blood-2006-11-055921.

Beerman, I. *et al.* (2010) ‘Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(12), pp. 5465–5470. Available at: https://doi.org/10.1073/pnas.1000834107.

Beerman, I. *et al.* (2014) ‘Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle’, *Cell stem cell*, 15(1), pp. 37–50. Available at: https://doi.org/10.1016/j.stem.2014.04.016.

Benveniste, P. *et al.* (2010) ‘Intermediate-term hematopoietic stem cells with extended but time-limited reconstitution potential’, *Cell stem cell*, 6(1), pp. 48–58. Available at: https://doi.org/10.1016/j.stem.2009.11.014.

Berget, S.M., Moore, C. and Sharp, P.A. (1977) ‘Spliced segments at the 5′ terminus of adenovirus 2 late mRNA’, *Proceedings of the* [Preprint]. Available at: https://www.pnas.org/doi/abs/10.1073/pnas.74.8.3171.

Bergmeier, W. *et al.* (2000) ‘Structural and functional characterization of the mouse von Willebrand factor receptor GPIb-IX with novel monoclonal antibodies’, *Blood*, 95(3), pp. 886–893. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10648400.

Bernitz, J.M. *et al.* (2016) ‘Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions’, *Cell*, 167(5), pp. 1296–1309.e10. Available at: https://doi.org/10.1016/j.cell.2016.10.022.

Besancenot, R. *et al.* (2010) ‘A senescence-like cell-cycle arrest occurs during megakaryocytic maturation: implications for physiological and pathological megakaryocytic proliferation’, *PLoS biology*, 8(9). Available at: https://doi.org/10.1371/journal.pbio.1000476.

Bettoni, S. *et al.* (2017) ‘Interaction between Multimeric von Willebrand Factor and Complement: A Fresh Look to the Pathophysiology of Microvascular Thrombosis’, *Journal of immunology* , 199(3), pp. 1021–1040. Available at: https://doi.org/10.4049/jimmunol.1601121.

Bianchi, E. *et al.* (2015) ‘MYB controls erythroid versus megakaryocyte lineage fate decision through the miR-486-3p-mediated downregulation of MAF’, *Cell death and differentiation*, 22(12), pp. 1906–1921. Available at: https://doi.org/10.1038/cdd.2015.30.

Bouilloux, F. *et al.* (2008) ‘EKLF restricts megakaryocytic differentiation at the benefit of erythrocytic differentiation’, *Blood*, 112(3), pp. 576–584. Available at: https://doi.org/10.1182/blood-2007-07-098996.

Bradley, H.L., Hawley, T.S. and Bunting, K.D. (2002) ‘Cell intrinsic defects in cytokine responsiveness of STAT5-deficient hematopoietic stem cells’, *Blood*, 100(12), pp. 3983–3989. Available at: https://doi.org/10.1182/blood-2002-05-1602.

Bray, P.F. *et al.* (2013) ‘The complex transcriptional landscape of the anucleate human platelet’, *BMC genomics*, 14, p. 1. Available at: https://doi.org/10.1186/1471-2164-14-1.

Brenner, S., Jacob, F. and Meselson, M. (1961) ‘An unstable intermediate carrying information from genes to ribosomes for protein synthesis’, *Nature*, 190, pp. 576–581. Available at: https://doi.org/10.1038/190576a0.

Brouze, M. *et al.* (2022) ‘DIS3L, cytoplasmic exosome catalytic subunit, is essential for development but not cell viability in mice’, *bioRxiv*. Available at: https://doi.org/10.1101/2022.12.14.520403.

Bruns, I. *et al.* (2014) ‘Megakaryocytes regulate hematopoietic stem cell quiescence through CXCL4 secretion’, *Nature medicine*, 20(11), pp. 1315–1320. Available at: https://doi.org/10.1038/nm.3707.

Buchsbaum, R.J. (2007) ‘Rho activation at a glance’, *Journal of cell science*, 120(Pt 7), pp. 1149–1152. Available at: https://doi.org/10.1242/jcs.03428.

Buenrostro, J.D. *et al.* (2015) ‘Single-cell chromatin accessibility reveals principles of

regulatory variation', *Nature*, 523(7561), pp. 486–490. Available at: https://doi.org/10.1038/nature14590.

Bujanover, N. *et al.* (2018) 'Identification of immune-activated hematopoietic stem cells', *Leukemia*, 32(9), pp. 2016–2020. Available at: https://doi.org/10.1038/s41375-018-0220-z.

Bulger, M. and Groudine, M. (2011) 'Functional and mechanistic diversity of distal transcription enhancers', *Cell*, 144(3), pp. 327–339. Available at: https://doi.org/10.1016/j.cell.2011.01.024.

Burns, C.E. *et al.* (2005) 'Hematopoietic stem cell fate is established by the Notch-Runx pathway', *Genes & development*, 19(19), pp. 2331–2342. Available at: https://doi.org/10.1101/gad.1337005.

Butler, A. *et al.* (2018) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature biotechnology*, 36(5), pp. 411–420. Available at: https://doi.org/10.1038/nbt.4096.

Byrne, A. *et al.* (2019) 'Realizing the potential of full-length transcriptome sequencing', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 374(1786), p. 20190097. Available at: https://doi.org/10.1098/rstb.2019.0097.

Cabezas-Wallscheid, N. *et al.* (2014) 'Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis', *Cell stem cell*, 15(4), pp. 507–522. Available at: https://doi.org/10.1016/j.stem.2014.07.005.

Cancelas, J.A. and Williams, D.A. (2009) 'Rho GTPases in hematopoietic stem cell functions', *Current opinion in hematology*, 16(4), pp. 249–254. Available at: https://doi.org/10.1097/MOH.0b013e32832c4b80.

Cantor, A.B. and Orkin, S.H. (2001) 'Hematopoietic development: a balancing act', *Current opinion in genetics & development*, 11(5), pp. 513–519. Available at: https://doi.org/10.1016/s0959-437x(00)00226-4.

Cantor, A.B. and Orkin, S.H. (2002) 'Transcriptional regulation of erythropoiesis: an affair involving multiple partners', *Oncogene*, 21(21), pp. 3368–3376. Available at: https://doi.org/10.1038/sj.onc.1205326.

Cao, J. *et al.* (2019) 'The single-cell transcriptional landscape of mammalian organogenesis', *Nature*, 566(7745), pp. 496–502. Available at: https://doi.org/10.1038/s41586-019-0969-x.

Carrelha, J. *et al.* (2018) 'Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells', *Nature*, 554(7690), pp. 106–111. Available at: https://doi.org/10.1038/nature25455.

Carter, R.N. *et al.* (2006) 'Molecular and electrophysiological characterization of transient receptor potential ion channels in the primary murine megakaryocyte', *The Journal of physiology*, 576(Pt 1), pp. 151–162. Available at: https://doi.org/10.1113/jphysiol.2006.113886.

Cazzola, M., Della Porta, M.G. and Malcovati, L. (2013) 'The genetic basis of myelodysplasia and its clinical relevance', *Blood*, 122(25), pp. 4021–4034. Available at: https://doi.org/10.1182/blood-2013-09-381665.

Cebrián, M. *et al.* (1988) 'Triggering of T cell proliferation through AIM, an activation inducer molecule expressed on activated human lymphocytes', *The Journal of experimental medicine*, 168(5), pp. 1621–1637. Available at: https://doi.org/10.1084/jem.168.5.1621.

Ceredig, R., Rolink, A.G. and Brown, G. (2009) 'Models of haematopoiesis: seeing the wood for the trees', *Nature reviews. Immunology*, 9(4), pp. 293–300. Available at:

https://doi.org/10.1038/nri2525.

Challen, G.A. *et al.* (2009) ‘Mouse hematopoietic stem cell identification and analysis’, *Cytometry. Part A: the journal of the International Society for Analytical Cytology*, 75(1), pp. 14–24. Available at: https://doi.org/10.1002/cyto.a.20674.

Challen, G.A. *et al.* (2010) ‘Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1’, *Cell stem cell*, 6(3), pp. 265–278. Available at: https://doi.org/10.1016/j.stem.2010.02.002.

Chambers, S.M. *et al.* (2007) ‘Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation’, *PLoS biology*, 5(8), p. e201. Available at: https://doi.org/10.1371/journal.pbio.0050201.

Chatterjee, S. *et al.* (2010) ‘Primitive Sca-1 Positive Bone Marrow HSC in Mouse Model of Aplastic Anemia: A Comparative Study through Flowcytometric Analysis and Scanning Electron Microscopy’, *Stem cells international*, 2010, p. 614395. Available at: https://doi.org/10.4061/2010/614395.

Chen, H. *et al.* (2019) ‘Long-Read RNA Sequencing Identifies Alternative Splice Variants in Hepatocellular Carcinoma and Tumor-Specific Isoforms’, *Hepatology* , 70(3), pp. 1011–1025. Available at: https://doi.org/10.1002/hep.30500.

Chen, J. *et al.* (2023) ‘Cell Cycle-Related Gene SPC24: A Novel Potential Diagnostic and Prognostic Biomarker for Laryngeal Squamous Cell Cancer’, *BioMed research international*, 2023, p. 1733100. Available at: https://doi.org/10.1155/2023/1733100.

Chen, L. *et al.* (2014) ‘Transcriptional diversity during lineage commitment of human blood progenitors’, *Science*, 345(6204), p. 1251033. Available at: https://doi.org/10.1126/science.1251033.

Chen, S. *et al.* (2018) ‘IGF-1 facilitates thrombopoiesis primarily through Akt activation’, *Blood*, 132(2), pp. 210–222. Available at: https://doi.org/10.1182/blood-2018-01-825927.

Chen, Z., Hu, M. and Shivdasani, R.A. (2007) ‘Expression analysis of primary mouse megakaryocyte differentiation and its application in identifying stage-specific molecular markers and a novel transcriptional target of NF-E2’, *Blood*, 109(4), pp. 1451–1459. Available at: https://doi.org/10.1182/blood-2006-08-038901.

Choi, K. *et al.* (1998) ‘A common precursor for hematopoietic and endothelial cells’, *Development* , 125(4), pp. 725–732. Available at: https://doi.org/10.1242/dev.125.4.725.

Cho, R.H., Sieburg, H.B. and Muller-Sieburg, C.E. (2008) ‘A new mechanism for the aging of hematopoietic stem cells: aging changes the clonal composition of the stem cell compartment but not individual stem cells’, *Blood*, pp. 5553–5561. Available at: https://doi.org/10.1182/blood-2007-11-123547.

Chou, S.T. *et al.* (2009) ‘Graded repression of PU.1/Sfpi1 gene transcription by GATA factors regulates hematopoietic cell fate’, *Blood*, 114(5), pp. 983–994. Available at: https://doi.org/10.1182/blood-2009-03-207944.

Chow, L.T. *et al.* (1977) ‘An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA’, *Cell*, 12(1), pp. 1–8. Available at: https://doi.org/10.1016/0092-8674(77)90180-5.

Cimmino, G. *et al.* (2015) ‘Activating stimuli induce platelet microRNA modulation and proteome reorganisation’, *Thrombosis and haemostasis*, 114(1), pp. 96–108. Available at: https://doi.org/10.1160/TH14-09-0726.

Cobb, M. (2015) ‘Who discovered messenger RNA?’, *Current biology: CB*, 25(13), pp. R526–32. Available at: https://doi.org/10.1016/j.cub.2015.05.032.

Cock, P.J.A. *et al.* (2009) ‘Biopython: freely available Python tools for computational molecular biology and bioinformatics’, *Bioinformatics* , 25(11), pp. 1422–1423. Available at: https://doi.org/10.1093/bioinformatics/btp163.

Cocquet, J. *et al.* (2006) ‘Reverse transcriptase template switching and false alternative transcripts’, *Genomics*, 88(1), pp. 127–131. Available at: https://doi.org/10.1016/j.ygeno.2005.12.013.

Coers, J., Ranft, C. and Skoda, R.C. (2004) ‘A truncated isoform of c-Mpl with an essential C-terminal peptide targets the full-length receptor for degradation’, *The Journal of biological chemistry*, 279(35), pp. 36397–36404. Available at: https://doi.org/10.1074/jbc.M401386200.

Collin, M., Dickinson, R. and Bigley, V. (2015) ‘Haematopoietic and immune defects associated with GATA2 mutation’, *British journal of haematology*, 169(2), pp. 173–187. Available at: https://doi.org/10.1111/bjh.13317.

Collins, T. *et al.* (1995) ‘Transcriptional regulation of endothelial cell adhesion molecules: NF-kappa B and cytokine-inducible enhancers’, *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 9(10), pp. 899–909. Available at: https://doi.org/10.1096/fasebj.9.10.7542214.

Comer, S.P. (2021) ‘Turning Platelets Off and On: Role of RhoGAPs and RhoGEFs in Platelet Activity’, *Frontiers in cardiovascular medicine*, 8, p. 820945. Available at: https://doi.org/10.3389/fcvm.2021.820945.

Condamine, T. *et al.* (2010) ‘Tmem176B and Tmem176A are associated with the immature state of dendritic cells’, *Journal of leukocyte biology*, 88(3), pp. 507–515. Available at: https://doi.org/10.1189/jlb.1109738.

Copley, M.R., Beer, P.A. and Eaves, C.J. (2012) ‘Hematopoietic stem cell heterogeneity takes center stage’, *Cell stem cell*, 10(6), pp. 690–697. Available at: https://doi.org/10.1016/j.stem.2012.05.006.

Cornejo, M.G. *et al.* (2011) ‘Crosstalk between NOTCH and AKT signaling during murine megakaryocyte lineage specification’, *Blood*, 118(5), pp. 1264–1273. Available at: https://doi.org/10.1182/blood-2011-01-328567.

Crick, F.H. (1958) ‘On protein synthesis’, *Symposia of the Society for Experimental Biology*, 12, pp. 138–163. Available at: https://www.ncbi.nlm.nih.gov/pubmed/13580867.

Crick, F.H.C. (no date) ‘On protein synthesis; 1957’, *Manuscript. Cold Spring Harbor Laboratory Archives* [Preprint].

Dahlin, J.S. *et al.* (2018) ‘A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice’, *Blood*, 131(21), pp. e1–e11. Available at: https://doi.org/10.1182/blood-2017-12-821413.

Dai, J., Sullivan, B.A. and Higgins, J.M.G. (2006) ‘Regulation of mitotic chromosome cohesion by Haspin and Aurora B’, *Developmental cell*, 11(5), pp. 741–750. Available at: https://doi.org/10.1016/j.devcel.2006.09.018.

Danecek, P. *et al.* (2021) ‘Twelve years of SAMtools and BCFtools’, *GigaScience*, 10(2). Available at: https://doi.org/10.1093/gigascience/giab008.

Darmanis, S. *et al.* (2016) ‘Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells’, *Cell reports*, 14(2), pp. 380–389. Available at:

https://doi.org/10.1016/j.celrep.2015.12.021.

Darnell, J.E., Jr (1978) ‘Implications of RNA-RNA splicing in evolution of eukaryotic cells’, *Science*, 202(4374), pp. 1257–1260. Available at: https://doi.org/10.1126/science.364651.

Davidson, E.H. (2010) *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Elsevier. Available at: https://play.google.com/store/books/details?id=F2ibJj1LHGEC.

Davis, E. *et al.* (1992) ‘Histologic studies of splenic megakaryocytes after bone marrow ablation with strontium 90’, *The Journal of laboratory and clinical medicine*, 120(5), pp. 767–777. Available at: https://www.ncbi.nlm.nih.gov/pubmed/1431506.

Debili, N. *et al.* (1991) ‘In vitro effects of hematopoietic growth factors on the proliferation, endoreplication, and maturation of human megakaryocytes’, *Blood*, 77(11), pp. 2326–2338. Available at: https://www.ncbi.nlm.nih.gov/pubmed/2039816.

Debili, N. *et al.* (1995) ‘The Mpl receptor is expressed in the megakaryocytic lineage from late progenitors to platelets’, *Blood*, 85(2), pp. 391–401. Available at: https://www.ncbi.nlm.nih.gov/pubmed/7529061.

Debili, N. *et al.* (1996) ‘Characterization of a bipotent erythro-megakaryocytic progenitor in human bone marrow’, *Blood*, 88(4), pp. 1284–1296. Available at: https://www.ncbi.nlm.nih.gov/pubmed/8695846.

Debili, N. *et al.* (2001) ‘Different expression of CD41 on human lymphoid and myeloid progenitors from adults and neonates’, *Blood*, 97(7), pp. 2023–2030. Available at: https://doi.org/10.1182/blood.v97.7.2023.

Delaney, C., Gutman, J.A. and Appelbaum, F.R. (2009) ‘Cord blood transplantation for haematological malignancies: conditioning regimens, double cord transplant and infectious complications’, *British journal of haematology*, 147(2), pp. 207–216. Available at: https://doi.org/10.1111/j.1365-2141.2009.07782.x.

Delgado, M.D. and León, J. (2010) ‘Myc roles in hematopoiesis and leukemia’, *Genes & cancer*, 1(6), pp. 605–616. Available at: https://doi.org/10.1177/1947601910377495.

Deng, P. *et al.* (2021) ‘Increased Expression of KNSTRN in Lung Adenocarcinoma Predicts Poor Prognosis: A Bioinformatics Analysis Based on TCGA Data’, *Journal of Cancer*, 12(11), pp. 3239–3248. Available at: https://doi.org/10.7150/jca.51591.

De Paoli-Iseppi, R., Gleeson, J. and Clark, M.B. (2021) ‘Isoform Age - Splice Isoform Profiling Using Long-Read Technologies’, *Frontiers in molecular biosciences*, 8, p. 711733. Available at: https://doi.org/10.3389/fmolb.2021.711733.

Desai, P. *et al.* (2018) ‘Somatic mutations precede acute myeloid leukemia years before diagnosis’, *Nature medicine*, 24(7), pp. 1015–1023. Available at: https://doi.org/10.1038/s41591-018-0081-z.

Desterke, C., Bennaceur-Griscelli, A. and Turhan, A.G. (2021) ‘EGR1 dysregulation defines an inflammatory and leukemic program in cell trajectory of human-aged hematopoietic stem cells (HSC)’, *Stem cell research & therapy*, 12(1), p. 419. Available at: https://doi.org/10.1186/s13287-021-02498-0.

Dias, C. *et al.* (2016) ‘BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription’, *American journal of human genetics*, 99(2), pp. 253–274. Available at: https://doi.org/10.1016/j.ajhg.2016.05.030.

Dobin, A. *et al.* (2013) ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics* , 29(1),

pp. 15–21. Available at: https://doi.org/10.1093/bioinformatics/bts635.

Dolznig, H. *et al.* (2006) ‘Erythroid progenitor renewal versus differentiation: genetic evidence for cell autonomous, essential functions of EpoR, Stat5 and the GR’, *Oncogene*, 25(20), pp. 2890–2900. Available at: https://doi.org/10.1038/sj.onc.1209308.

Donato, J.L. *et al.* (2002) ‘Human HTm4 is a hematopoietic cell cycle regulator’, *The Journal of clinical investigation*, 109(1), pp. 51–58. Available at: https://doi.org/10.1172/JCI14025.

Doré, L.C. *et al.* (2012) ‘Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis’, *Blood*, 119(16), pp. 3724–3733. Available at: https://doi.org/10.1182/blood-2011-09-380634.

Doré, L.C. and Crispino, J.D. (2011) ‘Transcription factor networks in erythroid cell and megakaryocyte development’, *Blood*, 118(2), pp. 231–239. Available at: https://doi.org/10.1182/blood-2011-04-285981.

Dorshkind, K. *et al.* (2020) ‘Do haematopoietic stem cells age?’, *Nature reviews. Immunology*, 20(3), pp. 196–202. Available at: https://doi.org/10.1038/s41577-019-0236-2.

Doulatov, S. *et al.* (2010) ‘Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development’, *Nature immunology*, 11(7), pp. 585–593. Available at: https://doi.org/10.1038/ni.1889.

Doulatov, S. *et al.* (2012) ‘Hematopoiesis: a human perspective’, *Cell stem cell*, 10(2), pp. 120–136. Available at: https://doi.org/10.1016/j.stem.2012.01.006.

Dütting, S. *et al.* (2017) ‘A Cdc42/RhoA regulatory circuit downstream of glycoprotein Ib guides transendothelial platelet biogenesis’, *Nature communications*, 8, p. 15838. Available at: https://doi.org/10.1038/ncomms15838.

Dykstra, B. *et al.* (2007) ‘Long-term propagation of distinct hematopoietic differentiation programs in vivo’, *Cell stem cell*, 1(2), pp. 218–229. Available at: https://doi.org/10.1016/j.stem.2007.05.015.

Dykstra, B. *et al.* (2011) ‘Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells’, *The Journal of experimental medicine*, 208(13), pp. 2691–2703. Available at: https://doi.org/10.1084/jem.20111490.

Early, P. *et al.* (1980) ‘An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH’, *Cell*, 19(4), pp. 981–992. Available at: https://doi.org/10.1016/0092-8674(80)90089-6.

Ebaugh, F.G., Jr and Bird, R.M. (1951) ‘The normal megakaryocyte concentration in aspirated human bone marrow’, *Blood*, 6(1), pp. 75–80. Available at: https://doi.org/10.1182/blood.V6.1.75.75.

Eckly, Anita et al. “Biogenesis of the demarcation membrane system (DMS) in megakaryocytes.” *Blood* vol. 123,6 (2014): 921-30. doi:10.1182/blood-2013-03-492330.

Edfors, F. *et al.* (2016) ‘Gene-specific correlation of RNA and protein levels in human cells and tissues’, *Molecular systems biology*, 12(10), p. 883. Available at: https://doi.org/10.15252/msb.20167144.

Edwards, C.R. *et al.* (2016) ‘A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages’, *Blood*, 127(17), pp. e24–e34. Available at: https://doi.org/10.1182/blood-2016-01-692764.

Eisenstaedt, R., Penninx, B.W.J.H. and Woodman, R.C. (2006) ‘Anemia in the elderly: current

understanding and emerging concepts', *Blood reviews*, 20(4), pp. 213–226. Available at: https://doi.org/10.1016/j.blre.2005.12.002.

Eliades, A., Papadantonakis, N. and Ravid, K. (2010) 'New roles for cyclin E in megakaryocytic polyploidization', *The Journal of biological chemistry*, 285(24), pp. 18909–18917. Available at: https://doi.org/10.1074/jbc.M110.102145.

Elizondo, D.M. *et al.* (2019) 'Allograft Inflammatory Factor-1 Governs Hematopoietic Stem Cell Differentiation Into cDC1 and Monocyte-Derived Dendritic Cells Through IRF8 and RelB in vitro', *Frontiers in immunology*, 10, p. 173. Available at: https://doi.org/10.3389/fimmu.2019.00173.

Eon Kuek, L. *et al.* (2016) 'The MS4A family: counting past 1, 2 and 3', *Immunology and cell biology*, 94(1), pp. 11–23. Available at: https://doi.org/10.1038/icb.2015.48.

Estevez, B. *et al.* (2021) 'RUNX-1 haploinsufficiency causes a marked deficiency of megakaryocyte-biased hematopoietic progenitor cells', *Blood*, 137(19), pp. 2662–2675. Available at: https://doi.org/10.1182/blood.2020006389.

Ewels, P. *et al.* (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics* , 32(19), pp. 3047–3048. Available at: https://doi.org/10.1093/bioinformatics/btw354.

Favara, D.M. *et al.* (2019) 'ADGRL4/ELTD1 Silencing in Endothelial Cells Induces ACLY and SLC25A1 and Alters the Cellular Metabolic Profile', *Metabolites*, 9(12). Available at: https://doi.org/10.3390/metabo9120287.

Federzoni, E.A. *et al.* (2012) 'PU.1 is linking the glycolytic enzyme HK3 in neutrophil differentiation and survival of APL cells', *Blood*, 119(21), pp. 4963–4970. Available at: https://doi.org/10.1182/blood-2011-09-378117.

Felsenfeld, G. and Groudine, M. (2003) 'Controlling the double helix', *Nature*, 421(6921), pp. 448–453. Available at: https://doi.org/10.1038/nature01411.

Ferreira, R. *et al.* (2005) 'GATA1 function, a paradigm for transcription factors in hematopoiesis', *Molecular and cellular biology*, 25(4), pp. 1215–1227. Available at: https://doi.org/10.1128/MCB.25.4.1215-1227.2005.

Fielder, P.J. *et al.* (1997) 'Human platelets as a model for the binding and degradation of thrombopoietin', *Blood*, 89(8), pp. 2782–2788. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9108396.

Flach, J. *et al.* (2014) 'Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells', *Nature*, 512(7513), pp. 198–202. Available at: https://doi.org/10.1038/nature13619.

Fliedner, T.M. *et al.* (2002) 'Structure and function of bone marrow hemopoiesis: mechanisms of response to ionizing radiation exposure', *Cancer biotherapy & radiopharmaceuticals*, 17(4), pp. 405–426. Available at: https://doi.org/10.1089/108497802760363204.

Flohr Svendsen, A. *et al.* (2021) 'A comprehensive transcriptome signature of murine hematopoietic stem cell aging', *Blood*, 138(6), pp. 439–451. Available at: https://doi.org/10.1182/blood.2020009729.

Florian, M.C. *et al.* (2012) 'Cdc42 activity regulates hematopoietic stem cell aging and rejuvenation', *Cell stem cell*, 10(5), pp. 520–530. Available at: https://doi.org/10.1016/j.stem.2012.04.007.

Florian, M.C. *et al.* (2018) 'Aging alters the epigenetic asymmetry of HSC division', *PLoS*

biology, 16(9), p. e2003389. Available at: https://doi.org/10.1371/journal.pbio.2003389.

Föger, N. *et al.* (2006) 'Requirement for coronin 1 in T lymphocyte trafficking and cellular homeostasis', *Science*, 313(5788), pp. 839–842. Available at: https://doi.org/10.1126/science.1130563.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic acids research*, 47(D1), pp. D766–D773. Available at: https://doi.org/10.1093/nar/gky955.

Frankish, A. *et al.* (2020) 'GENCODE 2021', *Nucleic acids research*, 49(D1), pp. D916–D923. Available at: https://doi.org/10.1093/nar/gkaa1087.

Freedman, J.E. (2011) 'A platelet transcriptome revolution', *Blood*, pp. 3760–3761. Available at: https://doi.org/10.1182/blood-2011-05-356600.

Frisch, B.J. *et al.* (2019) 'Aged marrow macrophages expand platelet-biased hematopoietic stem cells via Interleukin1B', *JCI insight*, 5(10). Available at: https://doi.org/10.1172/jci.insight.124213.

Frontelo, P. *et al.* (2007) 'Novel role for EKLF in megakaryocyte lineage commitment', *Blood*, 110(12), pp. 3871–3880. Available at: https://doi.org/10.1182/blood-2007-03-082065.

Gallagher, P.G. *et al.* (2010) 'Mutation of a barrier insulator in the human ankyrin-1 gene is associated with hereditary spherocytosis', *The Journal of clinical investigation*, 120(12), pp. 4453–4465. Available at: https://doi.org/10.1172/JCI42240.

Galloway, J.L. *et al.* (2005) 'Loss of gata1 but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos', *Developmental cell*, 8(1), pp. 109–116. Available at: https://doi.org/10.1016/j.devcel.2004.12.001.

Galloway, J.L. and Zon, L.I. (2003) 'Ontogeny of hematopoiesis: examining the emergence of hematopoietic cells in the vertebrate embryo', *Current topics in developmental biology*, 53, pp. 139–158. Available at: https://doi.org/10.1016/s0070-2153(03)53004-6.

Gazit, R. *et al.* (2014) 'Fgd5 identifies hematopoietic stem cells in the murine bone marrow', *The Journal of experimental medicine*, 211(7), pp. 1315–1331. Available at: https://doi.org/10.1084/jem.20130428.

Geddis, A.E. and Kaushansky, K. (2004) 'Megakaryocytes express functional Aurora-B kinase in endomitosis', *Blood*, 104(4), pp. 1017–1024. Available at: https://doi.org/10.1182/blood-2004-02-0419.

Geiger, H., de Haan, G. and Carolina Florian, M. (2013) 'The ageing haematopoietic stem cell compartment', *Nature Reviews Immunology*, pp. 376–389. Available at: https://doi.org/10.1038/nri3433.

Gekas, C. *et al.* (2009) 'Mef2C is a lineage-restricted target of Scl/Tal1 and regulates megakaryopoiesis and B-cell homeostasis', *Blood*, 113(15), pp. 3461–3471. Available at: https://doi.org/10.1182/blood-2008-07-167577.

Gekas, C. and Graf, T. (2013) 'CD41 expression marks myeloid-biased adult hematopoietic stem cells and increases with age', *Blood*, 121(22), pp. 4463–4472. Available at: https://doi.org/10.1182/blood-2012-09-457929.

Genest, D.S. *et al.* (2023) 'Renal Thrombotic Microangiopathy: A Review', *American journal of kidney diseases: the official journal of the National Kidney Foundation*, 81(5), pp. 591–605. Available at: https://doi.org/10.1053/j.ajkd.2022.10.014.

Geng, Y. *et al.* (2003) 'Cyclin E ablation in the mouse', *Cell*, 114(4), pp. 431–443. Available at: https://doi.org/10.1016/s0092-8674(03)00645-7.

Genovese, G. *et al.* (2014) 'Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence', *The New England journal of medicine*, 371(26), pp. 2477–2487. Available at: https://doi.org/10.1056/NEJMoa1409405.

George, J.N. and Nester, C.M. (2014) 'Syndromes of thrombotic microangiopathy', *The New England journal of medicine*, 371(7), pp. 654–666. Available at: https://doi.org/10.1056/NEJMra1312353.

Geue, Sascha et al. "Pivotal role of PDK1 in megakaryocyte cytoskeletal dynamics and polarization during platelet biogenesis." *Blood* vol. 134,21 (2019): 1847-1858. doi:10.1182/blood.2019000185.

Ghalloussi, D., Dhenge, A. and Bergmeier, W. (2019) 'New insights into cytoskeletal remodeling during platelet production', *Journal of thrombosis and haemostasis: JTH*, 17(9), pp. 1430–1439. Available at: https://doi.org/10.1111/jth.14544.

Giladi, A. *et al.* (2018) 'Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis', *Nature cell biology*, 20(7), pp. 836–846. Available at: https://doi.org/10.1038/s41556-018-0121-4.

Gilbert, W. (1978) 'Why genes in pieces?', *Nature*, 271(5645), p. 501. Available at: https://doi.org/10.1038/271501a0.

Gillespie, M. *et al.* (2022) 'The reactome pathway knowledgebase 2022', *Nucleic acids research*, 50(D1), pp. D687–D692. Available at: https://doi.org/10.1093/nar/gkab1028.

Ginsberg, D. (2002) 'E2F1 pathways to apoptosis', *FEBS letters*, 529(1), pp. 122–125. Available at: https://doi.org/10.1016/s0014-5793(02)03270-2.

Gnatenko, D.V. *et al.* (2003) 'Transcript profiling of human platelets using microarray and serial analysis of gene expression', *Blood*, 101(6), pp. 2285–2293. Available at: https://doi.org/10.1182/blood-2002-09-2797.

Goldenson, B. *et al.* (2015) 'Aurora kinase A is required for hematopoiesis but is dispensable for murine megakaryocyte endomitosis and differentiation', *Blood*, 125(13), pp. 2141–2150. Available at: https://doi.org/10.1182/blood-2014-12-615401.

Goldstein, O. *et al.* (2017) 'Mapping Whole-Transcriptome Splicing in Mouse Hematopoietic Stem Cells', *Stem cell reports*, 8(1), pp. 163–176. Available at: https://doi.org/10.1016/j.stemcr.2016.12.002.

Gong, Y. *et al.* (2018) 'Megakaryocyte-derived excessive transforming growth factor β1 inhibits proliferation of normal hematopoietic stem cells in acute myeloid leukemia', *Experimental hematology*, 60, pp. 40–46.e2. Available at: https://doi.org/10.1016/j.exphem.2017.12.010.

González, C. *et al.* (1964) 'Multiple molecular forms of ATP:hexose 6-phosphotransferase from rat liver', *Biochemical and biophysical research communications*, 16(4), pp. 347–352. Available at: https://doi.org/10.1016/0006-291x(64)90038-5.

Goyama, S. *et al.* (2008) 'Evi-1 is a critical regulator for hematopoietic stem cells and transformed leukemic cells', *Cell stem cell*, 3(2), pp. 207–220. Available at: https://doi.org/10.1016/j.stem.2008.06.002.

Graubert, T.A. *et al.* (2011) 'Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes', *Nature genetics*, 44(1), pp. 53–57. Available at:

https://doi.org/10.1038/ng.1031.

Greenberg, Z.J. *et al.* (2023) 'The tetraspanin CD53 protects stressed hematopoietic stem cells via promotion of DREAM complex-mediated quiescence', *Blood*, 141(10), pp. 1180–1193. Available at: https://doi.org/10.1182/blood.2022016929.

Grinenko, T. *et al.* (2018) 'Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice', *Nature communications*, 9(1), p. 1898. Available at: https://doi.org/10.1038/s41467-018-04188-7.

Gros, F. *et al.* (1961) 'Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli', *Nature*, 190, pp. 581–585. Available at: https://doi.org/10.1038/190581a0.

Grosveld, F. *et al.* (1987) 'Position-independent, high-level expression of the human beta-globin gene in transgenic mice', *Cell*, 51(6), pp. 975–985. Available at: https://doi.org/10.1016/0092-8674(87)90584-8.

Grover, A. *et al.* (2016) 'Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells', *Nature communications*, 7, p. 11075. Available at: https://doi.org/10.1038/ncomms11075.

Growney, J.D. *et al.* (2005) 'Loss of Runx1 perturbs adult hematopoiesis and is associated with a myeloproliferative phenotype', *Blood*, 106(2), pp. 494–504. Available at: https://doi.org/10.1182/blood-2004-08-3280.

Grozovsky, R. *et al.* (2015) 'Regulating billions of blood platelets: glycans and beyond', *Blood*, 126(16), pp. 1877–1884. Available at: https://doi.org/10.1182/blood-2015-01-569129.

Grzywa, T.M., Nowis, D. and Golab, J. (2021) 'The role of CD71+ erythroid cells in the regulation of the immune response', *Pharmacology & therapeutics*, 228, p. 107927. Available at: https://doi.org/10.1016/j.pharmthera.2021.107927.

Gupta, I. *et al.* (2018) 'Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells', *Nature biotechnology* [Preprint]. Available at: https://doi.org/10.1038/nbt.4259.

Gutiérrez, L. *et al.* (2020) 'Regulation of GATA1 levels in erythropoiesis', *IUBMB life*, 72(1), pp. 89–105. Available at: https://doi.org/10.1002/iub.2192.

Haas, S. *et al.* (2015) 'Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors', *Cell stem cell*, 17(4), pp. 422–434. Available at: https://doi.org/10.1016/j.stem.2015.07.007.

Hackl, H. *et al.* (2015) 'A gene expression profile associated with relapse of cytogenetically normal acute myeloid leukemia is enriched for leukemia stem cell genes', *Leukemia & lymphoma*, 56(4), pp. 1126–1128. Available at: https://doi.org/10.3109/10428194.2014.944523.

Haferlach, T. *et al.* (2014) 'Landscape of genetic lesions in 944 patients with myelodysplastic syndromes', *Leukemia*, 28(2), pp. 241–247. Available at: https://doi.org/10.1038/leu.2013.336.

Hagemann-Jensen, M. *et al.* (2020) 'Single-cell RNA counting at allele and isoform resolution using Smart-seq3', *Nature biotechnology*, 38(6), pp. 708–714. Available at: https://doi.org/10.1038/s41587-020-0497-0.

Haghverdi, L. *et al.* (2018) 'Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors', *Nature biotechnology*, 36(5), pp. 421–427. Available at: https://doi.org/10.1038/nbt.4091.

Haghverdi, L. and Ludwig, L.S. (2023) 'Single-cell multi-omics and lineage tracing to dissect

cell fate decision-making', *Stem cell reports*, 18(1), pp. 13–25. Available at: https://doi.org/10.1016/j.stemcr.2022.12.003.

Halene, S. *et al.* (2010) 'Serum response factor is an essential transcription factor in megakaryocytic maturation', *Blood*, 116(11), pp. 1942–1950. Available at: https://doi.org/10.1182/blood-2010-01-261743.

Hao, Y. *et al.* (2021) 'Integrated analysis of multimodal single-cell data', *Cell*, 184(13), pp. 3573–3587.e29. Available at: https://doi.org/10.1016/j.cell.2021.04.048.

Hassock, S.R. *et al.* (2002) 'Expression and role of TRPC proteins in human platelets: evidence that TRPC6 forms the store-independent calcium entry channel', *Blood*, 100(8), pp. 2801–2811. Available at: https://doi.org/10.1182/blood-2002-03-0723.

Hayashi, K. *et al.* (2008) 'Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states', *Cell stem cell*, 3(4), pp. 391–401. Available at: https://doi.org/10.1016/j.stem.2008.07.027.

Heaney, M.L. and Golde, D.W. (1998) 'Soluble receptors in human disease', *Journal of leukocyte biology*, 64(2), pp. 135–146. Available at: https://doi.org/10.1002/jlb.64.2.135.

Heib, T. *et al.* (2021) 'RhoA/Cdc42 signaling drives cytoplasmic maturation but not endomitosis in megakaryocytes', *Cell reports*, 35(6), p. 109102. Available at: https://doi.org/10.1016/j.celrep.2021.109102.

Hermiston, M.L., Xu, Z. and Weiss, A. (2003) 'CD45: a critical regulator of signaling thresholds in immune cells', *Annual review of immunology*, 21, pp. 107–137. Available at: https://doi.org/10.1146/annurev.immunol.21.120601.140946.

Heuston, E.F. *et al.* (2018) 'Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points', *Epigenetics & chromatin*, 11(1), p. 22. Available at: https://doi.org/10.1186/s13072-018-0195-z.

He, Z. *et al.* (2020) 'Hyaluronan Mediated Motility Receptor (HMMR) Encodes an Evolutionarily Conserved Homeostasis, Mitosis, and Meiosis Regulator Rather than a Hyaluronan Receptor', *Cells* , 9(4). Available at: https://doi.org/10.3390/cells9040819.

Hicks, S.C. *et al.* (2018) 'Missing data and technical variability in single-cell RNA-sequencing experiments', *Biostatistics* , 19(4), pp. 562–578. Available at: https://doi.org/10.1093/biostatistics/kxx053.

Higgins, J.M.G. (2010) 'Haspin: a newly discovered regulator of mitotic chromosome behavior', *Chromosoma*, 119(2), pp. 137–147. Available at: https://doi.org/10.1007/s00412-009-0250-4.

Hill, M. *et al.* (2022) 'The intracellular cation channel TMEM176B as a dual immunoregulator', *Frontiers in cell and developmental biology*, 10, p. 1038429. Available at: https://doi.org/10.3389/fcell.2022.1038429.

Hock, H. *et al.* (2004) 'Gfi-1 restricts proliferation and preserves functional integrity of haematopoietic stem cells', *Nature*, 431(7011), pp. 1002–1007. Available at: https://doi.org/10.1038/nature02994.

Hodge, D. *et al.* (2006) 'A global role for EKLF in definitive and primitive erythropoiesis', *Blood*, 107(8), pp. 3359–3370. Available at: https://doi.org/10.1182/blood-2005-07-2888.

Hojka-Osinska, A. *et al.* (2021) 'Landscape of functional interactions of human processive ribonucleases revealed by high-throughput siRNA screenings', *iScience*, 24(9), p. 103036. Available at: https://doi.org/10.1016/j.isci.2021.103036.

Hoppe, P.S. *et al.* (2016) 'Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios', *Nature*, 535(7611), pp. 299–302. Available at: https://doi.org/10.1038/nature18320.

de Hostos, E.L. (2008) 'A brief history of the coronin family', *Sub-cellular biochemistry*, 48, pp. 31–40. Available at: https://doi.org/10.1007/978-0-387-09595-0_4.

Houseley, J., LaCava, J. and Tollervey, D. (2006) 'RNA-quality control by the exosome', *Nature reviews. Molecular cell biology*, 7(7), pp. 529–539. Available at: https://doi.org/10.1038/nrm1964.

Houseley, J. and Tollervey, D. (2010) 'Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro', *PloS one*, 5(8), p. e12271. Available at: https://doi.org/10.1371/journal.pone.0012271.

Ho, Y.-H. *et al.* (2019) 'Remodeling of Bone Marrow Hematopoietic Stem Cell Niches Promotes Myeloid Cell Expansion during Premature or Physiological Aging', *Cell stem cell*, 25(3), pp. 407–418.e6. Available at: https://doi.org/10.1016/j.stem.2019.06.007.

Huang, H. and Cantor, A.B. (2009) 'Common features of megakaryocytes and hematopoietic stem cells: what's the connection?', *Journal of cellular biochemistry*, 107(5), pp. 857–864. Available at: https://doi.org/10.1002/jcb.22184.

Huang, M. *et al.* (2020) 'Genome-wide CRISPR screen uncovers a synergistic effect of combining Haspin and Aurora kinase B inhibition', *Oncogene*, 39(21), pp. 4312–4322. Available at: https://doi.org/10.1038/s41388-020-1296-2.

Huber, W. *et al.* (2015) 'Orchestrating high-throughput genomic analysis with Bioconductor', *Nature methods*, 12(2), pp. 115–121. Available at: https://doi.org/10.1038/nmeth.3252.

Hwang, Y. *et al.* (2017) 'Global increase in replication fork speed during a p57KIP2-regulated erythroid cell fate switch', *Science advances*, 3(5), p. e1700298. Available at: https://doi.org/10.1126/sciadv.1700298.

Ikonomi, P., Noguchi, C.T., *et al.* (2000) 'Levels of GATA-1/GATA-2 transcription factors modulate expression of embryonic and fetal hemoglobins', *Gene*, 261(2), pp. 277–287. Available at: https://doi.org/10.1016/s0378-1119(00)00510-2.

Ikonomi, P., Rivera, C.E., *et al.* (2000) 'Overexpression of GATA-2 inhibits erythroid and promotes megakaryocyte differentiation', *Experimental hematology*, 28(12), pp. 1423–1431. Available at: https://doi.org/10.1016/s0301-472x(00)00553-1.

International Human Genome Sequencing Consortium (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, 431(7011), pp. 931–945. Available at: https://doi.org/10.1038/nature03001.

Ishibashi, T. *et al.* (2016) 'ESAM is a novel human hematopoietic stem cell marker associated with a subset of human leukemias', *Experimental hematology*, 44(4), pp. 269–81.e1. Available at: https://doi.org/10.1016/j.exphem.2015.12.010.

Italiano, J.E., Jr *et al.* (1999) 'Blood platelets are assembled principally at the ends of proplatelet processes produced by differentiated megakaryocytes', *The Journal of cell biology*, 147(6), pp. 1299–1312. Available at: https://doi.org/10.1083/jcb.147.6.1299.

Iturri, L. *et al.* (2021) 'Megakaryocyte production is sustained by direct differentiation from erythromyeloid progenitors in the yolk sac until midgestation', *Immunity*, 54(7), pp. 1433–1446.e5. Available at: https://doi.org/10.1016/j.immuni.2021.04.026.

Ivanovs, A. *et al.* (2011) 'Highly potent human hematopoietic stem cells first emerge in the

intraembryonic aorta-gonad-mesonephros region', *The Journal of experimental medicine*, 208(12), pp. 2417–2427. Available at: https://doi.org/10.1084/jem.20111688.

Ivanovs, A. *et al.* (2014) 'Identification of the niche and phenotype of the first human hematopoietic stem cells', *Stem cell reports*, 2(4), pp. 449–456. Available at: https://doi.org/10.1016/j.stemcr.2014.02.004.

Jackson, B. *et al.* (2011) 'Update on the aldehyde dehydrogenase gene (ALDH) superfamily', *Human genomics*, 5(4), pp. 283–303. Available at: https://doi.org/10.1186/1479-7364-5-4-283.

Jansen, M. *et al.* (2005) 'Rac2-deficient hematopoietic stem cells show defective interaction with the hematopoietic microenvironment and long-term engraftment failure', *Stem cells*, 23(3), pp. 335–346. Available at: https://doi.org/10.1634/stemcells.2004-0216.

Joglekar, A. *et al.* (2021) 'A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain', *Nature communications*, 12(1), p. 463. Available at: https://doi.org/10.1038/s41467-020-20343-5.

Johnson, K.D. *et al.* (2012) 'Cis-element mutated in GATA2-dependent immunodeficiency governs hematopoiesis and vascular integrity', *The Journal of clinical investigation*, 122(10), pp. 3692–3704. Available at: https://doi.org/10.1172/JCI61623.

de Jonckheere, B., *et al.* (2023) 'Critical shifts in lipid metabolism promote megakaryocyte differentiation and proplatelet formation.' *Nat Cardiovasc Res* 2, 835–852 (2023). https://doi.org/10.1038/s44161-023-00325-8.

Jones, C.I. (2016) 'Platelet function and ageing', *Mammalian genome: official journal of the International Mammalian Genome Society*, 27(7-8), pp. 358–366. Available at: https://doi.org/10.1007/s00335-016-9629-8.

Kan, A. *et al.* (2018) 'ELTD1 Function in Hepatocellular Carcinoma is Carcinoma-Associated Fibroblast-Dependent', *Journal of Cancer*, 9(14), pp. 2415–2427. Available at: https://doi.org/10.7150/jca.24406.

Kanaji, T. *et al.* (2018) 'Tyrosyl-tRNA synthetase stimulates thrombopoietin-independent hematopoiesis accelerating recovery from thrombocytopenia', *Proceedings of the National Academy of Sciences of the United States of America*, 115(35), pp. E8228–E8235. Available at: https://doi.org/10.1073/pnas.1807000115.

Kanehisa, M. and Goto, S. (2000) 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic acids research*, 28(1), pp. 27–30. Available at: https://doi.org/10.1093/nar/28.1.27.

Kanitz, A. *et al.* (2015) 'Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data', *Genome biology*, 16(1), p. 150. Available at: https://doi.org/10.1186/s13059-015-0702-5.

Karlsson, K. and Linnarsson, S. (2017) 'Single-cell mRNA isoform diversity in the mouse brain', *BMC genomics*, 18(1), p. 126. Available at: https://doi.org/10.1186/s12864-017-3528-6.

Kato, Y. *et al.* (2005) 'Selective activation of STAT5 unveils its role in stem cell self-renewal in normal and leukemic hematopoiesis', *The Journal of experimental medicine*, 202(1), pp. 169–179. Available at: https://doi.org/10.1084/jem.20042541.

Kaufmann, K.B. *et al.* (2019) 'A stemness screen reveals C3orf54/INKA1 as a promoter of human leukemia stem cell latency', *Blood*, 133(20), pp. 2198–2211. Available at: https://doi.org/10.1182/blood-2018-10-881441.

Kaushansky, K. and Drachman, J.G. (2002) 'The molecular and cellular biology of thrombopoietin: the primary regulator of platelet production', *Oncogene*, 21(21), pp.

3359–3367. Available at: https://doi.org/10.1038/sj.onc.1205323.

Kavanagh, K.L. *et al.* (2008) 'Medium- and short-chain dehydrogenase/reductase gene and protein families : the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes', *Cellular and molecular life sciences: CMLS*, 65(24), pp. 3895–3906. Available at: https://doi.org/10.1007/s00018-008-8588-y.

Kelly, L.M. *et al.* (2000) 'MafB is an inducer of monocytic differentiation', *The EMBO journal*, 19(9), pp. 1987–1997. Available at: https://doi.org/10.1093/emboj/19.9.1987.

Kelly, K.L. *et al.* (2020) 'De novo lipogenesis is essential for platelet production in humans', Nat Metab. 2(10), pp.1163-1178. doi: 10.1038/s42255-020-00272-9.

Kent, D.G. *et al.* (2009) 'Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential', *Blood*, 113(25), pp. 6342–6350. Available at: https://doi.org/10.1182/blood-2008-12-192054.

Kieffer, N. *et al.* (1987) 'Biosynthesis of major platelet proteins in human blood platelets', *European journal of biochemistry / FEBS*, 164(1), pp. 189–195. Available at: https://doi.org/10.1111/j.1432-1033.1987.tb11010.x.

Kiel, M.J. *et al.* (2005) 'SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells', *Cell*, 121(7), pp. 1109–1121. Available at: https://doi.org/10.1016/j.cell.2005.05.026.

Kim, K.M. *et al.* (2022) 'Taz protects hematopoietic stem cells from an aging-dependent decrease in PU.1 activity', *Nature communications*, 13(1), p. 5187. Available at: https://doi.org/10.1038/s41467-022-32970-1.

Kimura, S. *et al.* (1998) 'Hematopoietic stem cell deficiencies in mice lacking c-Mpl, the receptor for thrombopoietin', *Proceedings of the National Academy of Sciences of the United States of America*, 95(3), pp. 1195–1200. Available at: https://doi.org/10.1073/pnas.95.3.1195.

Kimura, A. *et al.* (2010) 'The gene encoding the hematopoietic stem cell regulator CCN3/NOV is under direct cytokine control through the transcription factors STAT5A/B', *The Journal of biological chemistry*, 285(43), pp. 32704–32709. Available at: https://doi.org/10.1074/jbc.M110.141804.

Kingsley, P.D. *et al.* (2013) 'Ontogeny of erythroid gene expression', *Blood*, 121(6), pp. e5–e13. Available at: https://doi.org/10.1182/blood-2012-04-422394.

Kirkwood, T.B.L. (2005) 'Understanding the odd science of aging', *Cell*, 120(4), pp. 437–447. Available at: https://doi.org/10.1016/j.cell.2005.01.027.

Kishore, U. and Reid, K.B. (2000) 'C1q: structure, function, and receptors', *Immunopharmacology*, 49(1-2), pp. 159–170. Available at: https://doi.org/10.1016/s0162-3109(00)80301-x.

Klepin, H.D. (2016) 'Myelodysplastic Syndromes and Acute Myeloid Leukemia in the Elderly', *Clinics in geriatric medicine*, 32(1), pp. 155–173. Available at: https://doi.org/10.1016/j.cger.2015.08.010.

Klimchenko, O. *et al.* (2009) 'A common bipotent progenitor generates the erythroid and megakaryocyte lineages in embryonic stem cell-derived primitive hematopoiesis', *Blood*, 114(8), pp. 1506–1517. Available at: https://doi.org/10.1182/blood-2008-09-178863.

Koide, S. *et al.* (2021) 'Unraveling Heterogeneity of Aged Hematopoietic Stem Cells By Single-Cell RNA Sequence Analysis', *Blood*, 138(Supplement 1), pp. 4299–4299. Available at: https://doi.org/10.1182/blood-2021-149876.

Komorowska, K. *et al.* (2017) 'Hepatic Leukemia Factor Maintains Quiescence of Hematopoietic Stem Cells and Protects the Stem Cell Pool during Regeneration', *Cell reports*, 21(12), pp. 3514–3523. Available at: https://doi.org/10.1016/j.celrep.2017.11.084.

Kondo, M., Weissman, I.L. and Akashi, K. (1997) 'Identification of clonogenic common lymphoid progenitors in mouse bone marrow', *Cell*, 91(5), pp. 661–672. Available at: https://doi.org/10.1016/s0092-8674(00)80453-5.

Kowalczyk, M.S. *et al.* (2015) 'Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells', *Genome research*, 25(12), pp. 1860–1872. Available at: https://doi.org/10.1101/gr.192237.115.

Krishnaraju, K., Hoffman, B. and Liebermann, D.A. (2001) 'Early growth response gene 1 stimulates development of hematopoietic progenitor cells along the macrophage lineage at the expense of the granulocyte and erythroid lineages', *Blood*, 97(5), pp. 1298–1305. Available at: https://doi.org/10.1182/blood.v97.5.1298.

Kubota, Y. *et al.* (2009) 'Necdin restricts proliferation of hematopoietic stem cells during hematopoietic regeneration', *Blood*, 114(20), pp. 4383–4392. Available at: https://doi.org/10.1182/blood-2009-07-230292.

Kustikova, O.S. *et al.* (2006) 'Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways', *Blood*, 109(5), pp. 1897–1907. Available at: https://doi.org/10.1182/blood-2006-08-044156.

Kuter, D.J. and Begley, C.G. (2002) 'Recombinant human thrombopoietin: basic biology and evaluation of clinical studies', *Blood*, 100(10), pp. 3457–3469. Available at: https://doi.org/10.1182/blood.V100.10.3457.

Kuvardina, O.N. *et al.* (2015) 'RUNX1 represses the erythroid gene expression program during megakaryocytic differentiation', *Blood*, 125(23), pp. 3570–3579. Available at: https://doi.org/10.1182/blood-2014-11-610519.

Kwiatkowski, B.A. *et al.* (1998) 'The ets family member Tel binds to the Fli-1 oncoprotein and inhibits its transcriptional activity', *The Journal of biological chemistry*, 273(28), pp. 17525–17530. Available at: https://doi.org/10.1074/jbc.273.28.17525.

Kwiatkowski, B.A. *et al.* (2000) 'The ETS family member Tel antagonizes the Fli-1 phenotype in hematopoietic cells', *Blood cells, molecules & diseases*, 26(1), pp. 84–90. Available at: https://doi.org/10.1006/bcmd.2000.0282.

Lambert, M.P. *et al.* (2009) 'Platelet factor 4 regulates megakaryopoiesis through low-density lipoprotein receptor-related protein 1 (LRP1) on megakaryocytes', *Blood*, 114(11), pp. 2290–2298. Available at: https://doi.org/10.1182/blood-2009-04-216473.

Lancrin, C. *et al.* (2009) 'The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage', *Nature*, 457(7231), pp. 892–895. Available at: https://doi.org/10.1038/nature07679.

Laurenti, E. and Göttgens, B. (2018) 'From haematopoietic stem cells to complex differentiation landscapes', *Nature*, 553(7689), pp. 418–426. Available at: https://doi.org/10.1038/nature25022.

Lauzurica, P. *et al.* (2000) 'Phenotypic and functional characteristics of hematopoietic cell lineages in CD69-deficient mice', *Blood*, 95(7), pp. 2312–2320. Available at: https://doi.org/10.1182/blood.V95.7.2312.

Le Blanc, J. and Lordkipanidzé, M. (2019) 'Platelet Function in Aging', *Frontiers in cardiovascular medicine*, 6, p. 109. Available at: https://doi.org/10.3389/fcvm.2019.00109.

Lebrigand, K. *et al.* (2020) ‘High throughput error corrected Nanopore single cell transcriptome sequencing’, *Nature communications*, 11(1), p. 4025. Available at: https://doi.org/10.1038/s41467-020-17800-6.

Lecine, P. *et al.* (1998) ‘Mice lacking transcription factor NF-E2 provide in vivo validation of the proplatelet model of thrombocytopoiesis and show a platelet production defect that is intrinsic to megakaryocytes’, *Blood*, 92(5), pp. 1608–1616. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9716588.

Lee-Six, H. and Kent, D.G. (2020) ‘Tracking hematopoietic stem cells and their progeny using whole-genome sequencing’, *Experimental hematology*, 83, pp. 12–24. Available at: https://doi.org/10.1016/j.exphem.2020.01.004.

Lee, T.I. and Young, R.A. (2000) ‘TRANSCRIPTION OF EUKARYOTIC PROTEIN-CODING GENES’, *Annual review of genetics*, 34(1), pp. 77–137. Available at: https://doi.org/10.1146/annurev.genet.34.1.77.

Lefrançais, E. *et al.* (2017) ‘The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors’, *Nature*, 544(7648), pp. 105–109. Available at: https://doi.org/10.1038/nature21706.

Lever, J., Krzywinski, M. and Altman, N. (2017) ‘Principal component analysis’, *Nature methods*, 14(7), pp. 641–642. Available at: https://doi.org/10.1038/nmeth.4346.

Levin, J. *et al.* (1999) ‘Pathophysiology of thrombocytopenia and anemia in mice lacking transcription factor NF-E2’, *Blood*, 94(9), pp. 3037–3047. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10556187.

Lhoumeau, A.-C. *et al.* (2016) ‘Ptk7-Deficient Mice Have Decreased Hematopoietic Stem Cell Pools as a Result of Deregulated Proliferation and Migration’, *Journal of immunology* , 196(10), pp. 4367–4377. Available at: https://doi.org/10.4049/jimmunol.1500680.

Liang, K.L. *et al.* (2015) ‘Investigation of the role of TRIB2 in normal murine hematopoiesis’, *Experimental hematology*, 43(9, Supplement), p. S77. Available at: https://doi.org/10.1016/j.exphem.2015.06.181.

Liang, Y. and Tedder, T.F. (2001) ‘Identification of a CD20-, FcepsilonRIbeta-, and HTm4-related gene family: sixteen new MS4A family members expressed in human and mouse’, *Genomics*, 72(2), pp. 119–127. Available at: https://doi.org/10.1006/geno.2000.6472.

Liang, Y., Van Zant, G. and Szilvassy, S.J. (2005) ‘Effects of aging on the homing and engraftment of murine hematopoietic stem and progenitor cells’, *Blood*, 106(4), pp. 1479–1487. Available at: https://doi.org/10.1182/blood-2004-11-4282.

Liao, Y., Smyth, G.K. and Shi, W. (2014) ‘featureCounts: an efficient general purpose program for assigning sequence reads to genomic features’, *Bioinformatics* , 30(7), pp. 923–930. Available at: https://doi.org/10.1093/bioinformatics/btt656.

Liggett, L.A. and Sankaran, V.G. (2020) ‘Unraveling Hematopoiesis through the Lens of Genomics’, *Cell*, 182(6), pp. 1384–1400. Available at: https://doi.org/10.1016/j.cell.2020.08.030.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics* , 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Li, H. (2018) ‘Minimap2: pairwise alignment for nucleotide sequences’, *Bioinformatics* , 34(18), pp. 3094–3100. Available at: https://doi.org/10.1093/bioinformatics/bty191.

Li, L. *et al.* (2013) ‘Trib2 Pseudokinase Marks Early CMP Progenitors, and Promotes

Granulocytic Plus Erythroid Progenitor Cell Formation', *Blood*, 122(21), pp. 2448–2448. Available at: https://doi.org/10.1182/blood.V122.21.2448.2448.

Lin, S.-C. *et al.* (2022) 'Functional association of NR4A3 downregulation with impaired differentiation in myeloid leukemogenesis', *Annals of hematology*, 101(10), pp. 2209–2218. Available at: https://doi.org/10.1007/s00277-022-04961-1.

Li, Q. *et al.* (2002) 'Locus control regions', *Blood*, 100(9), pp. 3077–3086. Available at: https://doi.org/10.1182/blood-2002-04-1104.

Liu, T. *et al.* (2017) 'NF-κB signaling in inflammation', *Signal Transduction and Targeted Therapy*, 2(1), pp. 1–9. Available at: https://doi.org/10.1038/sigtrans.2017.23.

Lok, S. *et al.* (1994) 'Cloning and expression of murine thrombopoietin cDNA and stimulation of platelet production in vivo', *Nature*, 369(6481), pp. 565–568. Available at: https://doi.org/10.1038/369565a0.

López-Otín, C. *et al.* (2013) 'The hallmarks of aging', *Cell*, 153(6), pp. 1194–1217. Available at: https://doi.org/10.1016/j.cell.2013.05.039.

Losick, R. and Desplan, C. (2008) 'Stochasticity and cell fate', *Science*, 320(5872), pp. 65–68. Available at: https://doi.org/10.1126/science.1147888.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology*, 15(12), p. 550. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Luc, S. *et al.* (2008) 'Down-regulation of Mpl marks the transition to lymphoid-primed multipotent progenitors with gradual loss of granulocyte-monocyte potential', *Blood*, 111(7), pp. 3424–3434. Available at: https://doi.org/10.1182/blood-2007-08-108324.

Lurton, J. *et al.* (1999) 'Isolation of a gene product expressed by a subpopulation of human lung fibroblasts by differential display', *American journal of respiratory cell and molecular biology*, 20(2), pp. 327–331. Available at: https://doi.org/10.1165/ajrcmb.20.2.3368.

Lu, Y.-C., Krause, D.S., *et al.* (2018) 'Molecular Signature of Megakaryocyte-Erythroid Progenitors Reveals Role of Cell Cycle in Fate Specification', *Blood*, 132(Supplement 1), pp. 3828–3828. Available at: https://doi.org/10.1182/blood-2018-99-119105.

Lu, Y.-C., Sanada, C., *et al.* (2018) 'The Molecular Signature of Megakaryocyte-Erythroid Progenitors Reveals a Role for the Cell Cycle in Fate Specification', *Cell reports*, 25(8), pp. 2083–2093.e4. Available at: https://doi.org/10.1016/j.celrep.2018.10.084.

Maaten, L. and Hinton, G. (2008) 'Visualizing high-dimensional data using t-sne journal of machine learning research', *Journal of machine learning research: JMLR* [Preprint].

Macaulay, I.C. *et al.* (2015) 'G&T-seq: parallel sequencing of single-cell genomes and transcriptomes', *Nature methods*, 12(6), pp. 519–522. Available at: https://doi.org/10.1038/nmeth.3370.

Machlus, K.R. *et al.* (2016) 'CCL5 derived from platelets increases megakaryocyte proplatelet formation', *Blood*, 127(7), pp. 921–926. Available at: https://doi.org/10.1182/blood-2015-05-644583.

Machlus, K.R., Thon, J.N. and Italiano, J.E., Jr (2014) 'Interpreting the developmental dance of the megakaryocyte: a review of the cellular and molecular processes mediating platelet formation', *British journal of haematology*, 165(2), pp. 227–236. Available at: https://doi.org/10.1111/bjh.12758.

Ma, F. *et al.* (2001) 'Molecular cloning of Porimin, a novel cell surface receptor mediating oncotic cell death', *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp. 9778–9783. Available at: https://doi.org/10.1073/pnas.171322898.

Maglott, D. *et al.* (2007) 'Entrez Gene: gene-centered information at NCBI', *Nucleic acids research*, 35(Database issue), pp. D26–31. Available at: https://doi.org/10.1093/nar/gkl993.

Ma, I. and Allan, A.L. (2011) 'The role of human aldehyde dehydrogenase in normal and cancer stem cells', *Stem cell reviews and reports*, 7(2), pp. 292–306. Available at: https://doi.org/10.1007/s12015-010-9208-4.

Majno, G. and Joris, I. (1995) 'Apoptosis, oncosis, and necrosis. An overview of cell death', *The American journal of pathology*, 146(1), pp. 3–15. Available at: https://www.ncbi.nlm.nih.gov/pubmed/7856735.

Mann, M. *et al.* (2018) 'Heterogeneous Responses of Hematopoietic Stem Cells to Inflammatory Stimuli Are Altered with Age', *Cell reports*, 25(11), pp. 2992–3005.e5. Available at: https://doi.org/10.1016/j.celrep.2018.11.056.

Månsson, R. *et al.* (2007) 'Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors', *Immunity*, 26(4), pp. 407–419. Available at: https://doi.org/10.1016/j.immuni.2007.02.013.

Manz, M.G. *et al.* (2002) 'Prospective isolation of human clonogenic common myeloid progenitors', *Proceedings of the National Academy of Sciences of the United States of America*, 99(18), pp. 11872–11877. Available at: https://doi.org/10.1073/pnas.172384399.

Mao, Q. *et al.* (2015) 'SimplePPT: A Simple Principal Tree Algorithm', in *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics (Proceedings), pp. 792–800. Available at: https://doi.org/10.1137/1.9781611974010.89.

Mao, Q. *et al.* (2017) 'Principal Graph and Structure Learning Based on Reversed Graph Embedding', *IEEE transactions on pattern analysis and machine intelligence*, 39(11), pp. 2227–2241. Available at: https://doi.org/10.1109/TPAMI.2016.2635657.

Marinov, G.K. *et al.* (2014) 'From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing', *Genome research*, 24(3), pp. 496–510. Available at: https://doi.org/10.1101/gr.161034.113.

Maston, G.A., Evans, S.K. and Green, M.R. (2006) 'Transcriptional regulatory elements in the human genome', *Annual review of genomics and human genetics*, 7, pp. 29–59. Available at: https://doi.org/10.1146/annurev.genom.7.080505.115623.

Matsumura, I. *et al.* (2000) 'Increased D-type cyclin expression together with decreased cdc2 activity confers megakaryocytic differentiation of a human thrombopoietin-dependent hematopoietic cell line', *The Journal of biological chemistry*, 275(8), pp. 5553–5559. Available at: https://doi.org/10.1074/jbc.275.8.5553.

Matutes, E. *et al.* (2011) 'Mixed-phenotype acute leukemia: clinical and laboratory features and outcome in 100 patients defined according to the WHO 2008 classification', *Blood*, 117(11), pp. 3163–3171. Available at: https://doi.org/10.1182/blood-2010-10-314682.

McCarty, J.M. *et al.* (1995) 'Murine thrombopoietin mRNA levels are modulated by platelet count', *Blood*, 86(10), pp. 3668–3675. Available at: https://www.ncbi.nlm.nih.gov/pubmed/7579332.

McDonald, T.P. *et al.* (1987) 'High doses of recombinant erythropoietin stimulate platelet production in mice', *Experimental hematology*, 15(6), pp. 719–721. Available at:

https://www.ncbi.nlm.nih.gov/pubmed/3595770.

McInnes, L., Healy, J. and Melville, J. (2018) 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', *arXiv [stat.ML]*. Available at: http://arxiv.org/abs/1802.03426.

Mead, A. (2021) 'Clonal hematopoiesis: An abrupt and inevitable consequence of aging?', *The Hematologist*, 18(6). Available at: https://doi.org/10.1182/hem.v18.6.2021614.

Mead, A.J. *et al.* (2017) 'Niche-mediated depletion of the normal hematopoietic stem cell reservoir by Flt3-ITD-induced myeloproliferation', *The Journal of experimental medicine*, 214(7), pp. 2005–2021. Available at: https://doi.org/10.1084/jem.20161418.

van der Meijden, P.E.J. and Heemskerk, J.W.M. (2019) 'Platelet biology and functions: new concepts and clinical perspectives', *Nature reviews. Cardiology*, 16(3), pp. 166–179. Available at: https://doi.org/10.1038/s41569-018-0110-0.

Meng, Y. *et al.* (2023) Epigenetic programming defines haematopoietic stem cell fate restriction. Nat Cell Biol. 25(6):812-822. doi: 10.1038/s41556-023-01137-5.

Mercher, T. *et al.* (2008) 'Notch signaling specifies megakaryocyte development from hematopoietic stem cells', *Cell stem cell*, 3(3), pp. 314–326. Available at: https://doi.org/10.1016/j.stem.2008.07.010.

Merryweather-Clarke, A.T. *et al.* (2016) 'Distinct gene expression program dynamics during erythropoiesis from human induced pluripotent stem cells compared with adult and cord blood progenitors', *BMC genomics*, 17(1), p. 817. Available at: https://doi.org/10.1186/s12864-016-3134-z.

Mignotte, V. *et al.* (1994) 'Structure and transcription of the human c-mpl gene (MPL)', *Genomics*, 20(1), pp. 5–12. Available at: https://doi.org/10.1006/geno.1994.1120.

Mincarelli, L. *et al.* (2018) 'Defining Cell Identity with Single-Cell Omics', *Proteomics*, 18(18), p. e1700312. Available at: https://doi.org/10.1002/pmic.201700312.

Mincarelli, L. *et al.* (2023) 'Single-cell gene and isoform expression analysis reveals signatures of ageing in haematopoietic stem and progenitor cells', *Communications Biology*, 6(1), pp. 1–11. Available at: https://doi.org/10.1038/s42003-023-04936-6.

Min, I.M. *et al.* (2008) 'The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells', *Cell stem cell*, 2(4), pp. 380–391. Available at: https://doi.org/10.1016/j.stem.2008.01.015.

Mirantes, C., Passegué, E. and Pietras, E.M. (2014) 'Pro-inflammatory cytokines: emerging players regulating HSC function in normal and diseased hematopoiesis', *Experimental cell research*, 329(2), pp. 248–254. Available at: https://doi.org/10.1016/j.yexcr.2014.08.017.

Miyakawa, Y. *et al.* (1995) 'Recombinant thrombopoietin induces rapid protein tyrosine phosphorylation of Janus kinase 2 and Shc in human blood platelets', *Blood*, 86(1), pp. 23–27. Available at: https://doi.org/10.1182/blood.V86.1.23.bloodjournal86123.

Miyakawa, Y. *et al.* (1996) 'Thrombopoietin induces tyrosine phosphorylation of Stat3 and Stat5 in human blood platelets', *Blood*, 87(2), pp. 439–446. Available at: https://www.ncbi.nlm.nih.gov/pubmed/8555464.

Miyawaki, K. *et al.* (2017) 'Identification of unipotent megakaryocyte progenitors in human hematopoiesis', *Blood*, 129(25), pp. 3332–3343. Available at: https://doi.org/10.1182/blood-2016-09-741611.

Mjelle, R. *et al.* (2015) ‘Cell cycle regulation of human DNA repair and chromatin remodeling genes’, *DNA repair*, 30, pp. 53–67. Available at: https://doi.org/10.1016/j.dnarep.2015.03.007.

Mohan Rao, L.V., Esmon, C.T. and Pendurthi, U.R. (2014) ‘Endothelial cell protein C receptor: a multiliganded and multifunctional receptor’, *Blood*, 124(10), pp. 1553–1562. Available at: https://doi.org/10.1182/blood-2014-05-578328.

Montecino-Rodriguez, E. *et al.* (2019) ‘Lymphoid-Biased Hematopoietic Stem Cells Are Maintained with Age and Efficiently Generate Lymphoid Progeny’, *Stem cell reports*, 12(3), pp. 584–596. Available at: https://doi.org/10.1016/j.stemcr.2019.01.016.

Moran, P.A.P. (1950) ‘Notes on continuous stochastic phenomena’, *Biometrika*, 37(1-2), pp. 17–23. Available at: https://www.ncbi.nlm.nih.gov/pubmed/15420245.

Morita, Y., Ema, H. and Nakauchi, H. (2010) ‘Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment’, *The Journal of experimental medicine*, 207(6), pp. 1173–1182. Available at: https://doi.org/10.1084/jem.20091318.

Morodomi, Y. *et al.* (2020) ‘Mechanisms of anti-GPIbα antibody–induced thrombocytopenia in mice’, *Blood*, 135(25), pp. 2292–2301. Available at: https://doi.org/10.1182/blood.2019003770.

Morodomi, Y. *et al.* (2022) ‘Inflammatory platelet production stimulated by tyrosyl-tRNA synthetase mimicking viral infection’, *Proceedings of the National Academy of Sciences of the United States of America*, 119(48), p. e2212659119. Available at: https://doi.org/10.1073/pnas.2212659119.

Morrison, S.J. *et al.* (1996) ‘The aging of hematopoietic stem cells’, *Nature Medicine*, pp. 1011–1016. Available at: https://doi.org/10.1038/nm0996-1011.

Morrison, S.J. *et al.* (1997) ‘Identification of a lineage of multipotent hematopoietic progenitors’, *Development* , 124(10), pp. 1929–1939. Available at: https://doi.org/10.1242/dev.124.10.1929.

Morrison, S.J. and Weissman, I.L. (1994) ‘The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype’, *Immunity*, 1(8), pp. 661–673. Available at: https://doi.org/10.1016/1074-7613(94)90037-x.

Mousli, M. *et al.* (2003) ‘ICBP90 belongs to a new family of proteins with an expression that is deregulated in cancer cells’, *British journal of cancer*, 89(1), pp. 120–127. Available at: https://doi.org/10.1038/sj.bjc.6601068.

Müller-Sieburg, C.E. *et al.* (2002) ‘Deterministic regulation of hematopoietic stem cell self-renewal and differentiation’, *Blood*, 100(4), pp. 1302–1309. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12149211.

Muller-Sieburg, C.E. *et al.* (2004) ‘Myeloid-biased hematopoietic stem cells have extensive self-renewal capacity but generate diminished lymphoid progeny with impaired IL-7 responsiveness’, *Blood*, 103(11), pp. 4111–4118. Available at: https://doi.org/10.1182/blood-2003-10-3448.

Muller-Sieburg, C. and Sieburg, H.B. (2008) ‘Stem cell aging: survival of the laziest?’, *Cell cycle* , 7(24), pp. 3798–3804. Available at: https://doi.org/10.4161/cc.7.24.7214.

Mullican, S.E. *et al.* (2007) ‘Abrogation of nuclear receptors Nr4a3 and Nr4a1 leads to development of acute myeloid leukemia’, *Nature medicine*, 13(6), pp. 730–735. Available at: https://doi.org/10.1038/nm1579.

Munoz, J. *et al.* (2014) ‘Concise review: umbilical cord blood transplantation: past, present, and future’, *Stem cells translational medicine*, 3(12), pp. 1435–1443. Available at:

https://doi.org/10.5966/sctm.2014-0151.

Muntean, A.G. *et al.* (2007) 'Cyclin D-Cdk4 is regulated by GATA-1 and required for megakaryocyte growth and polyploidization', *Blood*, 109(12), pp. 5199–5207. Available at: https://doi.org/10.1182/blood-2006-11-059378.

Mussbacher, M. *et al.* (2019) 'Cell Type-Specific Roles of NF-κB Linking Inflammation and Thrombosis', *Frontiers in immunology*, 10, p. 85. Available at: https://doi.org/10.3389/fimmu.2019.00085.

Naik, S.H. *et al.* (2013) 'Diverse and heritable lineage imprinting of early haematopoietic progenitors', *Nature*, 496(7444), pp. 229–232. Available at: https://doi.org/10.1038/nature12013.

Nakamura-Ishizu, A. *et al.* (2014) 'Megakaryocytes are essential for HSC quiescence through the production of thrombopoietin', *Biochemical and biophysical research communications*, 454(2), pp. 353–357. Available at: https://doi.org/10.1016/j.bbrc.2014.10.095.

Nakamura-Ishizu, A. *et al.* (2018) 'Thrombopoietin Metabolically Primes Hematopoietic Stem Cells to Megakaryocyte-Lineage Differentiation', *Cell reports*, 25(7), pp. 1772–1785.e6. Available at: https://doi.org/10.1016/j.celrep.2018.10.059.

Nakamura, Y. *et al.* (1998) 'Impaired erythropoiesis in transgenic mice overexpressing a truncated erythropoietin receptor', *Experimental hematology*, 26(12), pp. 1105–1110. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9808048.

Nakorn, T.N., Miyamoto, T. and Weissman, I.L. (2003) 'Characterization of mouse clonogenic megakaryocyte progenitors', *Proceedings of the National Academy of Sciences of the United States of America*, 100(1), pp. 205–210. Available at: https://doi.org/10.1073/pnas.262655099.

Nam, D.K. *et al.* (2002) 'Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription', *Proceedings of the National Academy of Sciences of the United States of America*, 99(9), pp. 6152–6156. Available at: https://doi.org/10.1073/pnas.092140899.

Newman, P.J. *et al.* (1988) 'Enzymatic amplification of platelet-specific messenger RNA using the polymerase chain reaction', *The Journal of clinical investigation*, 82(2), pp. 739–743. Available at: https://doi.org/10.1172/JCI113656.

Nichogiannopoulou, A. *et al.* (1999) 'Defects in hemopoietic stem cell activity in Ikaros mutant mice', *The Journal of experimental medicine*, 190(9), pp. 1201–1214. Available at: https://doi.org/10.1084/jem.190.9.1201.

Nieswandt, B. *et al.* (2000) 'Identification of critical antigen-specific mechanisms in the development of immune thrombocytopenic purpura in mice', *Blood*, 96(7), pp. 2520–2527. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11001906.

Niinivirta, M. *et al.* (2020) 'Tumor endothelial ELTD1 as a predictive marker for treatment of renal cancer patients with sunitinib', *BMC cancer*, 20(1), p. 339. Available at: https://doi.org/10.1186/s12885-020-06770-z.

Nishikii, H. *et al.* (2015) 'Unipotent megakaryopoietic pathway bridging hematopoietic stem cells and mature megakaryocytes: Megakaryopoiesis bridging the HSC and meg', *Stem cells*, 33(7), pp. 2196–2207. Available at: https://doi.org/10.1002/stem.1985.

Nishikii, H., Kurita, N. and Chiba, S. (2017) 'The road map for megakaryopoietic lineage from hematopoietic stem/progenitor cells', *Stem cells translational medicine*, 6(8), pp. 1661–1665. Available at: https://doi.org/10.1002/sctm.16-0490.

Nishimura, S. *et al.* (2015) 'IL-1α induces thrombopoiesis through megakaryocyte rupture in response to acute platelet needs', *The Journal of cell biology*, 209(3), pp. 453–466. Available at: https://doi.org/10.1083/jcb.201410052.

Noetzli Leila J., French Shauna L. and Machlus Kellie R. (2019) 'New Insights Into the Differentiation of Megakaryocytes From Hematopoietic Progenitors', *Arteriosclerosis, thrombosis, and vascular biology*, 39(7), pp. 1288–1300. Available at: https://doi.org/10.1161/ATVBAHA.119.312129.

Noone, D.G. *et al.* (2016) 'Von Willebrand factor regulates complement on endothelial cells', *Kidney international*, 90(1), pp. 123–134. Available at: https://doi.org/10.1016/j.kint.2016.03.023.

Notta, F. *et al.* (2016) 'Distinct routes of lineage development reshape the human blood hierarchy across ontogeny', *Science*, 351(6269), p. aab2116. Available at: https://doi.org/10.1126/science.aab2116.

Okada, S. *et al.* (1999) 'Prolonged expression of c-fos suppresses cell cycle entry of dormant hematopoietic stem cells', *Blood*, 93(3), pp. 816–825. Available at: https://doi.org/10.1182/blood.V93.3.816.

Okuno, Y. *et al.* (2005) 'Potential autoregulation of transcription factor PU.1 by an upstream regulatory element', *Molecular and cellular biology*, 25(7), pp. 2832–2845. Available at: https://doi.org/10.1128/MCB.25.7.2832-2845.2005.

Ong, C.-T. and Corces, V.G. (2011) 'Enhancer function: new insights into the regulation of tissue-specific gene expression', *Nature reviews. Genetics*, 12(4), pp. 283–293. Available at: https://doi.org/10.1038/nrg2957.

Orkin, S.H. (2000) 'Diversification of haematopoietic stem cells to specific lineages', *Nature reviews. Genetics*, 1(1), pp. 57–64. Available at: https://doi.org/10.1038/35049577.

Orkin, S.H. and Zon, L.I. (2008) 'Hematopoiesis: an evolving paradigm for stem cell biology', *Cell*, 132(4), pp. 631–644. Available at: https://doi.org/10.1016/j.cell.2008.01.025.

Osawa, M. *et al.* (1996) 'Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell', *Science*, 273(5272), pp. 242–245. Available at: https://doi.org/10.1126/science.273.5272.242.

van den Oudenrijn, S. *et al.* (2000) 'Mutations in the thrombopoietin receptor, Mpl, in children with congenital amegakaryocytic thrombocytopenia', *British journal of haematology*, 110(2), pp. 441–448. Available at: https://doi.org/10.1046/j.1365-2141.2000.02175.x.

Palii, C.G. *et al.* (2019) 'Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate', *Cell stem cell*, 24(5), pp. 812–820.e5. Available at: https://doi.org/10.1016/j.stem.2019.02.006.

Pang, W.W. *et al.* (2011) 'Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age', *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp. 20012–20017. Available at: https://doi.org/10.1073/pnas.1116110108.

Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nature genetics*, 40(12), pp. 1413–1415. Available at: https://doi.org/10.1038/ng.259.

Papaemmanuil, E. *et al.* (2011) 'Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts', *The New England journal of medicine*, 365(15), pp. 1384–1395. Available at: https://doi.org/10.1056/NEJMoa1103283.

Papayannopoulou, T. *et al.* (1996) 'Insights into the cellular mechanisms of erythropoietin-thrombopoietin synergy', *Experimental hematology*, 24(5), pp. 660–669. Available at: https://www.ncbi.nlm.nih.gov/pubmed/8605971.

Passegué, E. *et al.* (2005) 'Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates', *The Journal of experimental medicine*, 202(11), pp. 1599–1611. Available at: https://doi.org/10.1084/jem.20050967.

Paul, F. *et al.* (2015) 'Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors', *Cell*, 163(7), pp. 1663–1677. Available at: https://doi.org/10.1016/j.cell.2015.11.013.

Paulus, J.-M. *et al.* (2004) 'Thrombopoietin responsiveness reflects the number of doublings undergone by megakaryocyte progenitors', *Blood*, 104(8), pp. 2291–2298. Available at: https://doi.org/10.1182/blood-2003-05-1745.

Pearen, M.A. and Muscat, G.E.O. (2010) 'Minireview: Nuclear hormone receptor 4A signaling: implications for metabolic disease', *Molecular endocrinology* , 24(10), pp. 1891–1903. Available at: https://doi.org/10.1210/me.2010-0015.

Peled, A. *et al.* (1999) 'Dependence of human stem cell engraftment and repopulation of NOD/SCID mice on CXCR4', *Science*, 283(5403), pp. 845–848. Available at: https://doi.org/10.1126/science.283.5403.845.

Pellin, D. *et al.* (2019) 'A comprehensive single cell transcriptional landscape of human hematopoietic progenitors', *Nature communications*, 10(1), p. 2395. Available at: https://doi.org/10.1038/s41467-019-10291-0.

Perié, L. *et al.* (2015) 'The Branching Point in Erythro-Myeloid Differentiation', *Cell*, 163(7), pp. 1655–1662. Available at: https://doi.org/10.1016/j.cell.2015.11.059.

Pernes, G. *et al.* (2019) 'Fat for fuel: lipid metabolism in haematopoiesis', *Clinical & translational immunology*, 8(12), p. e1098. Available at: https://doi.org/10.1002/cti2.1098.

Picelli, S. *et al.* (2014) 'Full-length RNA-seq from single cells using Smart-seq2', *Nature protocols*, 9(1), pp. 171–181. Available at: https://doi.org/10.1038/nprot.2014.006.

Piemontese, S. *et al.* (2015) 'A survey on unmanipulated haploidentical hematopoietic stem cell transplantation in adults with acute leukemia', *Leukemia*, 29(5), pp. 1069–1075. Available at: https://doi.org/10.1038/leu.2014.336.

Pietras, E.M. *et al.* (2015) 'Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions', *Cell stem cell*, 17(1), pp. 35–46. Available at: https://doi.org/10.1016/j.stem.2015.05.003.

Pimanda, J.E. *et al.* (2007) 'Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development', *Proceedings of the National Academy of Sciences of the United States of America*, 104(45), pp. 17692–17697. Available at: https://doi.org/10.1073/pnas.0707045104.

Pinho, S. *et al.* (2018) 'Lineage-Biased Hematopoietic Stem Cells Are Regulated by Distinct Niches', *Developmental cell*, 44(5), pp. 634–641.e4. Available at: https://doi.org/10.1016/j.devcel.2018.01.016.

Plé, H. *et al.* (2012) 'Alteration of the platelet transcriptome in chronic kidney disease', *Thrombosis and haemostasis*, 108(4), pp. 605–615. Available at: https://doi.org/10.1160/TH12-03-0153.

Pleines, I. *et al.* (2010) 'Multiple alterations of platelet functions dominated by increased

secretion in mice lacking Cdc42 in platelets', *Blood*, 115(16), pp. 3364–3373. Available at: https://doi.org/10.1182/blood-2009-09-242271.

Pleines, I. *et al.* (2013) 'Defective tubulin organization and proplatelet formation in murine megakaryocytes lacking Rac1 and Cdc42', *Blood*, 122(18), pp. 3178–3187. Available at: https://doi.org/10.1182/blood-2013-03-487942.

Pleines, I., Cherpokova, D. and Bender, M. (2019) 'Rho GTPases and their downstream effectors in megakaryocyte biology', *Platelets*, 30(1), pp. 9–16. Available at: https://doi.org/10.1080/09537104.2018.1478071.

Pop, R. *et al.* (2010) 'A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU.1 and S-phase progression', *PLoS biology*, 8(9). Available at: https://doi.org/10.1371/journal.pbio.1000484.

Portier, I. and Campbell, R.A. (2021) 'Role of Platelets in Detection and Regulation of Infection', *Arteriosclerosis, thrombosis, and vascular biology*, 41(1), pp. 70–78. Available at: https://doi.org/10.1161/ATVBAHA.120.314645.

Poscablo, D.M. *et al.* (2021) 'Megakaryocyte progenitor cell function is enhanced upon aging despite the functional decline of aged hematopoietic stem cells', *Stem cell reports*, 16(6), pp. 1598–1613. Available at: https://doi.org/10.1016/j.stemcr.2021.04.016.

Prebet, T. *et al.* (2010) 'The cell polarity PTK7 receptor acts as a modulator of the chemotherapeutic response in acute myeloid leukemia and impairs clinical outcome', *Blood*, 116(13), pp. 2315–2323. Available at: https://doi.org/10.1182/blood-2010-01-262352.

Preller, A. and Wilson, J.E. (1992) 'Localization of the type III isozyme of hexokinase at the nuclear periphery', *Archives of biochemistry and biophysics*, 294(2), pp. 482–492. Available at: https://doi.org/10.1016/0003-9861(92)90715-9.

Pronk, C.J.H. *et al.* (2007) 'Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy', *Cell stem cell*, 1(4), pp. 428–442. Available at: https://doi.org/10.1016/j.stem.2007.07.005.

Psaila, B. *et al.* (2016) 'Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways', *Genome biology*, 17, p. 83. Available at: https://doi.org/10.1186/s13059-016-0939-7.

Psaila, B. *et al.* (2020) 'Single-Cell Analyses Reveal Megakaryocyte-Biased Hematopoiesis in Myelofibrosis and Identify Mutant Clone-Specific Targets', *Molecular cell*, 78(3), pp. 477–492.e8. Available at: https://doi.org/10.1016/j.molcel.2020.04.008.

Psaila, B. and Mead, A.J. (2019) 'Single-cell approaches reveal novel cellular pathways for megakaryocyte and erythroid differentiation', *Blood*, 133(13), pp. 1427–1435. Available at: https://doi.org/10.1182/blood-2018-11-835371.

Qian, H. *et al.* (2007) 'Critical role of thrombopoietin in maintaining adult quiescent hematopoietic stem cells', *Cell stem cell*, 1(6), pp. 671–684. Available at: https://doi.org/10.1016/j.stem.2007.10.008.

Qiu, X. *et al.* (2017) 'Reversed graph embedding resolves complex single-cell trajectories', *Nature methods*, 14(10), pp. 979–982. Available at: https://doi.org/10.1038/nmeth.4402.

Rahman, N.-T. *et al.* (2018) 'MRTFA Augments Megakaryocyte Maturation By Enhancing the SRF Regulatory Axis', *Blood*, 132(Supplement 1), pp. 640–640. Available at: https://doi.org/10.1182/blood-2018-99-118954.

Raj, A. and van Oudenaarden, A. (2008) 'Nature, nurture, or chance: stochastic gene expression

and its consequences', *Cell*, 135(2), pp. 216–226. Available at: https://doi.org/10.1016/j.cell.2008.09.050.

Ramanathan, G. and Mannhalter, C. (2016) 'Increased expression of transient receptor potential canonical 6 (TRPC6) in differentiating human megakaryocytes', *Cell biology international*, 40(2), pp. 223–231. Available at: https://doi.org/10.1002/cbin.10558.

Ramirez-Herrick, A.M. *et al.* (2011) 'Reduced NR4A gene dosage leads to mixed myelodysplastic/myeloproliferative neoplasms in mice', *Blood*, 117(9), pp. 2681–2690. Available at: https://doi.org/10.1182/blood-2010-02-267906.

Reddi, D. and Belibasakis, G.N. (2012) 'Transcriptional profiling of bone marrow stromal cells in response to Porphyromonas gingivalis secreted products', *PloS one*, 7(8), p. e43899. Available at: https://doi.org/10.1371/journal.pone.0043899.

Regev, A. *et al.* (2017) 'The Human Cell Atlas', *eLife*, 6. Available at: https://doi.org/10.7554/eLife.27041.

Rekhtman, N. *et al.* (1999) 'Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells', *Genes & development*, 13(11), pp. 1398–1411. Available at: https://doi.org/10.1101/gad.13.11.1398.

Rhodes, J. *et al.* (2005) 'Interplay of pu.1 and gata1 determines myelo-erythroid progenitor cell fate in zebrafish', *Developmental cell*, 8(1), pp. 97–108. Available at: https://doi.org/10.1016/j.devcel.2004.11.014.

Rice, K.L. *et al.* (2008) 'Overexpression of stem cell associated ALDH1A1, a target of the leukemogenic transcription factor TLX1/HOX11, inhibits lymphopoiesis and promotes myelopoiesis in murine hematopoietic progenitors', *Leukemia research*, 32(6), pp. 873–883. Available at: https://doi.org/10.1016/j.leukres.2007.11.001.

Rieger, M.A. *et al.* (2009) 'Hematopoietic cytokines can instruct lineage choice', *Science*, 325(5937), pp. 217–218. Available at: https://doi.org/10.1126/science.1171461.

van de Rijn, M. *et al.* (1989) 'Mouse hematopoietic stem-cell antigen Sca-1 is a member of the Ly-6 antigen family', *Proceedings of the National Academy of Sciences of the United States of America*, 86(12), pp. 4634–4638. Available at: https://doi.org/10.1073/pnas.86.12.4634.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics* , 26(1), pp. 139–140. Available at: https://doi.org/10.1093/bioinformatics/btp616.

Roch, A., Trachsel, V. and Lutolf, M.P. (2015) 'Brief Report: Single-Cell Analysis Reveals Cell Division-Independent Emergence of Megakaryocytes From Phenotypic Hematopoietic Stem Cells', *Stem cells*, 33(10), pp. 3152–3157. Available at: https://doi.org/10.1002/stem.2106.

Rodrigues, C.P., Shvedunova, M. and Akhtar, A. (2020) 'Epigenetic Regulators as the Gatekeepers of Hematopoiesis', *Trends in genetics: TIG* [Preprint]. Available at: https://doi.org/10.1016/j.tig.2020.09.015.

Rodriguez-Fraticelli, A.E. *et al.* (2018) 'Clonal analysis of lineage fate in native haematopoiesis', *Nature*, 553(7687), pp. 212–216. Available at: https://doi.org/10.1038/nature25168.

Rodriguez-Fraticelli, A.E. *et al.* (2020) 'Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis', *Nature*, 583(7817), pp. 585–589. Available at: https://doi.org/10.1038/s41586-020-2503-6.

Rojnuckarin, P. and Kaushansky, K. (2001) 'Actin reorganization and proplatelet formation in

murine megakaryocytes: the role of protein kinase calpha', *Blood*, 97(1), pp. 154–161. Available at: https://doi.org/10.1182/blood.v97.1.154.

Rondina, M.T., Weyrich, A.S. and Zimmerman, G.A. (2013) 'Platelets as cellular effectors of inflammation in vascular diseases', *Circulation research*, 112(11), pp. 1506–1519. Available at: https://doi.org/10.1161/CIRCRESAHA.113.300512.

Rosmarin, A.G., Yang, Z. and Resendes, K.K. (2005) 'Transcriptional regulation in myelopoiesis: Hematopoietic fate choice, myeloid differentiation, and leukemogenesis', *Experimental hematology*, 33(2), pp. 131–143. Available at: https://doi.org/10.1016/j.exphem.2004.08.015.

Rossaint, J. *et al.* (2021) 'Platelets orchestrate the resolution of pulmonary inflammation in mice by T reg cell repositioning and macrophage education', *The Journal of experimental medicine*, 218(7). Available at: https://doi.org/10.1084/jem.20201353.

Rossi, D.J. *et al.* (2005) 'Cell intrinsic alterations underlie hematopoietic stem cell aging', *Proceedings of the National Academy of Sciences of the United States of America*, 102(26), pp. 9194–9199. Available at: https://doi.org/10.1073/pnas.0503280102.

Rossi, D.J., Bryder, D., *et al.* (2007) 'Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age', *Nature*, 447(7145), pp. 725–729. Available at: https://doi.org/10.1038/nature05862.

Rossi, D.J., Seita, J., *et al.* (2007) 'Hematopoietic stem cell quiescence attenuates DNA damage response and permits DNA damage accumulation during aging', *Cell cycle* , 6(19), pp. 2371–2376. Available at: https://doi.org/10.4161/cc.6.19.4759.

Roundtree, I.A. and He, C. (2016) 'RNA epigenetics--chemical messages for posttranscriptional gene regulation', *Current opinion in chemical biology*, 30, pp. 46–51. Available at: https://doi.org/10.1016/j.cbpa.2015.10.024.

Roy, A. *et al.* (2021) 'Transitions in lineage specification and gene regulatory networks in hematopoietic stem/progenitor cells over human development', *Cell reports*, 36(11), p. 109698. Available at: https://doi.org/10.1016/j.celrep.2021.109698.

Sadasivam, S. and DeCaprio, J.A. (2013) 'The DREAM complex: master coordinator of cell cycle-dependent gene expression', *Nature reviews. Cancer*, 13(8), pp. 585–595. Available at: https://doi.org/10.1038/nrc3556.

Sadler, J.E. (1998) 'Biochemistry and genetics of von Willebrand factor', *Annual review of biochemistry*, 67, pp. 395–424. Available at: https://doi.org/10.1146/annurev.biochem.67.1.395.

Saelens, W. *et al.* (2019) 'A comparison of single-cell trajectory inference methods', *Nature biotechnology*, 37(5), pp. 547–554. Available at: https://doi.org/10.1038/s41587-019-0071-9.

Sanada, C. *et al.* (2016) 'Adult human megakaryocyte-erythroid progenitors are in the CD34+CD38mid fraction', *Blood*, 128(7), pp. 923–933. Available at: https://doi.org/10.1182/blood-2016-01-693705.

Sanjuan-Pla, A. *et al.* (2013) 'Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy', *Nature*, 502(7470), pp. 232–236. Available at: https://doi.org/10.1038/nature12495.

Sansal, I. *et al.* (2000) 'NPDC-1, a regulator of neural cell proliferation and differentiation, interacts with E2F-1, reduces its binding to DNA and modulates its transcriptional activity', *Oncogene*, 19(43), pp. 5000–5009. Available at: https://doi.org/10.1038/sj.onc.1203843.

Sashida, G. and Iwama, A. (2012) 'Epigenetic regulation of hematopoiesis', *International*

journal of hematology, 96(4), pp. 405–412. Available at: https://doi.org/10.1007/s12185-012-1183-x.

de Sauvage, F.J. *et al.* (1996) 'Physiological regulation of early and late stages of megakaryocytopoiesis by thrombopoietin', *The Journal of experimental medicine*, 183(2), pp. 651–656. Available at: https://doi.org/10.1084/jem.183.2.651.

Schiöth, H.B. and Fredriksson, R. (2005) 'The GRAFS classification system of G-protein coupled receptors in comparative perspective', *General and comparative endocrinology*, 142(1-2), pp. 94–101. Available at: https://doi.org/10.1016/j.ygcen.2004.12.018.

Schroeder, T. (2010) 'Hematopoietic stem cell heterogeneity: subtypes, not unpredictable behavior', *Cell stem cell*, 6(3), pp. 203–207. Available at: https://doi.org/10.1016/j.stem.2010.02.006.

Schwer, H.D. *et al.* (2001) 'A lineage-restricted and divergent beta-tubulin isoform is essential for the biogenesis, structure and function of blood platelets', *Current biology: CB*, 11(8), pp. 579–586. Available at: https://doi.org/10.1016/s0960-9822(01)00153-1.

Seiler, K. *et al.* (2022) 'Hexokinase 3 enhances myeloid cell survival via non-glycolytic functions', *Cell death & disease*, 13(5), p. 448. Available at: https://doi.org/10.1038/s41419-022-04891-w.

Sengupta, A. *et al.* (2013) 'Regulation Of Erythro-Megakaryocytic Lineage Bifurcation By The Gfi1b Gene Target Rgs18', *Blood*, 122(21), p. 1191. Available at: https://doi.org/10.1182/blood.V122.21.1191.1191.

Shahrezaei, V. and Swain, P.S. (2008) 'The stochastic nature of biochemical networks', *Current opinion in biotechnology*, 19(4), pp. 369–374. Available at: https://doi.org/10.1016/j.copbio.2008.06.011.

Shalek, A.K. *et al.* (2013) 'Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells', *Nature*, 498(7453), pp. 236–240. Available at: https://doi.org/10.1038/nature12172.

Sharon, D. *et al.* (2013) 'A single-molecule long-read survey of the human transcriptome', *Nature biotechnology*, 31(11), pp. 1009–1014. Available at: https://doi.org/10.1038/nbt.2705.

Shattil, S.J., Kashiwagi, H. and Pampori, N. (1998) 'Integrin signaling: the platelet paradigm', *Blood*, 91(8), pp. 2645–2657. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9531572.

Shin, J.Y. *et al.* (2014) 'High c-Kit expression identifies hematopoietic stem cells with impaired self-renewal and megakaryocytic bias', *The Journal of experimental medicine*, 211(2), pp. 217–231. Available at: https://doi.org/10.1084/jem.20131128.

Shivdasani, R.A. *et al.* (1995) 'Transcription factor NF-E2 is required for platelet formation independent of the actions of thrombopoietin/MGDF in megakaryocyte development', *Cell*, 81(5), pp. 695–704. Available at: https://doi.org/10.1016/0092-8674(95)90531-6.

Shivdasani, R.A. *et al.* (1997) 'A lineage-selective knockout establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet development', *The EMBO journal*, 16(13), pp. 3965–3973. Available at: https://doi.org/10.1093/emboj/16.13.3965.

Shi, Z. *et al.* (2022) 'HIT-scISOseq: High-throughput and high-accuracy single-cell full-length isoform sequencing'. Available at: https://doi.org/10.21203/rs.3.rs-114035/v1.

Shlush, L.I. *et al.* (2017) 'Tracing the origins of relapse in acute myeloid leukaemia to stem cells', *Nature*, 547(7661), pp. 104–108. Available at: https://doi.org/10.1038/nature22993.

Siatecka, M. and Bieker, J.J. (2011) 'The multifunctional role of EKLF/KLF1 during erythropoiesis', *Blood*, 118(8), pp. 2044–2054. Available at: https://doi.org/10.1182/blood-2011-03-331371.

Sieburg, H.B. *et al.* (2006) 'The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets', *Blood*, 107(6), pp. 2311–2316. Available at: https://doi.org/10.1182/blood-2005-07-2970.

Signer, R.A.J. *et al.* (2007) 'Age-related defects in B lymphopoiesis underlie the myeloid dominance of adult leukemia', *Blood*, 110(6), pp. 1831–1839. Available at: https://doi.org/10.1182/blood-2007-01-069401.

Siminovitch, L., Mcculloch, E.A. and Till, J.E. (1963) 'The distribution of colony forming-cells among spleen colonies', *Journal of cellular and comparative physiology*, 62, pp. 327–336. Available at: https://doi.org/10.1002/jcp.1030620313.

Sim, X. *et al.* (2016) 'Understanding platelet generation from megakaryocytes: implications for in vitro-derived platelets', *Blood*, 127(10), pp. 1227–1233. Available at: https://doi.org/10.1182/blood-2015-08-607929.

Skoda, R.C. *et al.* (1993) 'Murine c-mpl: a member of the hematopoietic growth factor receptor superfamily that transduces a proliferative signal', *The EMBO journal*, 12(7), pp. 2645–2653. Available at: https://doi.org/10.1002/j.1460-2075.1993.tb05925.x.

Solar, G.P. *et al.* (1998) 'Role of c-mpl in early hematopoiesis', *Blood*, 92(1), pp. 4–10. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9639492.

Song, Q. *et al.* (2021) 'The functional landscape of Golgi membrane protein 1 (GOLM1) phosphoproteome reveal GOLM1 regulating P53 that promotes malignancy', *Cell death discovery*, 7(1), p. 42. Available at: https://doi.org/10.1038/s41420-021-00422-2.

Song, R. *et al.* (2021) 'IRF1 governs the differential interferon-stimulated gene responses in human monocytes and macrophages by regulating chromatin accessibility', *Cell reports*, 34(12), p. 108891. Available at: https://doi.org/10.1016/j.celrep.2021.108891.

Song, S.-H. *et al.* (2012) 'Ldb1 regulates carbonic anhydrase 1 during erythroid differentiation', *Biochimica et biophysica acta*, 1819(8), pp. 885–891. Available at: https://doi.org/10.1016/j.bbagrm.2012.05.001.

Song, Y. *et al.* (2017) 'Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation', *Molecular cell*, 67(1), pp. 148–161.e5. Available at: https://doi.org/10.1016/j.molcel.2017.06.003.

Spangrude, G.J., Heimfeld, S. and Weissman, I.L. (1988) 'Purification and characterization of mouse hematopoietic stem cells', *Science*, 241(4861), pp. 58–62. Available at: https://doi.org/10.1126/science.2898810.

Spies, N., Burge, C.B. and Bartel, D.P. (2013) '3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts', *Genome research*, 23(12), pp. 2078–2090. Available at: https://doi.org/10.1101/gr.156919.113.

Squair, J.W. *et al.* (2021) 'Confronting false discoveries in single-cell differential expression', *Nature communications*, 12(1), p. 5692. Available at: https://doi.org/10.1038/s41467-021-25960-2.

Starck, J. *et al.* (2003) 'Functional cross-antagonism between transcription factors FLI-1 and EKLF', *Molecular and cellular biology*, 23(4), pp. 1390–1402. Available at: https://doi.org/10.1128/MCB.23.4.1390-1402.2003.

Steensma, D.P. *et al.* (2015) 'Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes', *Blood*, 126(1), pp. 9–16. Available at: https://doi.org/10.1182/blood-2015-03-631747.

Steidl, U. *et al.* (2006) 'Essential role of Jun family transcription factors in PU.1 knockdown-induced leukemic stem cells', *Nature genetics*, 38(11), pp. 1269–1277. Available at: https://doi.org/10.1038/ng1898.

Steijger, T. *et al.* (2013) 'Assessment of transcript reconstruction methods for RNA-seq', *Nature methods*, 10(12), pp. 1177–1184. Available at: https://doi.org/10.1038/nmeth.2714.

Stoeckius, M. *et al.* (2017) 'Simultaneous epitope and transcriptome measurement in single cells', *Nature methods*, 14(9), pp. 865–868. Available at: https://doi.org/10.1038/nmeth.4380.

Storms, R.W. *et al.* (1999) 'Isolation of primitive human hematopoietic progenitors on the basis of aldehyde dehydrogenase activity', *Proceedings of the National Academy of Sciences of the United States of America*, 96(16), pp. 9118–9123. Available at: https://doi.org/10.1073/pnas.96.16.9118.

Stuart, T. *et al.* (2018) 'Comprehensive integration of single cell data', *bioRxiv*. Available at: https://doi.org/10.1101/460147.

Stuart, T. *et al.* (2019) 'Comprehensive Integration of Single-Cell Data', *Cell*, 177(7), pp. 1888–1902.e21. Available at: https://doi.org/10.1016/j.cell.2019.05.031.

Sudo, K. *et al.* (2000) 'Age-associated characteristics of murine hematopoietic stem cells', *The Journal of experimental medicine*, 192(9), pp. 1273–1280. Available at: https://doi.org/10.1084/jem.192.9.1273.

Sukhatme, V.P. *et al.* (1988) 'A zinc finger-encoding gene coregulated with c-fos during growth and differentiation, and after cellular depolarization', *Cell*, 53(1), pp. 37–43. Available at: https://doi.org/10.1016/0092-8674(88)90485-0.

Sun, D. *et al.* (2014) 'Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal', *Cell stem cell*, 14(5), pp. 673–688. Available at: https://doi.org/10.1016/j.stem.2014.03.002.

Sun, S. *et al.* (2013) 'Expression of plasma membrane receptor genes during megakaryocyte development', *Physiological genomics*, 45(6), pp. 217–227. Available at: https://doi.org/10.1152/physiolgenomics.00056.2012.

Sun, W. and Downing, J.R. (2004) 'Haploinsufficiency of AML1 results in a decrease in the number of LTR-HSCs while simultaneously inducing an increase in more mature progenitors', *Blood*, 104(12), pp. 3565–3572. Available at: https://doi.org/10.1182/blood-2003-12-4349.

Supernat, A. *et al.* (2021) 'Transcriptomic landscape of blood platelets in healthy donors', *Scientific reports*, 11(1), p. 15679. Available at: https://doi.org/10.1038/s41598-021-94003-z.

Su, Y.Y. *et al.* (2001) 'Human ERMAP: an erythroid adhesion/receptor transmembrane protein', *Blood cells, molecules & diseases*, 27(5), pp. 938–949. Available at: https://doi.org/10.1006/bcmd.2001.0465.

Suzuki, M. *et al.* (2013) 'GATA factor switching from GATA2 to GATA1 contributes to erythroid differentiation', *Genes to cells: devoted to molecular & cellular mechanisms*, 18(11), pp. 921–933. Available at: https://doi.org/10.1111/gtc.12086.

Svensson, V. *et al.* (2017) 'Power analysis of single-cell RNA-sequencing experiments', *Nature methods*, 14(4), pp. 381–387. Available at: https://doi.org/10.1038/nmeth.4220.

Tabula Muris Consortium *et al.* (2018) 'Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris', *Nature*, 562(7727), pp. 367–372. Available at: https://doi.org/10.1038/s41586-018-0590-4.

Talebian, L. *et al.* (2007) 'T-lymphoid, megakaryocyte, and granulocyte development are sensitive to decreases in CBFbeta dosage', *Blood*, 109(1), pp. 11–21. Available at: https://doi.org/10.1182/blood-2006-05-021188.

Tardaguila, M. *et al.* (2018) 'SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification', *Genome research*, 28(3), pp. 396–411. Available at: https://doi.org/10.1101/gr.222976.117.

Tedder, T.F. *et al.* (1988) 'Isolation and structure of a cDNA encoding the B1 (CD20) cell-surface antigen of human B lymphocytes', *Proceedings of the National Academy of Sciences of the United States of America*, 85(1), pp. 208–212. Available at: https://doi.org/10.1073/pnas.85.1.208.

Testi, R. *et al.* (1990) 'CD69 is expressed on platelets and mediates platelet activation and aggregation', *The Journal of experimental medicine*, 172(3), pp. 701–707. Available at: https://doi.org/10.1084/jem.172.3.701.

Thapa, R. *et al.* (2023) 'Rapid activation of hematopoietic stem cells', Stem Cell Res Ther. Jun 6;14(1):152. doi: 10.1186/s13287-023-03377-6.

Thiel, G. and Cibelli, G. (2002) 'Regulation of life and death by the zinc finger transcription factor Egr-1', *Journal of cellular physiology*, 193(3), pp. 287–292. Available at: https://doi.org/10.1002/jcp.10178.

Thon, J.N. and Italiano, J.E. (2012) 'Platelets: production, morphology and ultrastructure', *Handbook of experimental pharmacology*, (210), pp. 3–22. Available at: https://doi.org/10.1007/978-3-642-29423-5_1.

Tian, L. *et al.* (2021) 'Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing', *Genome biology*, 22(1), p. 310. Available at: https://doi.org/10.1186/s13059-021-02525-6.

Tijssen, M.R. *et al.* (2011) 'Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators', *Developmental cell*, 20(5), pp. 597–609. Available at: https://doi.org/10.1016/j.devcel.2011.04.008.

Tilgner, H. *et al.* (2013) 'Accurate identification and analysis of human mRNA isoforms using deep long read sequencing', *G3* , 3(3), pp. 387–397. Available at: https://doi.org/10.1534/g3.112.004812.

Till, J.E. and McCULLOCH, E.A. (1961) 'A direct measurement of the radiation sensitivity of normal mouse bone marrow cells', *Radiation research*, 14, pp. 213–222. Available at: https://www.ncbi.nlm.nih.gov/pubmed/13776896.

Tipping, A.J. *et al.* (2009) 'High GATA-2 expression inhibits human hematopoietic stem and progenitor cell function by effects on cell cycle', *Blood*, 113(12), pp. 2661–2672. Available at: https://doi.org/10.1182/blood-2008-06-161117.

Tirosh, I. *et al.* (2016) 'Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq', *Science*, 352(6282), pp. 189–196. Available at: https://doi.org/10.1126/science.aad0501.

Towler, B.P. *et al.* (2020) 'Dis3L2 regulates cell proliferation and tissue growth through a conserved mechanism', *PLoS genetics*, 16(12), p. e1009297. Available at: https://doi.org/10.1371/journal.pgen.1009297.

Trapnell, C. *et al.* (2010) ‘Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation’, *Nature biotechnology*, 28(5), pp. 511–515. Available at: https://doi.org/10.1038/nbt.1621.

Trapnell, C. *et al.* (2014) ‘The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells’, *Nature biotechnology*, 32(4), pp. 381–386. Available at: https://doi.org/10.1038/nbt.2859.

Trikha, P. *et al.* (2011) ‘E2F1–3 are critical for myeloid development’, *Journal of Biological* [Preprint]. Available at: https://www.jbc.org/article/S0021-9258(20)56180-2/abstract.

Tsai, S.F., Strauss, E. and Orkin, S.H. (1991) ‘Functional analysis and in vivo footprinting implicate the erythroid transcription factor GATA-1 as a positive regulator of its own promoter’, *Genes & development*, 5(6), pp. 919–931. Available at: https://doi.org/10.1101/gad.5.6.919.

Tseng, E. (no date) *cDNA_Cupcake: Miscellaneous collection of Python and R scripts for processing Iso-Seq data*. Github. Available at: https://github.com/Magdoll/cDNA_Cupcake (Accessed: 11 February 2023).

Uda, M. *et al.* (2008) ‘Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia’, *Proceedings of the National Academy of Sciences of the United States of America*, 105(5), pp. 1620–1625. Available at: https://doi.org/10.1073/pnas.0711566105.

Upadhaya, S. *et al.* (2018) ‘Kinetics of adult hematopoietic stem cell differentiation in vivo’, *The Journal of experimental medicine*, 215(11), pp. 2815–2832. Available at: https://doi.org/10.1084/jem.20180136.

Uxa, S. *et al.* (2021) ‘Ki-67 gene expression’, *Cell death and differentiation*, 28(12), pp. 3357–3370. Available at: https://doi.org/10.1038/s41418-021-00823-x.

Van Rossum, G. and Drake, F.L. (2009) *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. Available at: https://dl.acm.org/doi/book/10.5555/1593511.

Vas, V. *et al.* (2012) ‘Aging of the microenvironment influences clonality in hematopoiesis’, *PloS one*, 7(8), p. e42080. Available at: https://doi.org/10.1371/journal.pone.0042080.

Vitali, C. *et al.* (2015) ‘SOCS2 Controls Proliferation and Stemness of Hematopoietic Cells under Stress Conditions and Its Deregulation Marks Unfavorable Acute Leukemias’, *Cancer research*, 75(11), pp. 2387–2399. Available at: https://doi.org/10.1158/0008-5472.CAN-14-3625.

Vogel, C. and Marcotte, E.M. (2012) ‘Insights into the regulation of protein abundance from proteomic and transcriptomic analyses’, *Nature reviews. Genetics*, 13(4), pp. 227–232. Available at: https://doi.org/10.1038/nrg3185.

Wagers, A.J. and Weissman, I.L. (2006) ‘Differential expression of alpha2 integrin separates long-term and short-term reconstituting Lin-/loThy1.1(lo)c-kit+ Sca-1+ hematopoietic stem cells’, *Stem cells*, 24(4), pp. 1087–1094. Available at: https://doi.org/10.1634/stemcells.2005-0396.

Wahlestedt, M. *et al.* (2013) ‘An epigenetic component of hematopoietic stem cell aging amenable to reprogramming into a young state’, *Blood*, 121(21), pp. 4257–4264. Available at: https://doi.org/10.1182/blood-2012-11-469080.

Walker, M. *et al.* (2022) ‘An NFIX-mediated regulatory network governs the balance of hematopoietic stem and progenitor cells during hematopoiesis’, *Blood Advances* [Preprint]. Available at: https://doi.org/10.1182/bloodadvances.2022007811.

Walter, D. *et al.* (2015) 'Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells', *Nature*, 520(7548), pp. 549–552. Available at: https://doi.org/10.1038/nature14131.

Wang, E.T. *et al.* (2008) 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, 456(7221), pp. 470–476. Available at: https://doi.org/10.1038/nature07509.

Wang, P. *et al.* (2021) 'Loss of haspin suppresses cancer cell proliferation by interfering with cell cycle progression at multiple stages', *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 35(10), p. e21923. Available at: https://doi.org/10.1096/fj.202100099R.

Wang, Y. *et al.* (2002) 'Large scale identification of human hepatocellular carcinoma-associated antigens by autoantibodies', *Journal of immunology* , 169(2), pp. 1102–1109. Available at: https://doi.org/10.4049/jimmunol.169.2.1102.

Wan, Y. and Wu, C.J. (2013) 'SF3B1 mutations in chronic lymphocytic leukemia', *Blood*, 121(23), pp. 4627–4634. Available at: https://doi.org/10.1182/blood-2013-02-427641.

Watowich, S.S. (2011) 'The erythropoietin receptor: molecular structure and hematopoietic signaling pathways', *Journal of investigative medicine: the official publication of the American Federation for Clinical Research*, 59(7), pp. 1067–1072. Available at: https://doi.org/10.2310/JIM.0b013e31820fb28c.

Weinreb, C. *et al.* (2020) 'Lineage tracing on transcriptional landscapes links state to fate during differentiation', *Science*, 367(6479). Available at: https://doi.org/10.1126/science.aaw3381.

Weiskopf, K. *et al.* (2016) 'Myeloid Cell Origins, Differentiation, and Clinical Implications', *Microbiology spectrum*, 4(5). Available at: https://doi.org/10.1128/microbiolspec.MCHD-0031-2016.

Weissman, I.L. and Shizuru, J.A. (2008) '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), pp. 3543–3553. Available at: https://doi.org/10.1182/blood-2008-08-078220.

Wenger, A.M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature biotechnology*, 37(10), pp. 1155–1162. Available at: https://doi.org/10.1038/s41587-019-0217-9.

Wen, Q. *et al.* (2012) 'Identification of Regulators of Polyploidization Presents Therapeutic Targets for Treatment of AMKL', *Cell*, 150(3), pp. 575–589. Available at: https://doi.org/10.1016/j.cell.2012.06.032.

Wen, W.X., Mead, A.J. and Thongjuea, S. (2020) 'Technological advances and computational approaches for alternative splicing analysis in single cells', *Computational and structural biotechnology journal*, 18, pp. 332–343. Available at: https://doi.org/10.1016/j.csbj.2020.01.009.

Wessels, M.W. *et al.* (2021) 'Molecular analysis of the erythroid phenotype of a patient with BCL11A haploinsufficiency', *Blood advances*, 5(9), pp. 2339–2349. Available at: https://doi.org/10.1182/bloodadvances.2020003753.

Wickham, H. (2007) 'Reshaping Data with the reshape Package', *Journal of statistical software*, 21, pp. 1–20. Available at: https://doi.org/10.18637/jss.v021.i12.

Wickham, H. (2010) 'stringr: modern, consistent string processing', *The R Journal*, p. 38. Available at: https://doi.org/10.32614/rj-2010-012.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: https://play.google.com/store/books/details?id=bes-AAAAQBAJ.

Wickham, H. *et al.* (2019) 'Welcome to the tidyverse', *Journal of open source software*, 4(43), p. 1686. Available at: https://doi.org/10.21105/joss.01686.

Wilkinson, A.C. and Göttgens, B. (2013) 'Transcriptional regulation of haematopoietic stem cells', *Advances in experimental medicine and biology*, 786, pp. 187–212. Available at: https://doi.org/10.1007/978-94-007-6621-1_11.

Wilson, A. *et al.* (2008) 'Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair', *Cell*, 135(6), pp. 1118–1129. Available at: https://doi.org/10.1016/j.cell.2008.10.048.

Wilson, A. and Radtke, F. (2006) 'Multiple functions of Notch signaling in self-renewing organs and cancer', *FEBS letters*, 580(12), pp. 2860–2868. Available at: https://doi.org/10.1016/j.febslet.2006.03.024.

Wilson, J.E. (2003) 'Isozymes of mammalian hexokinase: structure, subcellular localization and metabolic function', *The Journal of experimental biology*, 206(Pt 12), pp. 2049–2057. Available at: https://doi.org/10.1242/jeb.00241.

Wilson, N.K. *et al.* (2010) 'Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators', *Cell stem cell*, 7(4), pp. 532–544. Available at: https://doi.org/10.1016/j.stem.2010.07.016.

Wolf, F.A. *et al.* (2019) 'PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells', *Genome biology*, 20(1), p. 59. Available at: https://doi.org/10.1186/s13059-019-1663-x.

Wu, A.M. *et al.* (1967) 'A cytological study of the capacity for differentiation of normal hemopoietic colony-forming cells', *Journal of cellular physiology*, 69(2), pp. 177–184. Available at: https://doi.org/10.1002/jcp.1040690208.

Wu, A.M. *et al.* (1968) 'Cytological evidence for a relationship between normal hemotopoietic colony-forming cells and cells of the lymphoid system', *The Journal of experimental medicine*, 127(3), pp. 455–464. Available at: https://doi.org/10.1084/jem.127.3.455.

Wu, T. *et al.* (2021) 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data', *The innovation journal: the public sector innovation journal*, 2(3), p. 100141. Available at: https://doi.org/10.1016/j.xinn.2021.100141.

Xavier-Ferrucio, J. *et al.* (2018) 'Hematopoietic defects in response to reduced Arhgap21', *Stem cell research*, 26, pp. 17–27. Available at: https://doi.org/10.1016/j.scr.2017.11.014.

Xavier-Ferrucio, J. and Krause, D.S. (2018) 'Concise Review: Bipotent Megakaryocytic-Erythroid Progenitors: Concepts and Controversies', *Stem cells*, 36(8), pp. 1138–1145. Available at: https://doi.org/10.1002/stem.2834.

Xiong, Y. *et al.* (2014) 'Erythrocyte-derived sphingosine 1-phosphate is essential for vascular development', *The Journal of clinical investigation*, 124(11), pp. 4823–4828. Available at: https://doi.org/10.1172/JCI77685.

Xu, J.-Z., Zhang, J.-L. and Zhang, W.-G. (2018) 'Antisense RNA: the new favorite in genetic research', *Journal of Zhejiang University. Science. B*, 19(10), pp. 739–749. Available at: https://doi.org/10.1631/jzus.B1700594.

Xu, S. *et al.* (2018) 'E2F1 Suppresses Oxidative Metabolism and Endothelial Differentiation of Bone Marrow Progenitor Cells', *Circulation research*, 122(5), pp. 701–711. Available at:

https://doi.org/10.1161/CIRCRESAHA.117.311814.

Xu, X. *et al.* (2015) ‘Aldehyde dehydrogenases and cancer stem cells’, *Cancer letters*, 369(1), pp. 50–57. Available at: https://doi.org/10.1016/j.canlet.2015.08.018.

Xu, Y. *et al.* (2022) ‘Cell Division Cycle-Associated Protein 3 (CDCA3) Is a Potential Biomarker for Clinical Prognosis and Immunotherapy in Pan-Cancer’, *BioMed research international*, 2022, p. 4632453. Available at: https://doi.org/10.1155/2022/4632453.

Yamada, M. *et al.* (1995) ‘Thrombopoietin induces tyrosine phosphorylation and activation of mitogen-activated protein kinases in a human thrombopoietin-dependent cell line’, *Biochemical and biophysical research communications*, 217(1), pp. 230–237. Available at: https://doi.org/10.1006/bbrc.1995.2768.

Yamamoto, R. *et al.* (2013) ‘Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells’, *Cell*, 154(5), pp. 1112–1126. Available at: https://doi.org/10.1016/j.cell.2013.08.007.

Yang, D. and de Haan, G. (2021) ‘Inflammation and Aging of Hematopoietic Stem Cells in Their Niche’, *Cells* , 10(8). Available at: https://doi.org/10.3390/cells10081849.

Yang, L.-X. *et al.* (2022) ‘C1Q labels a highly aggressive macrophage-like leukemia population indicating extramedullary infiltration and relapse’, *Blood* [Preprint]. Available at: https://doi.org/10.1182/blood.2022017046.

Yap, K. and Makeyev, E.V. (2016) ‘Functional impact of splice isoform diversity in individual cells’, *Biochemical Society transactions*, 44(4), pp. 1079–1085. Available at: https://doi.org/10.1042/BST20160103.

Yokota, T. *et al.* (2009) ‘The endothelial antigen ESAM marks primitive hematopoietic progenitors throughout life in mice’, *Blood*, 113(13), pp. 2914–2923. Available at: https://doi.org/10.1182/blood-2008-07-167106.

Yoshida, K. *et al.* (2011) ‘Frequent pathway mutations of splicing machinery in myelodysplasia’, *Nature*, 478(7367), pp. 64–69. Available at: https://doi.org/10.1038/nature10496.

Yoshihara, H. *et al.* (2007) ‘Thrombopoietin/MPL signaling regulates hematopoietic stem cell quiescence and interaction with the osteoblastic niche’, *Cell stem cell*, 1(6), pp. 685–697. Available at: https://doi.org/10.1016/j.stem.2007.10.020.

Young, M.D., Wakefield, M.J. and Smyth, G.K. (2012) ‘goseq: Gene Ontology testing for RNA-seq datasets’, *R Bioconductor* [Preprint]. Available at: https://www.andersvercelli.com/packages/3.8/bioc/vignettes/goseq/inst/doc/goseq.pdf.

Yu, C. *et al.* (2002) ‘Targeted deletion of a high-affinity GATA-binding site in the GATA-1 promoter leads to selective loss of the eosinophil lineage in vivo’, *The Journal of experimental medicine*, 195(11), pp. 1387–1395. Available at: https://doi.org/10.1084/jem.20020656.

Yu, G. *et al.* (2012) ‘clusterProfiler: an R package for comparing biological themes among gene clusters’, *Omics: a journal of integrative biology*, 16(5), pp. 284–287. Available at: https://doi.org/10.1089/omi.2011.0118.

Yu, G. and He, Q.-Y. (2016) ‘ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization’, *Molecular bioSystems*, 12(2), pp. 477–479. Available at: https://doi.org/10.1039/c5mb00663e.

Yu, H. *et al.* (2012) ‘Association between single nucleotide polymorphisms in ERCC4 and risk of squamous cell carcinoma of the head and neck’, *PloS one*, 7(7), p. e41853. Available at:

https://doi.org/10.1371/journal.pone.0041853.

Zahavi, J. *et al.* (1980) 'Enhanced in vivo platelet "release reaction" in old healthy individuals', *Thrombosis research*, 17(3-4), pp. 329–336. Available at: https://doi.org/10.1016/0049-3848(80)90067-5.

Zannettino, A.C. *et al.* (1998) 'The sialomucin CD164 (MGC-24v) is an adhesive glycoprotein expressed by human hematopoietic progenitors and bone marrow stromal cells that serves as a potent negative regulator of hematopoiesis', *Blood*, 92(8), pp. 2613–2628. Available at: https://doi.org/10.1182/blood.V92.8.2613.

Zhang, C. *et al.* (1998) 'A cell surface receptor defined by a mAb mediates a unique type of cell death similar to oncosis', *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp. 6290–6295. Available at: https://doi.org/10.1073/pnas.95.11.6290.

Zhang, C.C. and Lodish, H.F. (2008) 'Cytokines regulating hematopoietic stem cell function', *Current opinion in hematology*, 15(4), pp. 307–311. Available at: https://doi.org/10.1097/MOH.0b013e3283007db5.

Zhang, J. *et al.* (2003) 'Identification of the haematopoietic stem cell niche and control of the niche size', *Nature*, 425(6960), pp. 836–841. Available at: https://doi.org/10.1038/nature02041.

Zhang, L. *et al.* (2020) 'Molecular and cellular mechanisms of aging in hematopoietic stem cells and their niches', *Journal of hematology & oncology*, 13(1), p. 157. Available at: https://doi.org/10.1186/s13045-020-00994-z.

Zhang, P. *et al.* (2004) 'Enhancement of Hematopoietic Stem Cell Repopulating Capacity and Self-Renewal in the Absence of the Transcription Factor C/EBPα', *Immunity*, 21(6), pp. 853–863. Available at: https://doi.org/10.1016/j.immuni.2004.11.006.

Zhang, Y. *et al.* (2004) 'Aberrant quantity and localization of Aurora-B/AIM-1 and survivin during megakaryocyte polyploidization and the consequences of Aurora-B/AIM-1-deregulated expression', *Blood*, 103(10), pp. 3717–3726. Available at: https://doi.org/10.1182/blood-2003-09-3365.

Zhao, A. *et al.* (2023) 'Epigenetic regulation in hematopoiesis and its implications in the targeted therapy of hematologic malignancies', *Signal Transduction and Targeted Therapy*, 8(1), pp. 1–40. Available at: https://doi.org/10.1038/s41392-023-01342-6.

Zhao, M. *et al.* (2014) 'Megakaryocytes maintain homeostatic quiescence and promote post-injury regeneration of hematopoietic stem cells', *Nature medicine*, 20(11), pp. 1321–1326. Available at: https://doi.org/10.1038/nm.3706.

Zheng, G.X.Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nature communications*, 8, p. 14049. Available at: https://doi.org/10.1038/ncomms14049.

Zhu, F. *et al.* (2018) 'Screening for genes that regulate the differentiation of human megakaryocytic lineage cells', *Proceedings of the National Academy of Sciences of the United States of America*, 115(40), pp. E9308–E9316. Available at: https://doi.org/10.1073/pnas.1805434115.

Ziegenhain, C. *et al.* (2017) 'Comparative Analysis of Single-Cell RNA Sequencing Methods', *Molecular cell*, 65(4), pp. 631–643.e4. Available at: https://doi.org/10.1016/j.molcel.2017.01.023.

Zikherman, J. and Weiss, A. (2008) 'Alternative splicing of CD45: the tip of the iceberg', *Immunity*, pp. 839–841. Available at: https://doi.org/10.1016/j.immuni.2008.12.005.

Zink, F. *et al.* (2017) ‘Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly’, *Blood*, 130(6), pp. 742–752. Available at: https://doi.org/10.1182/blood-2017-02-769869.

Zuccolo, J. *et al.* (2010) ‘Phylogenetic analysis of the MS4A and TMEM176 gene families’, *PloS one*, 5(2), p. e9369. Available at: https://doi.org/10.1371/journal.pone.0009369.

Appendices

Supplementary Table 2.1. Illumina index sequences used in Nextera library preparation of Smart-seq2 cDNA (Chapters 3 and 4).

Index Set Plate	Well position	i7 Index Sequence	i5 Index Sequence
Plate 1	A1	TAAGGCGA	CTCTCTAT
Plate 1	B1	TAAGGCGA	TATCCTCT
Plate 1	C1	TAAGGCGA	GTAAGGAG
Plate 1	D1	TAAGGCGA	ACTGCATA
Plate 1	E1	TAAGGCGA	AAGGAGTA
Plate 1	F1	TAAGGCGA	CTAAGCCT
Plate 1	G1	TAAGGCGA	CGTCTAAT
Plate 1	H1	TAAGGCGA	TCTCTCCG
Plate 1	A2	CGTACTAG	CTCTCTAT
Plate 1	B2	CGTACTAG	TATCCTCT
Plate 1	C2	CGTACTAG	GTAAGGAG
Plate 1	D2	CGTACTAG	ACTGCATA
Plate 1	E2	CGTACTAG	AAGGAGTA
Plate 1	F2	CGTACTAG	CTAAGCCT
Plate 1	G2	CGTACTAG	CGTCTAAT
Plate 1	H2	CGTACTAG	TCTCTCCG
Plate 1	A3	AGGCAGAA	CTCTCTAT
Plate 1	B3	AGGCAGAA	TATCCTCT
Plate 1	C3	AGGCAGAA	GTAAGGAG
Plate 1	D3	AGGCAGAA	ACTGCATA
Plate 1	E3	AGGCAGAA	AAGGAGTA
Plate 1	F3	AGGCAGAA	CTAAGCCT
Plate 1	G3	AGGCAGAA	CGTCTAAT
Plate 1	H3	AGGCAGAA	TCTCTCCG
Plate 1	A4	TCCTGAGC	CTCTCTAT
Plate 1	B4	TCCTGAGC	TATCCTCT
Plate 1	C4	TCCTGAGC	GTAAGGAG
Plate 1	D4	TCCTGAGC	ACTGCATA
Plate 1	E4	TCCTGAGC	AAGGAGTA
Plate 1	F4	TCCTGAGC	CTAAGCCT
Plate 1	G4	TCCTGAGC	CGTCTAAT
Plate 1	H4	TCCTGAGC	TCTCTCCG
Plate 1	A5	GGACTCCT	CTCTCTAT
Plate 1	B5	GGACTCCT	TATCCTCT
Plate 1	C5	GGACTCCT	GTAAGGAG
Plate 1	D5	GGACTCCT	ACTGCATA
Plate 1	E5	GGACTCCT	AAGGAGTA
Plate 1	F5	GGACTCCT	CTAAGCCT
Plate 1	G5	GGACTCCT	CGTCTAAT
Plate 1	H5	GGACTCCT	TCTCTCCG
Plate 1	A6	TAGGCATG	CTCTCTAT
Plate 1	B6	TAGGCATG	TATCCTCT

Plate 1	C6	TAGGCATG	GTAAGGAG
Plate 1	D6	TAGGCATG	ACTGCATA
Plate 1	E6	TAGGCATG	AAGGAGTA
Plate 1	F6	TAGGCATG	CTAAGCCT
Plate 1	G6	TAGGCATG	CGTCTAAT
Plate 1	H6	TAGGCATG	TCTCTCCG
Plate 1	A7	CTCTCTAC	CTCTCTAT
Plate 1	B7	CTCTCTAC	TATCCTCT
Plate 1	C7	CTCTCTAC	GTAAGGAG
Plate 1	D7	CTCTCTAC	ACTGCATA
Plate 1	E7	CTCTCTAC	AAGGAGTA
Plate 1	F7	CTCTCTAC	CTAAGCCT
Plate 1	G7	CTCTCTAC	CGTCTAAT
Plate 1	H7	CTCTCTAC	TCTCTCCG
Plate 1	A8	CGAGGCTG	CTCTCTAT
Plate 1	B8	CGAGGCTG	TATCCTCT
Plate 1	C8	CGAGGCTG	GTAAGGAG
Plate 1	D8	CGAGGCTG	ACTGCATA
Plate 1	E8	CGAGGCTG	AAGGAGTA
Plate 1	F8	CGAGGCTG	CTAAGCCT
Plate 1	G8	CGAGGCTG	CGTCTAAT
Plate 1	H8	CGAGGCTG	TCTCTCCG
Plate 1	A9	AAGAGGCA	CTCTCTAT
Plate 1	B9	AAGAGGCA	TATCCTCT
Plate 1	C9	AAGAGGCA	GTAAGGAG
Plate 1	D9	AAGAGGCA	ACTGCATA
Plate 1	E9	AAGAGGCA	AAGGAGTA
Plate 1	F9	AAGAGGCA	CTAAGCCT
Plate 1	G9	AAGAGGCA	CGTCTAAT
Plate 1	H9	AAGAGGCA	TCTCTCCG
Plate 1	A10	GTAGAGGA	CTCTCTAT
Plate 1	B10	GTAGAGGA	TATCCTCT
Plate 1	C10	GTAGAGGA	GTAAGGAG
Plate 1	D10	GTAGAGGA	ACTGCATA
Plate 1	E10	GTAGAGGA	AAGGAGTA
Plate 1	F10	GTAGAGGA	CTAAGCCT
Plate 1	G10	GTAGAGGA	CGTCTAAT
Plate 1	H10	GTAGAGGA	TCTCTCCG
Plate 1	A11	GCTCATGA	CTCTCTAT
Plate 1	B11	GCTCATGA	TATCCTCT
Plate 1	C11	GCTCATGA	GTAAGGAG
Plate 1	D11	GCTCATGA	ACTGCATA
Plate 1	E11	GCTCATGA	AAGGAGTA
Plate 1	F11	GCTCATGA	CTAAGCCT
Plate 1	G11	GCTCATGA	CGTCTAAT
Plate 1	H11	GCTCATGA	TCTCTCCG
Plate 1	A12	ATCTCAGG	CTCTCTAT
Plate 1	B12	ATCTCAGG	TATCCTCT
Plate 1	C12	ATCTCAGG	GTAAGGAG
Plate 1	D12	ATCTCAGG	ACTGCATA
Plate 1	E12	ATCTCAGG	AAGGAGTA

Plate 1	F12	ATCTCAGG	CTAAGCCT
Plate 1	G12	ATCTCAGG	CGTCTAAT
Plate 1	H12	ATCTCAGG	TCTCTCCG
Plate 2	A1	TAAGGCGA	TCGACTAG
Plate 2	B1	TAAGGCGA	TTCTAGCT
Plate 2	C1	TAAGGCGA	CCTAGAGT
Plate 2	D1	TAAGGCGA	GCGTAAGA
Plate 2	E1	TAAGGCGA	CTATTAAG
Plate 2	F1	TAAGGCGA	AAGGCTAT
Plate 2	G1	TAAGGCGA	GAGCCTTA
Plate 2	H1	TAAGGCGA	TTATGCGA
Plate 2	A2	CGTACTAG	TCGACTAG
Plate 2	B2	CGTACTAG	TTCTAGCT
Plate 2	C2	CGTACTAG	CCTAGAGT
Plate 2	D2	CGTACTAG	GCGTAAGA
Plate 2	E2	CGTACTAG	CTATTAAG
Plate 2	F2	CGTACTAG	AAGGCTAT
Plate 2	G2	CGTACTAG	GAGCCTTA
Plate 2	H2	CGTACTAG	TTATGCGA
Plate 2	A3	AGGCAGAA	TCGACTAG
Plate 2	B3	AGGCAGAA	TTCTAGCT
Plate 2	C3	AGGCAGAA	CCTAGAGT
Plate 2	D3	AGGCAGAA	GCGTAAGA
Plate 2	E3	AGGCAGAA	CTATTAAG
Plate 2	F3	AGGCAGAA	AAGGCTAT
Plate 2	G3	AGGCAGAA	GAGCCTTA
Plate 2	H3	AGGCAGAA	TTATGCGA
Plate 2	A4	TCCTGAGC	TCGACTAG
Plate 2	B4	TCCTGAGC	TTCTAGCT
Plate 2	C4	TCCTGAGC	CCTAGAGT
Plate 2	D4	TCCTGAGC	GCGTAAGA
Plate 2	E4	TCCTGAGC	CTATTAAG
Plate 2	F4	TCCTGAGC	AAGGCTAT
Plate 2	G4	TCCTGAGC	GAGCCTTA
Plate 2	H4	TCCTGAGC	TTATGCGA
Plate 2	A5	GGACTCCT	TCGACTAG
Plate 2	B5	GGACTCCT	TTCTAGCT
Plate 2	C5	GGACTCCT	CCTAGAGT
Plate 2	D5	GGACTCCT	GCGTAAGA
Plate 2	E5	GGACTCCT	CTATTAAG
Plate 2	F5	GGACTCCT	AAGGCTAT
Plate 2	G5	GGACTCCT	GAGCCTTA
Plate 2	H5	GGACTCCT	TTATGCGA
Plate 2	A6	TAGGCATG	TCGACTAG
Plate 2	B6	TAGGCATG	TTCTAGCT
Plate 2	C6	TAGGCATG	CCTAGAGT
Plate 2	D6	TAGGCATG	GCGTAAGA
Plate 2	E6	TAGGCATG	CTATTAAG
Plate 2	F6	TAGGCATG	AAGGCTAT
Plate 2	G6	TAGGCATG	GAGCCTTA
Plate 2	H6	TAGGCATG	TTATGCGA

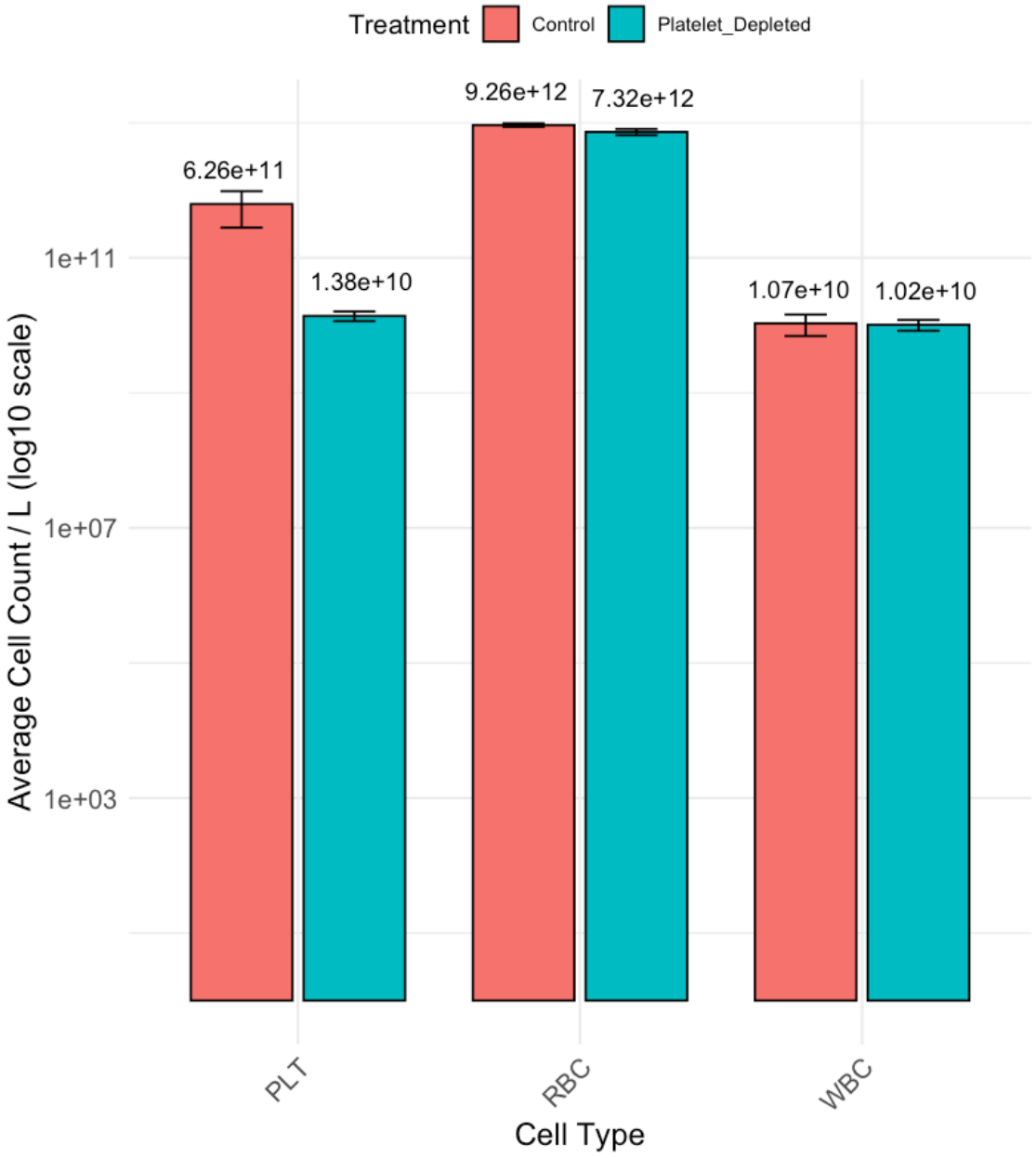
Plate 2	A7	CTCTCTAC	TCGACTAG
Plate 2	B7	CTCTCTAC	TTCTAGCT
Plate 2	C7	CTCTCTAC	CCTAGAGT
Plate 2	D7	CTCTCTAC	GCGTAAGA
Plate 2	E7	CTCTCTAC	CTATTAAG
Plate 2	F7	CTCTCTAC	AAGGCTAT
Plate 2	G7	CTCTCTAC	GAGCCTTA
Plate 2	H7	CTCTCTAC	TTATGCGA
Plate 2	A8	CGAGGCTG	TCGACTAG
Plate 2	B8	CGAGGCTG	TTCTAGCT
Plate 2	C8	CGAGGCTG	CCTAGAGT
Plate 2	D8	CGAGGCTG	GCGTAAGA
Plate 2	E8	CGAGGCTG	CTATTAAG
Plate 2	F8	CGAGGCTG	AAGGCTAT
Plate 2	G8	CGAGGCTG	GAGCCTTA
Plate 2	H8	CGAGGCTG	TTATGCGA
Plate 2	A9	AAGAGGCA	TCGACTAG
Plate 2	B9	AAGAGGCA	TTCTAGCT
Plate 2	C9	AAGAGGCA	CCTAGAGT
Plate 2	D9	AAGAGGCA	GCGTAAGA
Plate 2	E9	AAGAGGCA	CTATTAAG
Plate 2	F9	AAGAGGCA	AAGGCTAT
Plate 2	G9	AAGAGGCA	GAGCCTTA
Plate 2	H9	AAGAGGCA	TTATGCGA
Plate 2	A10	GTAGAGGA	TCGACTAG
Plate 2	B10	GTAGAGGA	TTCTAGCT
Plate 2	C10	GTAGAGGA	CCTAGAGT
Plate 2	D10	GTAGAGGA	GCGTAAGA
Plate 2	E10	GTAGAGGA	CTATTAAG
Plate 2	F10	GTAGAGGA	AAGGCTAT
Plate 2	G10	GTAGAGGA	GAGCCTTA
Plate 2	H10	GTAGAGGA	TTATGCGA
Plate 2	A11	GCTCATGA	TCGACTAG
Plate 2	B11	GCTCATGA	TTCTAGCT
Plate 2	C11	GCTCATGA	CCTAGAGT
Plate 2	D11	GCTCATGA	GCGTAAGA
Plate 2	E11	GCTCATGA	CTATTAAG
Plate 2	F11	GCTCATGA	AAGGCTAT
Plate 2	G11	GCTCATGA	GAGCCTTA
Plate 2	H11	GCTCATGA	TTATGCGA
Plate 2	A12	ATCTCAGG	TCGACTAG
Plate 2	B12	ATCTCAGG	TTCTAGCT
Plate 2	C12	ATCTCAGG	CCTAGAGT
Plate 2	D12	ATCTCAGG	GCGTAAGA
Plate 2	E12	ATCTCAGG	CTATTAAG
Plate 2	F12	ATCTCAGG	AAGGCTAT
Plate 2	G12	ATCTCAGG	GAGCCTTA
Plate 2	H12	ATCTCAGG	TTATGCGA
Plate 3	A1	ACTCGCTA	CTCTCTAT
Plate 3	B1	ACTCGCTA	TATCCTCT
Plate 3	C1	ACTCGCTA	GTAAGGAG

Plate 3	D1	ACTCGCTA	ACTGCATA
Plate 3	E1	ACTCGCTA	AAGGAGTA
Plate 3	F1	ACTCGCTA	CTAAGCCT
Plate 3	G1	ACTCGCTA	CGTCTAAT
Plate 3	H1	ACTCGCTA	TCTCTCCG
Plate 3	A2	GGAGCTAC	CTCTCTAT
Plate 3	B2	GGAGCTAC	TATCCTCT
Plate 3	C2	GGAGCTAC	GTAAGGAG
Plate 3	D2	GGAGCTAC	ACTGCATA
Plate 3	E2	GGAGCTAC	AAGGAGTA
Plate 3	F2	GGAGCTAC	CTAAGCCT
Plate 3	G2	GGAGCTAC	CGTCTAAT
Plate 3	H2	GGAGCTAC	TCTCTCCG
Plate 3	A3	GCGTAGTA	CTCTCTAT
Plate 3	B3	GCGTAGTA	TATCCTCT
Plate 3	C3	GCGTAGTA	GTAAGGAG
Plate 3	D3	GCGTAGTA	ACTGCATA
Plate 3	E3	GCGTAGTA	AAGGAGTA
Plate 3	F3	GCGTAGTA	CTAAGCCT
Plate 3	G3	GCGTAGTA	CGTCTAAT
Plate 3	H3	GCGTAGTA	TCTCTCCG
Plate 3	A4	CGGAGCCT	CTCTCTAT
Plate 3	B4	CGGAGCCT	TATCCTCT
Plate 3	C4	CGGAGCCT	GTAAGGAG
Plate 3	D4	CGGAGCCT	ACTGCATA
Plate 3	E4	CGGAGCCT	AAGGAGTA
Plate 3	F4	CGGAGCCT	CTAAGCCT
Plate 3	G4	CGGAGCCT	CGTCTAAT
Plate 3	H4	CGGAGCCT	TCTCTCCG
Plate 3	A5	TACGCTGC	CTCTCTAT
Plate 3	B5	TACGCTGC	TATCCTCT
Plate 3	C5	TACGCTGC	GTAAGGAG
Plate 3	D5	TACGCTGC	ACTGCATA
Plate 3	E5	TACGCTGC	AAGGAGTA
Plate 3	F5	TACGCTGC	CTAAGCCT
Plate 3	G5	TACGCTGC	CGTCTAAT
Plate 3	H5	TACGCTGC	TCTCTCCG
Plate 3	A6	ATGCGCAG	CTCTCTAT
Plate 3	B6	ATGCGCAG	TATCCTCT
Plate 3	C6	ATGCGCAG	GTAAGGAG
Plate 3	D6	ATGCGCAG	ACTGCATA
Plate 3	E6	ATGCGCAG	AAGGAGTA
Plate 3	F6	ATGCGCAG	CTAAGCCT
Plate 3	G6	ATGCGCAG	CGTCTAAT
Plate 3	H6	ATGCGCAG	TCTCTCCG
Plate 3	A7	TAGCGCTC	CTCTCTAT
Plate 3	B7	TAGCGCTC	TATCCTCT
Plate 3	C7	TAGCGCTC	GTAAGGAG
Plate 3	D7	TAGCGCTC	ACTGCATA
Plate 3	E7	TAGCGCTC	AAGGAGTA
Plate 3	F7	TAGCGCTC	CTAAGCCT

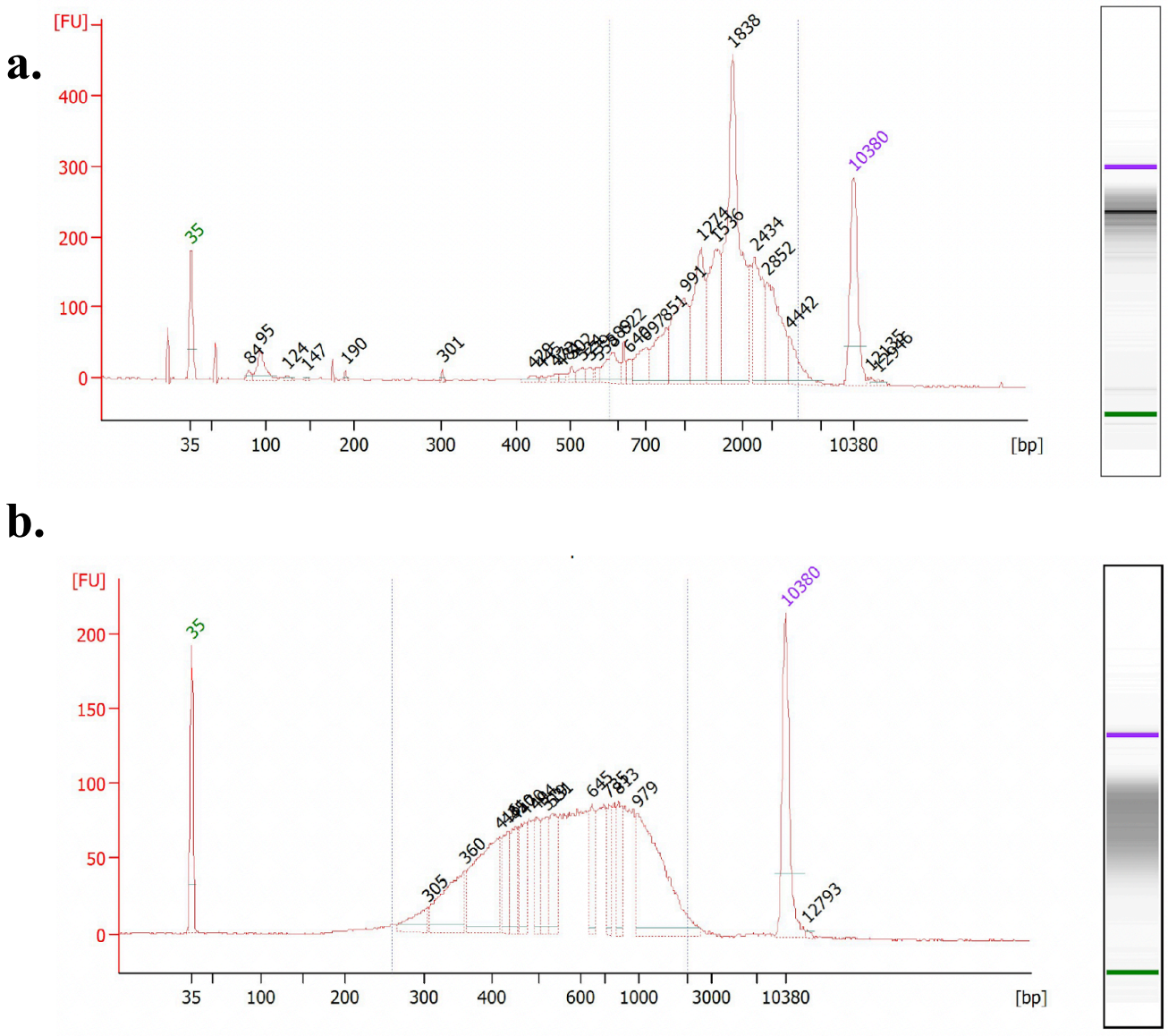
Plate 3	G7	TAGCGCTC	CGTCTAAT
Plate 3	H7	TAGCGCTC	TCTCTCCG
Plate 3	A8	ACTGAGCG	CTCTCTAT
Plate 3	B8	ACTGAGCG	TATCCTCT
Plate 3	C8	ACTGAGCG	GTAAGGAG
Plate 3	D8	ACTGAGCG	ACTGCATA
Plate 3	E8	ACTGAGCG	AAGGAGTA
Plate 3	F8	ACTGAGCG	CTAAGCCT
Plate 3	G8	ACTGAGCG	CGTCTAAT
Plate 3	H8	ACTGAGCG	TCTCTCCG
Plate 3	A9	CCTAAGAC	CTCTCTAT
Plate 3	B9	CCTAAGAC	TATCCTCT
Plate 3	C9	CCTAAGAC	GTAAGGAG
Plate 3	D9	CCTAAGAC	ACTGCATA
Plate 3	E9	CCTAAGAC	AAGGAGTA
Plate 3	F9	CCTAAGAC	CTAAGCCT
Plate 3	G9	CCTAAGAC	CGTCTAAT
Plate 3	H9	CCTAAGAC	TCTCTCCG
Plate 3	A10	CGATCAGT	CTCTCTAT
Plate 3	B10	CGATCAGT	TATCCTCT
Plate 3	C10	CGATCAGT	GTAAGGAG
Plate 3	D10	CGATCAGT	ACTGCATA
Plate 3	E10	CGATCAGT	AAGGAGTA
Plate 3	F10	CGATCAGT	CTAAGCCT
Plate 3	G10	CGATCAGT	CGTCTAAT
Plate 3	H10	CGATCAGT	TCTCTCCG
Plate 3	A11	TGCAGCTA	CTCTCTAT
Plate 3	B11	TGCAGCTA	TATCCTCT
Plate 3	C11	TGCAGCTA	GTAAGGAG
Plate 3	D11	TGCAGCTA	ACTGCATA
Plate 3	E11	TGCAGCTA	AAGGAGTA
Plate 3	F11	TGCAGCTA	CTAAGCCT
Plate 3	G11	TGCAGCTA	CGTCTAAT
Plate 3	H11	TGCAGCTA	TCTCTCCG
Plate 3	A12	TCGACGTC	CTCTCTAT
Plate 3	B12	TCGACGTC	TATCCTCT
Plate 3	C12	TCGACGTC	GTAAGGAG
Plate 3	D12	TCGACGTC	ACTGCATA
Plate 3	E12	TCGACGTC	AAGGAGTA
Plate 3	F12	TCGACGTC	CTAAGCCT
Plate 3	G12	TCGACGTC	CGTCTAAT
Plate 3	H12	TCGACGTC	TCTCTCCG
Plate 4	A1	ACTCGCTA	TCGACTAG
Plate 4	B1	ACTCGCTA	TTCTAGCT
Plate 4	C1	ACTCGCTA	CCTAGAGT
Plate 4	D1	ACTCGCTA	GCGTAAGA
Plate 4	E1	ACTCGCTA	CTATTAAG
Plate 4	F1	ACTCGCTA	AAGGCTAT
Plate 4	G1	ACTCGCTA	GAGCCTTA
Plate 4	H1	ACTCGCTA	TTATGCGA
Plate 4	A2	GGAGCTAC	TCGACTAG

Plate 4	B2	GGAGCTAC	TTCTAGCT
Plate 4	C2	GGAGCTAC	CCTAGAGT
Plate 4	D2	GGAGCTAC	GCGTAAGA
Plate 4	E2	GGAGCTAC	CTATTAAG
Plate 4	F2	GGAGCTAC	AAGGCTAT
Plate 4	G2	GGAGCTAC	GAGCCTTA
Plate 4	H2	GGAGCTAC	TTATGCGA
Plate 4	A3	GCGTAGTA	TCGACTAG
Plate 4	B3	GCGTAGTA	TTCTAGCT
Plate 4	C3	GCGTAGTA	CCTAGAGT
Plate 4	D3	GCGTAGTA	GCGTAAGA
Plate 4	E3	GCGTAGTA	CTATTAAG
Plate 4	F3	GCGTAGTA	AAGGCTAT
Plate 4	G3	GCGTAGTA	GAGCCTTA
Plate 4	H3	GCGTAGTA	TTATGCGA
Plate 4	A4	CGGAGCCT	TCGACTAG
Plate 4	B4	CGGAGCCT	TTCTAGCT
Plate 4	C4	CGGAGCCT	CCTAGAGT
Plate 4	D4	CGGAGCCT	GCGTAAGA
Plate 4	E4	CGGAGCCT	CTATTAAG
Plate 4	F4	CGGAGCCT	AAGGCTAT
Plate 4	G4	CGGAGCCT	GAGCCTTA
Plate 4	H4	CGGAGCCT	TTATGCGA
Plate 4	A5	TACGCTGC	TCGACTAG
Plate 4	B5	TACGCTGC	TTCTAGCT
Plate 4	C5	TACGCTGC	CCTAGAGT
Plate 4	D5	TACGCTGC	GCGTAAGA
Plate 4	E5	TACGCTGC	CTATTAAG
Plate 4	F5	TACGCTGC	AAGGCTAT
Plate 4	G5	TACGCTGC	GAGCCTTA
Plate 4	H5	TACGCTGC	TTATGCGA
Plate 4	A6	ATGCGCAG	TCGACTAG
Plate 4	B6	ATGCGCAG	TTCTAGCT
Plate 4	C6	ATGCGCAG	CCTAGAGT
Plate 4	D6	ATGCGCAG	GCGTAAGA
Plate 4	E6	ATGCGCAG	CTATTAAG
Plate 4	F6	ATGCGCAG	AAGGCTAT
Plate 4	G6	ATGCGCAG	GAGCCTTA
Plate 4	H6	ATGCGCAG	TTATGCGA
Plate 4	A7	TAGCGCTC	TCGACTAG
Plate 4	B7	TAGCGCTC	TTCTAGCT
Plate 4	C7	TAGCGCTC	CCTAGAGT
Plate 4	D7	TAGCGCTC	GCGTAAGA
Plate 4	E7	TAGCGCTC	CTATTAAG
Plate 4	F7	TAGCGCTC	AAGGCTAT
Plate 4	G7	TAGCGCTC	GAGCCTTA
Plate 4	H7	TAGCGCTC	TTATGCGA
Plate 4	A8	ACTGAGCG	TCGACTAG
Plate 4	B8	ACTGAGCG	TTCTAGCT
Plate 4	C8	ACTGAGCG	CCTAGAGT
Plate 4	D8	ACTGAGCG	GCGTAAGA

Plate 4	E8	ACTGAGCG	CTATTAAG
Plate 4	F8	ACTGAGCG	AAGGCTAT
Plate 4	G8	ACTGAGCG	GAGCCTTA
Plate 4	H8	ACTGAGCG	TTATGCGA
Plate 4	A9	CCTAAGAC	TCGACTAG
Plate 4	B9	CCTAAGAC	TTCTAGCT
Plate 4	C9	CCTAAGAC	CCTAGAGT
Plate 4	D9	CCTAAGAC	GCGTAAGA
Plate 4	E9	CCTAAGAC	CTATTAAG
Plate 4	F9	CCTAAGAC	AAGGCTAT
Plate 4	G9	CCTAAGAC	GAGCCTTA
Plate 4	H9	CCTAAGAC	TTATGCGA
Plate 4	A10	CGATCAGT	TCGACTAG
Plate 4	B10	CGATCAGT	TTCTAGCT
Plate 4	C10	CGATCAGT	CCTAGAGT
Plate 4	D10	CGATCAGT	GCGTAAGA
Plate 4	E10	CGATCAGT	CTATTAAG
Plate 4	F10	CGATCAGT	AAGGCTAT
Plate 4	G10	CGATCAGT	GAGCCTTA
Plate 4	H10	CGATCAGT	TTATGCGA
Plate 4	A11	TGCAGCTA	TCGACTAG
Plate 4	B11	TGCAGCTA	TTCTAGCT
Plate 4	C11	TGCAGCTA	CCTAGAGT
Plate 4	D11	TGCAGCTA	GCGTAAGA
Plate 4	E11	TGCAGCTA	CTATTAAG
Plate 4	F11	TGCAGCTA	AAGGCTAT
Plate 4	G11	TGCAGCTA	GAGCCTTA
Plate 4	H11	TGCAGCTA	TTATGCGA
Plate 4	A12	TCGACGTC	TCGACTAG
Plate 4	B12	TCGACGTC	TTCTAGCT
Plate 4	C12	TCGACGTC	CCTAGAGT
Plate 4	D12	TCGACGTC	GCGTAAGA
Plate 4	E12	TCGACGTC	CTATTAAG
Plate 4	F12	TCGACGTC	AAGGCTAT
Plate 4	G12	TCGACGTC	GAGCCTTA
Plate 4	H12	TCGACGTC	TTATGCGA

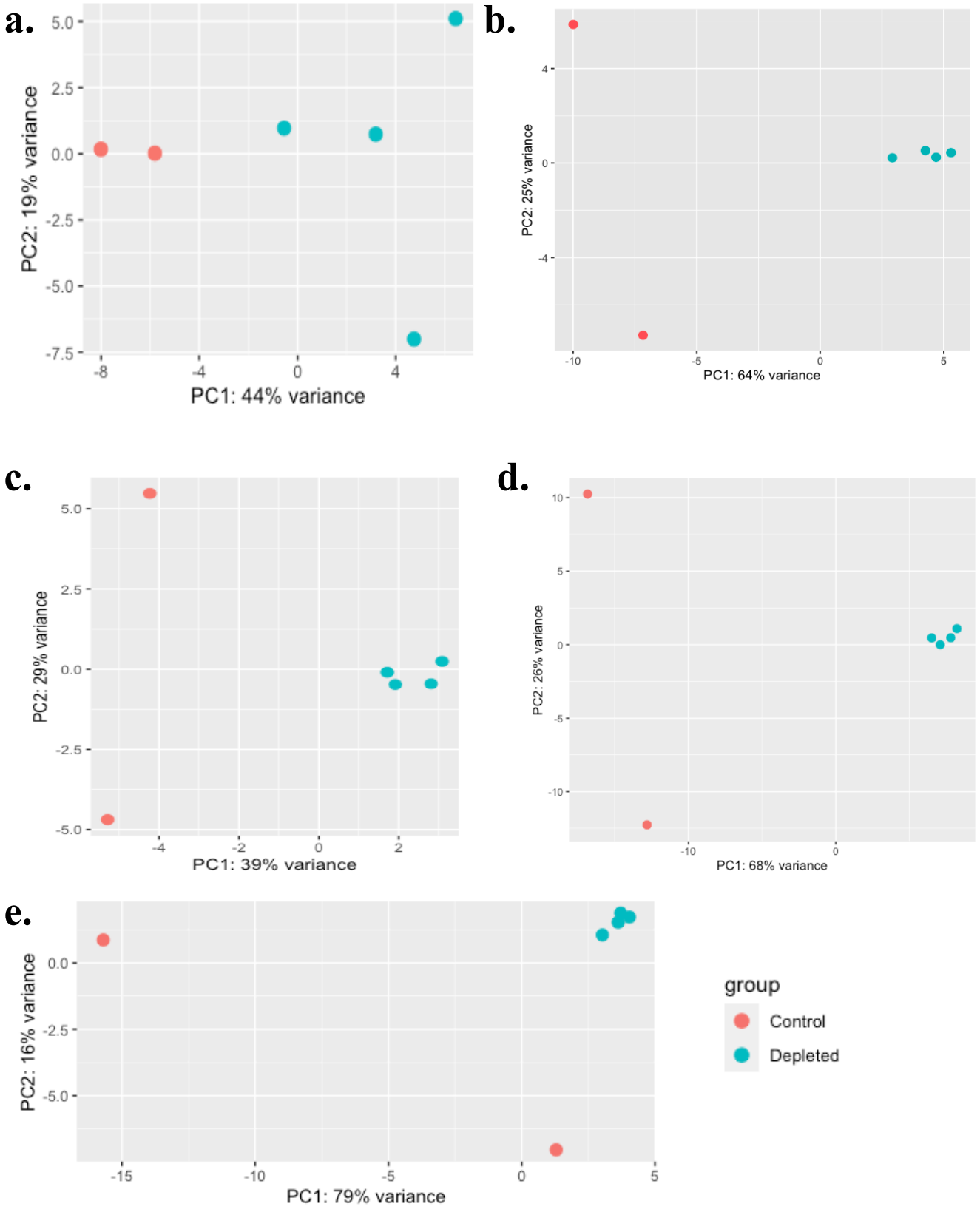


Supplementary Figure 3.1. The average count of platelets, RBCs and WBCs (Alinity HQ, Abbott Laboratories) from mouse peripheral blood samples collected 24 hours following treatment with platelet depletion or control antibody (n = 2 per treatment condition).

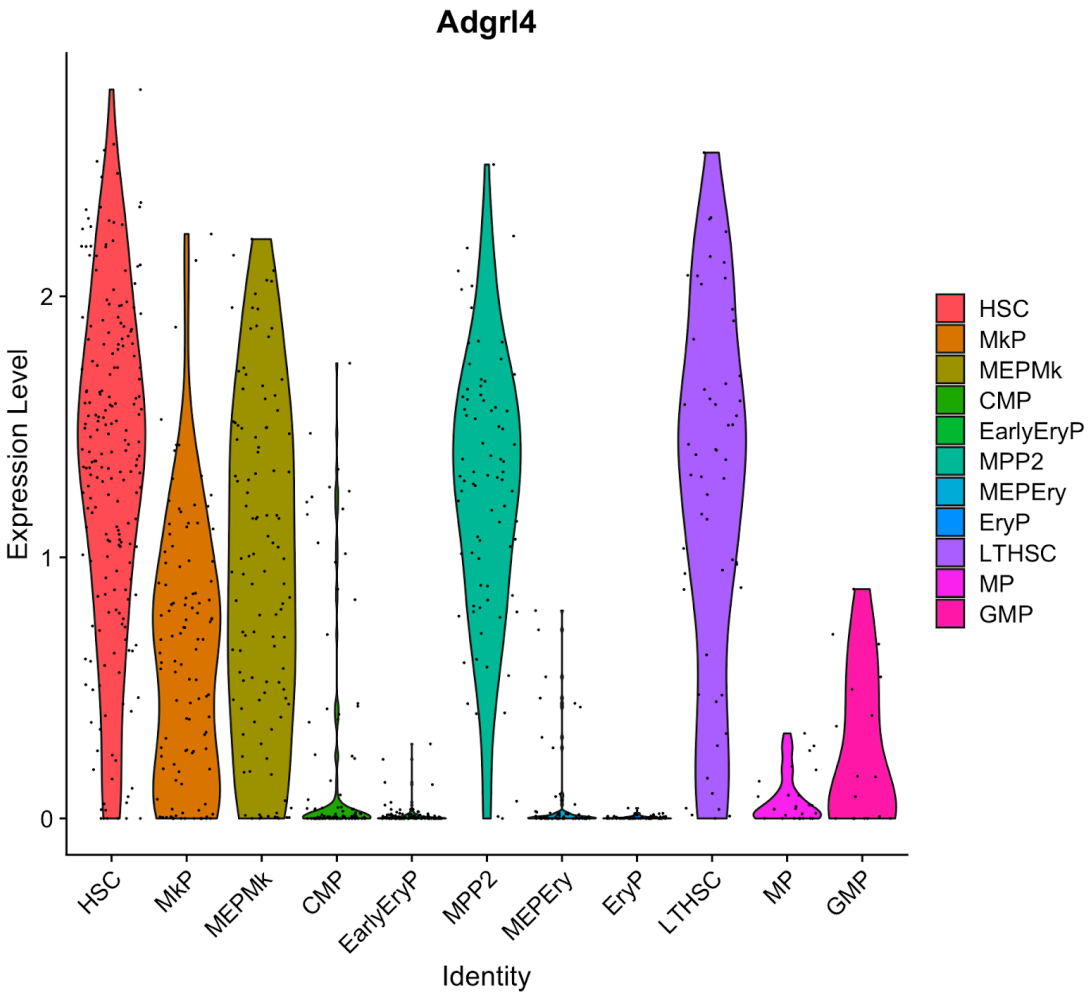


Supplementary Figure 3.2. Representative bioanalyzer size distribution traces during Smart-seq2 library preparation quality control. (a) Single-cell cDNA trace of post bead clean-up (b) Nextera library pool of 384 single-cells. All libraries were normalised to 1 ng/L.

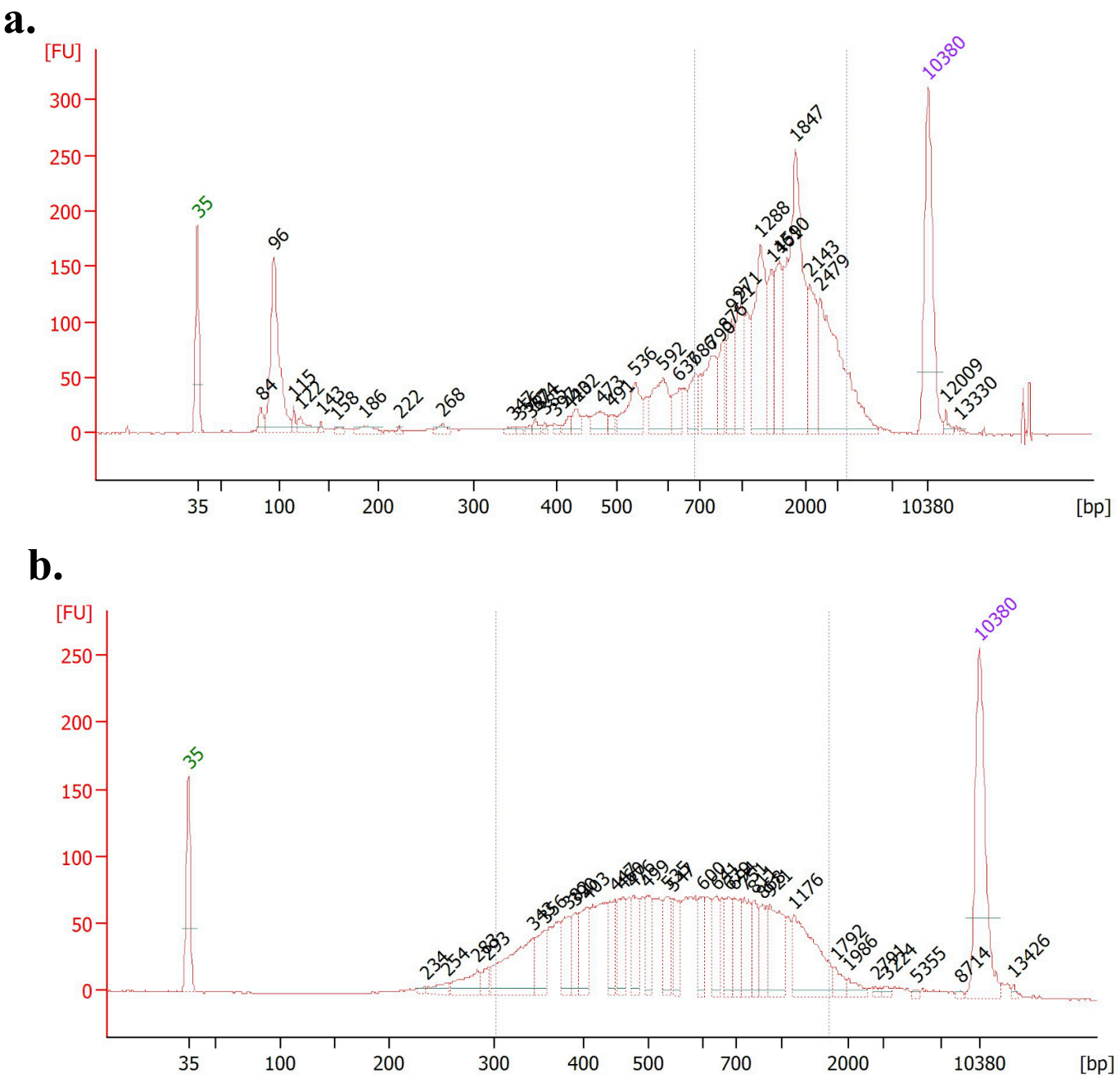
Supplementary Figure 3.3. MultiQC report summary statistics for one plate of Smart-seq2 libraries. (A) Percentage of uniquely assigned reads per cell using featureCounts (B) Percentage of reads uniquely mapping to the mouse genome per cell using STAR.



Supplementary Figure 3.4. Sample level variance of biological replicates across experimental conditions for each cell type. (a) LT-HSC (b) HSC (c) MPP2 (d) Mk-MEPs and (e) MkPs.



Supplementary Figure 3.5. Expression of *Aldgrl4* across cell-types showing a correlation between its expression and the Mk lineage.



Supplementary Figure 4.1. Representative bioanalyzer size distribution traces during Smart-seq2 library preparation quality control. (a) Single-cell cDNA trace of post bead clean-up (b) NextEra library pool of 384 single-cells. All libraries were normalised to 1ng/uL for loading.

a.

Module	Gene ID	Cell Type(s)
11	*Sult1a1*	LTHSCS
16	*Gm19590*	LTHSCS
16	*Prtn3*	LTHSCS
11	*Clec1a*	LTHSCS
16	*Tgm2*	LTHSCS
16	*Gm2830*	LTHSCS
16	*Procr*	LTHSCS
16	*Pglyrp2*	LTHSCS
16	*Mecom*	LTHSCS
16	*Myct1*	LTHSCS
16	*Tmem176a*	LTHSCS
16	*H2-Aa*	LTHSCS
11	*Cpne8*	LTHSCS
11	*Ltb*	LTHSCS
16	*Gm8292*	LTHSCS
16	*Cd63*	LTHSCS
16	*Pygm*	LTHSCS
11	*Iigp1*	LTHSCS
16	*Arhgef6*	LTHSCS
16	*Hacd4*	LTHSCS

b.

Module	Gene ID	Cell Type(s)
2	*Mpl*	Mk lineage
2	*Tmsb4x*	Mk lineage
2	*Pkm*	Mk lineage
2	*Cavin2*	Mk lineage
2	*Mmrn1*	Mk lineage
2	*Lyz2*	Mk lineage
2	*Fxyd5*	Mk lineage
2	*Meis1*	Mk lineage
2	*Tbxas1*	Mk lineage
2	*Ctla2a*	Mk lineage
2	*Pbx1*	Mk lineage
2	*Ptgs1*	Mk lineage
2	*Atp2a3*	Mk lineage
2	*Tmem176b*	Mk lineage
2	*Esam*	Mk lineage
2	*Gm6560*	Mk lineage
2	*Trpc6*	Mk lineage
2	*Angpt1*	Mk lineage
2	*Alox5ap*	Mk lineage
2	*Fli1*	Mk lineage

c.

Module	Gene ID	Cell Type(s)
5	*Car1*	Ery lineage
5	*Mfsd2b*	Ery lineage
5	*Ermap*	Ery lineage
5	*Atp1b2*	Ery lineage
5	*Klf1*	Ery lineage
5	*Tspo2*	Ery lineage
5	*Ugcg*	Ery lineage
5	*Trib2*	Ery lineage
5	*Gm15915*	Ery lineage
5	*Aqp1*	Ery lineage
5	*Optn*	Ery lineage
5	*Ces2g*	Ery lineage
5	*Mns1*	Ery lineage
5	*Abcb4*	Ery lineage
5	*Blvrb*	Ery lineage
5	*Arap3*	Ery lineage
5	*Cbx3*	Ery lineage
5	*Slc25a21*	Ery lineage
5	*S100a10*	Ery lineage
5	*Myb*	Ery lineage

d.

Module	Gene ID	Cell Type(s)
17	*Pf4*	MkP
17	*Plek*	MkP
17	*Dusp3*	MkP
17	*Itga2b*	MkP
17	*Itgb3*	MkP
17	*Vwf*	MkP
17	*Robo3*	MkP
17	*Pdcd4*	MkP
17	*Tuba8*	MkP
17	*Kalrn*	MkP
17	*Rap1b*	MkP
17	*Rab27b*	MkP
17	*Serpine2*	MkP
17	*Cd9*	MkP
17	*Treml1*	MkP
17	*Slc14a1*	MkP
17	*Sla*	MkP
17	*Pde10a*	MkP
17	*F2r*	MkP
17	*Ehd3*	MkP

e.

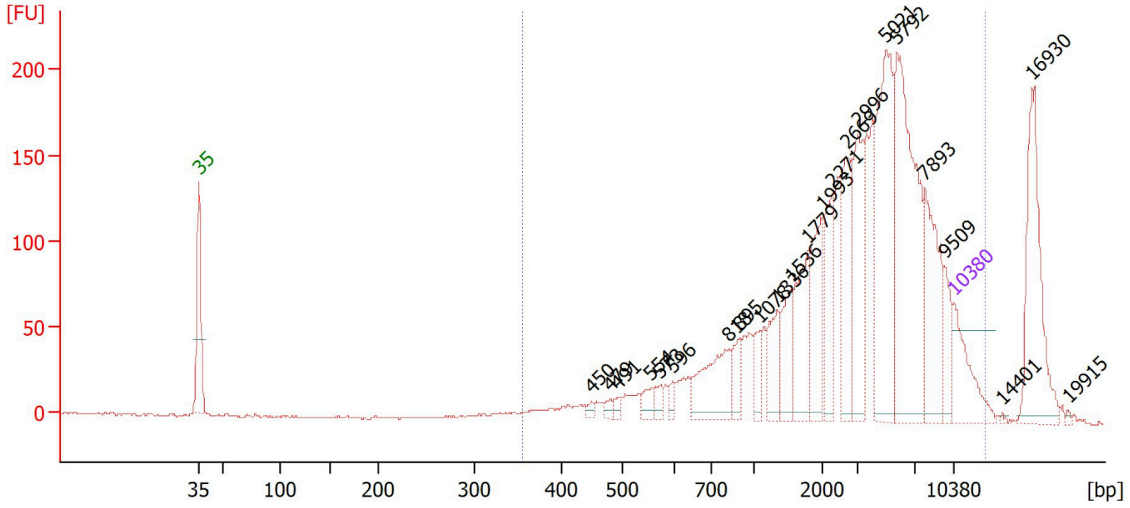
Module	Gene ID	Cell Type(s)
10	*Birc5*	MEP (cell cycling signature)
15	*Ube2c*	MEP (cell cycling signature)
10	*Aurkb*	MEP (cell cycling signature)
10	*Cdca3*	MEP (cell cycling signature)
10	*Ccna2*	MEP (cell cycling signature)
10	*Spc24*	MEP (cell cycling signature)
10	*Top2a*	MEP (cell cycling signature)
10	*Mki67*	MEP (cell cycling signature)
10	*Cdk1*	MEP (cell cycling signature)
15	*Nusap1*	MEP (cell cycling signature)
15	*Ccnb2*	MEP (cell cycling signature)
10	*Kif22*	MEP (cell cycling signature)
10	*Kif15*	MEP (cell cycling signature)
15	*Gm4739*	MEP (cell cycling signature)
10	*Cdca8*	MEP (cell cycling signature)
10	*Tpx2*	MEP (cell cycling signature)
15	*Hmmr*	MEP (cell cycling signature)
10	*Ankle1*	MEP (cell cycling signature)
10	*Gm14150*	MEP (cell cycling signature)
15	*Ckap2l*	MEP (cell cycling signature)

f.

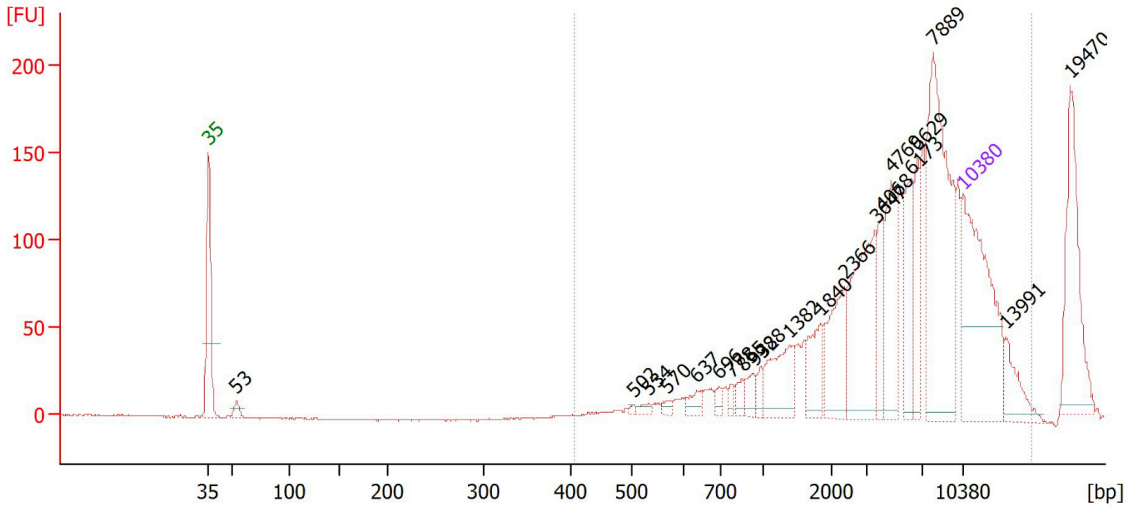
Module	Gene ID	Cell Type(s)
5	*Car1*	EryP
5	*Mfsd2b*	EryP
5	*Ermap*	EryP
5	*Atp1b2*	EryP
5	*Klf1*	EryP
5	*Tspo2*	EryP
5	*Ugcg*	EryP
5	*Trib2*	EryP
5	*Gm15915*	EryP
1	*Rhd*	EryP
5	*Aqp1*	EryP
5	*Optn*	EryP
5	*Ces2g*	EryP
1	*Sphk1*	EryP
5	*Mns1*	EryP
1	*Mt1*	EryP
5	*Abcb4*	EryP
5	*Blvrb*	EryP
5	*Arap3*	EryP
1	*Ank1*	EryP

Supplementary Figure 4.2. Top 20 genes within modules calculated from genes differentially expressed along pseudotime that correlated with cell cluster identity (a) LT-HSCS (b) throughout Mk lineage (c) throughout Ery lineage (d) MkPs (e) MEPs (cell cycling signature) (f) EryPs.

a.

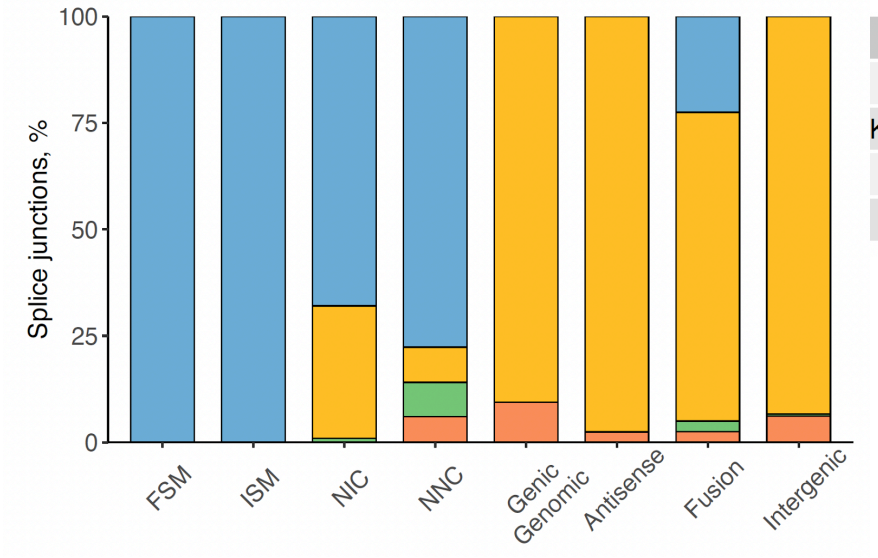


b.



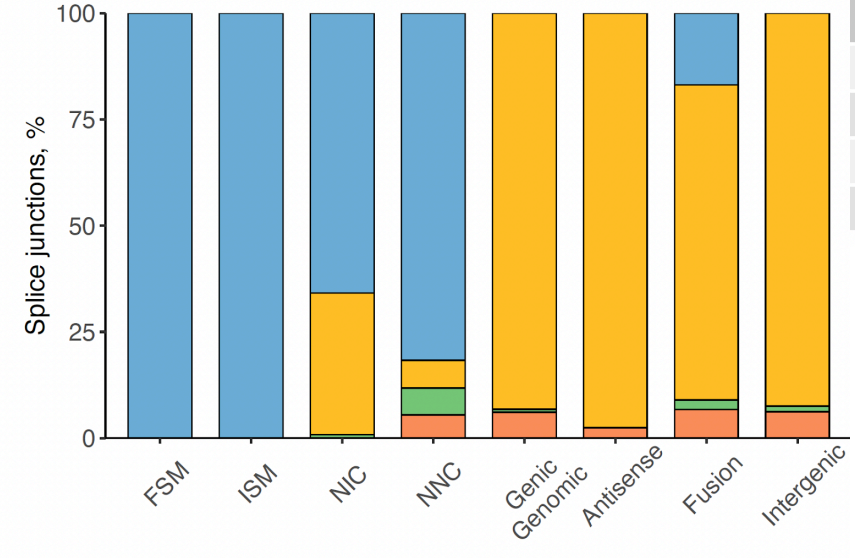
Supplementary Figure 5.1. **Size distribution bioanalyzer traces of Iso-Seq libraries.** (a) Size distribution of final young Iso-Seq library and (b) Final old Iso-Seq library sequenced in PacBio.

a.



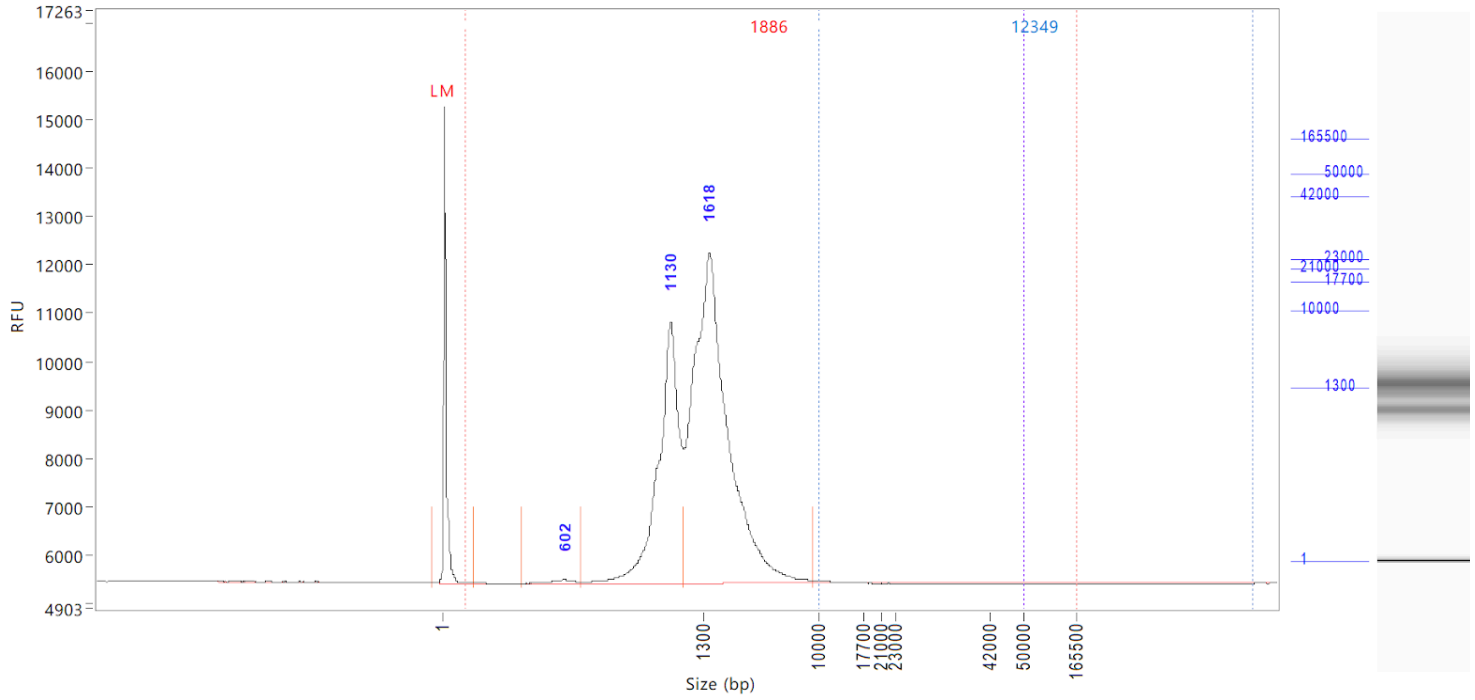
Category	SJs, count	Percent
Known canonical	3939	46.09
Known Non−canonical	4381	51.26
Novel canonical	60	0.70
Novel Non−canonical	166	1.94

b.



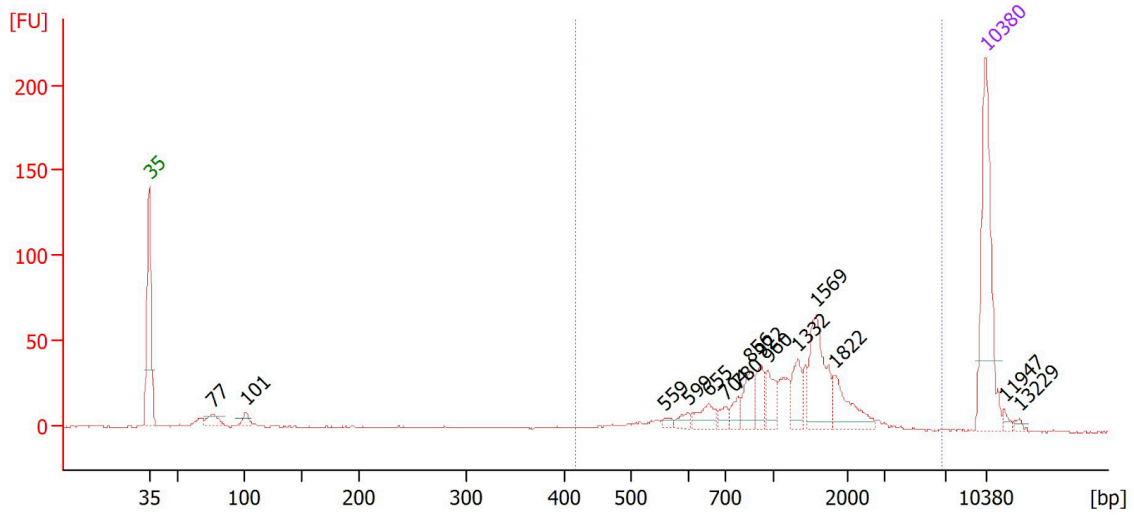
Category	SJs, count	Percent
Known canonical	11189	44.77
Known Non−canonical	12992	51.98
Novel canonical	224	0.90
Novel Non−canonical	588	2.35

Supplementary Figure 5.2. Splice junction coverage in young and aged IsoSeq libraries. (a) Distribution of splice junctions by structural classification and summary statistics for the library from young mouse HSC cDNA (b) Distribution of splice junctions by structural classification and summary statistics for the library from aged mouse HSC cDNA.

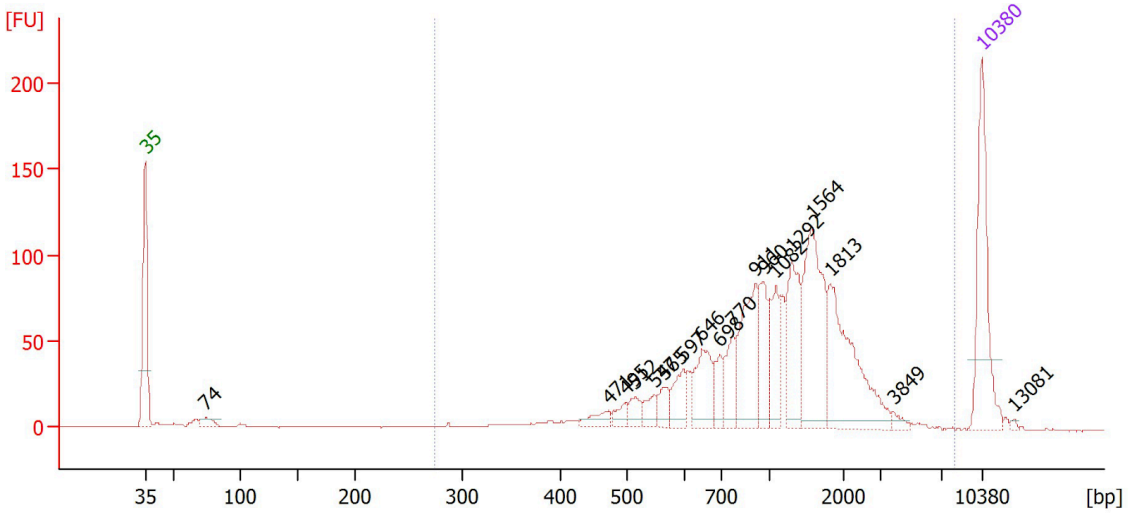


Supplementary Figure 5.3. Femto pulse size distribution trace of cDNA generated from mouse bone-marrow single cells using the 10X Genomics LT 3' GEM scRNA-seq kit.

a.

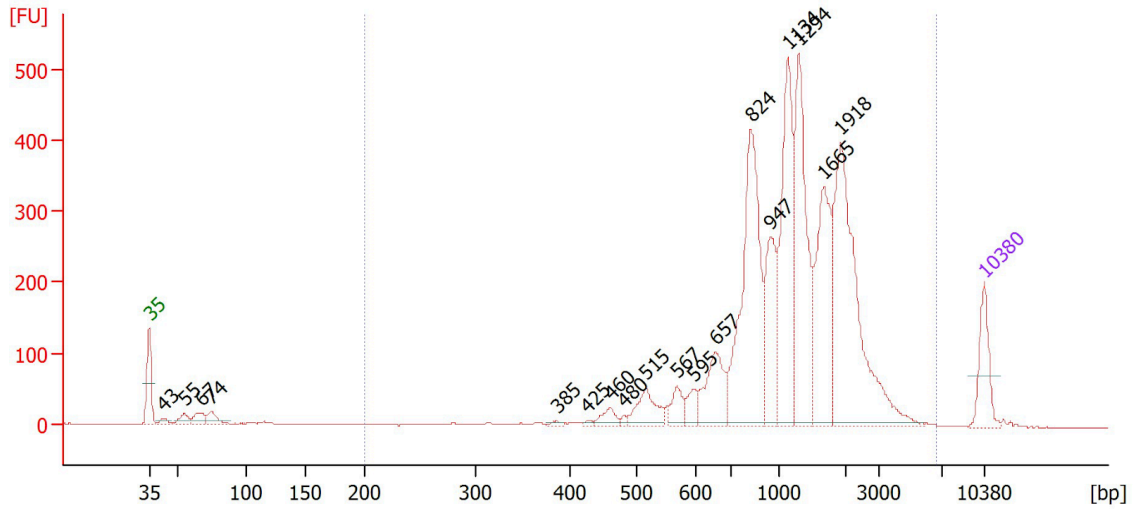


b.

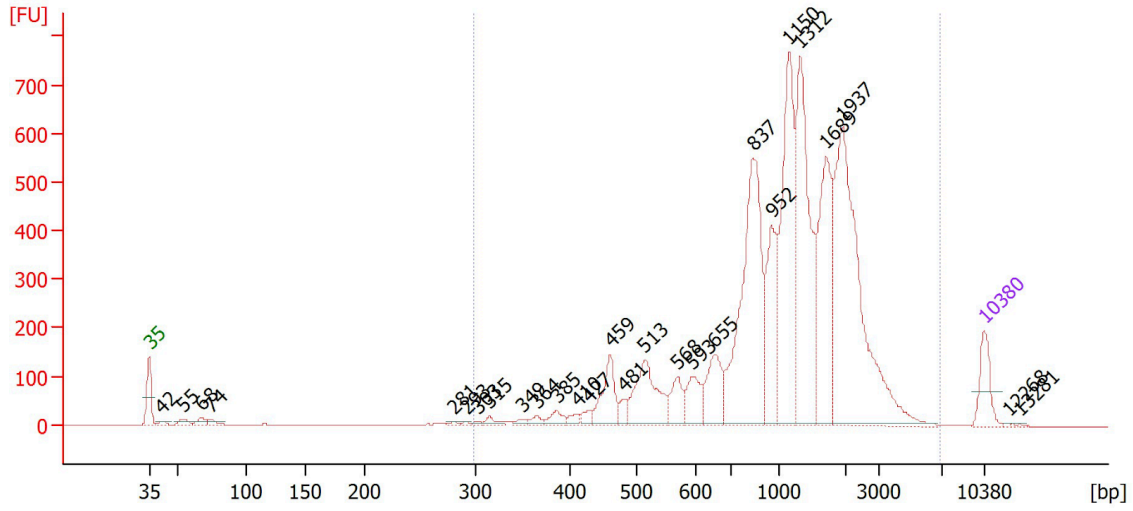


Supplementary Figure 5.4. Quality control of cDNA generated from human PBMCs using the 10X Genomics HT 3' GEM scRNA-seq kit. Bioanalyzer size distribution traces of cDNA from (a) Sample 1 ('1b') and (b) Sample 2 ('2a').

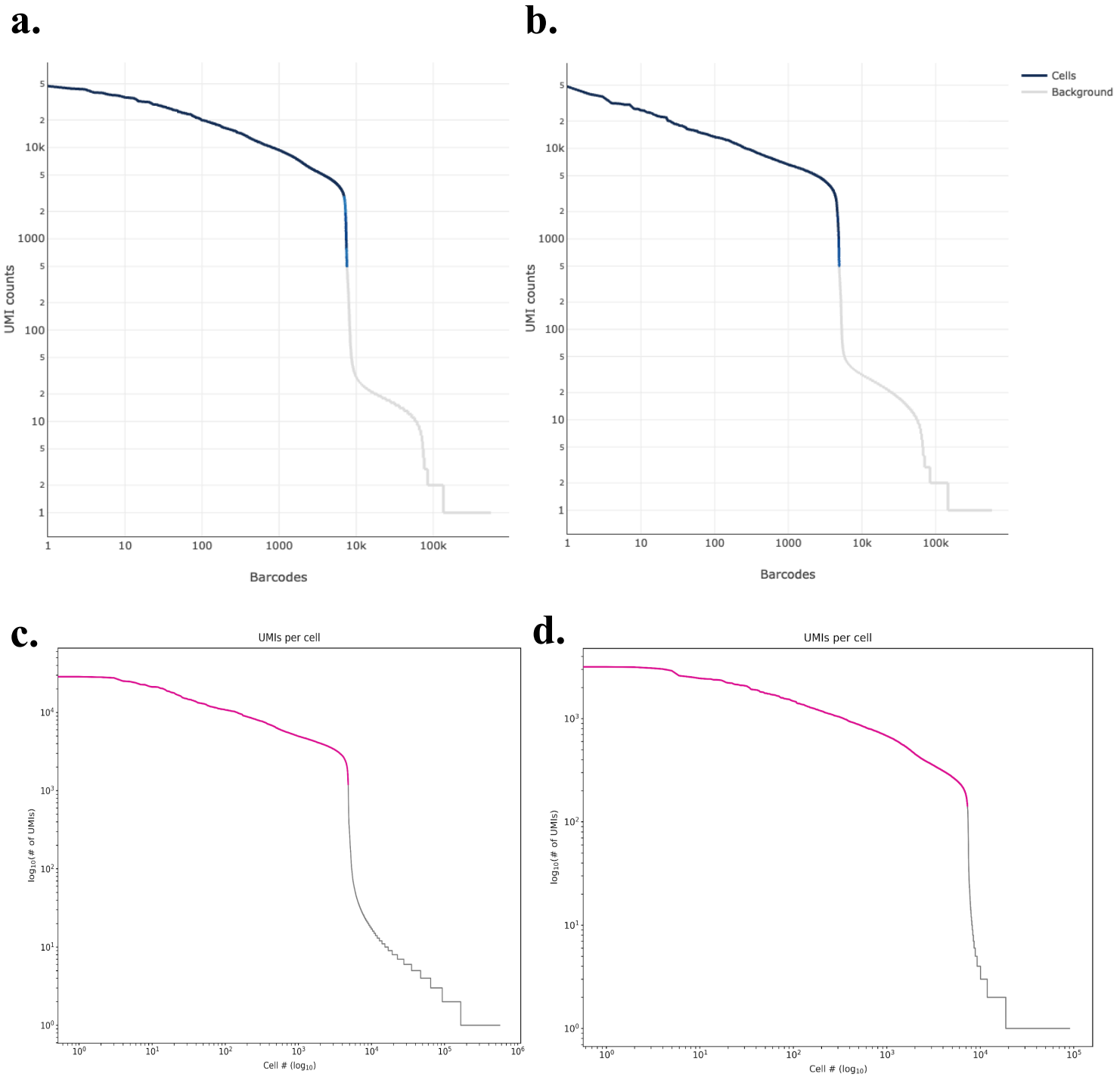
a.



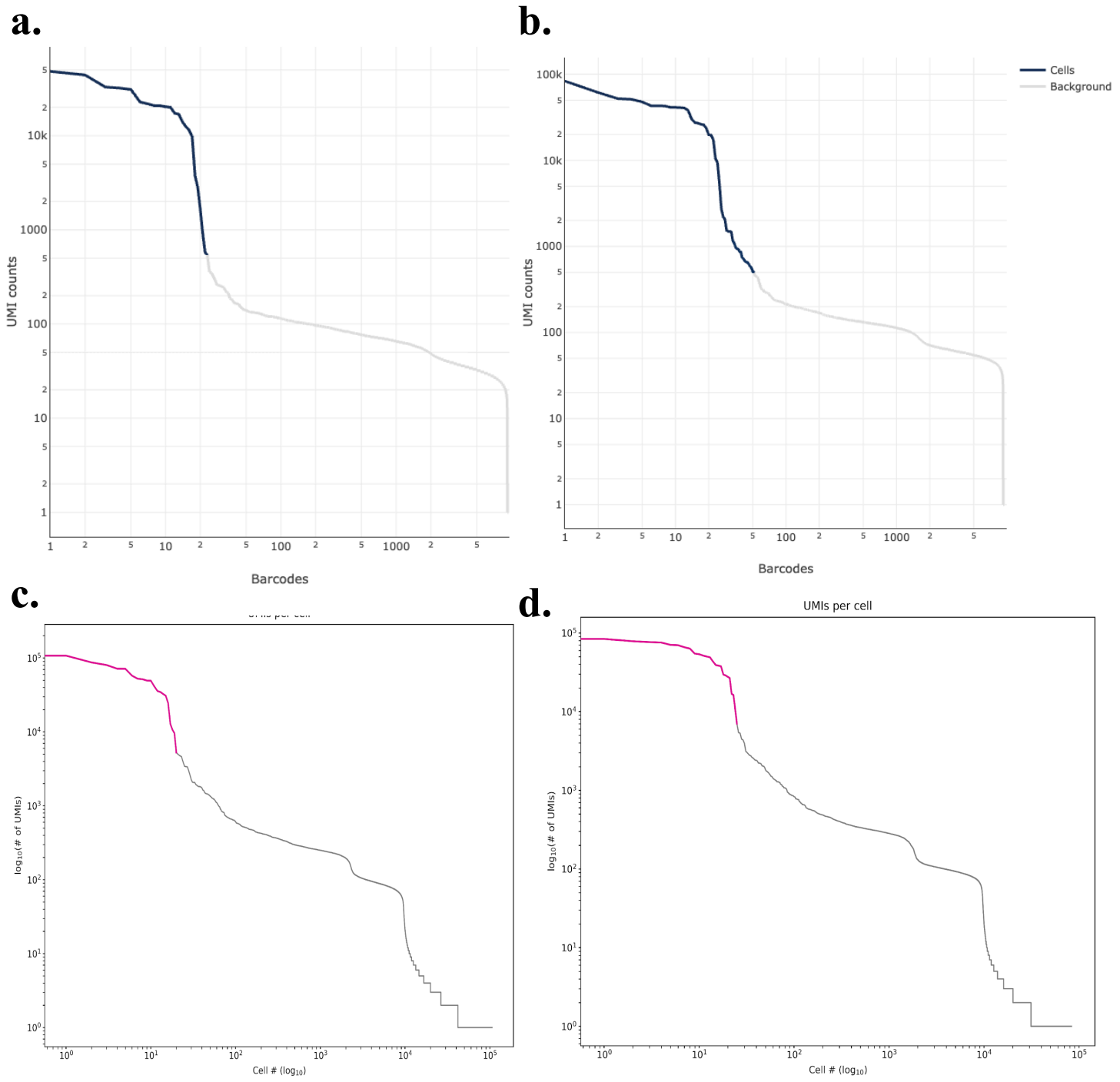
b.



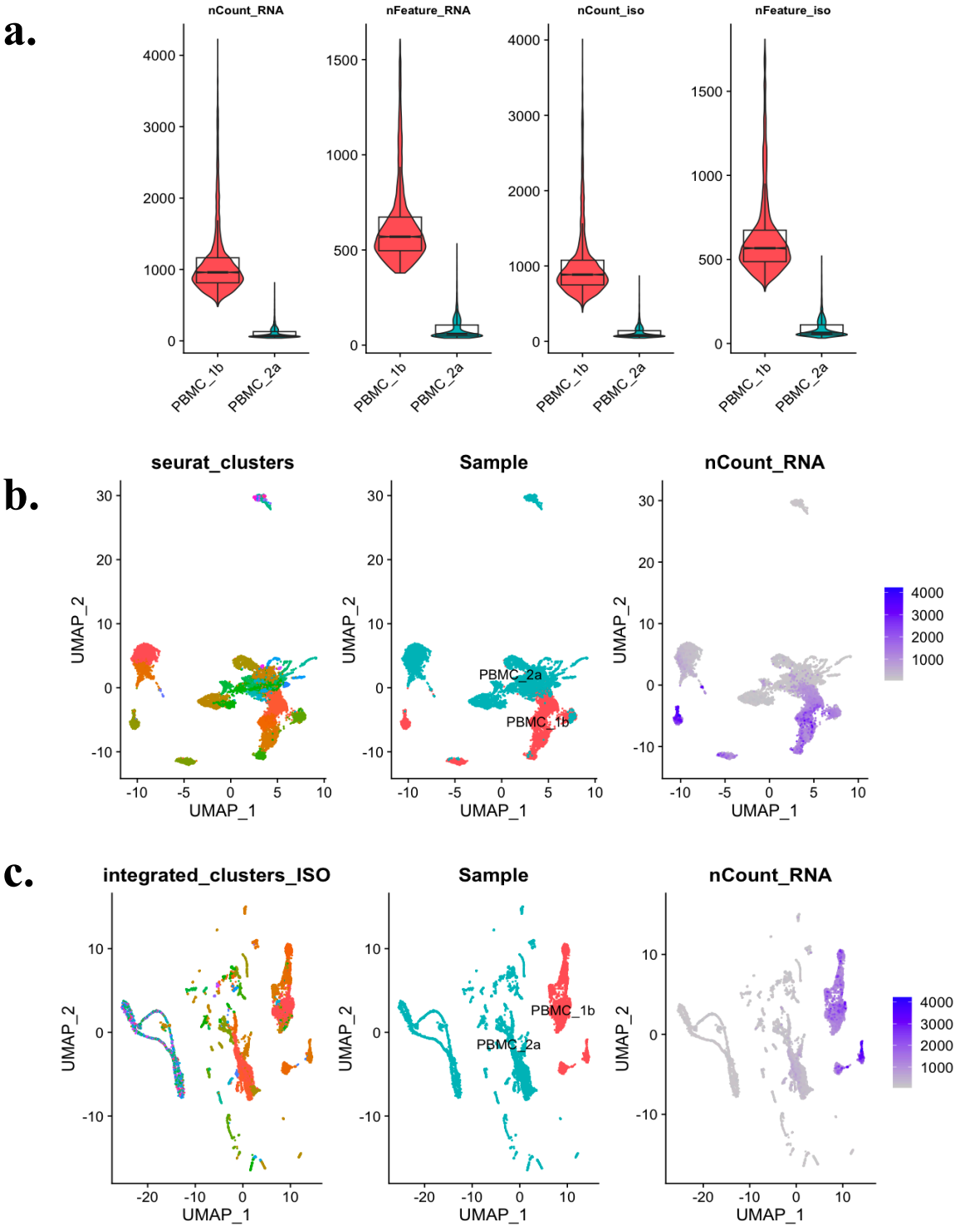
Supplementary Figure 5.5. Bioanalyzer size distribution of cDNA generated from FACS sorted LK Cd150+ single cells with 10X Genomics LT 3' Next GEM scRNA-seq (a) Size distribution from sample 1 (b) Size distribution from sample 2.



Supplementary Figure 5.6. MAS-seq PBMC barcode rank plot showing the distribution of 10X Genomics barcodes against UMI counts (a) Illumina PBMC sample 1: 4,875 cells (blue) with a median of 5,062 UMI counts per cell (b) Illumina PBMC sample 2: 7,384 cells (blue) with a median of 4,993 UMI counts per cell (c) PacBio sample 1: 4,773 cells (pink) with a median 3,788 UMI counts per cell (d) PacBio sample 2: 7,277 cells (pink) with a median of 327 UMI counts per cell.



Supplementary Figure 5.7. MAS-seq 10X Genomics barcode rank plot showing the distribution of barcodes against UMI counts (a) Illumina Sample 1: 23 cells with a median of 4,381 UMI counts per cell (b) Illumina Sample 2: 50 cells with a median of 4,030 UMI counts per cell (c) PacBio sample 1: 22 cells with a median 57,723 UMI counts per cell.(d) PacBio sample 2: 17 cell cells with a median 51,588 UMI counts per cell.



Supplementary Figure 5.8. Discrepancy in sequencing depth between MAS-seq PBMC libraries leads to batch-effect during downstream analysis at RNA and isoform level (a) Number of counts and features at the gene and isoform level coloured by sample ID (b) UMAP projection of single cell data prior to data integration coloured by seurat cluster (left) sample ID (centre) and count expression (right) (c) UMAP projection of single cell data post data integration coloured by seurat cluster (left) sample ID (centre) and count expression (right).