Deciphering Leukaemogenic Mechanisms through System-Scale Analysis of Single-Cell RNA Sequencing Data

Samuel Peter Watcham

Department of Haematology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College September 2020

To Mum, Dad, and Stephen, for always making me smile.

Declaration

I hereby declare that this dissertation 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 and in the declarations at the start of each chapter. This dissertation is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit of 60,000 words.

Samuel Peter Watcham
September 2020

Deciphering Leukaemogenic Mechanisms through System-Scale Analysis of Single-Cell RNA Sequencing Data

Samuel Peter Watcham

Haematopoietic stem cells are responsible for producing and sustaining the diverse array of cell types present in the adult blood system. This complex process requires the strict regulation of haematopoietic fate decisions and differentiation trajectories in order to maintain a healthy state. Haematological malignancies such as leukaemia are associated with various perturbations that disrupt this regulation and drive aberrant cell fate decisions, leading to disease. Much of this dysregulation is proposed to occur at the transcriptional level, and recent technological advancements in single-cell sequencing have made it possible to study the transcriptional effects of leukaemic perturbations at the scale of individual haematopoietic stem and progenitor cells. However, the mechanisms through which specific perturbations lead to dysregulation of the blood system remain poorly understood.

The primary aim of this work was to build an integrative computational framework for the analysis and comparison of leukaemic perturbations of the murine blood system as measured by single-cell RNA sequencing. Presented in Chapter 3, this framework aims to dissect the perturbation response across different scales – from individual genes to specific progenitor cell types to the entire blood system – and allow informative comparisons to be made about the similarities and differences between several perturbations. In total, eight genetic perturbations known to associate with leukaemia were analysed, resulting in novel biological insights concerning the behaviour of coordinated gene modules and the cellular abundance shifts driven by them.

As many leukaemic drivers act directly upon the most immature long-term haematopoietic stem cells, a highly targeted analysis of these cells was performed across the leukaemic perturbations. In Chapter 4 a novel computational pipeline was built to link FACS-sorted cell populations and single-cell transcriptional landscapes. Using this, the cellular and molecular responses of the perturbations were investigated, resulting in several novel hypotheses. For example, the data suggests that many leukaemic perturbations gain a competitive advantage against wild-type cells by pushing their MPP1 cells into more active states. Additionally the data suggests that increases in the transcriptional variability of blood stem cells is associated with pro-erythroid fate decision shifts and vice-versa.

Many different types of haematopoietic perturbations exist and can drive disease progression in the blood system. Chapter 5 focuses on single-cell RNA sequencing data from three further perturbations in various settings, including an infection model of Malaria and a model susceptible to endogenous DNA damage by aldehydes. These analyses have driven and validated bodies of experimental work, and comparing them to the previously described perturbation models highlighted both conserved changes and differences in the response of the haematopoietic system across different perturbation settings.

The final project aimed to improve upon current computational methods for cellular trajectory inference from single-cell data. Whilst high-throughput experiments allow for the sequencing of large cell numbers, this is balanced by the sparse and noisy nature of the returned data. Current methods perform poorly on such datasets and either cannot deal with large cell numbers or cannot extract enough relevant signal from sparse count matrices. A new computational tool was designed to work best on these large, sparse datasets, and infer the most likely cellular trajectories through snapshot sequencing data using an iterative process. In Chapter 6 this algorithm was applied to different systems including adult haematopoiesis, and was compared to state-of-the-art methods.

Overall, this thesis has investigated the transcriptional consequences of numerous preleukaemic perturbations on the haematopoietic stem and progenitor cell compartment at the single-cell level. New methods have been built for integration of single-cell perturbation experiments and their analysis across different biological scales. This has revealed novel biological insights regarding the mechanisms underpinning leukaemic transformation of the blood system.

Acknowledgements

I would first like to thank my supervisor, Bertie Göttgens, for his unwavering enthusiasm, optimism and encouragement throughout this PhD project. I am extremely grateful to have had such a perceptive and attentive mentor, and I have become a far better scientist as a result. I am equally indebted to Nicola Wilson for all her support, supervision, and advice over the past four years.

I would also like to thank all the members of the Göttgens lab, past and present, who helped make this PhD easier to complete. To the bioinformaticians - Xiaonan, Beccy, Fiona, Ivan, Hugo, Melania, Lila, Mariana and Kenny - thank you for all the laughter, the problem-solving, the memes, the code snippets, the cake and the constant distractions in both our tiny CIMR office and our shiny JCBC office. In addition, to all the wet lab workers - Iwo, Fernando, Sarah, Blanca, Winnie, Mai-Linh, Carolina and Kat - for their knowledge and support. I cannot imagine having met a better group of people anywhere else.

I am extremely grateful to Brian Hendrich and Austin Smith, for offering me a position at the Cambridge Stem Cell Institute and organising an excellent first year of rotations. I'd like to thank Lindsay Abbott and Jo Jack for all their help in running my PhD programme, as well as the Medical Research Council for funding it. I have also been lucky enough to work with a diverse group of excellent collaborators throughout my PhD, especially Myriam Haltalli and Dan Prins, as well as Meng Wang and Felix Dingler.

Most importantly, I want to thank my family and friends for all their fantastic support throughout this PhD. In particular to Lizzie, without whose love, support, and vivacity I would not have made it nearly as far as I have. Finally, I'd like to thank my Mum, my Dad, and Stephen, who have always believed in me, always been there for me, and who somehow managed to get me through five months of thesis writing during a global pandemic.

Table of contents

List of figures

List of tables

List of abbreviations

AML	Acute myeloid leukaemia
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
cDNA	Complementary DNA
CEL-seq	Cell expression by linear amplification and sequencing
CH	Clonal Haematopoiesis
CITE-seq	Cellular indexing of transcriptomes and epitopes by sequencing
CLP	Common lymphoid progenitor
CML	Chronic myeloid leukaemia
CMP	Common myeloid progenitor
CRISPR	Clustered regularly interspaced short palindromic repeats
DC	Diffusion component
dDEG	Dynamically differentially expressed gene
DEG	Differentially expressed gene
DPT	Diffusion pseudotime
DR	Dimensionality reduction
DTW	Dynamic time warping
DVG	Differentially variable gene

ET	Essential thrombocytosis
FACS	Fluorescence-activated cell sorting
GMP	Granulocyte-macrophage progenitor
HET	Heterozygous
HOM	Homozygous
HSC	Haematopoietic stem cell
HSPC	Haematopoietic stem and progenitor cell
HVGs	Highly variable genes
ICA	Independent component analysis
IPA	Ingenuity pathway analysis
IVT	In vitro transcription
kNN	k-nearest neighbour
KO	Knockout
LK	Lin- c-Kit+ sorting gate
LMPP	Lymphoid-primed multipotent progenitor
LSK	Lin- c-Kit+ Sca1+ sorting gate
LT-HSC	Long-term haematopoietic stem cell
MARS-seq	Massively parallel single-cell sequencing
MDS	Myelodysplastic syndrome
MEP	Megakaryocyte-erythroid progenitor
MF	Myelofibrosis
Mk	Megakaryocyte
MNN	Mutual nearest neighbour
MPN	Myeloproliferative neoplasm

MPP	Multipotent progenitor
mRNA	Messenger RNA
MST	Minimum spanning tree
PAGA	Partition-based graph abstraction
PBA	Population balance analysis
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PV	Polycythaemia Vera
qRT-PCR	Quantitative reverse-transcription polymerase chain reaction
sAML	Secondary AML
scNMT-seq	Single-cell nucleosome, methylation and transcription sequencing
scRNA-seq	Single-cell RNA sequencing
SM	Systemic mastocytosis
ST-HSC	Short-term haematopoietic stem cell
TITANS	Trajectory inference through iterative ancestral search
tSNE	t-distributed stochastic neighbour embedding
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
WT	Wild-type

Chapter 1

Introduction

Parts of this chapter have been adapted from Watcham et. al. (2019), which was written by Sam Watcham during this PhD with input from Iwo Kucinski.

1.1 Adult haematopoiesis

1.1.1 Emergence of the haematopoietic tree

Each day, the average human body produces an incredible one trillion (10^{12}) blood cells, in a process known as haematopoiesis (Doulatov et al., 2012). This enormous turnover is required to sustain the key functions of the blood system including oxygen transport, immune protection and wound healing. Associated with these functions are a myriad of blood cell types, each with their specific roles, processes and complexities.

Over 150 years, haematopoietic research has been significantly driven by new technological breakthroughs. The very earliest microscopy-based experiments in the 19th century established that blood was composed of two bone-marrow derived lineages: myeloid and lymphoid, potentially arising from a cell of common origin (Boisset and Robin, 2012). It was not until the 1950s - in the shadow of new, nuclear-weapon driven discoveries concerning the effects of ionising radiation - that this hypothesis was confirmed through bone marrow transplantation rescue of lethally irradiated mice (Jacobson et al., 1951; Urso and Congdon, 1956). This proved that within the bone marrow, there are cells capable of both self-renewal and differentiation towards all major blood lineages. Today these are known as haematopoietic stem cells (HSCs).

Subsequent *in vitro* assays suggested the existence of cells representing intermediate stages between HSCs and fully mature, differentiated blood cells that exhibited limited self

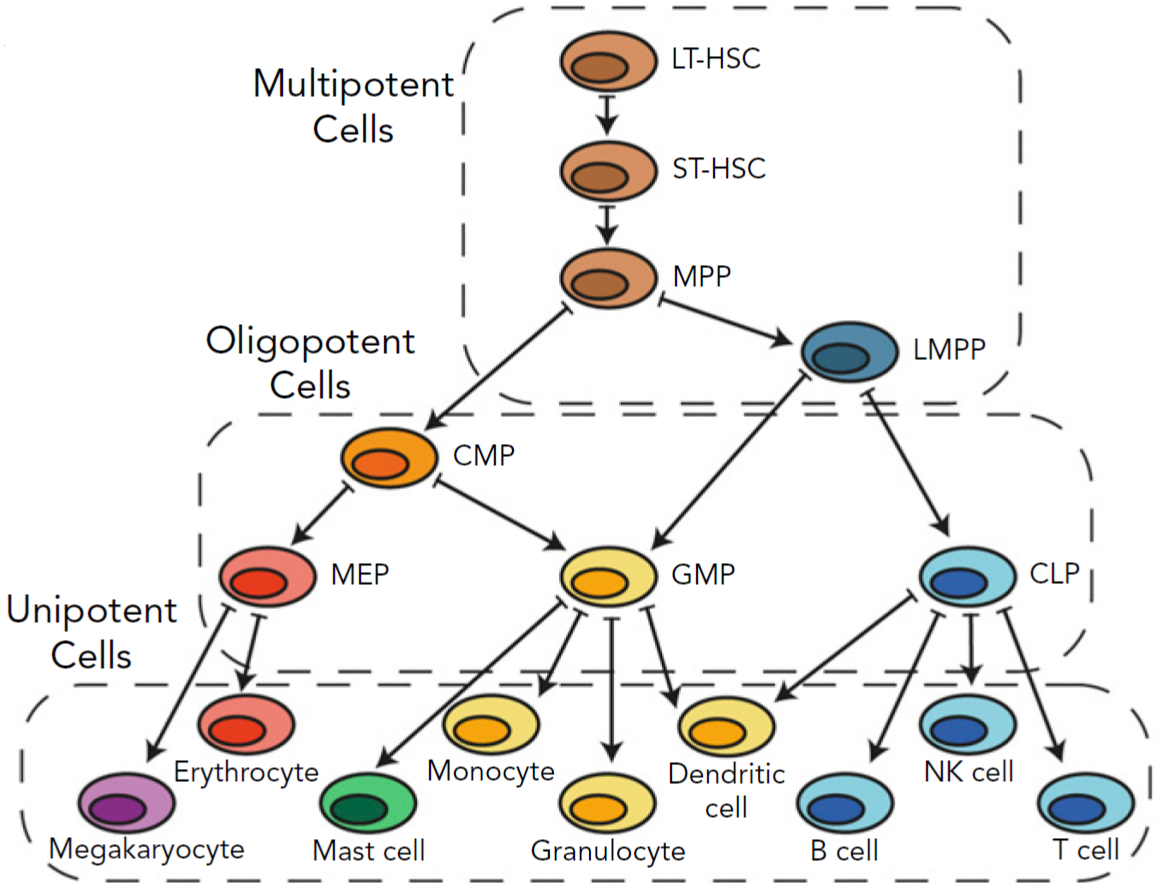


Fig. 1.1 The haematopoietic tree. Schematic showing one of the classic views of the haematopoietic cell hierarchy. Dashed boxes show 3 compartments encompassing cells of different potency: multipotent cells on top, bipotent/oligopotent cells in the middle, and terminally differentiated (unipotent) cells at the bottom. CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte-monocyte progenitor; LMPP, lymphoid-primed MPP; MEP, megakaryocyte-erythroid progenitor.

renewal or restricted potential. However, it was not until the development of cell sorting techniques using cell surface markers that a clear biological understanding of haematopoiesis began to form (Julius et al., 1972; Kohler and Milstein, 1975). By the earliest years of the 21st century, the concept of the 'haematopoietic tree' had become rooted in the field, with the self-renewing, multipotent HSC at the top of this cellular hierarchy (Eaves, 2015). Within this model, HSCs undergo stepwise differentiation along a number of possible bifurcations towards specific terminal fates, gradually losing both their self-renewal capacity and lineage potential (Figure 1.1). The aforementioned HSC state had been further subdivided into long-term HSCs (LT-HSC) capable of indefinite self-renewal in transplantation assays, short-term HSCs (ST-HSC) with limited self-renewal abilities, and multipotent progenitors (MPPs), which displayed minimal self-renewal but nonetheless maintained an ability to differentiate into all the main blood lineages (Nimmo et al., 2015; Orkin and Zon, 2008; Wilson et al., 2008). Below these a number of further progenitor states had been defined using specific surface marker combinations, including common myeloid progenitors (CMP), capable of producing both myeloid and erythroid but not lymphoid cells, and common lymphoid progenitors (CLP), which were restricted in the opposite manner (Akashi et al., 2000; Kondo et al., 1997).

Due to the simplistic appeal of the bifurcating tree model, the ease with which specific cell populations can be sorted and the sophisticated culture conditions applicable to blood cells, the haematopoietic system has become one of the most used model systems within stem cell biology. Over the past fifteen years, the haematopoietic tree has been subject to constant refinements, largely as a result of further technological breakthroughs. Chief amongst these is the development of high-throughput single-cell assays, the impact of which will be discussed in section 1.2. In parallel, the ability to barcode individual blood cells has led to the development of *in vivo* clonal tracking experiments, capable of gaining insight into unperturbed haematopoiesis in a native setting. Together these approaches have ushered in a new era of haematopoietic research.

1.1.2 Extrinsic and intrinsic control of cell fate decisions

During the process of differentiation from a HSC state towards a fully mature blood cell, multiple key fate decisions must be made. At the population level, this process must be tightly controlled in order to regulate the abundances of each mature cell type. Failure to do this can drive disease, and many haematological malignancies exhibit clear dysregulation of homeostatic fate decisions (Sperling et al., 2017). However, at the level of individual cells, the processes driving these fate decisions are believed to be stochastic, dependent on both cell-extrinsic and cell-intrinsic factors (Simons and Clevers, 2011).

Within the blood system, it is clear that the long-term self-renewal abilities of HSCs are highly dependent on cell-extrinsic signals mediated by their neighbours and their spatial location. Propagation of LT-HSCs *in vitro* is extremely challenging, and the existence of several types of HSC ‘niches’ within the bone marrow have been proposed (Kiel et al., 2005; Seita and Weissman, 2010; Zhang et al., 2003). It has been shown that HSCs require factors secreted by stromal and endothelial cells within the bone marrow to effectively maintain their HSC state, and that differences in the cytokines received by HSCs in different niches regulates their function and fate biases (Asada et al., 2017; Ding et al., 2012). For example, Pinho et. al described separate niches for myeloid-biased and lymphoid-biased HSCs occupied by megakaryocytes and arterioles respectively (Pinho and Frenette, 2019; Pinho et al., 2018). Furthermore, signalling between HSCs and bone marrow stromal cells has been implicated in the decreased erythro- and lympho-poiesis displayed by HSCs as they age (Valletta et al., 2020). The specific physical properties of a HSCs' microenvironment, such as oxygen levels, are also known to play an important role in regulating HSC quiescence and self-renewal (Gezer et al., 2014).

Alongside these extrinsic factors, numerous intrinsic processes are believed to play a role in cell fate decision-making. A large amount of work has gone into defining the key molecular drivers of haematopoietic fate decisions, with particular focus on regulation via the expression of key transcription factors for both HSCs and more mature progenitors (Menendez-gonzalez et al., 2019). Core regulatory networks underpinning the cellular identity of many blood lineages have been proposed, such as a pro-erythroid network involving GATA1, KLF1 and SCL (encoded by *Tal1*), a pro-megakaryocytic network involving FLI1 and PU.1, and a pro-granulocyte network involving GFI1 and GFI1B (Dore and Crispino, 2011; Meer et al., 2010; Wilson et al., 2010). The antagonistic roles of GATA1 and PU.1 in deciding between myeloid and erythroid fates has been particularly well-studied (Burda et al., 2010). The exact role that these and other transcription factors play in lineage commitment remains a topic of great debate, with more recent studies suggesting that their expression only reinforces fate decisions, rather than initiating them (Hoppe et al., 2016).

The current paradigm of haematopoietic fate decisions suggest that multipotent cells such as HSCs and MPPs display low-level expression of genes and/or transcription factors relating to distinct lineages, one of which 'wins out' due to intrinsic or extrinsic stochastic fluctuations (Moignard et al., 2013; Nimmo et al., 2015). Further cell-intrinsic mechanisms such as epigenetic regulation are increasingly thought to play a key role in lineage commitment by granting or withdrawing access to certain regulatory regions through DNA methylation, histone modification or chromatin conformation changes (Izzo et al., 2020). However, it is clear that despite the advancements brought about through single-cell techniques in the past fifteen years, the exact mechanisms through which blood cells make their fate decisions remain incompletely understood (Laurenti and Göttgens, 2018).

1.2 The single-cell revolution

1.2.1 Single cell experimental methods

To understand haematopoietic fate decisions at the level of single cells, methods were required that could measure some aspect of an individual cell's state, such as its transcriptome, proteome or epigenome. This is challenging given the minimal amount of material involved; a typical blood cell contains only a few picograms (10^{-12}g) of RNA (Livesey, 2003). Hence for many years, studies relied on bulk methods that captured population averages for a group of cells sorted on the basis of specific surface markers, leaving any heterogeneity amongst cells within the population obscured.

Some of the first single-cell transcriptional studies were performed using amplified RNA from haematopoietic cells to qualitatively compare transcript abundance (Brady et al., 1990). Technological improvements turned this idea into a quantitative single-cell approach known as single-cell quantitative reverse transcription polymerase chain reaction (qRT-PCR) (Hu et al., 1997). After reverse transcription of a cell's RNA, primers targeting specific genes are added and amplified using PCR alongside fluorescent reporter probes that allow quantification of the relative transcript levels of each targeted gene. Modern single-cell qRT-PCR techniques use microfluidics to optimise experimental throughput and multiplexing, allowing measurements of over 100 genes to be made in each cell with high signal to noise ratio (Guo et al., 2013; Ståhlberg and Bengtsson, 2010). Single-cell qRT-PCR has been widely applied in haematopoietic contexts (Moignard et al., 2013; Wilson et al., 2015), however the targeted nature of the method requires genes of interest to be selected beforehand, limiting its use for the novel identification of molecular drivers involved in haematopoiesis.

Parallel to these developments in transcriptional profiling, advancements in fluorescent-activated cell sorting (FACS) technologies led to the concept of index sorting, where the expression of up to 30 surface proteins can be measured for individual cells, prior to being sorted into individual wells for use in downstream functional or genomic assays (Osborne, 2010). This offers the opportunity to link surface marker profiles to functional output, transplantation ability or whole-transcriptome profiles (Nestorowa et al., 2016; Schulte et al., 2015; Wilson et al., 2015). Here the number of proteins measured is limited by the acquisition and delineation of different fluorophores to be used for sorting. Similarly, mass cytometry techniques - whereby protein antibodies are conjugated with heavy metal isotopes and quantified using mass spectrometry - have typically been limited to measuring the abundance of around 40 proteins per cell (Levy and Slavov, 2018). Excitingly, advancements in mass spectometry have very recently made possible the quantification of thousands of proteins in individual cells, suggesting that whole-proteome single-cell studies will be an area of great interest in the next decade (Yang et al., 2020).

In order to understand the molecular dynamics of cell fate decisions in a less biased manner, methods for measuring single-cell gene expression on a transcriptome-wide scale were needed. Initially, microarray techniques developed for bulk RNA-seq analysis were extended to single cells via the development of new amplification techniques (Kurimoto et al., 2006). In a microarray cDNA is reverse transcribed from mRNA and hybridised to a set of probes containing DNA matching genes from across the entire transcriptome. Subsequent labelling of the cDNA from each sample with a fluorescent dye allows quantification of gene expression across the transcriptome.

Following these advances, the first single-cell RNA sequencing (scRNA-seq) protocol was developed in 2009, just two years after the first applications of bulk RNA-seq techniques (Kolodziejczyk et al., 2015; Tang et al., 2009). In scRNA-seq protocols, reverse transcribed cDNA from an individual cell is amplified and subsequently prepared for sequencing via library preparation. These steps can be performed in a number of ways, and in the years following, many scRNA-seq protocols emerged that improved the accuracy, sensitivity and throughput which could be achieved. After the lysis of an individual cell, specific reverse transcription of mRNA molecules is achieved through poly(T) priming. Second strand synthesis of cDNA is then performed either by poly(A) tailing as in the CEL-seq/MARS-seq methods (Hashimshony et al., 2012; Jaitin et al., 2014), or by template switching as in the Smart-Seq (later optimised as Smart-Seq2) method (Picelli et al., 2014; Ramsköld et al., 2012). The latter method provided a key advantage in that full-length mRNA transcripts can be recovered, mitigating the bias towards 3' transcript fragments that arises from incomplete reverse transcription. This allows secondary mRNA features such as alternative splicing and allele-specific expression to be analysed. Finally, cDNA amplification can be performed using either a PCR reaction (Islam et al., 2011; Ramsköld et al., 2012), or in-vitro amplification (IVT) (Hashimshony et al., 2016)(Jaitin et al., 2014). PCR-based amplification is non-linear and sequence dependent, leading to technical biases from differential amplification of certain transcripts. IVT results in linear amplification, but requires a second round of reverse transcription; this also leads to the introduction of 3' coverage biases (Kolodziejczyk et al., 2015).

A clear problem with these approaches was the inability to distinguish between a) two sequencing reads that map to the same gene because they came from separate mRNA transcripts within the cell or b) two sequencing reads that map to the same gene because they are duplicates of the same transcript created during the cDNA amplification process. If all transcripts were amplified equally this would not be a problem; however this will not always be the case. To overcome this, methods were introduced that barcoded each mRNA molecule with a unique molecular identifier (UMI) during the initial reverse transcription step (Fu et al., 2011). Duplicated reads from the same mRNA molecule can then be identified, and the number of UMIs sequenced for each gene can simply be counted (Kivioja et al., 2012)(Islam et al., 2014). UMIs were first introduced in the MARS-seq and CEL-seq-UMI methods, and have since become widely used (Grün et al., 2014; Jaitin et al., 2014).

The other significant drawback associated with the initial wave of scRNA-seq protocols was the combination of high costs and low experimental throughput, since the entire protocol had to be performed for each cell individually. MARS-seq started to address this by adding a unique barcode to each well in a 384-well plate, which labelled transcripts during reverse transcription (Jaitin et al., 2014). Subsequent steps could then be performed on pooled material, greatly reducing the per-cell costs. However, the experimental throughput of this approach remained low. Shortly afterwards, a sweeping change in throughput arose from the introduction of droplet-based microfluidic protocols. For the first time, methods such as InDrops (Klein et al., 2015) and Drop-seq (Macosko et al., 2015) allowed the simultaneous profiling of thousands of single cells in a relatively low-cost manner. Each individual cell is encapsulated in a droplet along with a bead that contains both a cell-specific barcode and UMIs. Cells are lysed and reverse transcription occurs in each droplet individually, before the emulsion is broken and cDNA is amplified in a pooled manner (Figure 1.2). Droplet-based scRNA-seq became commercially available in 2017 using the 10X Chromium system, created by 10X Genomics (Zheng et al., 2017).

Whilst droplet-based methods (such as 10X) have huge advantages over plate-based methods (such as Smart-Seq2) in terms of throughput, this comes at the expense of sensitivity. Droplet-based methods typically have a lower mRNA capture efficiency and significant 3' coverage bias, leading to a reduction in the number of genes captured - around 1000-3000 compared with >5000 genes per cell. In addition, their high-throughput nature necessitates that less reads will be captured per cell for the same amount of sequencing power. Ultimately, the choice of protocol depends on the experimental questions at hand, with droplet-based approaches better for unbiased assessment of broad cell populations, and plate-based methods suited to answering focused questions about specific cell types.

In the past two years, further scRNA-seq approaches have appeared, including combinatorial indexing, whereby successive rounds of splitting, barcoding and repooling result in unique barcodes for each cell without ever requiring individual cells to be separated (Rosenberg et al., 2018). Furthermore, many other modalities have joined the single-cell

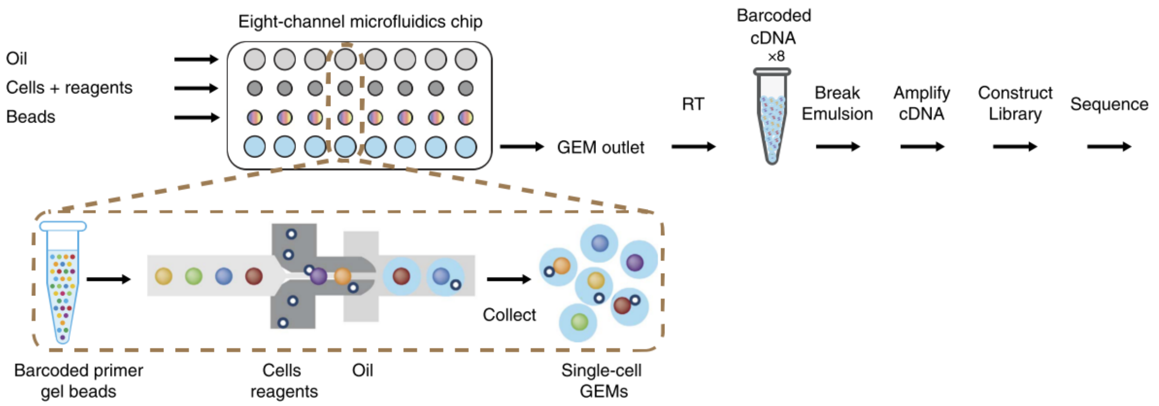


Fig. 1.2 Droplet-based scRNA-seq. Schematic of a typical droplet-based scRNA-seq protocol reproduced from Zheng et. al. (2017). Barcoded beads flow through a microfluidic chip alongside cells and reagents for RT. A single bead and cell are captured in an oil droplet, within which cell lysis and RT occur, before barcoded cells are pooled together. RT: reverse transcription; GEM: Gel bead in Emulsion.

revolution. Single-cell ATAC-seq for examining chromatin accessibility has made use of both droplet microfluidics and combinatorial indexing (Lareau et al., 2019; Preissl et al., 2018). Single-molecule RNA fluorescent in-situ hybridisation (smFISH) has made it possible to restore spatial information to whole-transcriptome profiling (Shah et al., 2016). In turn, this had led to the creation of a large number of 'multiomics' approaches, where two or more modalities are captured in single cells simultaneously (Zhu et al., 2020). Examples include scNMT-seq, capable of capturing whole-transcriptome, whole-methylome and chromatin accessibility measurements from a single cell (Clark et al., 2018), and CITE-seq, capable of simultaneous measurements of both the transcriptome and a targeted subset of cell surface markers (Stoeckius et al., 2017). These methods and their descendants, alongside new computational methods for integrating genomic data across separate modalities, will undoubtedly dominate the single-cell field in the years to come.

1.2.2 New insights into haematopoiesis from single-cell RNA sequencing

The development of experimental methods for scRNA-seq has led to an explosion of single-cell transcriptomic datasets that have transformed our understanding of haematopoiesis. In any given experiment, scRNA-seq offers a 'snapshot' of a group of cells and their expression states at a specific time point. The distribution of these states within a high-dimensional gene-expression space can be considered as a 'transcriptional landscape', which can offer detailed information on cellular differentiation trajectories. Behind this model lies two key assumptions; that differentiation is a continuous process (cells do not 'jump' their expression states without passing through intermediate states); and that a group of cells will differentiate asynchronously and therefore will be captured at multiple points along a differentiation trajectory (Weinreb et al., 2018a). Hence by combining a scRNA-seq experiment of a heterogeneous group of progenitor cells with a) prior biological knowledge and b) computational methods for visualising and ordering cell profiles, inferences about differentiation trajectories and their molecular drivers within the blood system can be made through these transcriptional landscapes.

Around 2013, haematopoietic single-cell studies started with the profiling of previously well-defined populations (such as those in Figure 1.1), and have subsequently shifted towards less biased selections of cells (Dahlin et al., 2018; Paul et al., 2015; Tusi et al., 2018; Wilson et al., 2015). Arguably the culmination of this shift has resulted in the Human Cell Atlas project, which aims to create reference maps for cells across >50 tissues in the human body (Benoist et al., 2017). This includes a study of >500,000 human bone marrow cells sequenced using droplet-based techniques. This dataset represents an excellent example of a large, unbiased transcriptional landscape (Figure 1.3). Starting from the haematopoietic stem and progenitor (HSPC) compartment containing the HSCs and other early progenitor states, cells can move along a number of possible trajectories, some of which are shown. Hence within this model of continuous transcriptional landscapes, cells differentiate probabilistically along one of numerous trajectories, passing through a myriad of possible transcriptomic states as they do. Fate decisions therefore occur at poorly-defined 'branchpoints' within the landscape that may or may not be preceded/followed by molecular signals within the data (Laurenti and Göttgens, 2018).

The idea of a continuous transcriptional landscape is clearly at odds with the discrete nature of the classical haematopoietic tree. Single-cell profiling of the CMP, GMP, and MEP compartments revealed high levels of heterogeneity, including at least 18 distinct cellular subtypes with various degrees of lineage priming (Paul et al., 2015). Additionally, the surface markers classically used to separate the three populations (FcgR and CD34) were shown to be poor predictors of cellular identity, suggesting these populations are not discrete at all, but rather arise from the same continuum. It was also shown that only a small fraction of these cells displayed expression profiles consistent with multilineage activity, suggesting that by this point in the haematopoietic hierarchy, the majority of cells had already made their key fate decisions (Olsson et al., 2016). These findings have been largely corroborated by single-cell functional assays, transplantations and barcoding experiments, further suggesting

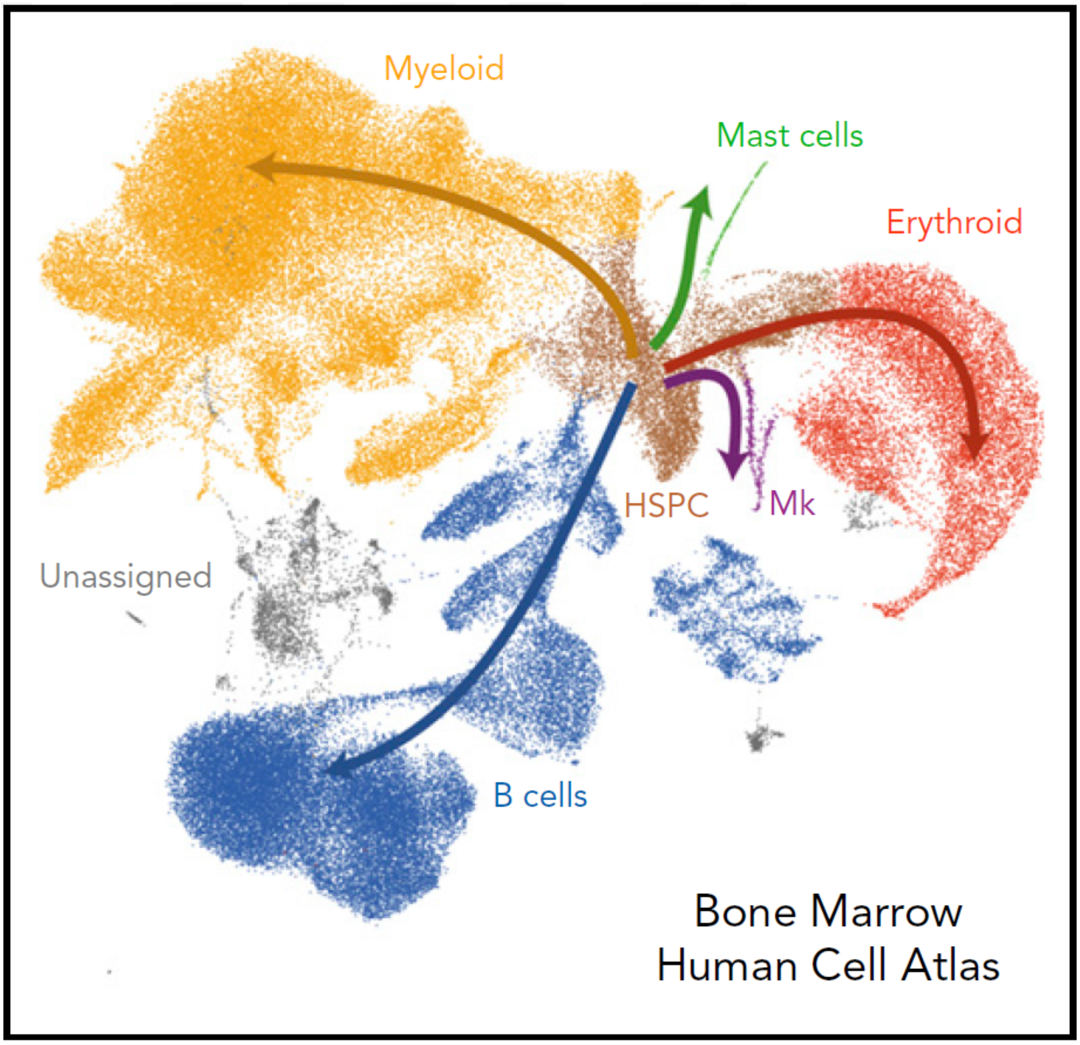


Fig. 1.3 The transcriptional landscape of the Human Cell Atlas. Each dot represents a single-cell transcriptome from the bone marrow mononuclear cell fraction, visualised in two dimensions using dimensionality reduction. Cells are coloured according to their likely identity. Arrows indicate the main directions of differentiation away from the most immature cells. HSPC, haematopoietic stem and progenitor cell; Mk, megakaryocyte.

that most key fate decisions occur in the HSC and MPP compartments (Adolfsson et al., 2005; Görgens et al., 2013; Naik et al., 2013; Notta et al., 2016; Rodriguez-Fraticelli et al., 2018).

Classically, the upper tier of the haematopoietic tree was split into LT-HSCs, ST-HSCs and MPPs, all capable of generating myeloid, erythroid and lymphoid cells. Prior to the advent of single-cell technology, the MPP compartment was thought to be composed of at least four subpopulations (MPP1-4) with distinct functional and molecular features (Cabezas-

Wallscheid et al., 2014; Oguro et al., 2013). Subsequent reanalysis of this model at the single-cell level suggested that in fact, the MPP (and HSC) compartment can be characterised by a continuous landscape in which the MPP2, MPP3 and MPP4 populations exhibited significant lineage biases towards the erythroid/Mk, myeloid and lymphoid lineages respectively (Nestorowa et al., 2016; Pietras et al., 2015). These findings further suggested that key fate decisions were occuring very early on, potentially in the MPP1 compartment. A landmark paper from Rodriguez-Fraticelli et. al. helped to confirm this hypothesis using a combination of scRNA-seq and cell barcoding experiments to study the MPP populations (Rodriguez-Fraticelli et al., 2018).

Single-cell profiling has helped to further characterise the true LT-HSC state at a molecular level. By combining various HSC isolation strategies, Wilson et. al. were able to isolate true functional LT-HSCs with 60% purity, describe their expression profiles and define genesets that correlate with HSC functional ability (Wilson et al., 2015). It was apparent that cell cycle is a major component in regulating HSC identity, confirming that the best functional LT-HSCs are marked by a dormancy signature (Wilson et al., 2008). Another landmark paper described how HSCs exist along a continuous dormancy spectrum, with the best functional HSCs characterised by deep quiescence and low biosynthetic activity (Cabezas-Wallscheid et al., 2017). Recent research has focused on the role of mitochondria in maintaining HSC quiescence (Luis et al., 2020). However, an aspect of HSCs that remains unclear is heterogeneity in functional lineage bias. Whilst it has been reported that HSCs exhibit no evidence of lineage priming towards specific fates (Kowalczyk et al., 2015), other studies have suggested the existence of a megakaryocyte-biased subset within the the HSC compartment (Carrelha et al., 2018; Grover et al., 2016; Sanjuan-Pla et al., 2013). Transplantation experiments have clearly suggested the existence of myeloid- and lymphoid-bias phenotypic HSCs (Benz et al., 2012; Muller-Sieburg et al., 2004), and reconciling this data with the lack of clear lineage priming in scRNA-seq HSC profiles remains an open question. Furthermore, it remains unclear to what extent LT-HSCs contribute to turnover of the blood system; it has been suggested that in steady-state conditions, the MPP1 compartment is responsible for most of the blood cell production, though other studies have concluded that the LT-HSC compartment also offers a significant contribution (Busch et al., 2015; Rodriguez-Fraticelli et al., 2018; Sawai et al., 2016; Sun et al., 2014).

Placing all these refinements to the haematopoietic tree in the context of a transcriptional landscape has required the production of very broad, unbiased single-cell datasets that profile the entire blood progenitor landscape. Initially, 1600 cells spanning 10 classical populations were profiled, with three broad trajectories towards the erythroid, myeloid and lymphoid lineages observed (Nestorowa et al., 2016). High throughout droplet-based studies have expanded this to tens of thousands of cells, revealing at least eight clear trajectories arising from HSCs, including neutrophils, monocytes, basophils, mast cells and eosinophils (Dahlin et al., 2018; Tusi et al., 2018). Reconstructing the hierarchy of haematopoietic fate decisions from scRNA-seq data is a considerable technical challenge (see section 1.4 and Chapter 6), and inferred hierarchies do not always agree with data from cellular barcoding experiments. For example, the erythroid and megakaryocytic trajectories appear closely linked in scRNA-seq data - consistent with the classical idea of an MEP state - whereas barcoding studies suggest there is little coupling of the megakaryocytic fate with others, leaving the erythroid lineage more closely coupled with neutrophils/monocytes (Pei et al., 2017; Rodriguez-Fraticelli et al., 2018). Similarly, it remains unclear whether basophils/mast cells originate alongside the neutrophil/monocyte trajectories (Wolf et al., 2019) or the erythroid trajectory (Grootens et al., 2018; Tusi et al., 2018). Evidence has arisen of multiple routes of differentiation for both megakaryocytes (Grover et al., 2016; Rodriguez-Fraticelli et al., 2018) and monocytes (Tusi et al., 2018). Further waves of barcoding and multiomics experiments are required to comprehensively resolve these questions.

Overall, the model of transcriptional landscapes has become an invaluable tool with which to study haematopoiesis. Whilst most work has been focused on the murine system, many human studies have reached similar conclusions about the early-onset of key fate decisions, the heterogeneous nature of classically homogeneous populations, and the continuous structure of the most immature compartments (Notta et al., 2016; Velten et al., 2017). Nonetheless there are bound to be disparities between mouse and human, and currently these can be difficult to resolve due to the 'flat' nature of the transcriptional landscape in the HSC and MPP region (Kowalczyk et al., 2015). It must be remembered that scRNA-seq data represents an incomplete view of a cell's state, and that further information from the proteome or epigenome may reveal great insight into haematopoiesis.

1.3 Haematological perturbations, malignancies and leukaemia

1.3.1 Genetic basis of pre-leukaemic states

There exists a wide range of blood diseases resulting from genetic alterations to haematopoietic cells. These are characterised by over/under-production of one or more blood lineages; the presence of abnormal, dysplastic cells with aberrant self-renewal or differentiation properties; and/or abnormal localisation of cells within the blood system (Taylor et al., 2017). The phenotype, treatment and prognosis of blood disorders is as heterogeneous as their genetic mechanisms, which are understood to varying degrees. Nevertheless, for many of the same reasons that haematopoiesis makes a good model system for stem cell biology, blood disorders have been at the forefront of research into cancer genetics.

Within the haematopoietic system, disease initiation and development occurs through dynamic genetic processes, potentially years or decades before the onset of clinical symptoms. If a genetic mutation within a differentiating progenitor cell confers a survival advantage, a clonal population will emerge within the blood that harbours the mutation. If the mutation also affects the cell's behaviour, disease can then be initiated, provided that the clone does not disappear due to exhaustion or terminal differentiation (Corces-Zimmerman and Majeti, 2014). A great deal of work has gone into understanding the 'cell of origin' for leukaemias and associated blood disorders. The first evidence connecting stem cells to cancer origin arose from a study of chronic myelogenous leukaemia (CML) in 1974 (Fialkow, 1974). Subsequent work looking at the self-renewal and repopulation abilities of leukaemic cells provided further evidence that for most acute myeloid leukaemia (AML) subtypes, transformation was occurring at the level of the most immature progenitors (Bonnet and Dick, 1997). The question of whether leukaemia is initiated by mutations in HSCs or by mutations in committed cells that then go on to exhibit HSC-like properties (such as self-renewal) remains a topic of debate, and it seems likely that both can occur depending on the specific subtype of leukaemia (Passegué et al., 2003).

Nonetheless, there is growing consensus that for many leukaemias, disease is initiated by the accumulation of numerous successive mutations within previously normal HSCs (Corces-Zimmerman and Majeti, 2014; Welch et al., 2012). This model is visualised in Figure 1.4. During this transformation, mutated HSCs exist in a 'pre-leukaemic' state defined by a small number (possibly only one) of genetic mutations, any of which can have distinct effects on the functional ability, differentiation capacity and molecular regulatory programmes of the HSC (Jan et al., 2012; Shlush et al., 2014). These pre-leukaemic states can manifest themselves clinically as less severe blood disorders such as clonal haematopoiesis, myeloproliferative neoplasms (MPNs) and myelodysplastic syndrome (MDS) (Grinfeld et al., 2017; Sperling et al., 2017). Further genetic alterations can then tip pre-leukaemic cells into a true leukaemic state, conferring large proliferative advantage and eventually overrunning the haematopoietic system. It has been shown that the mutational burden of HSCs increases with age, helping to explain the prevalence of AML-like leukaemias in elderly patients, where enough time has passed to acquire multiple mutations in a single HSC clone (Welch et al., 2012). This model also helps to explain why many leukaemia patients relapse after treatment; even if the dominant leukaemic clone is eradicated, pre-leukaemic cells may survive which then only require one further mutation event to cause a relapse. It is clear that the transformation from healthy to leukaemic stem cells is an accumulating stepwise process, and both the order of mutation acquisition and the exact identity of the cells that first acquire each mutation are likely to play a role in the development of the disease (Li et al., 2014; Shepherd et al., 2018).

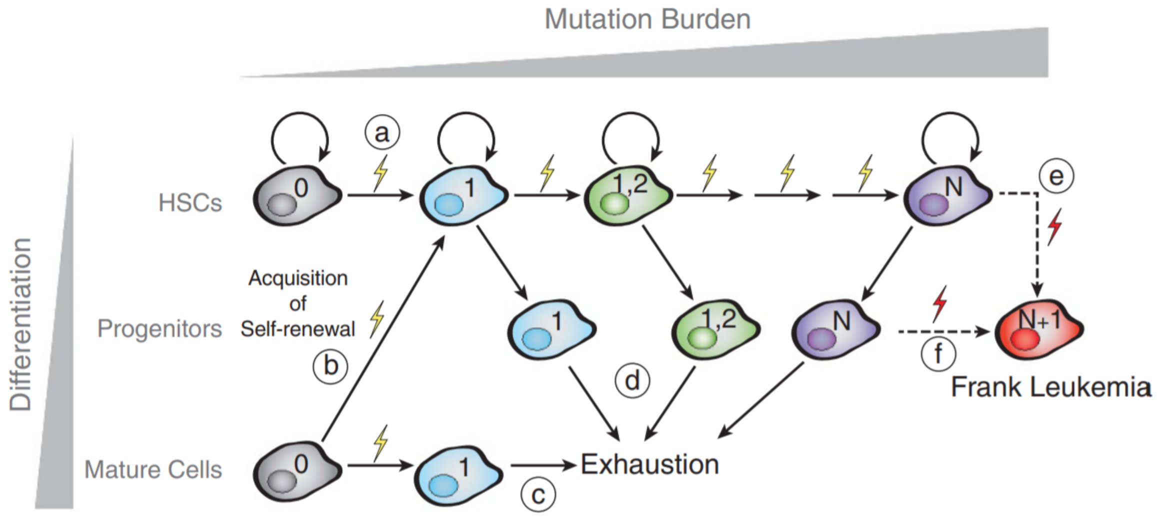


Fig. 1.4 A model for pre-leukaemic evolution of HSCs. Figure taken from Corces et. al. (2014). The successive accumulation of genetic mutations in a HSC is shown through changes in colour. (a) An initial mutation event occurs in a healthy HSC, conferring some self-renewal advantage. (b,c) A mutation event may occur in a differentiated cell: either this clone will exhaust, or it could confer HSC-like self-renewal properties onto the differentiated cell. (d) Further mutations in the HSC could lead to a loss of self-renewal ability and hence exhaustion. (e,f) Eventually, a final mutation will transform the pre-leukaemic HSC into a true leukaemic state, which starts proliferating rapidly and destroying any homeostasis within the blood system.

Pre-leukaemic haematopoietic states therefore have great potential to elucidate the mechanisms through which leukaemia is initiated. Results from high-throughput bulk sequencing of human leukaemia samples have identified many commonly-mutated genes associated with specific disease subsets (Chopra and Bohlander, 2019). Mouse models of these mutations have led to great insight into how they alter haematopoiesis and in particular the self-renewal and differentiation abilities of HSCs at a functional level (Moran-crusio et al., 2011; Ostrander et al., 2020). With the advent of high-throughout single-cell methods, it has now become possible to perform whole-transcriptome sequencing of the entire progenitor landscape within these mutated mice, and assess their perturbed haematopoietic behaviour at a systems-scale. If the cellular and molecular impacts of these pre-leukaemic states can be understood at a single-cell level, targeted therapies can be designed to counteract the specific impacts of the mutations, as opposed to using current drugs that target entire biological pathways. Here several of the most common pre-leukaemic mutations and their associated pre-leukaemic blood disorders are briefly reviewed.

1.3.2 Myeloproliferative Neoplasms

MPNs are themselves a diverse group of clonal myeloid haematopoietic malignancies characterised by the overproduction of terminally differentiated peripheral blood cells, an increased risk of thrombosis and predisposition to acute myeloid leukaemia (Grinfeld et al., 2017). They are distinguished from myelodysplastic syndromes (MDS) and leukaemia by the absence of clear morphological dysplasia (abnormalities) within the blood compartment. Typically, MPNs are split into two groups: Philadelphia-chromosome positive and negative neoplasms. The Philadelphia chromosome is a specific reciprocal translocation of chromosome 22 that contains a BCR-ABL1 fusion gene, leading to the constitutive activation of a tyrosine kinase signalling protein that causes high levels of proliferation and inhibits DNA repair (Kang et al., 2016). It is a hallmark of chronic myelogenous leukaemia (CML), characterised by the overproduction of mature granulocytes such as neutrophil, basophils or eosinophils (Hehlmann and Hochhaus, 2007). CML accounts for 15-25% of adult leukaemias, and if left untreated can transition into AML. The use of tyrosine-kinase inhibitors to treat CML became prevalent in the early 2000s and has dramatically increased patient survival (Gambacorti-passerini et al., 2011). Because of the relative severity of CML, it is usually not considered as a pre-leukaemic state in the same manner as other MPNs, but rather as a true leukaemia that can be fatal over a short time period if left untreated.

Philadelphia-chromosome negative MPNs are typically split into three distinct diseases: polycythaemia vera (PV), essential thrombocytosis (ET) and idiopathic myelofibrosis (MF). PV and ET are characterised by increases in red blood cell and platelet counts respectively, whilst MF is caused by collagen fibres slowly replacing the bone marrow microenvironment, leading to haematopoietic failure (Grinfeld et al., 2017). These disorders are relatively rare and progress slowly, but are nonetheless known to predispose to acute leukaemia (Campbell and Green, 2006). In 2005, several studies reported the existence of a single acquired point mutation in the Janus Kinase 2 (JAK2) gene of MPN patients (Baxter et al., 2005; Coue et al., 2005), which was subsequently found to exist in >95% of PV and 50-60% of ET and MF patients. The mutation is known as the V617F mutation since a valine is substituted by a phenylalanine at codon 617. JAK2 is a cytoplasmic kinase critical for instigating signal transduction from several surface receptors including those for erythropoietin (EPO), thrombopoietin (TPO) and granulocyte colony-stimulating factor (G-CSF) (Kisseleva et al., 2002). Acquisition of the V617F mutation leads to constitutive, cytokine-independent activation of multiple signalling pathways that are usually initiated when EPO binds to its receptor (Figure 1.5), including JAK-STAT, MAPK and ERK (Ugo et al., 2004; Witthuhn et al., 1993). Mouse models of the mutation have clearly recapitulated the phenotype of PV, as evidenced by increased red blood cell counts and subsequent transition to bone marrow fibrosis (Coue et al., 2005; Li et al., 2014).

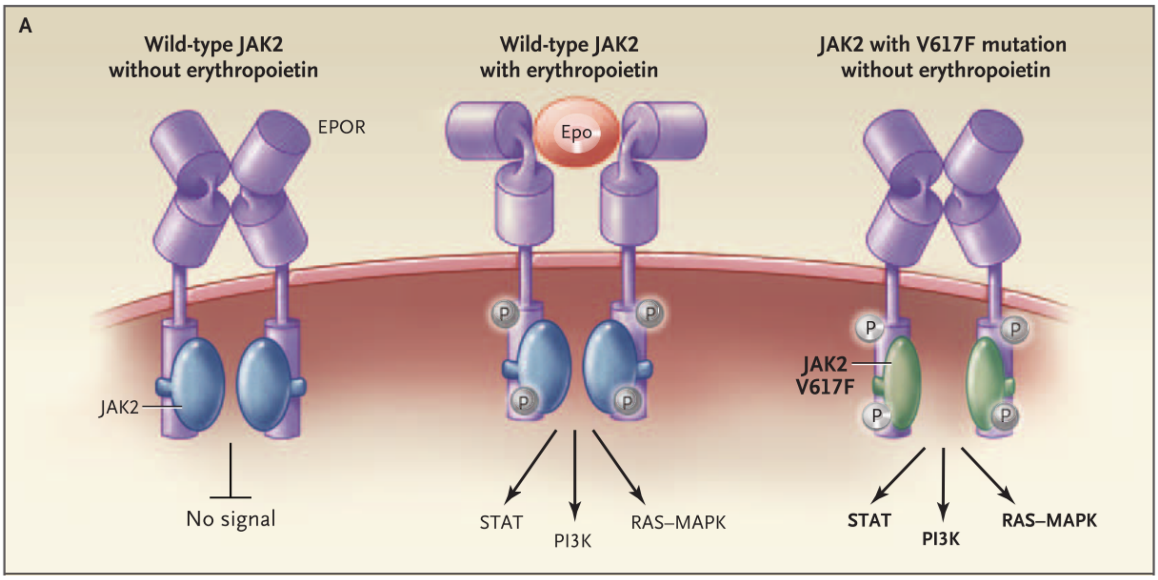

A
Wild-type JAK2 without erythropoietin
EPOR
JAK2
No signal
Wild-type JAK2 with erythropoietin
Epo
P
STAT
PI3K
RAS–MAPK
JAK2 with V617F mutation without erythropoietin
JAK2 V617F
P
STAT
PI3K
RAS–MAPK

Fig. 1.5 The JAK2 V617F mutation. Reproduced from Campbell et. al. (2006). WT JAK2 phosphorylates EPO receptor in the presence of EPO to intiate multiple signalling cascades. The V617F mutation causes this signalling to be constitutively active.

The JAK2 V617F mutation has been shown to occur within the HSC compartment of long-lived MPNs (Anand et al., 2011; James et al., 2008), predisposing towards erythroid differentiation (Jamieson et al., 2006). Results from multiple mouse models and single-cell assays have suggested that the V617F mutation does not in itself confer a self-renewal advantage on HSCs (Kent et al., 2013; Li et al., 2014; Mullally et al., 2010), suggesting that this mutation alone is insufficient to drive disease. It seems likely that cooperation between mutations is necessary to drive MPN disease progression; for example the self-renewal defect observed in JAK2 V617F HSCs was rescued by introducing a second mutation in TET2 (see section 1.3.3), leading to a stronger disease phenotype (Shepherd et al., 2018). TET2 is the gene most commonly comutated with JAK2 in MPN patients, highlighting how the accumulation and synergy between HSC mutations can start to drive a disease state. In the clinic, JAK2 kinase inhibitors such as ruxolitinib have been used to reduce the symptoms of JAK2-driven MPNs, but largely fail to induce remission (Bose and Verstovsek, 2017). As a result, other biological processes such as hypoxia response have been touted as potential therapeutic targets (Baumeister et al., 2020).

The majority of JAK2 V617F-negative ET and MF patients exhibit acquired mutations in either Myeloproliferative leukaemia (MPL) or CALR, which manifest themselves clinically in extremely similar phenotypes. MPL is the cell surface receptor for TPO, and missense mutations in MPL have been associated with increased JAK-STAT, ERK and AKT signalling in a thrombopoietin-independent manner (Klampfl et al., 2013; Rampal et al., 2014). Similarly, insertion/deletion mutations in Calreticulin (CALR) are found in 25-35% of ET and 35-40% of MF patients (Grinfeld et al., 2017). The discovery of CALR in MPNs was unexpected, given its WT role as a chaperone involved with protein folding, but mutant CALR has been shown to promote TPO-independent growth in a MPL-like manner by binding to MPL and acting as a rogue ligand, subsequently activating the same signalling cascades (Chachoua et al., 2016; Elf et al., 2016). Notably, it is very rare for an MPN patient to have more than one of the three main driver mutations (JAK2, MPL, or CALR), suggesting they must have similar mechanistic backgrounds; in all three cases, constitutive cytokine signalling in a JAK2-mediated manner clearly plays a major role. However, this belies the phenotypic heterogeneity of MPNs; whilst the JAK2 V617F mutation can lead to PV, ET or MF, MPL and CALR mutations are limited to ET and MF. There is evidence to suggest that high allelic burden and/or homozygosity of JAK2 V617F predisposes towards a PV phenotype (pro-erythroid) compared to an ET phenotype (pro-megakaryocyte), potentially suggesting that JAK2-driven ET or MF only arises if there is not a large enough mutational burden to drive erythrocytosis (Godfrey et al., 2013; Li et al., 2014; Tiedt et al., 2008). The identity of advantage-conferring comutations, the order in which these occur and the specifics of bone marrow microenvironment may all also play a role in driving MPN heterogeneity (Prick et al., 2014; Walkley et al., 2007).

The final MPN of relevance is systemic mastocytosis (SM), characterised by an overproduction of mature mast cells. Over 90% of patients presenting with SM exhibit constitutive gain-of-function mutations in KIT, a receptor kinase that in a WT setting is tightly controlled by the abundance of its ligand, stem cell factor (SCF) (Chatterjee et al., 2015). Mutated KIT leads to uncontrolled proliferation and enhanced survival of mast cells, which can infiltrate into a diverse array of organs including the bone marrow itself, resulting in organ failure (Verstovsek, 2012). Recent work again concluded that KIT mutations are likely to originate within the HSC/MPP pool in SM patients (Grootens et al., 2019). Much like other MPNs, patients with SM often exhibit comutations that are likely to have distinct impacts on progenitor function and differentiation, and therefore contribute to the diverse phenotypes observed among different SM patients. This will be discussed further in the following section.

1.3.3 Clonal Haematopoiesis

Clonal haematopoiesis (CH) is the generic term for haematological malignancies where a clonal population of cells - harbouring one or more somatic mutations - has expanded within a patient due to a fitness advantage (Bowman et al., 2018). Often, patients with CH display no symptoms or phenotype, but are nonetheless predisposed to further malignancies including MPN, MDS and AML (Jaiswal et al., 2014). This fits into the model of these diseases being caused by successive accumulation of somatic mutations in HSCs over time, where CH is likely to be the first step towards fully malignant transformation (Corces-Zimmerman and Majeti, 2014; Shlush et al., 2014; Welch et al., 2012). Therefore CH can be thought of as a basal pre-leukaemic state, and provides a critical window into how a mutation can perturb haematopoiesis and prime it towards disease.

The two most commonly mutated genes in CH are DNA methyltransferase 3A (DNMT3A) and ten-eleven translocation 2 (TET2) (Buscarlet et al., 2017; Genovese et al., 2014). Both of these are epigenetic regulators; DNMT3A is responsible for establishing *de novo* methylation marks (Chaudry and Chevassut, 2017; Smith and Meissner, 2013), whilst TET2 catalyses the removal of methylation marks from DNA (Ko et al., 2011). DNA methylation is known to play an important role in altering DNA accessibility and hence impacts upon gene regulation, self-renewal and differentiation processes. However the exact role specific epigenetic regulators play in HSC functionality remains unclear. Despite their antagonistic functions, mouse models of both TET2 and DNMT3A mutations have shown that they lead to both increased HSC self-renewal and expansion of the most immature HSC/MPP compartment (Bowman and Levine, 2017; Challen et al., 2014; Moran-crusio et al., 2011). Furthermore, both mutations are thought to cause repression of lineage-specific transcription factors in HSCs (Zhang et al., 2016), although the mechanistic reasons behind this apparent synergy remain unclear (Jeong et al., 2013).

Recently, a landmark study used scRNA-seq to profile the entire blood progenitor compartment of TET2 KO and DNMT3A KO mice (Izzo et al., 2020). This suggested divergent impacts of the two mutations upon downstream progenitors, with TET2 KO generating a pro-monocytic differentiation bias and DNMT3A KO leading to a pro-erythroid differentiation bias. In addition, TET2 KO HSCs were shown to be more quiescent than WT counterparts. This complemented recent functional data that further elucidated the divergent effects of the two mutations (Ostrander et al., 2020). Whilst both mutations lead to increased self-renewal, this increase was shown to be transient in TET2 KO HSCs, whilst being long-lasting in DNMT3A mutant cells as assayed via transplantation. These studies have suggested that previously-observed increases in HSC fitness are where the similarities between TET2 and DNMT3A mutations end.

Izzo et. al. went on to propose a mechanistic link between epigenetic alterations and HSC differentiation skewing by performing methylation and scATAC-seq using the same mouse models. Pro-erythroid transcription factor genes such as TAL1, KLF1 and GATA1 were shown to be relatively denser in CpG (DNA base pairs that can be methylated) content compared to myeloid factors such as IRF8 and SPI1. Therefore if the function of DNMT3A is impaired, relatively more methylation will be removed from erythroid transcription factors, potentially leading to their upregulation and subsequent erythroid-skewing of uncommitted progenitors. The opposite would then be true of HSCs with impaired TET2 function. Whilst such a hypothesis requires refining and confirming, it highlights how current single-cell techniques are paving the way for new mechanistic understanding of pre-leukaemic states.

Many other genes have been shown to drive CH. One of the next most commonly mutated genes is JAK2, again highlighting the heterogeneity within blood disorders and the continuum of phenotypes observed across CH, MPNs and MDS (Sperling et al., 2017). As discussed above, the JAK2 V617F mutation may require cooperating mutations to exhibit a fitness advantage, though why the mutation should sometimes present as CH and sometimes as an MPN remains unclear. Recently it has been shown that JAK2 is an upstream regulator of TET2 in haematopoiesis, potentially linking JAK2 mutations to epigenetic changes (Jeong et al., 2019). Double mutant HSCs for JAK2 and TET2 have been shown to promote a strong disease phenotype that pushes towards an MPN-like state, albeit with clear differences depending on whether JAK2 or TET2 was mutated first (Ortmann et al., 2015; Shepherd et al., 2018). In a similar manner, CH mutations in the isocitrate dehydrogenase enzymes IDH1 and IDH2 are known to have an inhibitory effect on TET2 function, and are largely mutually exclusive with TET2 mutations, suggesting similar mechanistic origins (Figueroa et al., 2010). Mutation of polycomb repressive complex genes such as additional sex combs-like 1 (ASXL1) further implicates post-translational histone modifications as playing a role in regulating HSC self-renewal and differentiation (Fujino and Kitamura, 2020).

In addition to a number of epigenetic modifiers, two other major biological pathways have been implicated with CH. The first of these is the spliceosome, with mutations in SF3B1, SRSF2 and U2AF1 (amongst others) all found in patients with CH and to a greater extent, MDS (Inoue et al., 2016; Sperling et al., 2017). However, mechanistic understanding of which spliced isoforms are responsible for driving clonal expansion has proved elusive. Similarly, several genes that form part of the cohesin complex were found to be mutated in CH, including STAG2, SMC1A and RAD21 (Mazumdar et al., 2015). Current hypotheses suggest that these may drive disease through their dysregulation of long-range chromatin interactions, potentially placing further emphasis on the role of higher-order epigenetic modifications in driving aberrant haematopoiesis (Thota et al., 2014).

Whilst clonal haematopoiesis is a relatively benign disease, it has an important role to play in understanding leukaemic transformation. Disease progression is likely to involve a complex interplay between a) the initial driver mutation, b) the synergistic effects of subsequent mutations and c) bidirectional interactions between mutant HSCs and their niches (Raaijmakers et al., 2010; Schepers et al., 2013). Notably, many patients that progress to more serious malignancies exhibit mutations from several different pathways (e.g. one epigenetic mutation, one spliceosome mutation, one cohesin mutation etc), potentially suggesting that accumulating 'orthogonal' mutations primes HSCs towards more aggressive transformations. An improved understanding of the impacts of the key driving genes associated with CH is essential to produce better therapies across a range of haematopoietic malignancies in the future.

1.3.4 Myelodysplastic syndrome, sAML and *de novo* AML

In contrast to MPNs and CH, myelodysplastic syndrome (MDS) is characterised by clear morphological dysplasia of a patient's blood cells, as well as the presence of cytopenias (decreased blood counts) amongst terminally differentiated cells. Whilst this defines the clinical distinction between MDS, CH and MPNs, the underlying genetics of these disorders lie on a continuum, with largely the same sets of somatic mutations found in each. MDS is considered to arise from the progression of clonal haematopoiesis into a more severe disease, with the majority of MDS patients exhibiting at least two - and sometimes many - somatic mutations, often with large clone sizes (>10%) (Papaemmanuil et al., 2013). Around one third of patients diagnosed with MDS progress to some form of AML as a result of further mutations (Sperling et al., 2017). Compared with CH and MPNs, MDS patients exhibit a higher proportion of spliceosome mutations, suggesting these mutations are linked to morphological dysplasia. Nevertheless, DNMT3A and TET2 remain the most commonly mutated genes in MDS. Conversely, mutations in signalling pathway genes such as JAK2 and CALR occur at relatively low frequencies in MDS compared to MPNs; if they do occur, it is often a sign of disease progression towards AML (Haferlach et al., 2014).

Secondary AML (sAML) is defined as an AML that arises due to an existing blood disorder such as MDS or an MPN. The threshold for diagnosis of AML is the presence of >20% hyperproliferative leukaemic cells (blasts) within the haematopoietic system (Vardiman et al., 2009). In contrast, *de novo* AML arises rapidly without any such background - most typically in children or young adults - and whilst it is associated with better prognosis than sAML, it is more common and leads to more deaths (Siegel et al., 2015). *de novo* AML is associated with large chromosomal translocations, but also with mutations in many of the same genes as sAML, such as DNMT3A, TET2 and ASXL1 (Patel et al., 2012). Furthermore, there are a number of genes found mutated in *de novo* AML that are rarely seen in sAML or predecessor neoplasms. These include nucleophosmin 1 (NPM1), which has roles in multiple cellular processes including ribosome biogenesis and centrosome duplication, and is found mutated in around 30% of *de novo* AML patients, often alongside mutations in DNMT3A (Network, 2013). NPM1 mutant mouse models exhibit increased HSC self-renewal and expanded myeloid differentiation (Vassiliou et al., 2011). Other genes, such as the transcriptional coactivator CREBBP, are often impaired and/or fused with other genes as a result of chromosomal translocations found in *de novo* AML, and impaired CREBBP has been shown to associate with increased HSC quiescence (Chan et al., 2011). A wide range of other mutations/translocations exist in AML subgroups, and have been reviewed extensively (Kouchkovsky and Abdul-Hay, 2016).

The differences in the genetic underpinnings of *de novo* and sAML are still debated, though it seems likely that *de novo* AML is generally associated with a ‘stronger’ set of perturbations than sAML. For example, whilst a DNMT3A mutation may not be able to generate an AML phenotype by itself, it can synergise with an NPM1 mutation to drive a *de novo* AML phenotype, or with spliceosome or JAK2 mutations to drive an MDS/MPN phenotype that may then evolve into sAML. It has been reported that *de novo* AMLs exhibit greater microenvironmental dysregulation compared to sAMLs, further highlighting the importance of cooperation between driver mutations and the bone marrow niche (Lopes et al., 2017). Mouse models of these mutations, combined with single-cell profiling and functional assays, offer the opportunity to investigate their affects within the context of the entire blood system at high resolution. Integrating together information about many different mutations will therefore help to understand the genetic, phenotypic and prognostic heterogeneity observed across haematological malignancies, as well as highlighting conserved biological pathways and mechanisms across different diseases.

1.4 Computational tools for single-cell biology

1.4.1 Addressing technical and biological noise in scRNA-seq data

The raw gene counts returned by a scRNA-seq experiment are subject to confounding noise that can inhibit downstream analyses and biological inference unless dealt with sensibly. This noise can be technical or biological in origin, and whilst sources of biological noise can often be of interest despite their confounding effects (e.g. cell cycle, apoptosis), technical noise does little other than obscure relevant signal. Unfortunately, the deconvolution and removal of technical and biological noise from scRNA-seq data remains a significant challenge.

Sources of technical noise occur at every stage of a scRNA-seq protocol, although some will be more important than others. Only a small proportion of mRNA molecules within a given cell will be reverse-transcribed after lysis, with the exact proportion dependent on the protocol being used (Kolodziejczyk et al., 2015). For lowly-expressed genes, this Poisson sampling can cause no mRNA molecules to be captured despite their existence within the cell (Islam et al., 2014). The efficiency of the PCR amplification used in most protocols is sequence dependent, introducing further technical variation if UMIs are not used. Stochastic sampling of counts from the sequencing library then adds another layer of technical variation, regardless of whether UMIs are used or not. The end result of these technical noise sources is that biologically identical cells will return different count distributions - and importantly, different numbers of a) total UMIs/reads and b) number of expressed genes - due to intrinsically technical effects.

The process of correcting each cell's count distribution to correct for technical effects - whilst preserving biological variation - is referred to as performing a 'normalisation' of the data. A wide variety of techniques have been suggested to do this. Initially, the concept of 'size-factors' was used to scale each gene in a cell by a cell-specific constant factor that represented the amount of material captured from that cell. Therefore technical variation in the amount of mRNA captured from each cell would be mitigated. This idea was originally implemented using a method introduced for bulk RNA-seq in the DESeq method (Anders and Huber, 2010), and was extended for heterogeneous single-cell datasets by pooling information across cells in the SCRAN method (Lun et al., 2016). For plate-based protocols such as Smart-Seq2, external RNA spike-ins to each cell can be used to quantify and correct for technical variability using the BASiCS method, which estimates size-factors based on the abundance of these spike-ins (Vallejos et al., 2015).

Underlying size-factor based normalisations is the assumption that the total mRNA content of each cell in the dataset is of roughly the same magnitude, and that the counts of all genes can be modified by a single factor, regardless of their abundance levels. More recently, these assumptions have been questioned and new, probabilistic models of technical variation have been proposed. These have centred on using the negative binomial distribution to model count data from scRNA-seq experiments (Grün et al., 2014; Kharchenko et al., 2014). Methods such as sctransform use negative binomial regression to correct count values in a gene-specific manner, thus overcoming some of the limitations of previous approaches (Hafemeister and Satija, 2019). Other studies have additionally modelled the presence of 'zero-inflation' in scRNA-seq data (extra zero counts caused by technical undersampling) as part of new normalisation methods, such as the ZINB-WaVE approach (Risso et al., 2018). However the existence of zero-inflation has been credibly disproved (Svensson, 2020), highlighting the importance of model assumptions on normalisation techniques. Currently, size-factor based normalisation methods are still used for the majority of scRNA-seq workflows, although the consensus is beginning to shift towards more generalised probabilistic approaches. It has been suggested that normalisation methods for plate- and UMI-based experiments should be different, as empirically the data produced by them is best fit by slightly different models (Luecken and Theis, 2019).

Droplet-based protocols incur an orthogonal layer of technical noise due to the existence of doublets (droplets with two cells in rather than one) that get barcoded and sequenced as if they were a single cell, and are believed to occur at a rate of several percent (Zheng et al., 2017). Methods have been introduced to account for this by simulating synthetic doublets and comparing these to transcriptomes within a dataset (Wolock et al., 2018). Whilst this has been shown to effectively remove doublets containing two disparate cell types, removing doublets containing two highly similar cells remains difficult.

Biological 'noise' within scRNA-seq counts can also originate from a wide variety of sources, many of which are related to areas of current stem cell research. These include the effects of transcriptional bursting kinetics, whereby a gene with exactly the same activity in two cells can record different counts due to the periodical bursting nature of mRNA transcription (Ochiai et al., 2020). Additionally, cell-specific extrinsic signals will certainly contribute to gene expression; this idea is now at the forefront of single cell research as new tools are developed for spatial transcriptomics (Shah et al., 2016). The most tractable sources of biological variation come from well studied cell intrinsic processes such as cell cycle and apoptosis. Prior knowledge can be used to either remove cells in a specific cellular state (e.g. apoptotic cells with a high proportion of reads mapping to mitochondrial genes) or regress out the source of variation, for example using one of the many methods designed to remove confounding cell cycle effects (Liang et al., 2020; Scialdone et al., 2015; Tirosh et al., 2016). Whilst in some circumstances cell cycle may be of primary interest, it typically obscures the analysis of single-cell differentiation trajectories by causing cells at similar stages of differentiation to have very different transcriptomic profiles.

The identification of genes that exhibit high levels of variation across a dataset (the ‘highly variable genes’) is an important step for many downstream applications including clustering. Empirically, the variance of a gene amongst a homogeneous group of single cells is found to be a quadratic function of its mean expression (hence it can be suitably modelled using a negative binomial distribution), such that highly-expressed genes display the largest variations across a dataset due to both technical and biological noise (Grün, 2020). Methods have been proposed to identify genes exceeding the expected level of variation given their mean expression; the method of Brennecke et. al. incorporates external RNA spike-in values and has been widely used in plate-based experiments (Brennecke et al., 2013). Similarly, the method from Macosko et. al. identifies genes with extreme variances compared to other genes with similar expression values, and has been adopted for many droplet-based experiments (Macosko et al., 2015). The importance of per-gene variation will be discussed and extended in Chapter 4.

1.4.2 Dimensionality reduction and clustering

The great advantage of single cell genomic profiling over bulk techniques is that a) the heterogeneity within a cellular population can be resolved, and b) the structure and relationships between these heterogeneous sub-populations can be analysed. However the task of uncovering this structure in a dataset containing many thousands of cells, each with many thousands of genes (or other modalities) is computationally challenging. Dimensionality reduction (DR) embeds a high-dimensional count matrix into a lower-dimensional space, whilst aiming to preserve as much of the biological structure within the data as possible. This low-dimensional representation can then be better comprehended and visualised (Figure 1.6A). Both the method used to perform DR and the dimensionality of the subsequent representation play an important role in determining the possible downstream applications of the transformed dataset (Watcham et al., 2019). For example, performing DR to a 2 or 3-dimensional representation will allow the dataset to be easily visualised, but will likely remove too much of the biological structure to perform accurate cell clustering or inference of differentiation trajectories (Figure 1.6B, see section 1.4.3). On the other hand, performing DR to achieve a representation with 10^1-10^2 dimensions is likely to capture and preserve the vast majority of the biological structure, allowing more difficult tasks to be performed on a suitably complex yet tractable representation of the data (Figure 1.6C). DR is used in almost all scRNA-seq workflows, as it helps to mitigate against the sparsity of the data; whilst a single gene is susceptible to both technical and biological zeros, a dimension within the transformed dataset will use information from across many (correlated) genes, leading to better resolved biological signals.

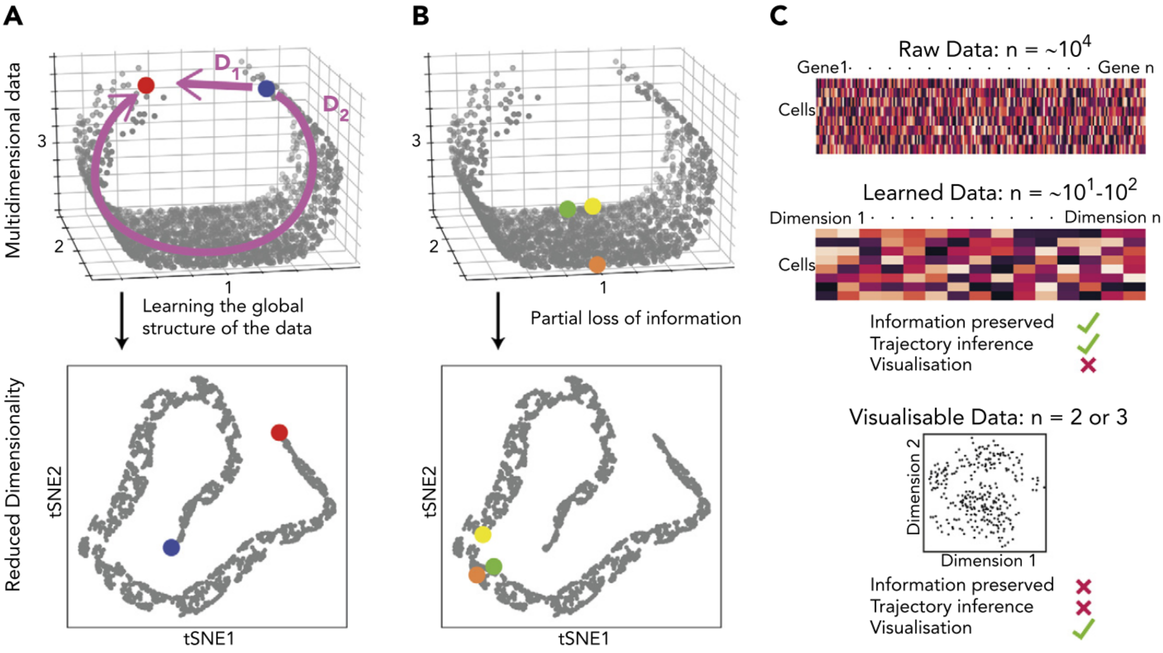


Fig. 1.6 Different levels of dimensionality reduction. (A) A set of single cells expressing 3 genes arranged along a curved shape has been simulated. There are 2 measures of distance between the blue and red cells. Whilst D_1 represents the shortest possible distance, D_2 is the distance between the cells through the structure of the data (manifold). The two arms of the curved shape may represent continuous transition processes (e.g, cell differentiation); therefore, distance D2 is the important distance measure. A dimensionality reduction technique (here tSNE) should capture such features. (B) Excessive reduction in dimensionality causes important information to be lost. In this case, a 2-dimensional representation of the data incorrectly suggests that the green cell is farther from the yellow cell than the orange cell, because information has been lost about axis 2. (C) To infer cellular trajectories from scRNA-seq data, dimensionality reduction is used to learn the structure of the data (learned data), which captures the important distances between cells in a suitable number of dimensions, typically 10 to 100. Trajectory inference can then be attempted from this learned data. For visualisation, the dimensionality of the data needs to be reduced to either 2 or 3, but this will inevitably result in the loss of some of the important biological information, rendering data unsuitable for trajectory inference.

A number of DR methods have been widely applied to single-cell transcriptomics. These include linear methods such as principal component analysis (PCA), which calculates a new coordinate system for the data where the first coordinate (principal component 1 or PC1) explains the greatest amount of variance possible, and subsequent components explain the largest possible amount of variance whilst remaining orthogonal to all previous PCs. Hence the PCs are ordered by their 'importance' in explaining the variation within the data, and therefore taking the top *n* PCs - *n*=50 is typically used - as a low-dimensional representation can explain the majority of the total observed variance (Tsuyuzaki et al., 2020). The closely related independent component analysis (ICA) method is another linear technique that aims to decompose a dataset into a set of independent signals that each capture a different source of variation (Stein-O'Brien et al., 2018). In a scRNA-seq setting these may correspond to processes such as differentiation, cell cycle or apoptosis (Ocasio et al., 2019).

Whilst PCA and ICA are widely used, they are general statistical methods that are not necessarily suited to dissecting the stochastic biological processes involved with differentiation. To address this, Haghverdi et. al. proposed the use of non-linear diffusion maps as a DR technique that better accounts for the continuous nature of cell trajectories in scRNA-seq data (Haghverdi et al., 2015). Distances between cells are calculated by simulating diffusive random walks in the original high-dimensional space, in a manner similar to how an individual cell may explore gene expression space before committing to a particular lineage. Similar to PCA, the top *n* diffusion components (DCs) calculated through this method can be used as a low-dimensional representation of the data, with each component capturing some aspect of a continuous process. For this reason, diffusion map representations have been successfully used as a basis for creating temporal orderings of cells along a differentiation trajectory (see section 1.4.3), as well as for reconstruction of gene regulatory networks (Haghverdi et al., 2016; Ocone et al., 2015).

One drawback of using PCA or diffusion maps to visualise single cell data is that whilst the top 50 components capture nearly all of the variance, the top 2 or 3 components by themselves do not. Therefore if a large, unbiased scRNA-seq dataset contains several different differentiation trajectories, some of them may be completely ignored when just the top two PCs or DCs are visualised. Therefore specific methods have been developed that aim to capture as much of the variance as possible in a very small number of components. Two of the most widely used are the t-distributed stochastic neighbour embedding (tSNE) and uniform manifold approximation (UMAP) methods; these aim to find a distribution of pairwise cell distances in 2 or 3 dimensions that accurately reflects the distribution in the original high-dimensional space (Maaten and Hinton, 2008; Mcinnes and Healy, 2018).

Both have been widely applied to a range of scRNA-seq datasets (Ibarra-Soria et al., 2018; Scialdone et al., 2016; Wang et al., 2018; Wilson et al., 2015).

A different approach to visualisation has been developed in the form of force-directed graphs (FDG), which are calculated using k-nearest-neighbour (kNN) graphs. In a kNN-graph, each cell is connected to its *k* nearest neighbours, as defined by a distance metric (which could be measured in the high-dimensional space, or in the PC or DC representations). An algorithm is then applied which simulates repelling forces between all cells, such that cells are only held together by the nearest neighbour connections. The resulting arrangement of cells in 2 or 3 dimensions is used as the FDG embedding (Weinreb et al., 2018a). Since being first introduced, they have been widely used to visualise the branching structure of differentiation trajectories within the blood system, as they excel at extracting all trajectories into a single visualisation (Dahlin et al., 2018; Giladi et al., 2018; Tusi et al., 2018).

Clustering single cells into discrete populations can often be a useful tool for analysing the structure of a dataset. Not only can it highlight groups of cells with the most similar profiles, but it also defines populations for differential expression analysis (or relevant features in other modalities). Whilst supervised clustering of scRNA-seq data using prior knowledge of highly-expressed marker genes is sometimes possible, it is often biased by the incompleteness of the prior knowledge relating to either the marker genes themselves or the cell types present within the data. Unsupervised clustering is a more data-driven approach, and different approaches have been used in a myriad of studies to investigate haematopoiesis at the single cell level (Guo et al., 2013; Jaitin et al., 2014). These include standard hierarchical clustering, probabilistic models and graph-based approaches (Drissen et al., 2016; Levine et al., 2015; Paul et al., 2015).

As scRNA-seq datasets have tended towards large, unbiased sampling of continuous transcriptomic landscapes, kNN-graph-based clustering approaches have been increasingly appropriated for use on single cell data. Inferring sensible clusters along a well-sampled continuous differentiation trajectory relies on maximising both intra-cluster similarities and inter-cluster differences, regardless of the steadily changing expression profiles along the continuous path. Graph-based approaches do this by optimising the ‘modularity’ of the assigned clusters; this works by maximising the difference between the number of nearest-neighbour connections within a cluster compared to what would be expected if nearest-neighbour connections were assigned randomly (Newman and Girvan, 2004). Hence the resulting clusters are tightly intra-connected and have the minimum possible overlap with each other. Furthermore, modularity-based algorithms are scalable and quick to run on large datasets; one of the most popular, the Louvain algorithm, has been widely applied to single cell data (Andrews and Hemberg, 2018; Blondel et al., 2008). More recently, the Leiden algorithm has been shown to improve upon the Louvain approach in terms of maximising the intra-cluster connectivity (Traag et al., 2019). Typically, the kNN-graphs used for clustering are calculated from a low-dimensional representation of the data such as PCA, as it has been suggested that performing nearest-neighbour calculations directly from a sparse, high-dimensional space can introduce significant biases to the resulting clusters (Radovanovic et al., 2010).

1.4.3 Inferring dynamic processes from transcriptomic landscapes

A blood cell's differentiation trajectory from HSC to a terminal fate is a continuous, dynamic process. When a scRNA-seq experiment is performed, all temporal information about a cell is lost, leaving only a snapshot of its transcriptional state at some moment in time. However if an entire trajectory (or set of trajectories) is sampled in an unbiased fashion and at a high enough density - as is now possible with droplet-based methods - these dynamic processes can start to be unravelled. This relies on the assumption that differentiation is asynchronous, leading to the idea that the temporal behaviour of a single cell along a trajectory can be approximately recreated by connecting the transcriptomic states of many different cells captured at different stages along the same trajectory (Weinreb et al., 2018b). This raises two key computational challenges. Firstly, given a set of cells belonging to a trajectory, how can they be ordered into the configuration that most accurately reflects the true temporal development of a cell along that trajectory? This is essential if the true transcriptional dynamics of differentiation are to be found. Secondly, in systems such as haematopoiesis where stem/progenitor cells are making fate decisions, can the potential terminal fates or a particular transcriptomic state be defined? This is equivalent to finding the locations of the 'branchpoints', where fate decisions are being made within the transcriptional landscape. Achieving this would allow the genetic drivers of specific fate decisions and the hierarchy between diverging lineages to be elucidated (Laurenti and Göttgens, 2018). The haematopoietic system has often been the model of choice for trying to answer these questions.

Ordering single cells along a differentiation trajectory encompasses the idea that cells with the most similar transcriptional profiles are likely to exist temporally close to each other during a biological process such as differentiation. In addition, these cells are likely to be closely related on a transcriptomic landscape. Initial attempts to order cells in this way simply used PCA to identify a principal component that looked like it corresponded to differentiation progress based on prior biological knowledge (Guo et al., 2010). Cells were then ordered simply by their location on that PC. However it is unlikely that a single component will accurately capture all aspects of differentiation whilst not being influenced by other processes. In 2014, landmark papers by Trapnell et. al. (2014) and Bendall et. al. (2014) introduced the concept of 'pseudotime', a quantitative measure of each cell's progress through a continuous differentiation trajectory. The former study used minimum spanning trees (MSTs) to identify the shortest path through a low-dimensional representation of the data. Cells were then ordered based on their position along the MST. This approach could further search for the location of fate decisions by looking for the largest separate branches within the MST, and assigning these to different trajectories based on user-defined information about the number of trajectories within the data. In practice, it was found that such a MST approach was often unstable for large datasets, and later versions of the same algorithm introduced completely new techniques for learning the structure of the data, such as reverse graph embedding, to improve the stability of pseudotime estimates (Qiu et al., 2017). Alternatively Bendall et. al. (2014) first calculated a kNN-graph on the data, and using ensemble methods calculated the average 'distance' on the kNN-graph between each cell and a user-defined starting cell. The authors beautifully demonstrated the efficacy of this approach by reconstructing a single cell trajectory of B cells that correctly ordered the known landmarks of B cell development, highlighting the potential ability of such methods to identify new drivers of differentiation. Later improvements to this approach allowed for the identification of branchpoints within the kNN-graph (Setty et al., 2016).

The concept of low-dimensional diffusion maps introduced in section 1.4.2 has been extended to calculate pseudotimes for continuous trajectories by constructing a random-walk based distance metric within the diffusion map space of a dataset (Haghverdi et al., 2016). This metric, known as diffusion pseudotime (DPT), is then calculated between a user-defined root cell (such as a HSC) and every other cell, before being used for ordering. DPT has been widely used, largely as a result of the robustness and efficiency of diffusion maps at capturing complex continuous structures (Nestorowa et al., 2016). However whilst it is capable of identifying bifurcation points, the DPT algorithm struggles when confronted with a dataset containing more than two trajectories (Qiu et al., 2017). Although pseudotime algorithms such as DPT have proved extremely effective at ordering single cells along a trajectory, they represent only a small fraction of the recent methods developed to infer differentiation trajectories (and by extension, lineage hierarchy and fate decision branchpoints) from scRNA-seq data. A wide array of approaches have been used to do this, many of which do not explicitly calculate pseudotimes; naturally, however, once a method has been used to identify trajectories a separate pseudotime algorithm can then be applied to achieve an ordering of cells.

Amongst methods designed to infer differentiation trajectories and branchpoints, there is enormous heterogeneity in the approaches used, the prior knowledge required and differentiation topologies that can be interrogated. Saelens et. al. (2019) provide a comprehensive review of over 40 different methods using various metrics including scalability and ease of use. One of the more common approaches is to rely on a clustering of the data to create a graph through which trajectories can be traced. For example, the partition-based graph abstraction (PAGA) algorithm calculates a connectivity score between each of the input clusters using a kNN graph, before creating a new graph based on a threshold for connectivity. Applied to a dataset of murine haematopoietic progenitor cells, PAGA suggested multiple routes of differentiation to a basophil fate through either a neutrophil/monocyte or an erythroid/megakaryocyte progenitor state (Wolf et al., 2019). Other graph-based methods such as Slingshot and StemID take similar approaches (Grün et al., 2016; Street et al., 2018). Nonetheless, the efficacy of this approach is heavily reliant on the quality and resolution of the clusters used for inference. It is also not an approach that particularly lends itself to continuous trajectories, given the need to discretise a dataset before inference can occur. However graph-based approaches do allow for easy identification of clusters near where fate decisions may be occurring, which can then be further interrogated at a molecular level.

As an alternative to this approach, other methods aim to build a classifier that calculates the probability of each terminal fate in the dataset for every cell. A highly supervised approach was introduced with the STEMNET algorithm, which uses the expression of predetermined marker genes for each terminal fate as input to a generalised linear model, which then predicts the degree of lineage priming towards each fate for the entire dataset. In human bone marrow HSPCs, STEMNET strikingly predicted a lack of oligopotent progenitor populations, with committed unipotent cells emerging from a pool of multipotent HSCs and MPPs that were equally primed in all directions (Velten et al., 2017). This appeared to support results from functional transplantation and *in vitro* assays, and raised important questions about the nature of the haematopoietic hierarchy in humans compared to mice, many of which remain unanswered (Laurenti and Göttgens, 2018; Notta et al., 2016). Building on this idea, the FateID algorithm classifies single cells using a semi-supervised iterative process (Herman et al., 2018). Provided with the identity of genes relating to each terminal fate, a random forest classifier is used to move backwards from the end of each trajectory, classifying a small number of adjacent cells before building a new classifier using the transcriptomes of cells added in the previous iteration. In this way, the features used for classification are constantly updated along each trajectory. The authors suggested that this iterative nature allowed them to observe a greater degree of lineage priming in MPPs compared to STEMNET, although it must be considered that FateID was benchmarked on murine haematopoietic cells rather than human. Nevertheless results once again suggested that key lineage decisions were being made early on in the haematopoietic hierarchy in contrast with the classical model structure. For each cell, FateID returns a vector of probabilities associated with commitment to each terminal fate, allowing the location of successive branchpoints in the data to be located wherever the probabilities for multiple fates are equivalent. One limitation of this approach is that if the magnitude of molecular changes are vastly different between two trajectories, classification of a bipotent progenitor population within both trajectories can be heavily biased (Saelens et al., 2019).

More complex trajectory inference methods have also been proposed, though these typically require more starting information. The Waddington-OT method uses the mathematical framework of optimal transport to map high-dimensional regions within a transcriptional landscape across different timepoints, allowing trajectories to be inferred (Schiebinger et al., 2019). However this approach relies on time course experiments, rendering it unsuitable to study steady state haematopoiesis. Similarly, population balance analysis (PBA) takes a physically-motivated approach to trajectory inference by attempting to solve a drift-diffusion equation for the potential underlying a transcriptomic landscape (Weinreb et al., 2018b). However this requires knowledge of the exit rates of cells from each terminal fate, which is not known *a priori* despite recent work attempting to quantify the flux of cells through specific HSPC populations (Busch et al., 2015).

Overall, pseudotime and trajectory inference methods have played an integral part in uncovering hitherto unknown routes of differentiation and coupling between lineages in the haematopoietic system. Combined with lineage barcoding experiments, trajectories inferred from scRNA-seq data allow the transcriptomic state of a cell to be linked to its functional fate biases, allowing questions about the hierarchy of haematopoiesis to be answered in a way that is not possible using single cell profiling alone (Rodriguez-Fraticelli et al., 2018; Weinreb et al., 2020). However many of the methods described here have clear drawbacks, and do not work well with the large, complex and multimodal datasets that are now being produced using multiomics approaches. Hence further innovation will be required to better define the cell states involved in haematopoietic fate decisions.

1.5 Single-cell perturbation experiments

1.5.1 Integrating scRNA-seq data across conditions

Profiling perturbed states of haematopoiesis, such as those associated with pre-leukaemic malignancies and AML, is an important step towards understanding the mechanisms driving disease states at the scale of the entire blood system. The effects of even a single driver mutation can be complex and alter the wild-type transcriptional landscape in a myriad of ways (Figure 1.7A). For example, aberrant fate decisions can result in downstream abundance shifts due to a greater/lesser proportion of cells entering a certain trajectory (Figure 1.7B) (Laurenti and Göttgens, 2018). Changes to the self-renewal abilities of immature populations can also alter their apparent abundances within the landscape, as can changes to the proliferation or apoptosis rates of a population or its upstream progenitors (Nimmo et al., 2015). Equally, aberrant rates of differentiation along particular trajectories, due to either differentiation blocks or over-active signalling cascades can cause a population to appear under- or over-represented, even in the absence of any changes to the actual fate decisions (Akinduro et al., 2018). This 'cellular' information (i.e. relating to differential abundances in specific regions of a landscape) must then be considered and interpreted in the context of the 'molecular' information (i.e. relating to the expression of specific genes) contained within the over/under-abundant regions, as well as in the immature progenitor populations where aberrant fate decisions may actually be occurring (Povinelli et al., 2018).

Before these signals can be compared between WT and perturbed states, data from the two conditions must be integrated effectively. As previously discussed, technical variation can play a large role in scRNA-seq datasets, and wildly different levels of variation can be observed between conditions even when they are subject to identical protocols and sequenced at the same time (Büttner et al., 2019). This can cause biological differences between conditions at either cellular or molecular scales to be largely obscured by technical effects. There are two main approaches towards dealing with these 'batch effects' in scRNA-seq data. The first and most common of these is to integrate each condition into a single dataset, before using one of a number of published methods that attempt to correct for technical variation using the conditions as a covariate. These include ComBat - which was originally designed for microarray data and subsequently applied to scRNA-seq data (Johnson and Li, 2007) - alongside more recent, nonlinear approaches designed specifically for single cell data, such as canonical correlation analysis (CCA), mutual nearest neighbours (MNN) and batch-balanced k-nearest neighbours (BB-KNN) (Haghverdi et al., 2018; Polanski et al., 2020; Stuart et al., 2019). The key assumption behind these later integration methods is that

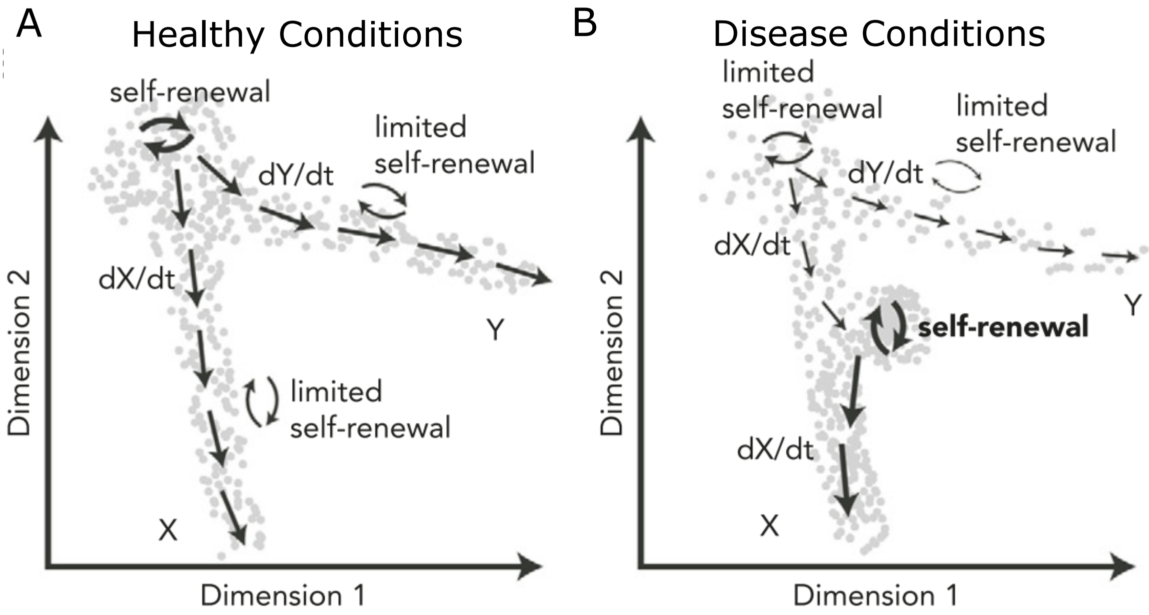


Fig. 1.7 Disease states initiate shifts in a transcriptional landscape. (A) A schematic of a wild-type transcriptional landscape where a pool of progenitors gives rise to 2 differentiated populations X and Y. Arrows indicate directions of differentiation throughout the landscape and the degree of self-renewal. (B) An example of a disease state, where the stem cell pool is exhausted (low self-renewal), compromising production of Y cells, whereas there is increased production of X cells because of increased self-renewal in a downstream population of X progenitors.

at least some cells of the same type exist in multiple conditions, and that differences between the same cell type across conditions are typically smaller than the differences between cell types within a condition. This allows for the identification of the same cell type across conditions using computationally efficient graph-based methods, and the (assumed technical) differences between them removed. Batch correction methods have been applied to integrate scRNA-seq data across different genders, species and sequencing protocols (Butler et al., 2018). Nonetheless, the simple empirical Bayes framework of ComBat has been suggested to outperform more recent methods on smaller datasets (Büttner et al., 2019), and it has been further suggested that many current batch correction methods are prone to overcorrection, thereby removing much of the biological variation inherent to a dataset (Hie et al., 2019; Korsunsky et al., 2019). In particular, many integration methods struggle to successfully integrate a large number of different conditions into a single dataset, limiting their practicality in certain situations.

The second approach for mitigating batch effects is to compare data from each condition with respect to a 'reference' dataset that acts as an anchor for the entire analysis. This reference can be either the data from one of the conditions in a multi-condition experiment (most typically the wild-type or control sample), or an external cellular atlas generated separately. Ideally, the reference data will be large, densely sampled and cover a similar (or larger) region of the transcriptomic landscape compared with the other conditions, allowing cells from each condition to be 'mapped' to the reference via a sensible classifier. This approach allows cellular information to be integrated across conditions in an easily comprehensible way and without the risk of overcorrection; it has therefore been used by a number of studies analysing the top of the haematopoietic hierarchy (Dahlin et al., 2018; Tusi et al., 2018). However this approach cannot be used to integrate perturbations containing cell populations that are not present in the reference. Published methods such as scmap attempt to perform this mapping robustly based on cell-to-cell distances in gene expression space, assigning each cell from one condition (e.g. perturbed state) onto a reference dataset; cells that cannot be sensibly mapped to any part of the reference landscape are left unassigned (Kiselev et al., 2018). Overall, using a reference landscape is arguably a more powerful and scalable approach to integrating a large number scRNA-seq perturbation experiments - provided that a suitable reference exists - compared with integrating all conditions into a single dataset. Data integration across conditions and modalities is likely to be one of the key tasks in the single-cell field moving forward, and both approaches will likely prove useful in extracting information from large and varied sources, such as the Human Cell Atlas (Benoist et al., 2017).

1.5.2 Identifying mechanisms driving perturbed haematopoietic states

Once data integration has been performed, perturbed transcriptional landscapes can be analysed and compared to their wild-type counterparts, with great potential for novel biological insight. In the past two to three years, a small number of studies have begun to exploit this potential by profiling broad haematopoietic populations from perturbed states, rather than remaining restricted to a single cell type (Shepherd and Kent, 2019). Dahlin et. al. profiled HSPCs (using 10X Genomics) from a mouse model with defective c-Kit signalling and compared them to wild-type cells from the same broad sorting gate (Dahlin et al., 2018). The study highlighted how the mast cell fate is completely absent in these mice, consistent with current knowledge of mastocytosis and the phenotype of the mice (see section 1.3.2). The perturbed cells were mapped onto clusters defined on the WT reference, and the relative abundances of each cluster were calculated, highlighting an increase in erythroid progenitors. Differential expression between conditions suggested a reduction in pro-apoptotic genes amongst the perturbed cells. Whilst informative, this study was only able to scratch the surface of the data in terms of the molecular drivers that were shifting the perturbed transcriptional landscape. Similarly, Tusi et. al. compared WT HSPCs to erythropoietin (EPO) stimulated cells in order to understand the effects of stress on the erythroid differentitation trajectory (Tusi et al., 2018). Unsurprisingly, it was found that the relative abundance of erythroid progenitors increased relative to WT. More interestingly, it was suggested that different regions of the erythroid trajectory were perturbed with different magnitudes, in terms of the number of differentially expressed genes in each region. Pro-myeloid (and hence anti-erythroid) transcription factors such as CEBPB were found to be downregulated in MPPs compared to WT, potentially linking transcriptional changes to altered fate bias.

In addition to mouse models and *in-vitro* perturbations, scRNA-seq has also been combined with targeted mutational detection in order to directly analyse human patient samples. Giustacchini et. al. successfully used this approach to characterise CML cells, revealing clear differences between a) healthy HSCs and CML HSCs, b) CML leukaemic cells before and after therapy and c) CML leukaemic HSCs and CML healthy HSCs (Giustacchini et al., 2017). One of the most interesting results was that genetically healthy HSCs from CML patients upregulated genes associated with HSC microenvironmental factors such as IL-6 and TNF-α, suggesting that the presence of leukaemic HSCs might be suppressing the activity of nonmutant HSCs by causing inflammation in the HSC niche (Shepherd and Kent, 2019). This approach could be revolutionary for understanding perturbed haematopoiesis in leukaemia patients, and recent work has attempted to optimise the technology such that even single point mutations can be confidently called whilst simultaneously capturing whole-transcriptome information (Rodriguez-Meira et al., 2019). This new technology, TARGET-seq, may play an exciting role in elucidating the interplay between healthy and malignant clones within pre-leukaemic patients over the next few years.

Most recently, Izzo et. al. performed scRNA-seq on the blood progenitor landscape for mouse models harbouring TET2 and DNMT3A mutations. As discussed in section 1.3.3, this study attempted to link the observed abundance shifts in the transcriptional landscape to the underlying epigenetic changes occurring in the most immature HSCs and MPPs (Izzo et al., 2020). Combined with single-cell work looking specifically at HSC self-renewal, transplantation ability and functional potential in these models (Ostrander et al., 2020; Shepherd et al., 2018), mechanistic models of how the most common mutations driving CH, MPNs and MDS lead to a pre-leukaemic state are starting to emerge. Nevertheless, our understanding of pre-leukaemic states at the scale of the whole blood system remains in its infancy. In particular, comparisons between different pre-leukaemic states are currently difficult to perform, and there is no global framework in which to interpret the cellular and molecular shifts observed as a result of individual mutations. This is essential if the insights obtained from different mouse models of pre-leukaemic states are going to be transformed into future therapeutic strategies.

1.6 Research goals

Somatic mutations within the HSC compartment can drive haematological malignancies and leukaemia in human patients. To decipher the mechanisms behind these processes, an integrated understanding of the effects of these pre-leukaemic perturbations at the scale of the entire blood system is required. High-throughput single-cell RNA sequencing combined with mutational mouse models now offers the opportunity to analyse the entire blood progenitor compartment of these perturbed mice at single cell resolution. Hence by comparing perturbed transcriptional landscapes to healthy haematopoiesis, insights about the aberrant fate decisions and differentiation trajectories occurring as a result of pre-leukaemic mutations can be obtained. However, current computational methods to achieve this are limited in their scope.

The work presented in this PhD thesis presents a range of computational methods and approaches designed to extract hypotheses from broad, unbiased scRNA-seq perturbation experiments. Furthermore, this thesis discusses a variety of novel biological insights into pre-leukaemic haematopoietic states. More specifically, the aims of this thesis can be summarised into four main goals:

1. To build a computational framework for the global integration and analysis of multiple single-cell perturbation experiments at both the cellular and molecular scales
2. To elucidate the mechanisms through which pre-leukaemic perturbations drive aberrant haematopoiesis within the most immature blood progenitor cells, and correlate these findings with the transcriptional shifts observed in more mature populations
3. To investigate the combinatorial effects of multiple mutations upon the blood progenitor landscape
4. To construct new methods for the inference of cellular differentiation trajectories from large, sparse, scRNA-seq datasets

Chapter 2

Materials and Methods

All methods were carried out by Sam Watcham unless explicitly stated in the text. Parts of this chapter describing data generation and methods have been adapted from Haltalli et. al. (2020), Dingler et. al. (2020) and Prins et. al. (2020, in review).

2.1 Isolation and profiling of murine cells using droplet-based scRNA-seq

For a given experiment, bone marrow cells were isolated from the femurs, tibiae and iliac crest of relevant mice and were lineage depleted using the EasySep Mouse Hematopoietic Progenitor Cell Enrichment Kit (STEMCELL Technologies) alongside red blood cell lysis. All mice were 12 weeks (3 months) old with the following exceptions: the WT and perturbed mice used for the p53 experiment were 9 weeks old; the WT and perturbed mice used for the second Calr experiment were 24 weeks old (6 months). All mice were either wild-type C57BL/6 mice or were bred from these as part of a perturbation model. Cells were then sorted into either the LK (Lin-c-Kit+) gate or the LSK (Lin-Sca1+c-Kit+) gate. The latter was used only in the WT reference and Calr experiments.

Droplet-based scRNA-seq was then performed using the 10X Genomics Chromium Single Cell 3' Reagent Kits v2 in all cases. Cells were processed according to the manufacturer's protocol. For each experiment, each sample (i.e. from a WT or perturbed mouse) was barcoded and subsequently sequenced alongside the other samples from the experiment. The exception to this was the WT reference experiment, in which cells from six wild-type mice were pooled first and then were sorted into three LK and three LSK samples, which were then barcoded and sequenced together. Sample demultiplexing, barcode processing and gene counting were all performed using the count command within the CellRanger pipeline. The only additional input required was the expected number of cells for each sample. This was determined experimentally. This experimental work was performed largely by Nicola Wilson, with help in specific experiments from Myriam Haltalli (Malaria), Daniel Prins and June Park (Calr), Meng Wang and Felix Dingler (Aldh2/Adh5). The CellRanger pipeline was run by Rebecca Hannah in the majority of cases and by Sam Watcham in all other cases.

2.2 Quality control and normalisation of droplet-based scRNA-seq data

The default CellRanger pipeline was applied to all wild-type and perturbed samples from either the LK or LSK (Calr experiment only) gate. The default CellRanger cell filtering step ranks the unique cell barcodes within the data by the number of UMI counts associated with each barcode, before calculating a cutoff for the minimum number of UMIs per barcode. All cells below this cutoff are excluded. The cutoff is calculated using the expected number of cells, N, which is given as input to the CellRanger pipeline. The 99th percentile of the top N barcodes is taken as an estimate, m, of the maximum number of UMIs for a cell. The cutoff is then calculated as $m/10$.

The majority of downstream analysis of the data was performed using the python module Scanpy (Wolf et al., 2018) with additional visualisation using the python modules Matplotlib and Seaborn. For initial quality control, the CellRanger output was assessed for potential doublets using code that is now freely available as the Scrublet method (Wolock et al., 2018). This simulates synthetic doublets from pairs of cells within the data, and then scores each cell based on its similarity to the synthetic profiles via a k-nearest-neighbour classifier. The distribution of these scores was analysed for each sample to ensure sensible results. For each sample, the top 1% of cells with the highest scores from each sample were removed, based on known rates of doublet occurrence within the 10X platform (Zheng et al., 2017).

Further quality control followed. Any cells with more than 10% of their reads mapping to mitochondrial genes were excluded as these were likely apoptotic. Any cells expressing fewer than 500 genes were also excluded. Any cells with a total number of UMIs that were greater than three standard deviations away from the mean for a given sample were further excluded.

Normalisation was then performed on a per-cell basis, such that every cell had the same number of UMI counts in total (10,000). All analysis downstream of this point (including differential expression) was performed on the normalised counts.

Within each experiment, highly variable genes (HVGs) were selected by concatenating all samples within an experiment and then following the procedure of Macosko et. al.(Macosko et al., 2015) implemented with Scanpy. An expression cutoff = 0.001 and dispersion = 0.05 were used in all cases. This typically resulted in around ~5000 HVGs for each experiment. No experiment had less than 4000 or more than 6500 HVGs. Count matrices were then log-transformed with an added pseudocount, i.e. $x \Rightarrow ln(x+1)$.

Log-transformed count matrices subset to HVGs were used for visualisation and clustering (see below) and for some mapping procedures (see below). Additionally for these tasks, each gene was scaled such that it was zero-centred (mean=0, SD=1), so that each gene was equally weighted.

2.3 Visualisation and Clustering of droplet-based scRNA-seq

For visualisation and clustering purposes only, a further list of 380 cell-cycle related genes were removed from the relevant dataset. This list was downloaded from Reactome (http:www.reactome.org/). This reduced the impact of cell-cycle on visualisations and allowed clustering to occur via cell-types rather than cell-cycle stage.

To visualise the WT reference data (LK+LSK) as a force-directed graph, principal component analysis (PCA) was performed on the normalised, log-transformed, scaled data subset to the HVGs. This was done using the pca function in Scanpy. From this, a k-nearest-neighbour graph with k=7 was produced for the data using the top 50 principal components and a Euclidean distance metric. This graph was used as input to the ForceAtlas2 algorithm within Gephi 0.9.1 (http://gephi.org/), which produced the resulting force-directed graph coordinates of each cell. Force-directed graph visualisations for the Malaria samples (WT, perturbed and combined) in Section 5.2 were produced in the same manner as above.

When plotting gene expression of the WT reference data on this visualisation, points were plotted in order of magnitude. Hence the cells with the greatest expression were always plotted on top.

To re-visualise the ‘Stem Cells’ cluster from the WT reference data as a UMAP, the `umap()` function within Scanpy was applied to these cells (this takes as input the PCA coordinates calculated as described above). Similarly, this method was used to visualise the entire Aldehyde dataset and the subset Aldehyde dataset (i.e. only the ‘HSC’ cluster) in Section 5.3.

All clustering was performed using the Louvain algorithm through the `louvain()` method within Scanpy (Blondel et al., 2008). For example for the WT reference data, the scaled HVGs were used to construct a k=15 nearest-neighbour graph from the PCA space as above. This graph was then used as input for the Louvain algorithm. Using a resolution=0.175 produced 13 clusters. In other cases where clustering was performed, different resolutions were used. To cluster the WT Malaria cells, a resolution=1.0 was used. To cluster the Aldehyde samples, a resolution=0.4 was used. To cluster the Aldehyde ‘HSCs’, a resolution=1.0 was used. To cluster the WT reference data into fine clusters for PAGA analysis in Section 5.4, a resolution=2.0 was used.

2.4 Mapping samples across experiments

To compare cellular abundances across pre-leukaemic models, each sample from each experiment was mapped back onto the WT reference data. This data had already been clustered as described above. To assign cells from a sample to these clusters, each cell was projected (after being scaled) into the PCA space of the WT reference data. The k=15 nearest neighbours of the cell within the reference data were located using the top 50 principal components with a Euclidean distance metric. The cell was then assigned to the same cluster that the majority of its 15 neighbours belonged to.

The same procedure was used to map the perturbed malaria samples onto the WT malaria samples in Section 5.2.

2.5 Measuring differential abundances using Z-scores/paired voting

All the pre-leukaemic models analysed in Chapter 3 (three repeats of the Jak2 model, plus one experiment for each of the W41, Dnmt3a, Npm1, Crebbp, Tet2 HET, Tet2 HOM, Jak/Tet Cross and p53 models) were mapped to the WT reference data as described above. The mean proportional cellular abundance for each of the 13 clusters was calculated across all

WT samples (excluding the p53 WT sample, which was not age-matched with the other WT samples), along with its standard deviation. The number of cells in each sample were used as weights for these calculations. For experiments where two samples from separate mice were sequenced for a given genotype (e.g. the Dnmt3a, Npm1 and Crebbp experiments had two WT and two perturbed mice), each genotype was considered as a single sample for the purposes of this and later sections. It was later statistically justified that the WT sample from the second Jak2 experiment and the WT sample from the Tet2 HET experiment were significantly different from the other WT samples. Therefore the mean and standard deviation across WT samples were recalculated with these samples removed. Z-scores (number of standard deviations from the mean) for the abundances of each cluster were then calculated for every sample (WT and perturbed) compared to these average values. Heatmaps displaying these values were created using the Python module Seaborn.

To test whether the set of 13 Z-scores obtained for a sample in Section 3.2 was significantly different from the WT average, the Z-scores were converted to p-values using the python module Scipy with the `stats.distributions.norm.isf()` method. These 13 p-values were then combined using Fischer's method, implemented in the `stats.combine_pvalues()` method. This analysis assumes that the 13 initial Z-scores for a sample are independent; this is not strictly true, since the total abundances must sum to 100%. Hence the computed test-statistic will be an underestimate of the true value. However with 13 clusters, the co-variance between any two clusters will be low, and Fischer's method will provide a good approximation.

A proxy for differential abundances was visualised on the WT reference force-directed graph/UMAP using paired voting. Two matched samples (i.e. WT and perturbed samples from the same experiment) were chosen. Each cell from the first sample was projected into the PCA space of the reference dataset as described above. Each of its k nearest neighbours in the reference data was then given a single vote. k was chosen so that the total number of votes given out by any sample was as close to 100,000 as possible. Hence k differed between samples in order that the total number of votes from any sample was the same. After this was done for each cell, the votes were smoothed over the reference dataset by sharing the number of votes each reference cell had equally amongst its k=100 nearest neighbours (for the entire WT reference dataset) or k=20 nearest neighbours (for just the 'Stem Cells' cluster). This procedure was repeated for the second (matched) sample. For each reference cell, the difference in the number of votes received from the first and second samples was visualised. Cells were plotted in order of their absolute magnitude, so that areas of both increased and decreased abundance could be observed.

Three experiments (W41, Tet2 HOM, and Jak/Tet Cross) did not have a matched WT sample alongside the perturbed sample. For these experiments, the Npm1 WT sample was used as the matched sample. This sample was chosen because its cellular abundances across all clusters were the closest to the average values out of all the WT samples. To compare the results of the paired voting method with the Z-score method, the voting differences for each experiment were averaged across each of the 13 clusters and displayed as a heatmap.

2.6 Performing differential expression

Differential expression (DE) between two sets of cells was performed using the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) using the SciPy method `stats.mannwhitneyu()`. For a given gene, this tests the likelihood that the expression values for each set of cells came from the same distribution and returns the result as a p-value. Normalised counts were used as input for all DE tests as suggested in reviews of single-cell DE methods (Soneson and Robinson, 2018). After performing this test for each gene, the returned p-values were corrected for multiple testing using the Benjamini-Hochberg correction. Genes were then called as significant differentially expressed genes (DEGs) if they had an adjusted p-value<0.05 and an absolute fold-change>1.2. Results were assessed visually using MA plots (fold-change vs mean expression) to check for spurious patterns.

In one specific experiment (the third repeat of the Jak2 model), differing sequencing saturations across genotypes introduced technical biases to the DE process and caused spurious results (see Figure 3.5B). If a shallowly-sequenced sample A is compared to a deeply-sequenced sample B, then a given gene in sample A may have many zero counts (due to the shallow sequencing) and a few relatively high counts (because each cell in sample A has few UMIs overall and hence if even one UMI is recorded, this is a relatively large proportion of the total for that cell). The same gene in sample B may have few zero counts (due to the deep sequencing) but also relatively small counts (because each cell has many UMIs overall). Hence the DE testing identifies this gene as coming from different distributions in the two samples, whereas in reality this is a technical effect rather than a biological one.

To correct for this, the raw counts for each cell in the homozygous sample of the third Jak2 experiment were randomly downsampled. For each cell, a random cell size (total UMIs) was chosen from the distribution of wild-type cell sizes in the cluster that the homozygous cell belonged to. The homozygous cell was then downsampled to this cell size. This had to be done on a per-cluster basis because different clusters had very different distributions of cell sizes (even in a wild-type setting). This procedure was effective at removing the spurious technical effects observed.

In Section 3.4, the experiments without a matched WT sample (W41, Tet2 HOM, and Jak/Tet Cross) had differential expression performed with both the Dnmt3a and Npm1 samples, as these were the WT samples that had closest-to-average cellular abundances. For each experiment these two lists were then intersected, with the smallest adjusted p-value and mean log_2(fold-change) being used for downstream molecular analysis.

2.7 Geneset enrichment analysis/IPA analysis

Where mentioned in Chapters 3 and 4, geneset enrichment analysis/gene ontology of differentially expressed or differentially variable genes was performed using the online tool EnrichR (Chen et al., 2013). Only terms with an adjusted p-value<0.05 were considered significant. The Ingenuity Pathway Analysis from Section 5.4 was performed by Michele Vacca using the QIAGEN IPA framework (https://digitalinsights.qiagen.com/).

2.8 Integrating differential expression results within a specific cell type

332 individual genes were identified as being significantly differentially expressed (either up or down) within the 'Erythroid – Middle/Late' cluster in at least two of the perturbation models analysed in Chapter 3.4. Note that these models must be distinct; a gene is not included if it is dysregulated in two or more of the Jak2 experimental repeats but not it any other models. The Tet2 HET and Tet2 HOM models were treated as distinct.

The log_2(fold-changes) for each experiment in each of the 332 genes was condensed into a matrix. A UMAP visualisation of this data was created with the second and third Jak2 experimental repeats excluded so that the UMAP would not be unduly influenced by the Jak2 model. Four genes had an 'infinite' fold-change in one of the models due to a sample within that model having zero expression of the gene. These infinite values were replaced with a log_2(fold-change) of 3 for positive infinities and -3 for negative infinities to allow the UMAP to be created.

In an identical manner, a UMAP was created in section 4.3 using 231 genes that were identified as being dysregulated in two or more models within the 'Stem Cells' cluster.

2.9 Integrating differential expression results across a whole trajectory

To integrate differential expression results across an entire trajectory, the trajectory of interest was first chosen through partition-based graph abstraction (PAGA, for full details see below). The shortest path between the 'Stem Cells' cluster and the most mature cluster in the trajectory (chosen manually) was calculated based on a connectivity threshold of 0.3, and cells belonging to all clusters on this shortest path were taken to belong to the trajectory of interest.

For each experiment, the normalised and logged data for the matched WT and perturbed samples was concatenated. Separately, the hscScore method (see below) was applied to the WT sample, and the cell with the highest hscScore was identified as the root cell for the experiment (i.e. the cell that is transcriptionally most similar to the transcriptome of functionally validated HSCs). Pseudotime was then performed on the concatenated samples using the `dpt()` method within Scanpy and the identified root cell. This method implements the diffusion-based pseudotime approach taken by Bendall et. al. (Bendall et al., 2014), giving each cell in the concatenated dataset a score between 0 (root cell) and 1. The pseudotime analysis of the concatenated dataset was checked visually via dimensionality reduction to ensure that a) no clear batch effects were visible between genotypes and b) the pseudotime results were sensible. In some experiments, a small number of outlier cells that were not removed by quality control caused problems, as they had wildly different transcriptomic profiles to anything else and therefore were given the large pseudotime scores whilst all other cells were given very small pseudotime scores. In these cases, the offending cells were removed and the analysis repeated.

A sliding window approach was then implemented to calculate the expression changes of the WT and perturbed cells over pseudotime for a gene of interest. This window had a size of 500 cells and was shifted by 100 cells at each step. Within the window, the mean expression of each genotype was calculated at each step, with the requirement that at least 100 cells from both genotypes were present within the window. This requirement helped to reduce edge effects near the end of a trajectory. The mean pseudotime value of the cells within the window was also recorded at each step.

The difference between the wild-type and perturbed expression levels was then plotted as a function of pseudotime for the gene of interest. For the Jak2 model, which had three experimental repeats, the results for a given gene were averaged over the three experiments.

For models with only a single experiment but two mice per genotype (W41, Dnmt3a, Npm1, Crebbp, Tet2 HOM and Jak/Tet Cross), the pseudotime alignment and sliding window process was repeated independently by first using cells from one WT mouse and one perturbed mouse, and then using cells from the other WT mouse and the other perturbed mouse. The results were then averaged over the two biological repeats. Models without a matched WT used the Npm1 WT mice as the matched sample.

A 4th-order polynomial spline was then fitted to the results for each model to smooth the visualisation and further reduce the impact of edge effects. This was done in SciPy using the `interpolate.univariatespline()` method. This spline acted as a 'pseudotime-series' for the gene along the trajectory of interest.

2.10 Unsupervised dDEG method

An unsupervised approach for the identification of dynamically differentially expressed genes (dDEGs) was extended from the pseudotime approach described above. As input, this method requires only the two matched samples to be compared, and the identity of the starting and ending clusters for the trajectory of interest. It will then automatically identify the cells belonging to the trajectory through a PAGA analysis, identify the root cell and perform pseudotime alignment. The sliding window approach is then performed for every gene in the dataset, followed by the fitting of a 4th-order polynomial spline to the results. To score each gene, the mean of the absolute values of the differences between the WT and perturbed expression along pseudotime was recorded. Simultaneously, the mean of the absolute values of the derivative of the 4th-order polynomial spline along pseudotime was calculated. If the mean difference was greater than 0.2, or the mean derivative was greater than 0.5, the gene passed the threshold for being called a dDEG and was carried forward for further analysis. These thresholds were designed to be relatively strict. The difference in expression was used as a metric rather than the fold-change in expression in order to penalise lowly-expressed genes. The absolute values were used so as not to penalise genes that switched between positive and negative regulation across pseudotime.

Within this framework, experimental/technical repeats are handled by requiring a gene to pass the thresholds in all repeats to be considered a dDEG.

The pseudotime-series of the dDEGs can then be clustered via K-means clustering. Initially, each gene's pseudotime series was scaled so that its maximum value was 1 and its minimum value was -1. A user-inputted number of clusters was required and initial cluster centroids were selected at random. The distance between a gene's pseudotime-series and each centroid was then calculated using dynamic time warping (DTW), which stretches and squeezes the pseudotime-series such that they can be clustered by shape (Berndt and Clifford, 1994). Genes were assigned to their nearest centroid. Centroids were then recalculated based on these assignments and this process was iterated 30 times before returning the cluster assignments of each gene and the centroid of each cluster. Empirically it was found that the returned centroids did not differ greatly between different initialisations as long as the number of iterations was greater than 10.

Centroids were then inspected visually and their components assessed for biological tractability. The number of clusters found to be best for the K-means/DTW algorithm was around 8-10, with potentially less clusters required if only a small number of dDEGs were identified.

2.11 Visualising differential expression across a transcriptomic landscape

Using a similar approach to the paired voting for differential abundances method (Section 2.5), a proxy for differential expression was visualised on the WT reference force-directed graph using a kNN classifier. Two matched samples (i.e. WT and perturbed samples from the same experiment) were chosen. Cells from both samples were projected into the PCA space of the reference dataset. For each cell in the WT reference data, the mean expression of its k=50 nearest neighbours in each of the matched samples was recorded. For each reference cell, the difference in the mean expression of its neighbours from the first and second samples was visualised. Cells were plotted in order of their absolute magnitude, so that areas of both increased and decreased expression could be observed. The log_2(fold-change) of the expression between the two matched samples could also be visualised.

2.12 Processing of Smart-Seq2 scRNA-seq data

Data from 1,654 mouse HSPCs sequenced using the Smart-Seq2 platform was published in Nestorowa et. al. (Nestorowa et al., 2016). Cells were collected from ten 12-week old C57BL/6 WT mice, lineage depleted and then sorted whilst capturing index measurements for c-Kit, Sca1, CD16/32, CD34, Flt3, CD150 and CD48 and EPCR. The Smart-Seq2 protocol was performed as described previously (Picelli et al., 2014) using an Illumina HiSeq 4000 system before reads were aligned using GSNAP (Wu and Nacu, 2010). Quality control removed cells with <200,000 reads mapping to nuclear genes, <4,000 genes detected or >10% of reads mapping to mitochondrial genes. Normalisation was then performed using the scran R package (Lun et al., 2016). This led to 1,654 cells with 40,587 detected genes being taken forward for downstream analysis.

2.13 Mapping Smart-Seq2 cells onto droplet-based scRNA-seq landscapes

To project each of the index-data defined cell populations from the Smart-Seq2 data (MPP1, ST-HSC etc) onto the WT reference droplet-based landscape, the normalised count matrices of the 1,654 Smart-Seq2 cells were concatenated with the 44,802 10X cells. Note that these count matrices had not been log-transformed or scaled and highly variable genes (HVGs) had not been chosen. 17,007 genes were present in both datasets. After concatenation, each cell was normalised using Scanpy to have the same number of counts (effectively, this step scaled the Smart-Seq2 cells to match the 10X cells). HVGs were calculated as for the droplet-based datasets. This retained 3,289 HVGs. All cells were then log-transformed and all genes were scaled (mean=0, SD=1).

Principal component analysis was then performed on the combined dataset. Batch correction was then performed using MNN-Correct (Haghverdi et al., 2018) as implemented in python. This method searches for mutual nearest-neighbours between batches (i.e. pairs of cells in different batches that are nearest-neighbours of each other) based on cosine distances between cells. Correction vectors are calculated between each MNN pair and cell-specific correction vectors are then calculated as a weighted average of the pair-specific vectors, ensuring that cells close together in gene-expression space are batch corrected by similar vectors. Hence this method can deal with non-constant batch effects within the high-dimensional gene expression space.

This batch-corrected PCA space containing both sets of cells was carried forward. For each index-data defined population in the Smart-Seq2 data, cells belonging to that population were subset and their k=15 nearest-neighbours in the 10X data were identified and given a vote. A score for each population was then defined on the 10X reference data as the number of votes for each reference cell divided by the total number of Smart-Seq2 cells in that population. This normalised the scores to account for the fact that some cell populations were more or less abundant within the Smart-Seq2 data. These scores were then smoothed over the reference data by sharing each cell's score for each population equally amongst its k=100 nearest neighbours.

A set of nine mutually exclusive and collectively exhaustive index populations were considered (CMP, MEP, GMP, LMPP, MPP1, MPP2, MPP3, ST-HSC and LT-HSC), and each WT reference cell was annotated as belonging to the population with the highest score in that cell.

2.14 Calculating hscScores/G2M scores/Module scores

hscScores were calculated using the hscScore algorithm published by Hamey et. al. (Hamey and Gottgens, 2019). This method uses a multi-layer perceptron neural net trained on data from Wilson et. al. (Wilson et al., 2015), where 92 HSC transcriptomes were assigned a score based on their surface expression of 11 index markers. These markers were in turn measured for HSCs that were placed in single-cell transplantation assays and their functional output measured. Hence, the score for each of the 92 cells was correlated with their functional potential in a transplantation setting. By training a model on the transcriptomes of these cells, new single-cell transcriptomes could be inputted into the model to retrieve their hscScore, which in turn is also a measure of their functional ability.

The hscScore algorithm uses as input the normalised counts matrix for a sample, and returns a score between zero and one for each cell. It also requires that all the genes used to train the model are present in the sample.

For the comparison of hscScores across leukaemic perturbations in Section 4.4, hscScores were calculated independently for each model. The WT scores were scaled so the highest-scoring WT cell was given a hscScore of one and the lowest-scoring WT cell remained the same. This scaling factor applied to the WT cells was then applied to each of the leukaemic models, so that scores would be comparable between models with a score of one indicating the highest-scoring WT cell. The only model to contain any cells scoring greater than one through this procedure was the Crebbp model, and this only occurred for two cells.

For the comparison of WT and perturbed hscScores in Sections 5.2 and 5.3, the same procedure was applied; the WT scores were interpolated between the minimum value returned by the algorithm and one, such that the highest-scoring WT cell had a hscScore of one. The same scaling factor was then applied to the perturbed sample(s) in each case.

G2M cell-cycle scores were calculated using the expression values of a set of 198 Hallmark G2/M checkpoint genes downloaded from the Molecular Signatures Database

(Liberzon et al., 2016). For each cell within a sample, the score was given by

$$\sum_{g} \frac{ln(x_g + 1)}{n} \tag{2.1}$$

with n equal to the number of G2/M checkpoint genes that passed the initial gene filtering QC step, and x_g the normalised expression of each gene. n was identical for each cell in an individual sample.

The module scores in Section 5.3 were calculated in an identical manner using specific gene lists. For the DNA Repair score, the gene ontology (GO) term GO:006281 – 'DNA Repair' was downloaded from http://informatics.jax.org/. For the p53 Targets score, a list of verified p53 targets published in (Tanikawa et al., 2017) was used. For the Intrinsic Apoptotic Signalling Pathway score, the GO term GO:0097193 – 'Intrinsic Apoptotic Signalling Pathway' was used.

2.15 Assigning cell-cycle stages

Cell-cycle stage assignment in Section 5.3 was performed following the method established in Tirosh et. al. (Tirosh et al., 2016) for scoring cycling cells and implemented within the Scanpy method `score_cell_cycle()`. Lists of 43 genes associated with S-phase and 55 genes associated with G2/M phases from Tirosh et. al. were used to quantify the relative expression of these cell-cycle stages compared with a randomly chosen set of reference genes. Cells with high relative expression levels of either program were assigned to be in S-phase or G2/M phase respectively, whilst cells with no clear expression of either program were assigned to the G1 phase. No cells expressed relatively high levels of both S and G2/M phase programs.

2.16 Resolving cell-cycle shifts across whole trajectories

To assess the changes in the proportional abundance of different cell-cycle states across the erythroid differentiation trajectory, the sample of interest was subset to only the cells belonging to the four clusters annotated as erythroid. Density plots of the G2M score within these erythroid cells were created using the `kdeplot()` method within the Seaborn python module.

A pseudotime analysis was then performed using Scanpy's `dpt()` method to calculate an 'erythroid pseudotime' for the sample. The root cell for this analysis was chosen to be the erythroid cell with the lowest pseudotime value when the same pseudotime analysis was performed on the entire sample prior to subsetting (e.g. for use in the dDEG pipeline). Therefore the root cell represented the most immature erythroid cell within the sample and was given a pseudotime score of 0. To visualise shifts along the erythroid trajectory in Figure 4.14, a sliding window of size=300 cells was shifted by 60 cells at each step in order of increasing pseudotime. The G2M scores of each cell within the window were recorded at each step, alongside the mean pseudotime score of the cells within the window. These G2M scores were then sorted in ascending order for each step and plotted as a grid of size (300 x numbers of steps), coloured by each cell's G2M score. The width of each column was scaled to be proportional to the difference between its mean pseudotime and that of the next column, such that the same x-coordinate on each plot represents the same pseudotime value.

To compare the observed shifts across samples in Figure 4.15, the percentage of cells in a G2M-high state (G2M score > 0.5) within each window for each sample was calculated. This percentage was then plotted as a function of pseudotime for each sample. Results for the WT samples (including the WT reference, and all WT samples from the pre-leukaemic perturbation models excluding the p53 WT) were averaged without using sample weights, and their standard deviation calculated.

2.17 Performing differential variability testing (VarID)

Single-cell variability calculations were based on the VarID method introduced in Grun et. al. (Grün, 2020). The core of the VarID approach was rewritten in Python (the original package was coded in R) with alterations to allow the calculated variabilities to be compared across experiments. In addition, the speed and scalability of the method was greatly increased. The altered version of the method is described here, with specific mention of the steps that were altered from the original publication.

To begin, the normalised expression counts of a sample (e.g. the first Jak2 WT sample) were filtered gene-wise, such that only genes expressed in greater than n=5 cells were retained. The remaining genes were then used to model the gene-wise mean-variance relationship as a quadratic polynomial in log-space. This returned an empirical relationship between the mean expression of a gene and its variance that was unique to each sample, but in practice differed very little between different samples or experiments. A k=16 nearest-neighbour graph was then obtained for the sample from the top 50 principal components, calculated using scaled data that had been subset to highly variable genes. For each cell in the sample, these nearest-neighbours will be used to calculate variabilities for all genes. In the original method, k was calculated independently for each gene in each cell, by requiring (for a given gene) that a cell's neighbours had expression values which fit a negative binomial model described by the neighbours' mean expression. It was found that this method led to highly variable values for k both within and across samples that led to severe batch effects in the calculated final variabilities.

For each cell, the mean expression and variance of all genes across their k nearest-neighbours (including the cell itself) was calculated. These local variances were then scaled for each gene by a factor calculated through the empirical mean-variance relationship described above. The scaled variances for each cell were then concatenated into a matrix of single-cell variabilities for the entire sample. Overall, this step allowed the variability values for different genes to be sensibly compared within the same sample.

Quantile normalisation was then performed on the sample following the method of Bolstad et. al. (Bolstad et al., 2003) to standardise the distribution of variabilities within each cell. Overall, applying this method of normalisation to all samples was found to stabilise the results of differential variability testing downstream, and allowed cross-sample and cross-experiment comparisons of the single-cell variabilities to be made. This step was not performed in the original method, as the author did not test any across-sample comparisons of variability.

Differential variability (DV) testing between two sets of cells was then performed using the Wilcoxon rank-sum test, as for standard differential expression. For a given gene, this tests the likelihood that the variability values for each set of cells came from the same distribution and returns the result as a p-value. Quantile normalised variabilities were used as input for all DV tests. After performing this test for each gene, the returned p-values were corrected for multiple testing using the Benjamini-Hochberg correction. Genes were then called as significant differentially variable genes (DVGs) if they had an adjusted p-value<0.05 and an absolute fold-change>1.2. Results were assessed visually using MA plots (fold-change vs mean expression) to check for spurious patterns. For two experiments – the third Jak2 experimental repeat and the W41 experiment – no sensible variability results could be calculated. The third Jak2 experiment observed extremely different cell numbers and sequencing saturation between the WT and perturbed samples. The W41 experiment (which did not have a matched WT sample) also observed a lower sequencing saturation than any of the WT samples that were unsuccessfully matched with it.

In Section 4.4, the experiments without a matched WT sample (Tet2 HOM and Jak/Tet Cross) had differential variability performed with both the Dnmt3a and Npm1 samples, as these were the WT samples that had closest-to-average cellular abundances. For each experiment these two lists were then intersected, with the smallest adjusted p-value and average log_2(fold-change) being used for downstream analysis.

2.18 Identifying correlated genes driving malarial response

In section 5.2, genes that were at least 4-fold upregulated in the infected malaria cells compared to the WT cells in all six clusters were chosen as the initial set of driving genes involved in the infection. This list contained 20 genes. The entire set of filtered genes (i.e. those that passed initial QC) calculated using the WT and infected cells combined was analysed, and all genes whose normalised expression correlated highly with any of the initial set of 20 genes were identified. Genes that had a spearman correlation >0.3 with any of the initial 20 genes were then added to the list of driving genes. This procedure added 89 genes for a total of 109 driving genes.

2.19 Partition-based Graph Abstraction (PAGA) Analyses

The partition-based graph abstraction (PAGA) method was introduced by Wolf et. al. (Wolf et al., 2019). Given a single-cell dataset and an associated partitioning of the dataset (e.g. into Louvain clusters), the method calculates a connectivity between each pair of clusters, and from these creates a graph with the clusters as nodes, based on a threshold for connectivity. Initially a k=15 nearest-neighbour graph is calculated for the data based on the Euclidean distance amongst the top 20 principal components. The connectivity is then calculated through

$$c_{ij} = \frac{1}{2}(e_{ij} + e_{ji})/\sqrt{k^2 n_i n_j} \tag{2.2}$$

where e_{ij} is the number of edges in the nearest neighbour graph originating in cluster i and leading to cluster j. k is the parameter associated with the nearest-neighbour graph, and n_i and n_j are the total number of cells in clusters i and j respectively. Hence c_{ij} measures the degree to which clusters i and j are interconnected.

To calculate trajectories for the dDEG method in Chapter 3, a connectivity threshold of 0.3 was applied to 13 clusters of the WT reference dataset. The shortest path from the ‘Stem Cells’ cluster to the end cluster of interest was calculated using the filtered edgelist (i.e. with all connections <0.3 removed) and the `get_shortest_paths()` method within the igraph python module. To calculate the fine-resolution abstracted graph in Section 5.4, a connectivity threshold of 0.3 was applied to the 60 fine Louvain clusters from the WT reference dataset. The filtered edgelist was used to calculate the positions of each node for visualisation using the ForceAtlas2 algorithm in Gephi as for the standard force-directed graphs.

The expression of specific genes were visualised on the abstracted graph in Section 5.4 by calculating the mean of the normalised expression values within each node. Different scalebars were used for each gene. The differential abundances were visualised by calculating the mean voting difference for all cells within each node.

2.20 TITANS: trajectory inference through iterative ancestral search

2.20.1 Preprocessing steps

Starting from a scRNAseq dataset of size $m \times n$ containing m samples (cells) and n features (genes), a diffusion map representation of the data was calculated following the method of Haghverdi et. al. (Haghverdi et al., 2016). The resulting representation contained d diffusion components, reducing the dataset to size $m \times d$. Each row of size $1 \times d$ now represents a vector of length d that stores the location of each cell in the diffusion-map space. The value of d needs to be at least greater than the maximum number of trajectories expected within the dataset, and in general should be >10 to efficiently capture the manifold in which the transcriptional landscape of the dataset resides. For all the results presented in Chapter 6, d=15.

Following the production of the diffusion components, several of TITANS preprocessing functions are used to prepare the dataset for trajectory inference. The only initial information that is required is either A) a list of genes known to associate with the most primitive cells in the dataset or B) the identity of a group of cells believed to represent the most primitive state in the dataset. If A) is provided, each cell in the dataset is scored based on its cumulative scaled expression of the genes provided, and the highest scoring 0.1% of cells are used to identify the 'root space' by calculating the average position of these cells within the diffusion-map representation of the data. If B) is provided, these cells are used instead.

In either case, the location of the root space in the diffusion map space is now known. Next, the location of every cell in the diffusion map space is translated (shifted) such that the root space is now located at the origin (the location of the root space is now a vector containing d zeros). This shifted diffusion-map representation still has size $m \times d$. Each row of size $1 \times d$ now represents a vector of length d that stores the location of each cell with respect to the location of the root space in the diffusion-map representation. A pseudotime value between 0-1 inclusive is now calculated for each cell in the dataset using the cell closest to the root space as the initial cell following the method of Bendall et. al. (Bendall et al., 2014).

Finally, TITANS identifies the terminal cell states ('endpoints') in the dataset by looking for cells which have a pseudotime value greater than any of its k nearest neighbours, where $k = m \times 0.01$. To account for the effect of cell intrinsic programs that may not be related to differentiation (such as cell cycle), pseudotime values were smoothed across the 25 nearest neighbours of each cell prior to identification of endpoints. Smoothed pseudotime values are not used at any other point in the algorithm. The identified endpoints can be sanity checked by visualisation on a low-dimensional representation of the data (such as tSNE or UMAP) to check that they coincide with the expected location of trajectory endpoints (in datasets where this is known *a priori*).

2.20.2 Main algorithm

Each cell is now represented by a vector of length d that stores its position with regards to the root space. For cell i denote this 'cell vector' by $\mathbf{x}_i$. Each cell also has an associated pseudotime value, for cell i denote this by dpt_i. Starting from each identified endpoint with cell vector $\mathbf{x}_i$, an iterative process containing three main steps is performed:

1. Calculate the cosine of the angle θ_{ij} between the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ for all cells with $j \neq i$ using the equation
$$cos\theta_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{|\mathbf{x}_i|\,|\mathbf{x}_j|} \tag{2.3}$$

2. Add to the trajectory all the cells $\mathbf{x}_j$ not already in the trajectory which satisfy the conditions:
$$cos\theta_{ij} > 1 - \varepsilon \tag{2.4}$$
and
$$dpt_j < dpt_i \tag{2.5}$$
with ε representing a user defined 'threshold parameter' in the range
$$\varepsilon \in [0, 0.01] \tag{2.6}$$

3. If the number of cells added in step 2 is greater than zero, start a new iteration in which steps 1 and 2 are repeated for each cell added in the current iteration. If the number of

cells added in an iteration is zero, check whether the minimum pseudotime value of all the cells currently in the trajectory is <0.01. If it is, assume that the trajectory is near the root space and stop trajectory inference. If it is not, increase ε by a set amount (0.002) and repeat the current iteration, whilst displaying a warning message to the user.

This process is then repeated for each identified endpoint. Importantly, the trajectories identified from each endpoint are independent; each cell can belong to more than one trajectory. Equally, a cell need not belong to any of the identified trajectories.

2.20.3 Data

Mouse haematopoietic data was taken from Dahlin et. al. (Dahlin et al., 2018). This is the same dataset used for the WT reference used in Chapters 3, 4 and 5. After generation of the count matrices using CellRanger, the three LK (Lin-c-Kit+) samples from six pooled WT mice were extracted and run through quality control and preprocessing steps independent of the LSK samples. This returned 21,836 transcriptomes expressing a median of 2,842 genes. Data was normalised, log-transformed, subset to HVGs and scaled before being used in the TITANS algorithm. The total number of HVGs used was 5,004. The force-directed graph representation of the data was calculated in Gephi using a k=7 nearest-neighbour graph, itself calculated using a Euclidean distance metric on the top 50 principal components.

TITANS was run on this data using the list of MolO genes from Wilson et. al. as input (Wilson et al., 2015). A threshold parameter of $\varepsilon = 0.002$ was used for all trajectories.

Mouse gastrulation data was taken from Pijuan-Sala et. al. (Pijuan-Sala et al., 2019). Normalised log-counts were downloaded for 116,312 transcriptomes. 5,876 HVGs were selected before being scaled. Cell metadata was downloaded to provide cluster annotation. A UMAP representation of the data was constructed using the scaled, subset data by running the `umap()` function within Scanpy. The genes associated with epiblast cells and used to define the root space were *Pou5f1*, *Utf1* and *Dnmt3b*.

TITANS was run on this data with these genes as input. A threshold parameter of $\varepsilon = 0.0005$ was used for all trajectories.

2.20.4 Other methods

FateID was run in the R programming language using the *FateID* R library (Herman et al., 2018). The two free parameters of the method, *m* (test set size at each iteration) and *h* (number of cells from each fate used for training at each iteration). The values used in the original publication of $m = 5$, $h = 20$ were tried, but this meant the algorithm took longer than 24 hours on the mouse haematopoietic data. The results presented in Chapter 6 were produced using $m = 30$, $h = 50$.

Population Balance Analysis (PBA) was run using the python scripts provided with the original publication (Weinreb et al., 2018b). Cells corresponding to the most mature cell for each of the seven clear lineages within the data (erythroid, MK, mast, basophil, neutrophil, monocyte and lymphoid) were identified using the force-directed graph representation of the haematopoietic data. The vector of cell sink rates from each fate was empirically fine-tuned to give the most sensible looking results. The final vector used to create the results presented in Chapter 6 was (-.12, -.10, -.09, -.09, -.11, -.12, -.12). These numbers are in the same order as the list of trajectories above. For each cell, the fate with the highest probability was chosen as its potential fate for visualisation.

Slingshot was run using the *slingshot* R library (Street et al., 2018). A Louvain clustering of the LK haematopoietic data was created using the Scanpy method `louvain()` on the scaled expression data with resolution=1.0. Slingshot was then run with default parameters on a 15-dimensional diffusion-map representation of the data. This representation was identical to that used to run TITANS on the same dataset.

Chapter 3

Integrated Global Analysis of Pre-Leukaemic Murine Haematopoiesis Models

Experimental work for this project was carried out by Nicola Wilson (isolation of primary bone marrow cells, scRNA-seq profiling). Initial preprocessing of the resulting data was carried out by Rebecca Hannah (running 10X Genomics CellRanger pipeline). After production of the raw count matrices from the CellRanger pipeline, all computational work was carried out by Sam Watcham, except where explicitly stated in the text.

3.1 Background

The advent of single cell RNA sequencing has led to an explosion of studies and datasets that have provided comprehensive new insights into native haematopoiesis (Watcham et al., 2019). The general trend has been towards larger and less biased datasets that rely only on the broadest of sorting gates, and which provide a systems-scale view of the blood system as a whole (Dahlin et al., 2018; Tusi et al., 2018). The ability of single cell RNA profiling to resolve the continuous and heterogeneous nature of haematopoietic differentiation has allowed the relationship between different blood lineages and their associated fate decisions to be dissected in great detail (Laurenti and Göttgens, 2018). However relatively less work has been performed to look at the effects of haematopoietic perturbations using a systems-scale approach, and in particular to integrate the effects of many different perturbations into a single framework that allows them to be sensibly and systematically compared.

There are a number of perturbations that are of direct clinical relevance to understanding the onset of haematopoietic malignancies - including clonal haematopoiesis, myeloproliferative neoplasms and different types of leukaemia - in human patients (Grinfeld et al., 2017; Sperling et al., 2017). Many of these are loss-of-function mutations which negate the production of a specific enzyme or protein relevant for native haematopoiesis. Others include mutations responsible for inducing aberrant intra- or inter-cell signalling pathways, leading to anomalous differentiation or proliferative behaviour. The perturbations of the greatest interest are those that confer some form of selective advantage to blood cells with long-term self renewal abilities, such as HSCs, and can therefore cause lasting effects to the haematopoietic system (Chen et al., 2015). Analysis of data from patients with haematopoietic malignancies has revealed the most common perturbations in many diseases, and their subsequent effect on prognosis (Taylor et al., 2017). Many of these perturbations do not immediately lead to the onset of leukaemia by themselves, but are known to lead to less serious haematological malignancies and predispose to leukaemia over a period of many years (Sperling et al., 2017). Hence a thorough understanding of their effects and the mechanisms by which these effects are mediated is essential in trying to treat these conditions quicker and more effectively in the clinic.

Until recently, mouse models of these pre-leukaemic perturbations have been studied with regards to specific blood cell types, such as phenotypic HSCs, to assess their functional behaviour both *in vitro* and *in vivo* using transplantation assays (Ostrander et al., 2020; Shepherd et al., 2018). In the past couple of years, single cell profiling has allowed the blood system of these pre-leukaemic mice to be analysed at a far broader scale (Izzo et al., 2020). The primary goals of these analyses are typically to assess the effects of a single perturbation at the cellular level (abundances of specific cell types, emergence of new cell types) and the molecular level (behaviour of specific genes in terms of RNA expression, chromatin accessibility etc., depending on the modality of profiling). Secondary goals include linking these effects in the context of aberrant differentiation trajectories or fate decisions. Loftier goals may include trying to identify the specific cell types driving the perturbation and how a mutation in a specific enzyme or protein is causing changes in these perturbation-initiating cells. Within the bounds of these goals there is ample opportunity for novel hypotheses to be uncovered, given the broad and unsupervised nature of the data and the current lack of understanding about how these perturbations act at a systems-scale.

The focus of this work was to analyse scRNA-seq datasets from several pre-leukaemic perturbations in an integrated manner, with the aim of producing novel biological insights that would not have been possible by analysing each dataset separately. This chapter describes the integration of the data at both cellular and molecular levels, and showcases a range of computational tools for the exploration and visualisation of single cell perturbation data. In addition, this chapter describes novel biological insights about the molecular drivers of the perturbations, their relationship to observed differentiation skews, and the interactions between combinations of perturbations at the scale of the entire blood system.

3.2 A large wild-type dataset as a reference for pre-leukaemic perturbation models

Prior to the generation of the perturbation datasets discussed below, a large wild-type dataset containing almost 45,000 transcriptional profiles from six pooled female C57BL/6 mice was generated and published in 2018 (Dahlin et al., 2018). The droplet-based 10X Genomics pipeline was used to capture mouse bone marrow HSPCs from both the Lin- c-Kit+ (LK) gate and the Lin- c-Kit+ Sca1+ (LSK) gate (Zheng et al., 2017). The LK sorting gate contains the majority of haematopoietic progenitor populations, whilst the LSK gate contains only the more immature populations such as HSCs and MPPs (Figure 3.1B). Since these immature cells are found at very low frequencies in the bone marrow, the inclusion of LSK cells ensured a high density of profiling across the most immature blood progenitor states. After quality control (see methods), a total of 21,809 LK cells with an average of 2,800 genes and 22,993 LSK cells with an average of 2,500 genes were detected for a total of 44,802 expression profiles.

Starting from this expression data, the blood progenitor transcriptomic landscape was visualised using a force-directed graph calculated from the k-nearest neighbour matrix of expression profiles (Weinreb et al., 2018a). Cells with similar expression profiles are necessarily closely positioned in the two-dimensional embedding, whilst cells from separate lineages are likely to be pulled apart (Figure 3.1A). This visualisation method has previously been shown to successfully capture complex differentiation structures from single cell expression data (Weinreb et al., 2018a). To highlight this, the expression of known lineage marker genes were visualised on the force-directed graph, demonstrating that cells of similar lineage are highly localised within the embedding (Figure 3.1C). Specifically, *Procr* (EPCR) expression marks HSCs at the top of the blood hierarchy (Balazs et al., 2006), with *Pf4* and *Klf1* marking megakaryocyte and erythroid progenitor cells respectively (Dzierzak and Philipsen, 2013; Rowley et al., 2011). In addition, *Gzmb* and *Mcpt8* mark mast cell and basophil progenitors (Dwyer et al., 2016; Ugajin et al., 2009), whilst *Elane* and *Ms4a6c* mark the closely related neutrophil and monocyte lineages respectively (Olsson et al., 2016).

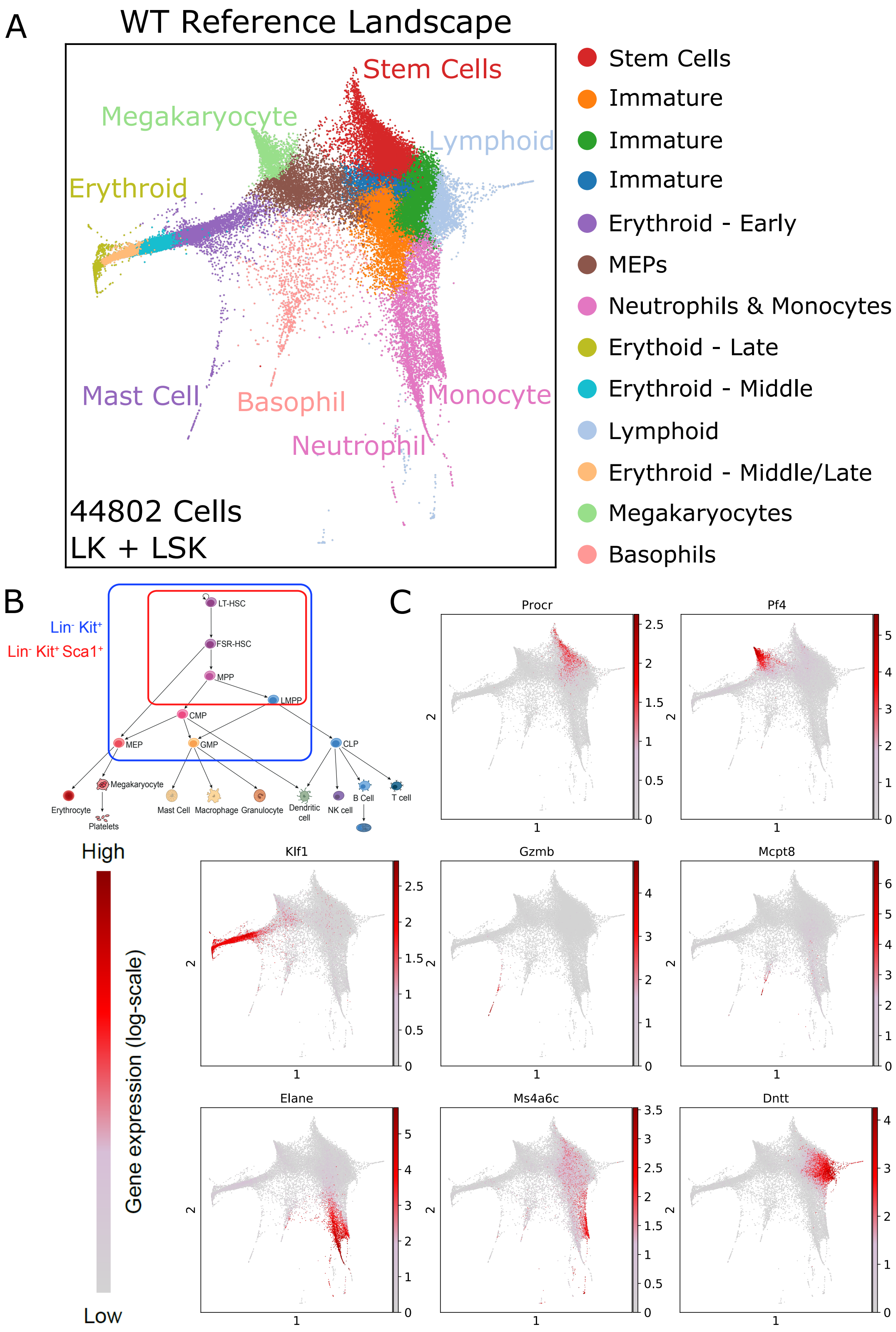


Fig. 3.1 A large wild-type reference of murine haematopoiesis containing over 44,000 transcriptomic profiles. (A) Force-directed graph visualisation of the entire reference dataset, coloured by cluster identity. (B) Diagram showing which classically-defined haematopoietic populations fall within the LK and LSK sorting gates. (C) Log-transformed gene expression of eight marker genes corresponding to different haematopoietic lineages. Panel B was adapted from Wilson et al. (2015).

Finally, *Dntt* highlights early lymphoid progenitors (Rothenberg, 2014). In total, the entry point into eight separate blood lineages can be found within the dataset, including an extremely small population of eosinophil progenitors marked by *Prg2* and *Prg3* (genes not shown) (Olsson et al., 2016).

To take advantage of the high density sampling achieved by this wild-type dataset, and to begin building an integrative framework for the analysis of perturbation models, it was reasoned that this dataset could be leveraged as a 'reference' against which all the perturbation models could be mapped and compared to at a cellular level. To facilitate this, the reference dataset was clustered using Louvain clustering, which is designed to maximise both intra-cluster similarities and inter-cluster differences (Blondel et al., 2008). This resulted in 13 large clusters, which were annotated based on the expression of known lineage marker genes (Figure 3.1A). Notably, the erythroid lineage is split into four separate clusters, with the most immature of these also containing the mast cell lineage. Additionally the neutrophil and monocyte lineages are clustered together in a large myeloid cluster, highlighting how similar the development of these progenitors may be. The most immature regions of the landscape split into four clusters which do not express clear lineage markers and hence are difficult to annotate; the most immature cluster that likely contains the true LT-HSCs is marked by the expression of known HSC genes such as *Procr* and *Fgd5*. This cluster was annotated as 'Stem Cells' but these will not all be true LT-HSCs, which occur at a frequency of ~3% within the LSK gate (see Chapter 4) (Challen et al., 2009). Overall, the reference dataset clearly supports the model of a continuous differentiation landscape in which blood cells pass through a myriad of poorly-defined states before settling into one of many possible terminal lineages (Laurenti and Göttgens, 2018). Within this model, the reference landscape can be used to define how specific perturbations alter murine haematopoiesis across the entire range of progenitor states.

In total, eight different perturbation models were analysed for this project across a total of eleven distinct datasets, as summarised in **Table 3.1**. This includes three biological repeats of the Jak2 V617F model (Li et al., 2014) performed on different days. Each of these repeats had matched WT and perturbed samples containing cells from a single mouse each. The Dnmt3a R882H, Npm1 KO (Vassiliou et al., 2011) and Crebbp KO (Chan et al., 2011) models each had matched WT and perturbed samples containing cells from two mice each. These three models are inducible, with multiple pIpC injections given at around 6 weeks of age to induce the perturbation. Samples from the W41 V831M (Nocka et al., 1990) and Jak2/Tet2 Cross (Shepherd et al., 2018) models did not have matching WT samples, and again each contained cells from two mice each. Two separate Tet2 KO (Ko et al., 2011) experiments

Model	Samples	Total Cell Number	Brief Description	Notes
Jak2 V617F (Li et al., 2014)	WT, HOM	60041	The V617F mutation is the most common mutation in MPNs, which can evolve into AML.	Three distinct experiments, each with matched WTs.
W41 V831M (Nocka et al., 1990)	HOM	13957	The W41 point mutation causes defective c-kit signalling.	No matched WT. Published as part of (Dahlin et al., 2018).
Dnmt3a R882H (Unpublished)	WT, HET	29718	Dnmt3a mutations are common in clonal haematopoiesis and AML patients, and are associated with poor prognosis.	Matched WT.
Npm1 KO (Vassiliou et al., 2011)	WT, HET	30440	Npm1 mutations are common in AML patients.	Matched WT.
Crebbp KO (Chan et al., 2011)	WT, HOM	31273	Crebbp mutations are common in AML and ALL patients.	Matched WT.
Tet2 KO (Ko et al., 2011)	WT, HET, HOM	28138	Loss of Tet2 is associated with clonal haematopoiesis and increased risk of malignancies including AML and CMML.	Two distinct experiments. HET experiment has a matched WT, HOM experiment does not.
Jak2/Tet2 Cross (Shepherd et al., 2018)	Jak2 HOM/Tet2 HET	16723	A cross that is homozygous for Jak2 V617F, and heterozygous for Tet2 KO.	No matched WT.
p53 KO (Tekippe et al., 2003)	WT, HOM	14098	p53 is a tumour suppressor whose function in haematopoiesis is largely unknown.	Matched WT. Younger mice than all other samples.

Table 3.1 List of datasets analysed in this project. Total cell number refers to the number of cells retained after preprocessing steps (see methods). All samples were taken from the Lin-c-Kit+ (LK) sorting gate. WT, wild-type; HET, heterozygous mutation; HOM, homozygous mutation; MPNs, myeloproliferative neoplasms; AML, acute myeloid leukaemia; CMML, chronic myelomonocytic leukaemia; ALL, acute lymphoblastic leukaemia.

were performed; the first experiment contained a homozygous Tet2 KO sample comprising cells from two mice without a matched WT sample; the second comprised of matching WT and heterozygous Tet2 KO samples containing cells from one mouse each. Finally, the p53 KO model (Tekippe et al., 2003) experiment had matched WT and perturbed samples containing cells from one younger mouse (∼9 weeks) each. This approach resulted from a lack of understanding early on in the project about how important matched WT samples were for downstream analysis of the perturbation models. In all cases, samples were taken from roughly 12-week old female C57BL/6 mice to achieve concordance with the WT reference dataset. Additionally, all samples were sorted from the Lin- c-Kit+ (LK) gate, so that each experiment covered the same range of blood progenitor states sampled at representative frequencies and could be compared directly with the reference dataset. An identical 10X Genomics pipeline (version 2 chemistry) was used for every experiment, to minimise the technical variation occurring due to initial preprocessing steps.

3.3 Integration of pre-leukaemic models at a cellular level reveals large-scale perturbation effects

As a first step towards an integrated framework for single-cell perturbation analysis, the perturbation models were characterised on a cellular level to understand how specific progenitor populations change in abundance as a result of different perturbations. Following quality control and filtering steps that were applied identically to each experiment (see methods), all WT and perturbed samples appeared to contain all the expected haematopoietic lineages, as evidenced by marker gene expression and the similar layouts of force-directed graphs calculated separately for each sample. The exceptions to this were the mast cell trajectories from the perturbed Jak2 V617F and W41 V831M models, which were notably absent (Ingram et al., 2000). The general similarity across samples highlighted how the perturbations described here are relatively small in magnitude, and still allow a homeostasis to be achieved within the blood system of these mice. Nevertheless it was clear that certain populations occurred at different frequencies across different perturbations, potentially as a result of differentiation skews/blocks, changes to proliferation/apoptosis, or faster differentiation through specific states.

To quantify these abundance changes in an integrative manner, each sample from each perturbation experiment was mapped back onto the reference dataset, and each individual cell was assigned to one of the 13 clusters identified within the reference dataset previously. It was reasoned that since all the samples - including the reference - were contained within

Sample	Combined p-value	Sample	Combined p-value
WT Reference	0.51	-	-
Jak2 WT (1)	0.34	Jak2 HOM (1)	4.4×10^{-6}
Jak2 WT (2)	0.01	Jak2 HOM (2)	8.5×10^{-20}
Jak2 WT (3)	0.55	Jak2 HOM (3)	2.1×10^{-8}
Dnmt3 WT	0.48	Dnmt3a	0.14
Npm1 WT	0.59	Npm1	0.05
Crebbp WT	0.20	Crebbp	0.02
Tet2 HET WT	0.01	Tet2 HET	0.01
-	-	Tet2 HOM	0.03
-	-	W41	1.0×10^{-8}
p53 WT (Young)	5.8×10^{-8}	p53 (Young)	2.0×10^{-8}

Table 3.2 Statistical testing reveals highly significant abundance changes in pre-leukaemic samples. For each sample, Fischer's method was used to calculate a combined p-value which measures the probability of the observed abundances assuming the null hypothesis is true. For each of the WT samples, the null hypothesis was calculated as the weighted mean of all the other WT samples excluding the the Jak2 WT (2) and Tet2 HET samples. For all the perturbed samples, the null hypothesis was calculated as the weighted mean of all the WT samples excluding the the Jak2 WT (2) and Tet2 HET samples. Values coloured red indicate significant results at a p-value of 0.05.

the LK gate, it was unlikely that there were completely new cell populations occurring in one or more of the perturbed states that were not present in the reference. The mapping itself was performed by calculating a PCA on the reference dataset, and then projecting each sample into this PCA space. For each cell in the sample being considered, its 15 nearest neighbours and their cluster identities in the reference were identified, and the cell was assigned to its most similar reference cluster according to the majority vote of these neighbours. This method of mapping had an accuracy of $> 98\%$ when mapping the reference back onto itself, and its performance in perturbation settings was superior to specialised tools such as *scmap*, *BB-KNN* or *Seurat's* integration tools (Kiselev et al., 2018; Polanski et al., 2020; Stuart et al., 2019), which either failed to assign a large proportion of cells to the reference or clearly removed some of the biological variability inherent to the perturbed samples.

For each sample (excluding the Jak2/Tet2 Cross, which will be the subject of section 3.6), the Z-score between the proportion of cells in each cluster and the average proportion of

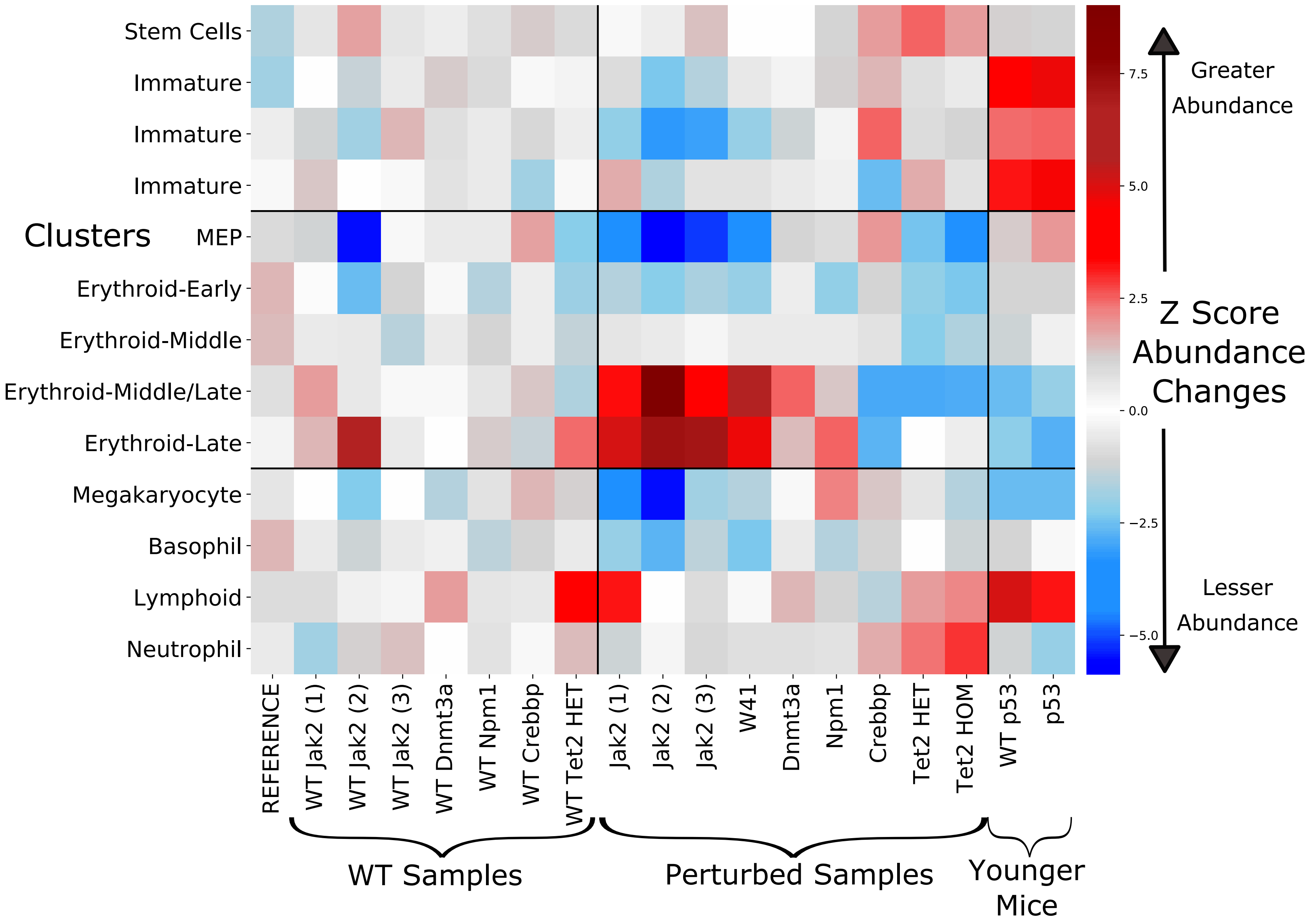


Fig. 3.2 Global integration reveals large cell-type specific abundance changes across perturbations. Heatmap showing the Z-score abundance deviation from the wild-type average in each experiment. The average abundance is calculated as the weighted mean of the wild-type abundances (left 8 columns). A Z-score of 2 indicates the abundance of that cluster is two standard deviations above the WT average.

cells in that cluster across all WT samples (including the reference) was calculated (Figure 3.2). It is immediately clear that there is greater variation across the perturbed samples than across the WT samples, and therefore that the perturbations are causing significant changes in progenitor proportions across the haematopoietic landscape. Since the WT samples are age, gender and background-matched they should contain minimal systematic biological noise, and the remaining variation should be due to a combination of A) the random biological noise inherent to murine haematopoiesis and B) the technical noise from the sequencing and mapping procedures. To quantify the relative variability of each sample, Fischer's method was used to compute a p-value from the set of Z-scores and test whether the observed progenitor frequencies are statistically different from the average WT frequencies (**Table 3.2**). All but two of the WT samples were non-significant, whereas all but one of the perturbed samples were significant to varying degrees.

Across the perturbed samples, divergent populations shifts were observed between different models. The three biological repeats of the Jak2 model show clear and reproducible changes in abundance (Figure 3.3A) with an increase of later erythroid cells and a reduction of MEPs and megakaryocytes, as has been described previously using bulk transcriptomics and phenotypic characterisation of the mouse model (Li et al., 2014). This reproducibility clearly highlights how the Jak2 perturbation induces shifts that are greater in magnitude than the biological and technical noise associated with these experiments. Interestingly, the W41 model shows a strikingly similar shift to the Jak2 model, suggesting a potentially shared underlying mechanisms driving these perturbations. The Dnmt3a and Npm1 models both show lesser but similar shifts towards later erythroid cells. Strikingly, the Crebbp and Tet2 models show the opposite shift in abundances, with an increase in neutrophils and monocytes opposed by a reduction along the majority of the erythroid trajectory. For the samples that contained cells from two mice, the intra-experiment variability could also be assessed (Figure 3.3B), suggesting that whilst abundance shifts are generally reproducible across mice, there is significant intrinsic biological noise as a result of these perturbations. Finally, the p53 experiment highlights how population frequencies within the LK shift with age, as the younger p53 WT sample shows significantly more immature and lymphoid cells compared with its older counterparts. There does not, however, appear to be large abundance shifts due to the p53 perturbation itself.

To take further advantage of the single-cell resolution in these datasets, a new approach for analysing differential abundances was conceived with the aim of providing extremely fine-grained information that could be visually compared across perturbations with ease. For each pair of matched samples (i.e. WT and perturbed samples from the same experiment), the

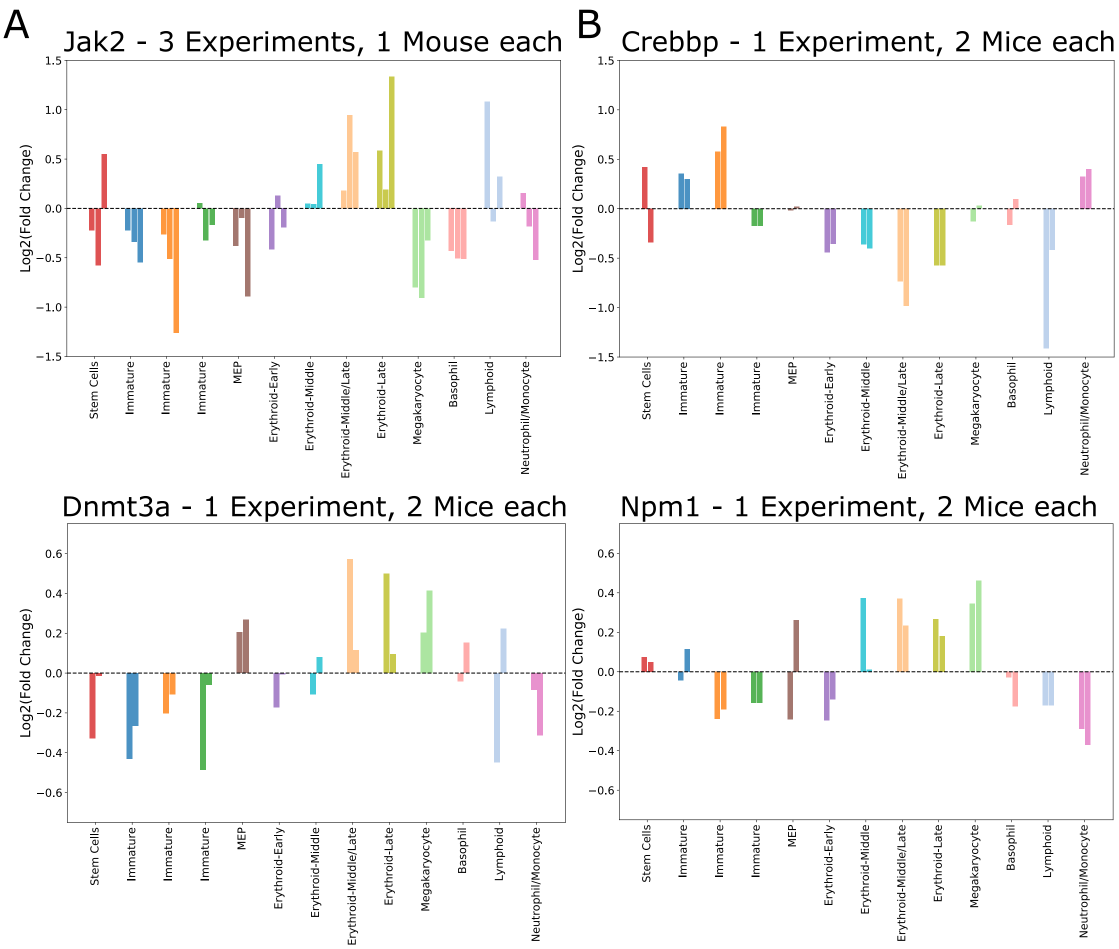


Fig. 3.3 Inter- and intra-experiment variability across perturbations. (A) The fold change within each cluster for each of the Jak2 experimental repeats. Each perturbed sample is compared to its own matched WT sample. (B) The fold change within each cluster for each mouse in the experiments where two mice were sequenced per sample. Each mouse from the perturbed sample was compared to the average of the two mice from the WT sample.

nearest neighbours of each cell in the reference were located and a single vote was given to each of them, such that the total number of votes given out from each sample was equal (see methods). For each cell in the reference, the difference between the number of votes received from each sample acts as a proxy for the differential abundance of that cell state across the matched samples. This therefore provides abundance information at a similar resolution to the sampling density of the reference dataset. Additionally, since the information can be displayed using only the reference, visual comparison across different perturbations was straightforward (Figure 3.4A). At this resolution, detail was visible that was intractable using a cluster-based approach. The Jak2 experiments appear to show an abundance decrease

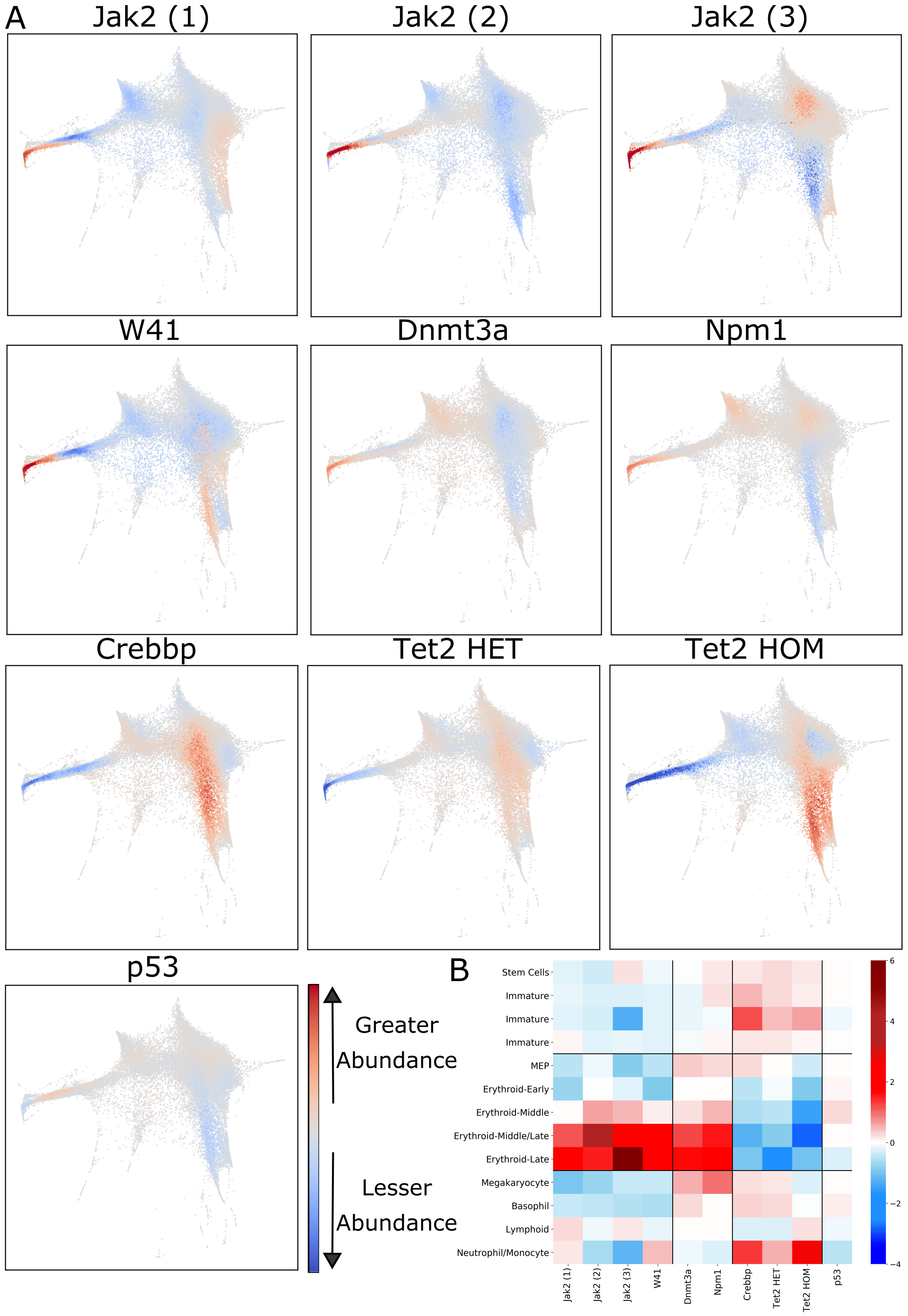


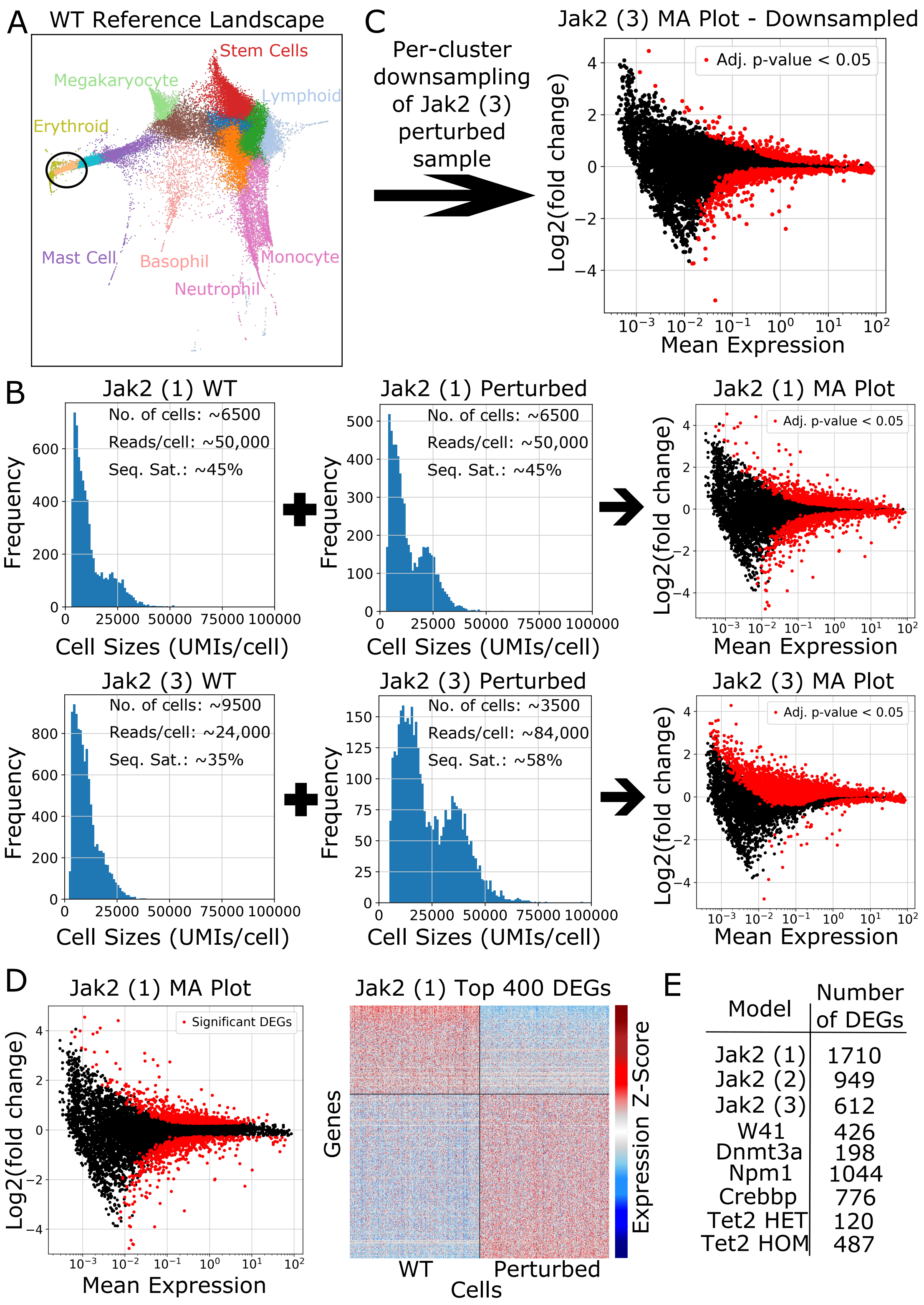
Fig. 3.4 High-resolution abundance changes using paired voting. (A) For each experiment, difference between the votes received from perturbed and matched WT samples is plotted for each reference cell. Red indicates greater abundance due to the perturbation, blue lesser abundance. For those experiments without a matched WT, the Npm1 WT sample was used. (B) The voting differences averaged over the 13 clusters of the reference dataset.

that is specific to the neutrophil but not the monocyte region of the landscape, compared with the W41 model which notably has more neutrophil and less monocyte progenitors, despite looking very similar to the Jak2 model overall. In addition the overall similarities and differences between the three Jak2 repeats are clearer, with each repeat showing at least one feature that the others do not. Strikingly, the Tet2 HET and Tet2 HOM experiments show a clear progression in their disparity from native haematopoiesis, with both the myeloid expansion and erythroid reduction increasing in strength between the heterozygous and homozygous models. By averaging the voting differences across each reference cluster, this method of assessing differential abundance was shown to be congruent with the Z-score approach taken previously (Figure 3.4B). Because the voting method visualises the differences between matched samples, it must be remembered that the visible differences between models could also be driven by differences in the WT samples. It has already been shown that the variability within the WT samples is far smaller than the perturbation effects, suggesting this variability can be discounted when discussing Figure 3.4. To test this further, the voting differences between two WT samples were visualised on the reference for several different WT pairs; in all cases, the scale of the resulting abundance deviations was small compared to the changes observed in Figure 3.4.

3.4 Global molecular integration highlights gene modules with coordinated patterns of perturbation response

Attempting to explain the mechanisms behind the abundance changes observed in the perturbation models requires an integrated analysis of perturbation response at the level of individual genes. RNA-seq is the ideal modality with which to perform this integration, as it is capable of capturing mRNA from across the transcriptome, and therefore allows the molecular perturbation response to be interpreted without biases towards specific genes. The disadvantage of droplet-based scRNA-seq technologies is that the low depth of sequencing leads to an abundance of false negatives within the data, though this is counterbalanced by the ability to sample the transcriptomic space at a far higher density than plate-based methods. Regardless of the experimental design, performing molecular analyses such as differential expression (DE) from single-cell experiments requires careful consideration of the underlying data.

To test whether significant abundance changes occur alongside significant and coordinated molecular changes, DE was performed in all the perturbation models for cells mapping to the cluster annotated as ‘Erythroid-Middle/Late’, since large abundance shifts were observed in



Model	Number of DEGs
Jak2 (1)	1710
Jak2 (2)	949
Jak2 (3)	612
W41	426
Dnmt3a	198
Npm1	1044
Crebbp	776
Tet2 HET	120
Tet2 HOM	487

Fig. 3.5 Single-cell differential expression is confounded by differential sequencing saturation. (A) Reference landscape with the 'Erythroid-Middle/Late' circled. (B) Cell size histograms and MA plots of the first and third Jak2 experiments. Genes that achieved an adj. p-value of <0.05 from the Wilcoxon test are coloured in red. (C) MA plot for the third Jak2 experiment after downsampling the perturbed cells. (D) MA plot from the first Jak2 experiment with significant DEGs in red, based on an adj. p-value cutoff of 0.05 and a fold change cutoff of 1.2. Heatmap of the 400 genes with the smallest adj. p-value from the first Jak2 experiment. (E) Number of DEGs in this cluster for each experiment.

this cluster across several different models (Figure 3.5A). DE was performed on a per-gene basis using the Wilcoxon rank-sum test, as this has been shown to outperform standard tools for bulk RNA-seq such as *EdgeR* and *DESeq2* in a single-cell setting (see methods) (Soneson and Robinson, 2018). Indeed, when testing *EdgeR* to assess its suitability for analysing droplet-based scRNA-seq data, highly spurious p-values were observed as a result of each cell being treated as an individual sample. However the largest problem in performing any method of DE was the effect of differential sequencing saturation, which occurs when different samples within the same experiment are sequenced at different depths (sequencing saturation is a measure of how many more reads must be sequenced to achieve proportional increases in the total number of UMIs). This can lead to highly biased DE results regardless of the method used, due to vastly different per-gene count distributions that are technical artefacts of the sequencing depth (Figure 3.5B). Where it was necessary to correct for this - such as in the third Jak2 experiment - the sample with the higher sequencing saturation was randomly downsampled such that its resulting cell-size distribution matched that of the sample with lower sequencing saturation. In addition, this cell-size based downsampling had to be performed on a per-cluster basis to achieve a sensible correction (Figure 3.5C, see methods).

Despite the limitations of the data, sensible results were obtained highlighting a varying number of significantly differentially expressed genes (DEGs) for each perturbation in these erythroid cells (Figure 3.5D,E). To integrate these results and assess which perturbations held overlapping molecular signatures, each pair of experiments was taken and the intersection of their DEG lists was found. Subsequently, for each pair of experiments the proportion of these overlapping genes that were differentially expressed in the same direction (i.e. either were both upregulated, or were both downregulated) was calculated (Figure 3.6A). This proportion represents a measure of the molecular similarity between different perturbation's responses within this erythroid cluster, where a proportion of 0.5 indicates no likely correlation (i.e. as many genes disagree between models as those that agree), and a score of 0 or 1 indicates complete (anti-)correlation between models. In conjunction, our confidence in these similarities depends largely on the size of the overlapping DEG lists between experiments, which varies greatly across the perturbation models (Figure 3.6B).

Reassuringly, the three biological repeats of the Jak2 experiment show very high similarity and large overlap; of the 383 genes differentially expressed in both the first and second Jak2 experiments, 98% of them move in the same direction (either up- or down-regulated) as would be expected (Figure 3.6A). In addition, the W41 model and to a lesser extent the Npm1 model both show high similarity and reasonable overlap with the Jak2 model. Notably

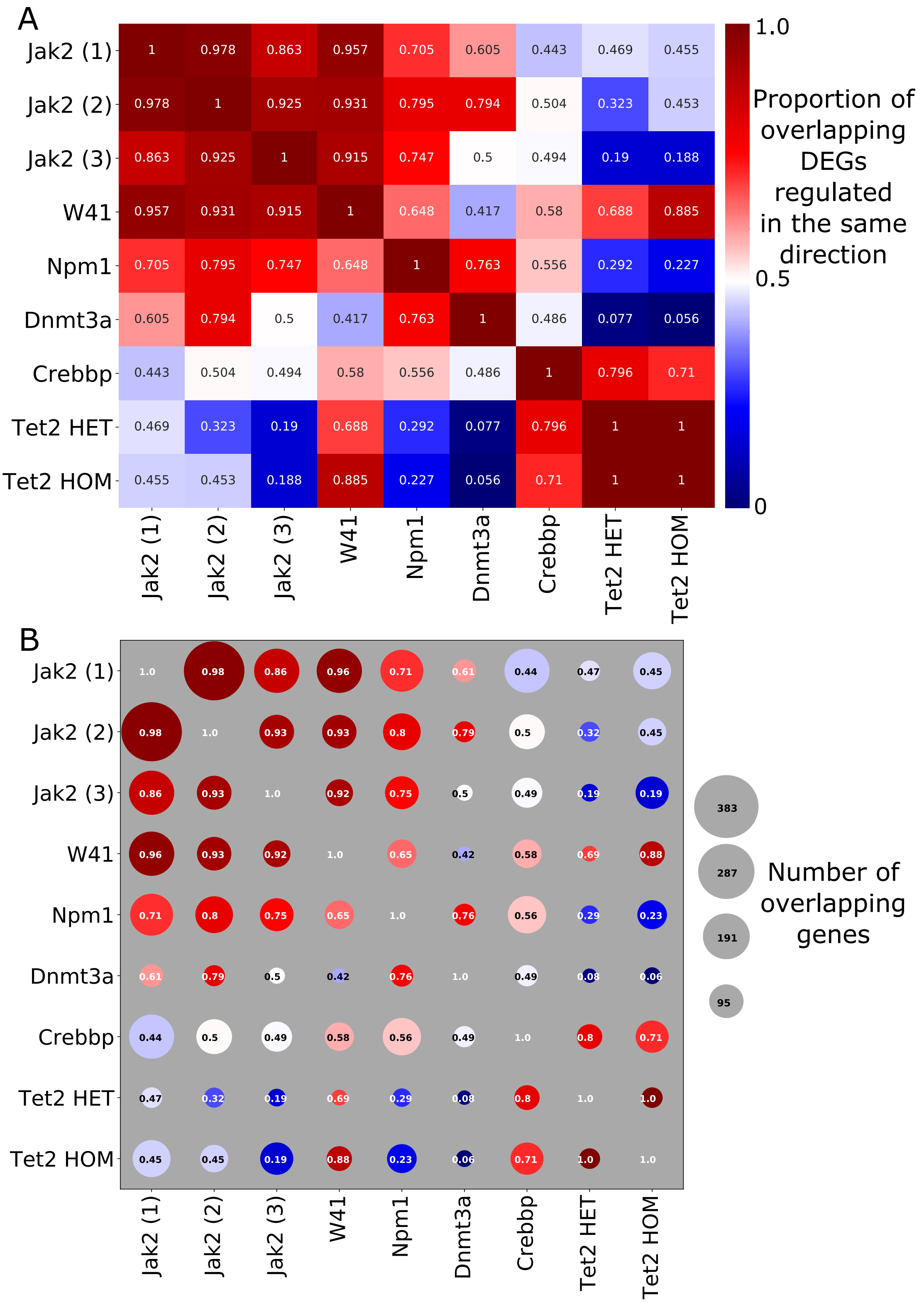


Fig. 3.6 Integration of molecular perturbation response reveals links to abundance changes (A) For each pair of perturbations, the heatmap is coloured by the proportion of overlapping DEGs that are regulated in the same direction in both models. A proportion of 0.5 indicates no correlation betwen models, whereas a proportion of 1 or 0 indicates high molecular correlation or anticorrelation respectively. (B) The size of the DEG overlap between each pair of models.

the Dnmt3a shows very minimal overlap with other models, even accounting for the low number of DEGs (198) observed for the model (Figure 3.6B). The Crebbp and Tet2 models both show anti-correlation with the Jak2 repeats but a high degree of similarity amongst themselves. For example, of the 120 DEGs identified in the Tet2 HET experiment, all 54 that overlap with the Tet2 HOM model are regulated in the same direction. It is unsurprising that the Tet2 HET experiment shows a lesser molecular response than the Tet2 HOM, given its reduced phenotype. In addition, whilst the high degree of similarity between the W41 and Tet2 HOM models is striking, it is worth remembering that these two experiments were those without a matched wild-type sample. As a result, for these experiments their list of DEGs was created by performing differential expression with both the Dnmt3a and Npm1 WT samples and intersecting the results, to mitigate the impact of technical variations across the samples. Nonetheless some technical artefacts are likely to remain and contribute to the observed similarity, which should be treated cautiously.

The most striking result of this analysis is how well the observed molecular similarities in the 'Erythroid-Middle/Late' cluster across models equate to the observed differential abundances in this cluster shown previously. For example, both the 'pro-erythroid' abundance models such as Jak2, W41 and Npm1 and the 'pro-myeloid' abundance models such as Crebbp and Tet2 forming distinct clusters based on their molecular similarities. This suggests there may be groups of genes whose perturbation responses are directly linked to the observed abundance changes and thus linked to haematopoietic fate decisions. To attempt to identify these groups, the fold changes in the 'Erythroid-Middle/Late' cluster for all 332 genes that were dysregulated (either up- or down-) in at least two distinct perturbation models were calculated for each experiment. Dimensionality reduction was then performed on this set of fold changes (Figure 3.7), to visualise their behaviour across different perturbations (Figure 3.8). There are several distinct clusters of genes that co-localise on the dimensionality reduction plot, and therefore have the same patterns of dysregulation across the perturbations. Most notably, there is a cluster of genes comprised of pro-neutrophil and pro-monocyte enzymes and transcription factors including *Mpo*, *Elane*, *Ctsg* and *Irf8* that are strongly downregulated in the Jak2, W41 and Npm1 models but strongly upregulated in the Crebbp and Tet2 HOM models, with smaller upregulation in the Tet2 HET model. These genes are most highly expressed in myeloid progenitor cells, but they maintain a reasonable level of expression along the erythroid trajectory in wild-type (Figure 3.1). The finding that these pro-myeloid genes are upregulated in the erythroid cells of the models with decreased erythroid abundance and downregulated in the erythroid cells of the models with increased erythroid abundance is strong evidence that the behaviour of these genes is playing a role in the aberrant fate decisions occurring due to these perturbations. Whilst the behaviour of

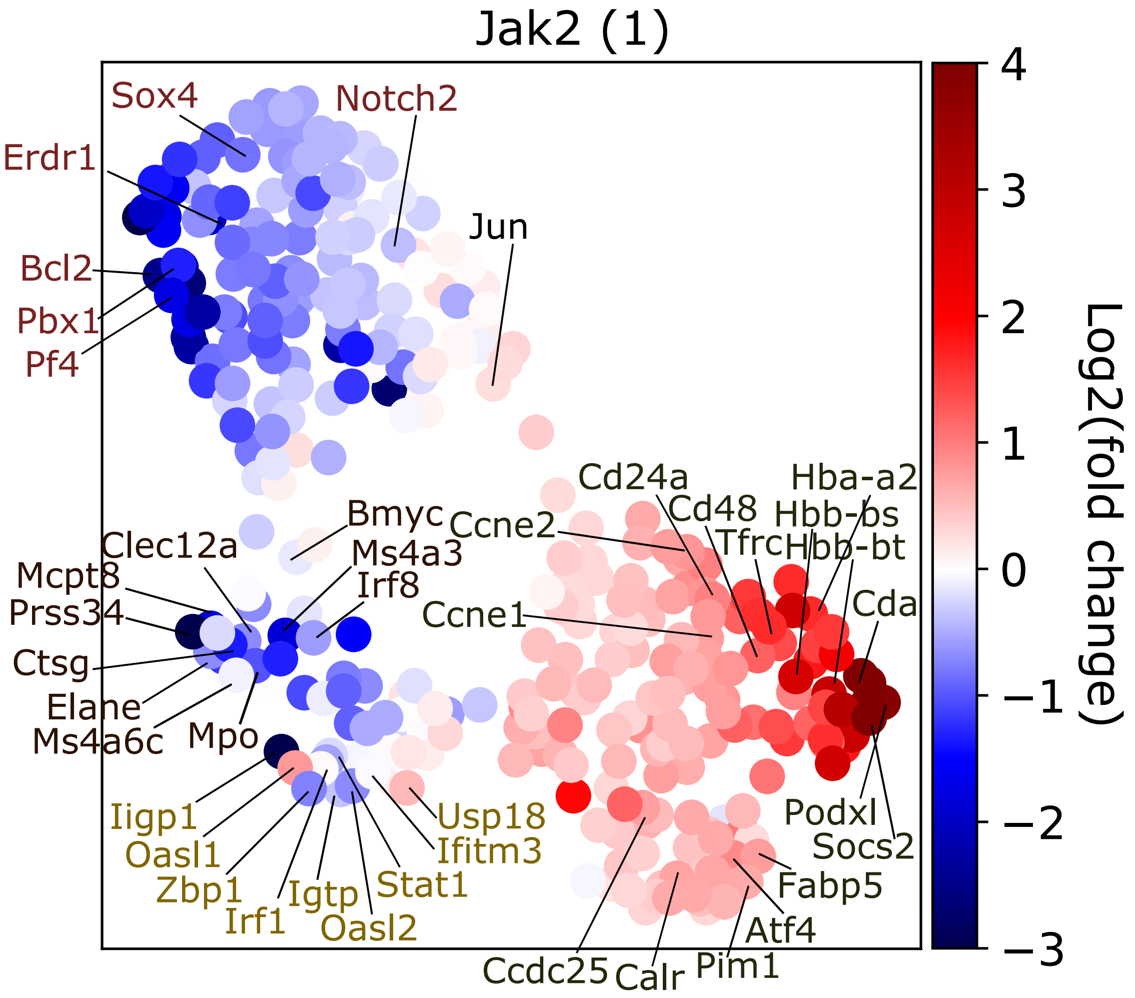


Fig. 3.7 Dimensionality reduction of gene fold changes. UMAP representation of the 332 genes identified as being dysregulated within the 'Erythroid-Middle/Late' cluster in two or more distinct perturbations. Each dot is a gene, coloured in by its fold change in the 'Erythroid-Middle/Late' cluster of the first Jak2 experiment. A number of specific genes are highlighted and coloured according to their locational and biological similarities.

these genes during myelopoiesis has been investigated extensively, the exact degree to which they play a role in erythropoiesis and HSC fate decisions is unknown (Olsson et al., 2016).

Similarly, there is a cluster of interferon-related genes including *Igtp*, *Zbp1*, *Stat1* and *Irf1* which show strong upregulation in the Crebbp and Tet2 models, but downregulation in the Jak2 and Npm1 models. Interferon-gamma signalling is thought to cause myeloid-skewing in HSCs, though its role in later erythroid cells is less clear (Morales-mantilla and King, 2018). Interestingly the W41 model also shows upregulation of these genes, which may contribute to the observed similarities between the W41 and Tet2 models in Figure 3.6A. It is unlikely

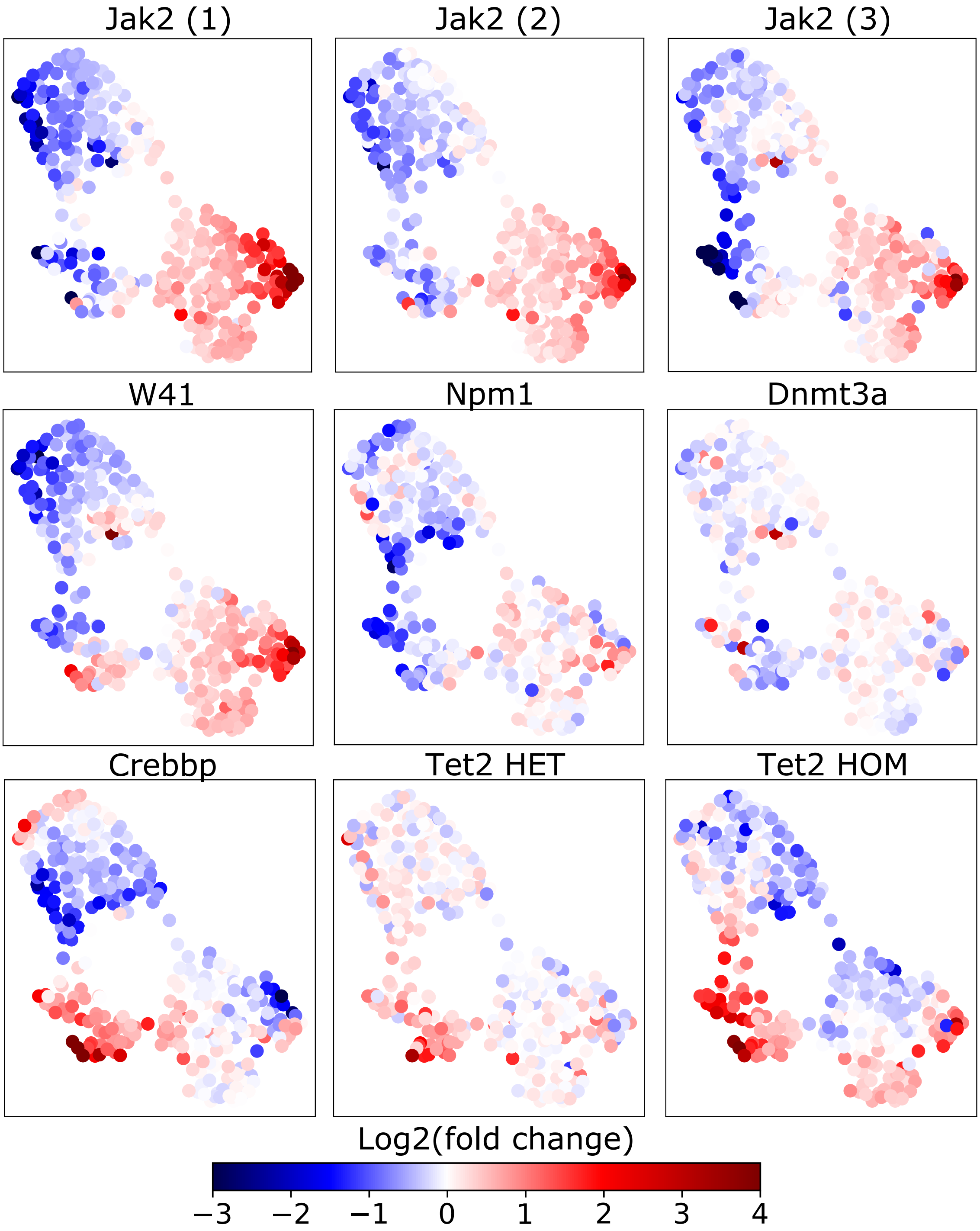


Fig. 3.8 Molecular integration reveals modules of genes that are dysregulated across many perturbations. UMAP representations of the 332 genes identified as being dysregulated in two or more distinct perturbations. For each model the genes are coloured by their fold change within the 'Erythroid-Middle/Late' cells in that model.

that this signature is a biological artefact caused by certain mice having a viral infection, since once again its pattern of dysregulation closely mirrors the observed abundance changes. Hence the interferon pathway may also play a significant role in aberrant pre-leukaemic fate decisions, and could potentially be a notable difference between the Jak2 and W41 models despite their similarity at a cellular level. In the top-left of the visualisation there is a large cluster of genes whose expression pattern is less clear, including pro-megakaryocytic genes such as *Pf4* and *Pbx1* alongside anti-apoptotic genes such as *Bcl2* and *Erdr1* (Bincoletto et al., 1999; Mango et al., 2014). All of these genes are strongly downregulated in the Jak2 and W41 models. The former makes sense given that the Jak2 and W41 models have a decreased abundance of megakaryocytes. However the downregulation of anti-apoptotic genes is surprising given the observed abundance increase of these erythroid cells in the Jak2 and W41; this suggests that despite their overabundance, they are experiencing a stress response and an associated increase in apoptosis. The dysregulation of these genes in the other models is less clear, but interestingly they appear to be downregulated in the Crebbp model, where potentially increased apoptosis or a decrease in cycling activity is driving the observed abundance reduction (see chapter 4).

Finally, there is a large cluster of genes in the lower right corner of the visualisation which are upregulated in the Jak2 and W41 models, but many of which are downregulated or unchanged in the Crebbp and Tet2 models. This cluster includes many known pro-erythroid genes such as the globins (*Hba-a2*, *Hbb-bs*, *Hbb-bt*), *Tfrc* (*CD71*), and *Fabp5* (Tusi et al., 2018), as well as a number of cell-cycle genes such as *Ccne1*, *Ccne2* and *Ccdc25*. In addition this cluster contains known targets of the JAK/STAT signalling pathway such as *Socs2* and *Pim1* (Letellier and Haan, 2015; Schwemmers et al., 2007). Notably, however, many of these genes are also upregulated in the Tet2 HOM model, including a number of heat shock proteins *Hsp90*, *Hspa5*, *Hspa8* and *Hsph1*, which are linked to cellular stress response (Chen et al., 2012), and have been shown previously to anti-correlate with genes such as *Erdr1* mentioned above (Jung et al., 2011). Hence whilst some genes in this cluster may contribute to cellular expansion in erythroid cells, some of them may also be playing different roles in the perturbation response that do not influence the proportional abundances of the progenitor landscape.

3.5 Extending perturbation analyses across whole cellular trajectories

In order to understand the molecular mechanisms driving skewed cellular abundances, it is not sufficient to look at the molecular changes occurring at one specific cellular stage of maturation, as in section 3.4. Instead, the molecular perturbation response across entire cellular trajectories must be analysed. This will provide a far broader and more dynamic understanding of the haematopoietic system's response to perturbation. The largest barrier to characterising perturbation response across entire trajectories is disentangling the gene expression changes due to the perturbation from those due to normal differentiation. Comparing WT cells from one stage of a trajectory to perturbed cells from another stage will reveal many differences, but these will not be useful (Hamey and Göttgens, 2018). A coarse-grained approach could be taken where many clusters are defined along the trajectory and comparisons are restricted to WT and perturbed cells mapping to the same cluster. However here an extremely fine-grained approach was desirable, and therefore using the concept of pseudotime was preferred.

Firstly, the clusters corresponding to the trajectory of interest (e.g. the erythroid trajectory) were chosen by performing Partition-Based Graph Abstraction (PAGA) on the reference dataset (Figure 3.9A). PAGA calculates the relative connection strengths between each of the 13 clusters and hence infers the most likely pathway between a starting cluster (Stem Cells) and an ending cluster (e.g. 'Erythroid - Late') (Wolf et al., 2019). Whilst this is a relatively crude method of trajectory inference (see chapter 6), it sensibly defines a set of cells that can be labelled as 'the erythroid trajectory', moving from the stem cell cluster to MEPs and then through the erythroid clusters. For each perturbation, the WT and perturbed cells that belong to this trajectory were then aligned using pseudotime (Figure 3.9B) (Bendall et al., 2014; Trapnell et al., 2014). Briefly, pseudotime is a measure of the distance between each cell and a root cell, defined to be the WT cell with the highest hscScore. The hscScore score is calculated using the expression of a number of genes known to identify the most HSC-like cells within a scRNA-seq dataset, and hence the root cell represents the best estimate of a true WT HSC within the data (see chapter 4) (Hamey and Gottgens, 2019). The pseudotime score of each WT and perturbed cell is then calculated. On the assumption that gene expression changes due to the perturbation are small compared to those due to maturation, this score should represent each cell's progression along normal differentiation. Hence by comparing WT and perturbed cells with the same pseudotime scores, any remaining differences should be solely due to the perturbation response. To do this, a sliding window approach was used to calculate the WT and perturbed expression levels of a given gene as a function

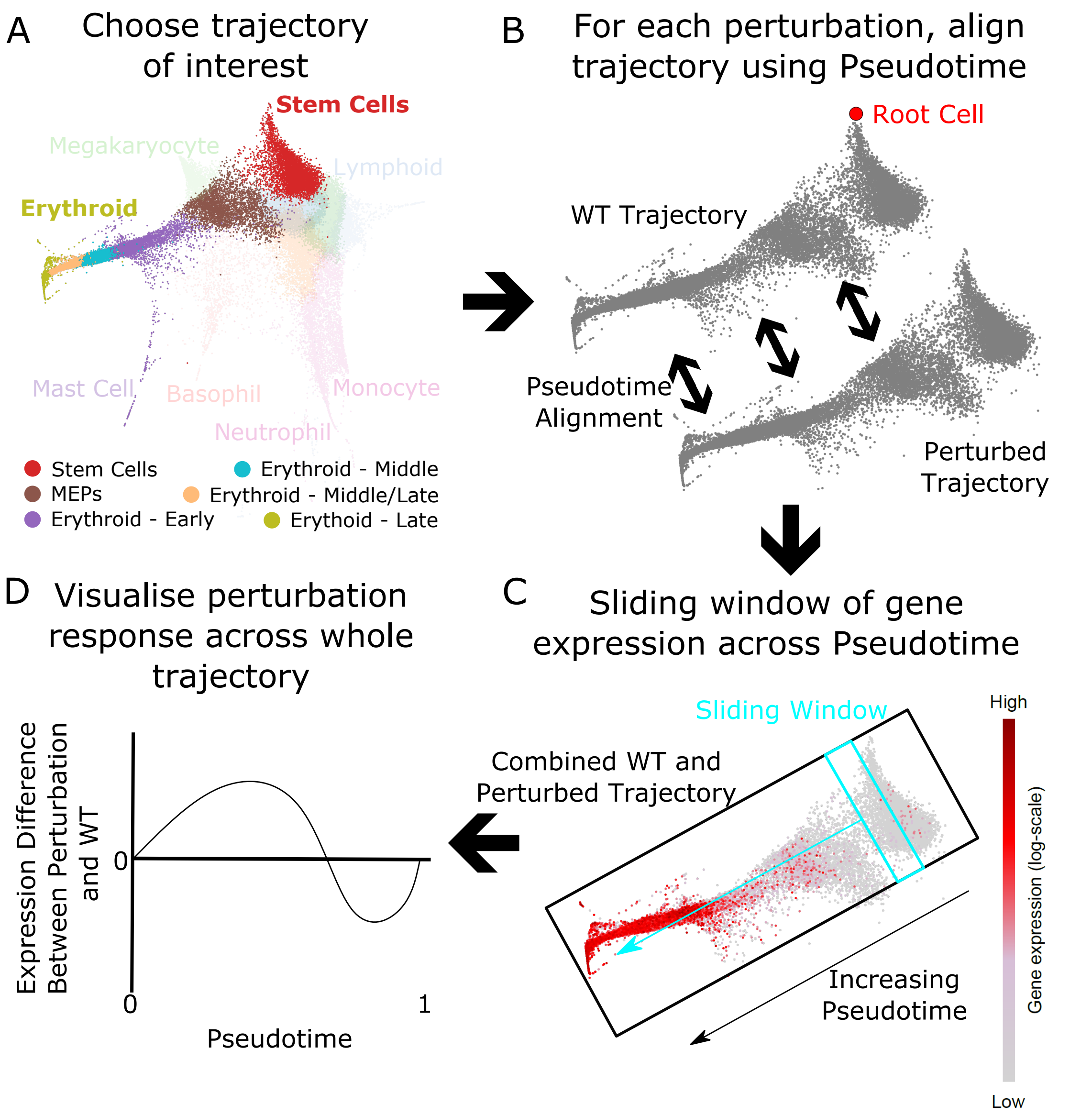

A
Choose trajectory of interest
Stem Cells
Megakaryocyte
Lymphoid
Erythroid
Mast Cell
Basophil
Monocyte
Neutrophil
● Stem Cells
● MEPs
● Erythroid - Early
● Erythroid - Middle
● Erythroid - Middle/Late
● Erythroid - Late
B
For each perturbation, align trajectory using Pseudotime
● Root Cell
WT Trajectory
Pseudotime Alignment
Perturbed Trajectory
C
Sliding window of gene expression across Pseudotime
Sliding Window
Combined WT and Perturbed Trajectory
Increasing Pseudotime
High
Gene expression (log-scale)
Low
D
Visualise perturbation response across whole trajectory
Expression Difference Between Perturbation and WT
0
0
1
Pseudotime

Fig. 3.9 Pseudotime aligns cellular trajectories to calculate dynamic differential expression. (A) Clusters corresponding to the desired trajectory are located in the reference dataset. (B) For a given perturbation, a root cell corresponding to the most HSC-like transcriptome is located in the WT data. Pseudotime is then calculated on the combined WT and perturbed trajectory to align them. (C) A sliding window is used to calculate the WT and perturbed expression of a specific gene along pseudotime. (D) The difference between the WT and perturbed expression levels are visualised as a function of pseudotime.

of pseudotime (Figure 3.9C). The smoothed difference between these expression levels can then be visualised between a pseudotime score of 0 (root cell) to 1 (end of trajectory) (Figure 3.9D). Hence the region above the x-axis represents upregulation of the gene at that point along the trajectory and vice versa. For the perturbation with technical repeats (Jak2), the mean perturbation response over the three experiments was calculated. For those with biological repeats (W41, Dnmt3a, Npm1, Crebbp and Tet2 HOM) the mean response over the two mice was calculated.

Calculating the expression increase due to each perturbation for the group of pro-myeloid enzymes and transcription factors identified previously highlights how these genes are dysregulated across the entire erythroid trajectory, and not simply in the single cluster analysed in section 3.4 (Figure 3.10A). The pro-neutrophil trio of *Mpo*, *Elane* and *Ctsg* are upregulated at a roughly constant level across the entire erythroid trajectory in the Crebbp and Tet2 models, and downregulated in a similar fashion in the Jak2, W41 and Npm1 models, with the Dnmt3a model showing little to no response. The pro-monocytic trio of *Irf8*, *Ms4a6c* and *Ms4a3* show more dynamic behaviour but the same general pattern, despite these genes being expressed at lower level in erythroid cells. This analysis suggests that even in very immature blood cells, these transcription factors have patterns of dysregulation that are highly coordinated with the observed differentiation skewing. Since it is believed that the major haematopoietic fate decisions occur early on in the hierarchy, this is further evidence that a crucial effect of these perturbations is to induce abnormal expression levels of these myeloid genes, which is then strikingly maintained along the entire erythroid trajectory. Similarly, the effects of the interferon-related genes identified previously are prominent across the entire erythroid trajectory, with the W41, Crebbp and Tet2 models upregulating these genes and the Dnmt3a and Npm1 models downregulating them, whilst the Jak2 model shows minimal response (Figure 3.10B). The pattern of upregulation in the Crebbp and Tet2 models suggests that the amplitude of the dysregulation rises to a peak midway along the trajectory before diminishing towards the end.

Analysis of the group of pro-erythroid genes upregulated in the Jak2 and W41 models in Figure 3.8 reveals two distinct modules of genes, with differing patterns of dynamic perturbation response when considering the entire erythroid trajectory. The first group appears to predominantly affect the Jak2 and W41 models towards the end of the trajectory (Figure 3.11A). Many of these genes, including *Arhgdig*, *Cda*, *Fabp5* and *Podxl* have previously been implicated in the response of late-stage erythroid progenitors to EPO over-stimulation (Tusi et al., 2018). Given that the Jak2 V617F mutation acts to constitutively activate EPO receptors (amongst others), it is unsurprising that the same behaviour is seen here (Silvennoinen and Hubbard, 2015). Of interest is how similar the response of these genes in the W41 model is, mirroring the abundance similarities seen previously. Whilst the synergism between c-Kit receptor and Epo-R for producing erythroid progenitors is well

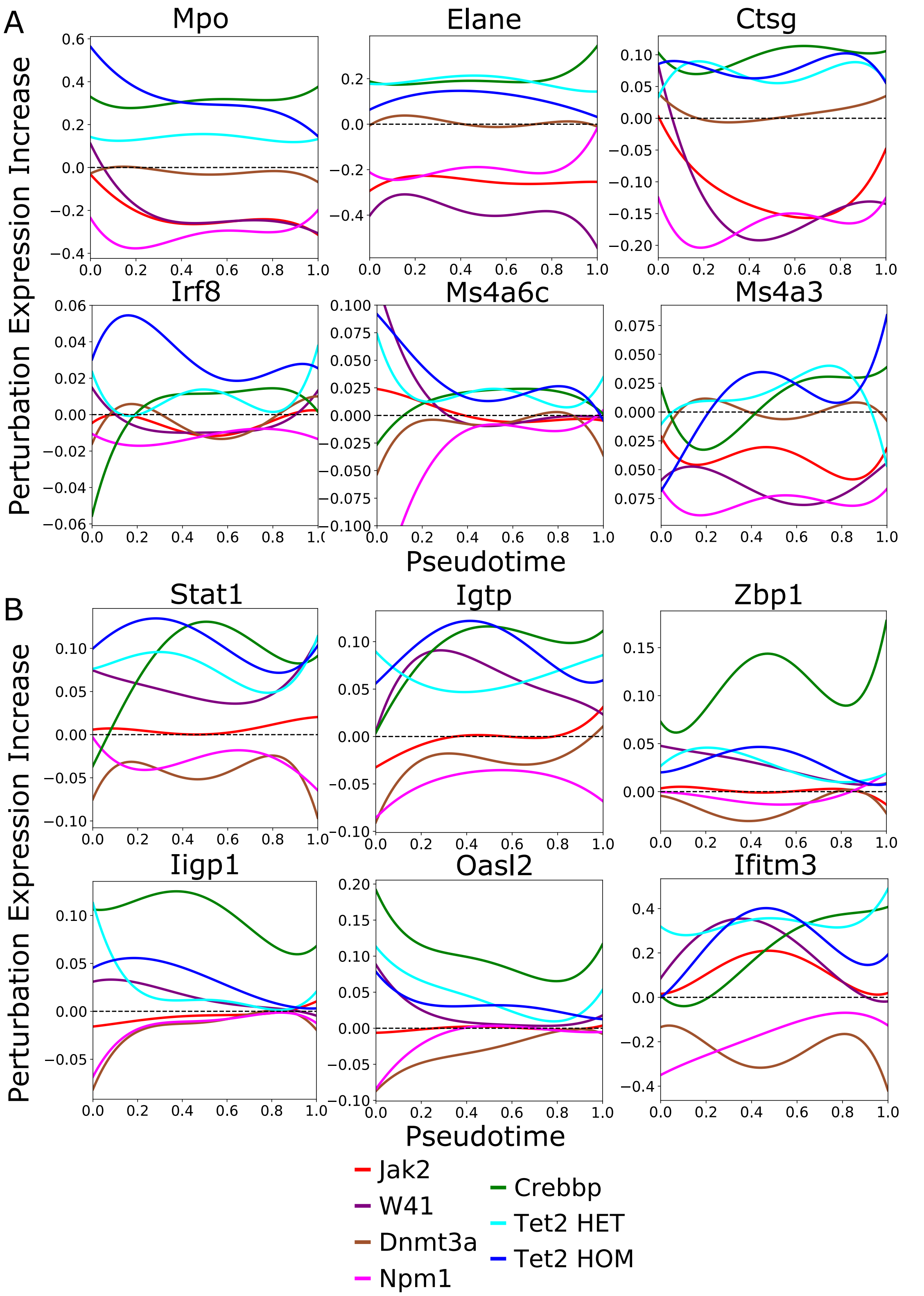


Fig. 3.10 Pro-myeloid and interferon-related genes show dynamic differential expression across the entire erythroid trajectory. (A) Expression difference between perturbed and WT cells (y-axis) as a function of pseudotime (x-axis) from 0 (stem cells) to 1 (end of erythroid trajectory) for 6 pro-myeloid transcription factors. Each curve corresponds to a specific perturbation. (B) Expression difference as a function of pseudotime along the erythroid trajectory for 6 interferon-related genes.

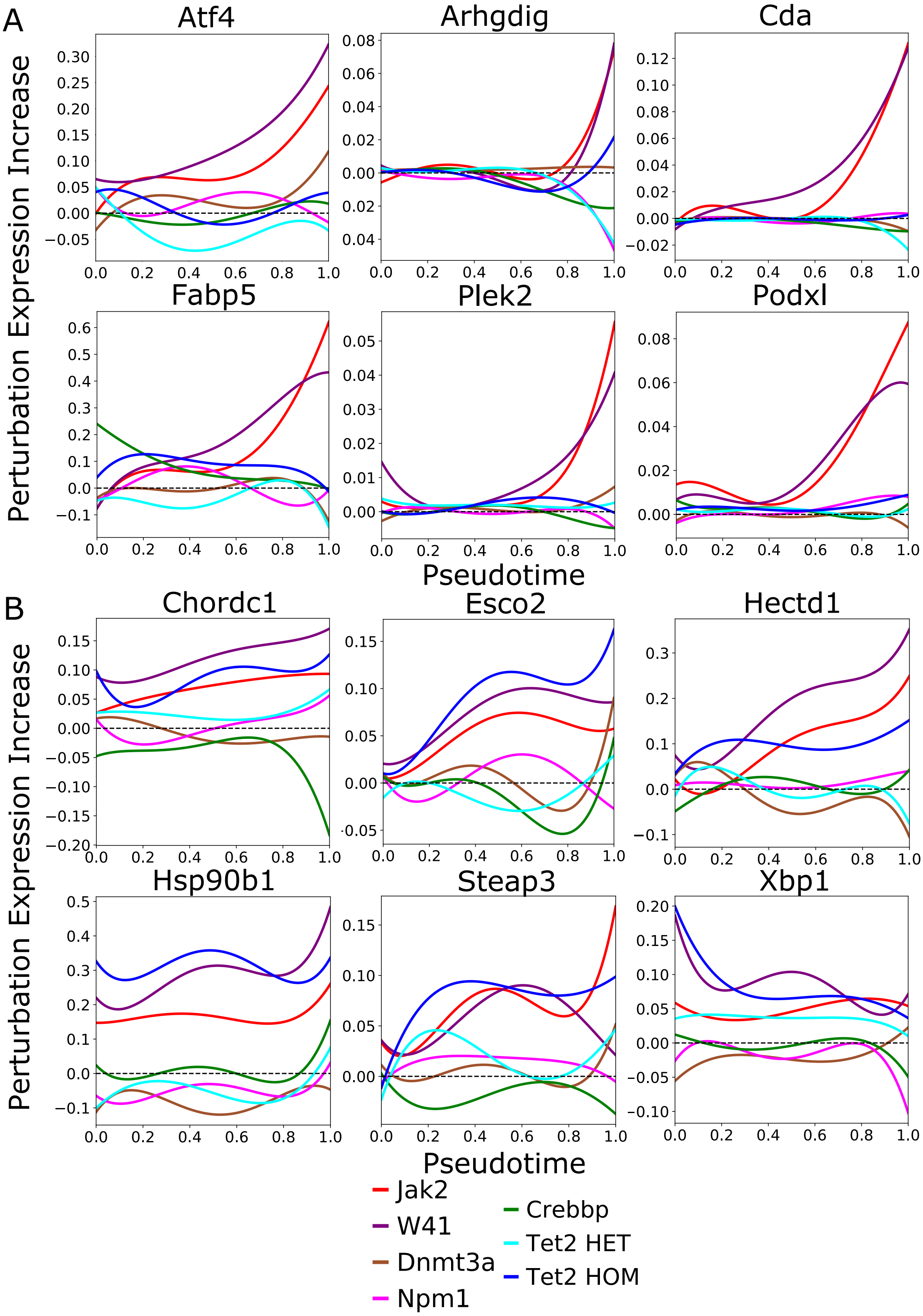


Fig. 3.11 Two groups of pro-erythroid genes show different patterns of dynamic expression across the erythroid trajectory. (A) Expression difference between perturbed and WT cells as a function of pseudotime from 0 (stem cells) to 1 (end of erythroid trajectory) for 6 genes affecting the Jak2 and W41 models. Each curve corresponds to a specific perturbation. (B) Expression difference as a function of pseudotime along the erythroid trajectory for 6 genes affecting the Jak2, W41 and Tet2 models.

established (Munugalavadla and Kapur, 2005), it is unclear why the W41 c-Kit mutation - which is believed to be a leaky loss-of-function mutation (Nocka et al., 1990) - should lead to an extremely similar perturbation response. However recent evidence has suggested that acute anaemia, a phenotype of the W41 mice, leads to compensatory EPO production outside of the bone marrow, potentially explaining the observed similarities (Bennett et al., 2019; Millot et al., 2010). No evidence of differential expression of *Epor* was observed in either perturbation. The second group appears to affect the Jak2, W41 and Tet2 HOM models, with dysregulation seen to be generally increasing over the course of the trajectory (Figure 3.11B). Whilst a clear biological interpretation of these genes is difficult, many of them have been implicated in haematological malignancies previously. For example, *Chordc1* expression is known to inhibit the ROCK signalling pathway, and *Chordc1*-deficient mice develop lethal myeloproliferative malignancies resembling chronic myeloid leukaemia (Rocca et al., 2018; Savino et al., 2015). *Hectd1* loss is known to result in reduced self-renewal properties of mouse HSPCs and interestingly is thought to inhibit the expression of interferon-related genes (Brennan, 2019). Both of these genes have been shown to interact with the heat shock protein *Hsp90*, suggesting the stress response occurring in the Jak2, W41 and Tet2 models may be tightly linked in a way that is independent of their observed differentiation skewing (Gano and Simon, 2010; Sarkar and Zohn, 2012). *Esco2* is a gene coding for cohesin establishment, which has been suggested as a master regulator of transcriptional reconfiguration in haematopoiesis (Panigrahi and Pati, 2012; Sasca et al., 2019), whilst *Steap3* is a ferrireductase that facilitates iron uptake in erythroid precursors (similarly to *Tfrc*) (Lambe et al., 2009); it has been linked with increased lymphoma proliferation and as a pathway for increased oxidative stress in red blood cells (Howie et al., 2019; Isobe et al., 2011). *Xbp1* is a transcription factor known to protect HSCs from stress-induced apoptosis through unfolded protein response, but its role in erythroid cells is unknown (Liu et al., 2019). Nevertheless there appears to be a clear stress-related link between the Jak2, W41 and Tet2 models that is not present in the others. Notably, the Tet2 HET model deviates from the Tet2 HOM model in this group, despite mirroring the latter closely in the pro-myeloid and interferon-related gene modules discussed above. This may suggest that whilst the Tet2 HET model elicits similar differentiation skewing to the HOM model, it is not capable of invoking a large-scale stress response in the erythroid trajectory.

This method of analysing perturbation response across entire cellular trajectories can also be performed in an unsupervised manner, without requiring prior identification of genes of interest. This novel unsupervised approach was named the dDEG method, as it is capable of directly extracting dynamically differentially expressed genes (dDEGs) from a scRNA-seq perturbation dataset consisting of a WT and perturbed sample. As input, the dDEG method

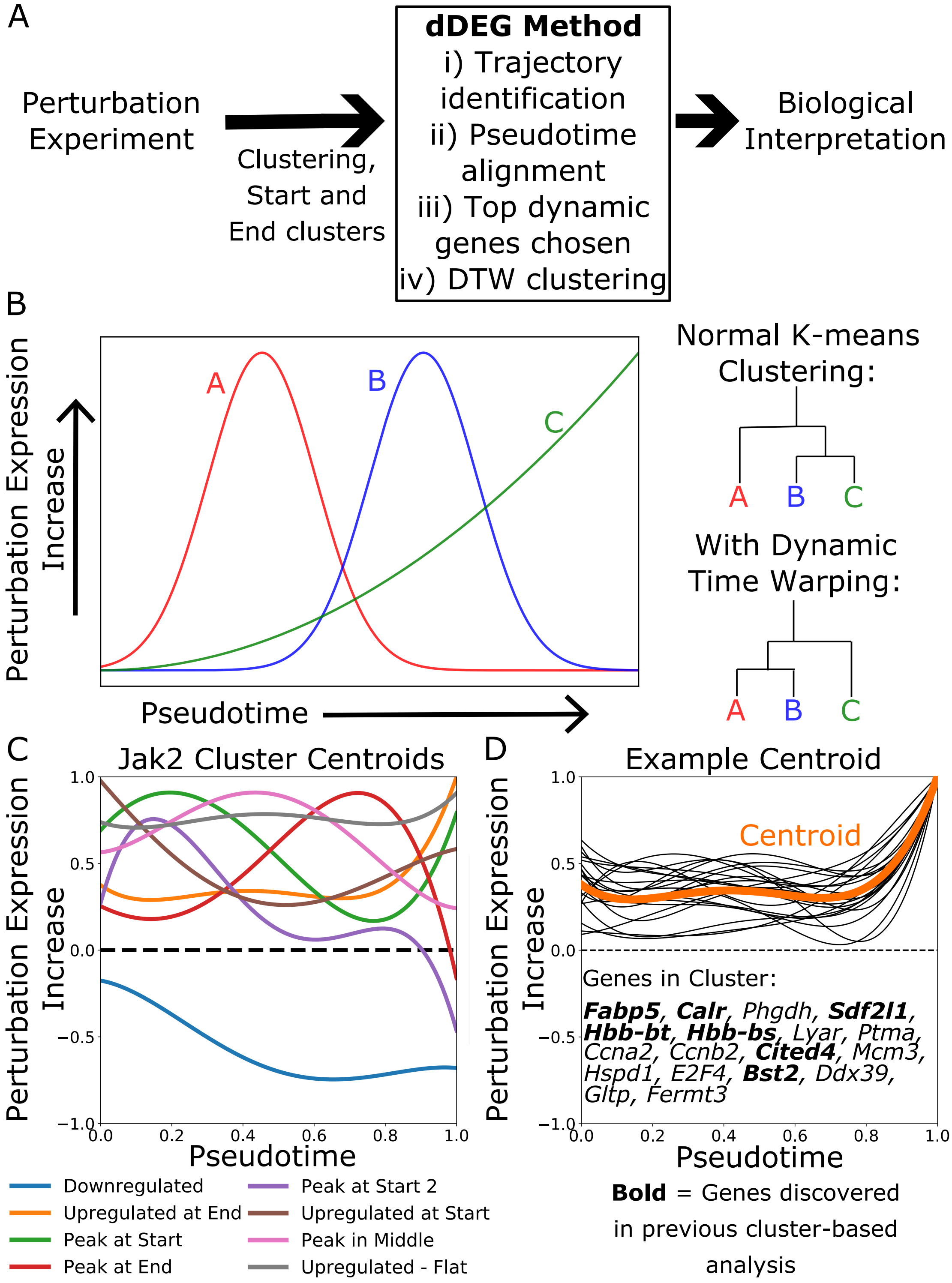


Fig. 3.12 The dDEG method identifies and clusters dynamically differentially expressed genes in an unsupervised manner. (A) Outline of the steps performed by the dDEG method. (B) Schematic of how dynamic time warping (DTW) clusters dDEGs based on the shape of their dysregulation, as opposed to a standard k-means clustering algorithm. (C) Results of DTW clustering on the 97 dDEGs identified across the erythroid trajectory of the three Jak2 experimental repeats. Each coloured line represents the scaled centroid of a gene cluster, which have been annotated according to their behaviour across the trajectory. (D) A closer view of the orange centroid from (C), showing the individual dDEG result for each gene in the cluster.

requires only a clustering of the dataset and the identity of the starting and ending clusters for the trajectory of interest. From this, it will follow a similar procedure as in Figure 3.9 to identify the cells belonging to the trajectory, followed by an integrated pseudotime analysis, including the calculation of the relevant root cell. The dDEG method will then analyse every gene in the dataset and rank them according to their degree of differential expression across the whole trajectory (see methods, Figure 3.12A). After extracting the top dDEGs based on a predefined threshold, dynamic-time-warping (DTW) is used to cluster the top dDEGs into groups of genes with the same patterns of dynamic differential expression in an unsupervised manner (Figure 3.12B) (Berndt and Clifford, 1994). Experimental repeats are handled naturally by requiring a gene to pass the threshold in each of the repeats. Each cluster can then be analysed further to understand its biological significance or relationship to other clusters.

As an example of this, the dDEG method was applied to the three experimental repeats of the Jak2 model for the erythroid trajectory. 97 genes were found to pass the threshold in all three repeats, which were grouped into 8 clusters (Figure 3.12C). This included clusters of genes whose magnitude of upregulation due to the perturbation peaked at different points along the trajectory; in Figure 3.12C, the green and purple clusters peak at the start of the trajectory, before the pink cluster peaks in the middle and the red cluster towards the end. In addition the orange cluster contained genes that began their upregulation at the end of the trajectory (Figure 3.12D). These included genes such as *Fabp5*, *Calr* and *Hbb-bt* that were identified using the previous supervised approach, but also included several other genes with the same pattern of dysregulation. Therefore whilst the supervised cluster-based approach excels at integrating information across different perturbation models, the dDEG method is designed to perform in-depth whole-trajectory analyses of a single model in a completely unbiased manner. Once the dynamics of the clusters identified using the dDEG method have been visualised, biological interpretation of the results can be attempted. Gene ontology results for clusters from the Jak2 erythroid trajectory largely focus on cell-cycle effects, particularly in clusters that are upregulated due to the perturbation towards the end of the trajectory. However the clusters upregulated towards the start of the trajectory were also linked to positive regulation of metabolic processes and protein-DNA complex assembly, suggesting that early erythroid cells such as MEPs may be more metabolically active as a result of the Jak2 mutation. This may link to the observed abundance skewing in the Jak2 model, where there are fewer MEPs and an overabundance of later erythroid cells; if the mutant MEPs are more metabolically active than their WT counterparts they may stay in the MEP state for less time and hence appear to be undersampled in the perturbed dataset.

The later erythroid cells may then expand by cycling and dividing quicker leading to an overabundance within the LK gate.

In addition to the approaches for trajectory-wide molecular analyses presented in this section, a new visualisation method for differential gene expression was also developed. In a similar approach to the paired-voting visualisation of abundance differences in section 3.2, the nearest neighbours of each reference cell were located in the WT and perturbed samples of a perturbation experiment. The mean expression of these neighbours for a given gene was then calculated, and the difference between the expression of WT and perturbed neighbours plotted on the reference. This provides a global view of a gene's differential expression profile at extremely fine resolution. Applying this method to the pro-myeloid genes identified previously recapitulates their behaviour along the erythroid trajectory and highlights similar behaviour across the entire LK landscape (Figure 3.13A). In the Jak2 model, *Mpo* was downregulated in all the immature clusters and megakaryocytes as well as the erythroid trajectory, and only reaches WT levels near to the end of the neutrophil trajectory. Meanwhile *Elane* exhibits even greater downregulation in the neutrophil and monocyte trajectories compared to the erythroid branch. This behaviour was almost exactly reversed in the Tet2 HET model (Figure 3.13B). On the other hand, upregulation of *Fabp5* in the Jak2 model was specific to the end of the erythroid trajectory as seen previously, with downregulation being observed in the most immature clusters. In addition whilst in the Tet2 HET model *Stat1* shows its largest upregulation midway along the erythroid trajectory, a general increase was observed across the entire LK landscape; this pattern was also seen in the other interferon-related genes in the Tet2 and Crebbp models.

3.6 Combining mutations reveals that perturbation interactions are complex and trajectory-specific

Patients with pre-leukaemic haematological malignancies such as MPNs or clonal haematopoiesis often possess more than one driver mutation, and hence discerning how different perturbations combine is of direct clinical interest (Sperling et al., 2017). However relatively little work has been performed in this area, especially at the scale of the whole blood system. The Jak2 V617F mutation is the most common mutation reported in MPN patients, and mutations in Tet2 are the most common co-mutations in Jak2 V617F-positive MPNs (Shepherd et al., 2018). Therefore incorporating a Jak2/Tet2 double mutant into the integrated perturbation framework developed in this chapter should allow highly relevant comparisons to be made across the single- and double-mutant Jak2 and Tet2 perturbed states.

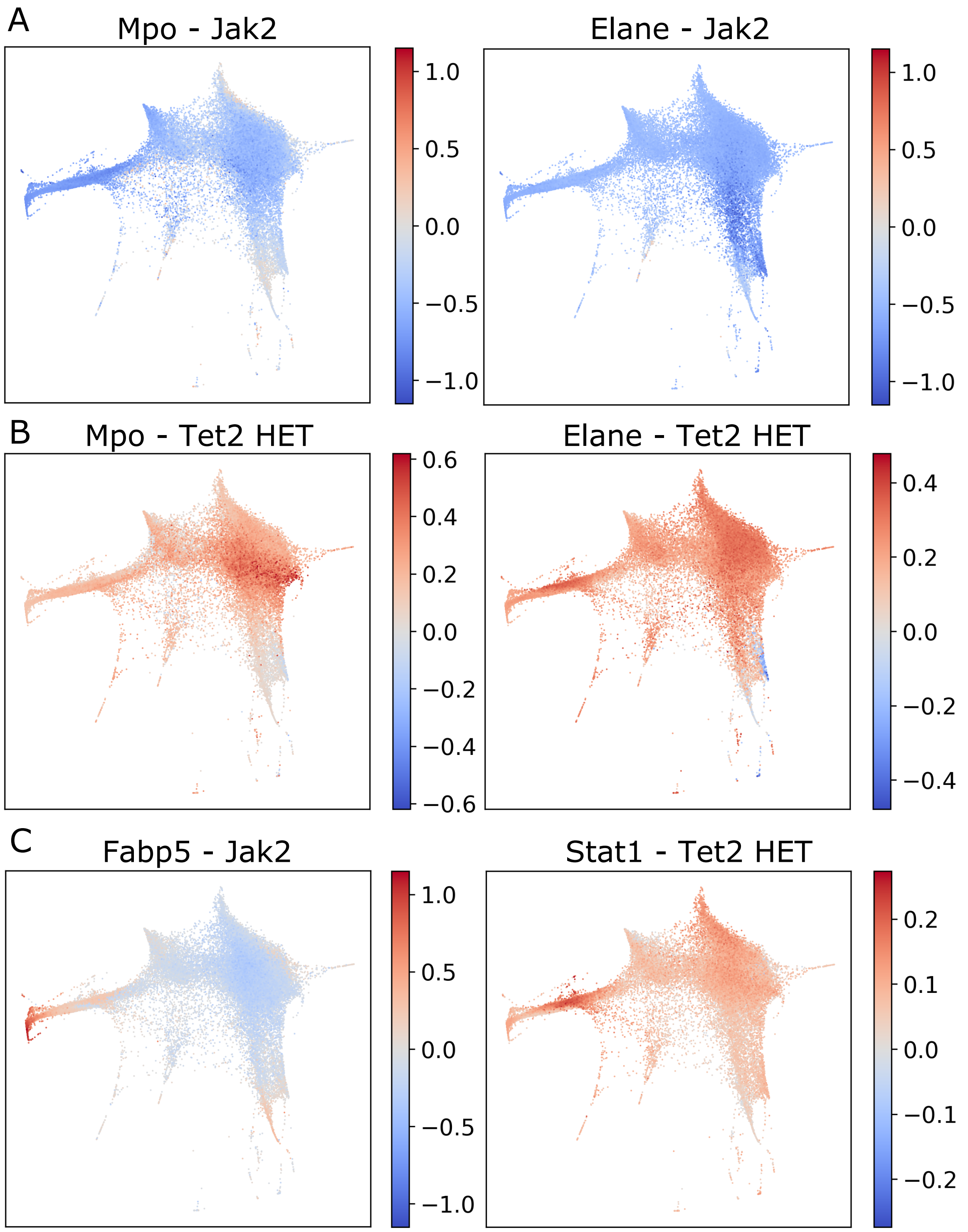


Fig. 3.13 Visualisation of differential gene expression at single-cell resolution. (A) For each reference cell, the difference between the mean expression of its nearest WT and perturbed neighbours is plotted for the gene and experiment shown. (B) Differential expression of pro-myeloid transcription factors in the Tet2 HET model differs completely from the Jak2 model. (C) Visualisation of differential gene expression recapitulates results from the dDEG method in the erythroid trajectory (Fig. 3.10B, 3.11A).

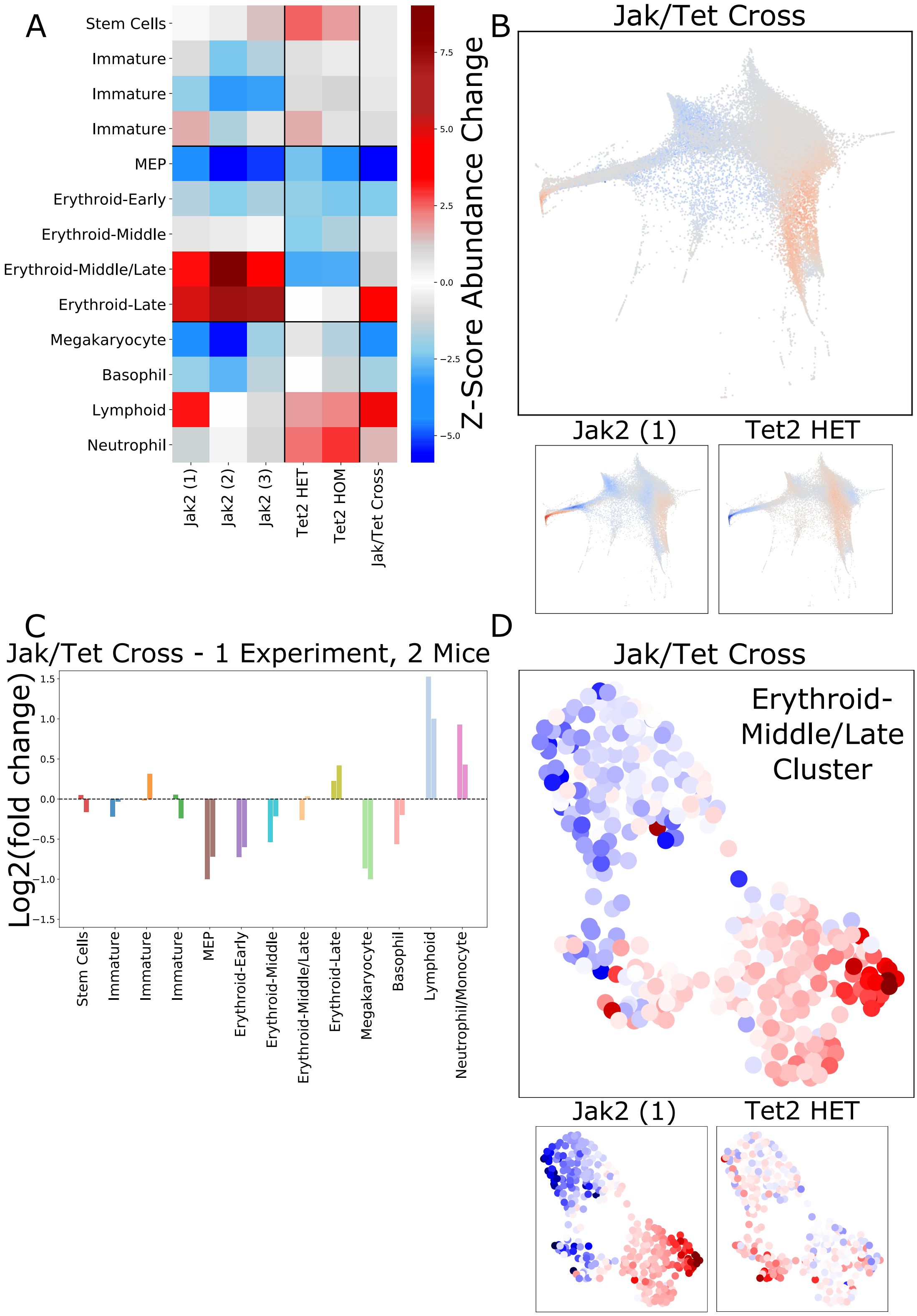


Fig. 3.14 A Jak2/Tet2 double mutant perturbation experiment reveals dissonant cellular and molecular responses. (A) Heatmap showing Z-score abundance deviations from the average WT abundance for the Jak2, Tet2 and Jak/Tet cross experiments. (B) Visualisation of the reference dataset coloured by the difference in votes received from the Jak/Tet Cross perturbed sample and the Npm1 WT sample. (C) The fold-change for each cluster in each mouse for the Jak/Tet Cross model. The Npm1 WT was used for comparison. (D) Fold changes of the 332 genes identified in section 3.4 in the 'Erythroid-Middle/Late' cluster.

In HSCs, the Jak2 mutation is believed to drive increased proliferation at the expense of self-renewal ability, whilst loss of Tet2 is thought to confer a self-renewal advantage whilst decreasing proliferation (Chen et al., 2015). The co-mutation of Tet2 with Jak2 is thought to rescue the self-renewal phenotype, leading to highly proliferative and competitive clones that can initiate myeloproliferative disease (Shepherd et al., 2018). Recently, increasing evidence for genetic interaction between the two molecules has been discovered, with Jak2 found to be responsible for phosphorylating Tet2 in early erythroid progenitors and hence directly impacting DNA methylation (Jeong et al., 2019). The Jak2 V617F mutation leads to an increase in Tet2 activity, resulting in decreased methylation levels which may lead to upregulation of pro-erythroid transcription factors and oncogenic transcripts associated with haematological malignancies (Shokouhian et al., 2020). The concomitant loss of Tet2 function in these cells would then negate this interaction and potentially result in an increase in DNA methylation, inhibiting the binding activity of pro-erythroid transcription factors due to their proportionally more CpG-rich binding motifs (Izzo et al., 2020). Therefore based on the established literature, it might be expected that the double-mutant would negate the effects of the Jak2 model. However its global impact on the blood system at both cellular and molecular scales is still unclear.

To try and understand the combinatorial effects of the Jak2 and Tet2 mutations, a perturbation experiment was performed on two mice that were homozygous for the Jak2 V617F mutation (as in the Jak2 model discussed in this chapter) and heterozygous for Tet2 loss-of-function (as in the Tet2 HET model), known as the ‘Jak/Tet Cross’. After integrating this model into the perturbation framework already described, it was clear that in terms of differential abundances the Jak2/Tet2 Cross looked like the Tet2 model in the myeloid region of the LK landscape, and like a milder version of the Jak2 model in the erythroid region (Figure 3.14A). Visualising the abundance changes on the reference dataset confirmed this (Figure 3.14B), with an overabundance of erythroid cells only occuring in the most mature erythroid cluster. These results were consistent across both of the mice sequenced (Figure 3.14C). This would suggest that the loss of Tet2 in Jak2 V617F mice does negate the amplification of erythroid progenitors, consistent with the hypothesis that Tet2 is a downstream target of Jak2. Nonetheless the later erythroid clusters are still overabundant compared to the Tet HET model alone, suggesting that the impact of the Jak2 mutation is not completely removed along the erythroid trajectory. Further evidence for this was seen in the loss of megakaryocytes observed in the Jak/Tet Cross, which was also seen in the Jak2 model. In the neutrophil/monocyte region of the landscape, where the effects of the Jak2 V617F mutation should be minimal due to a lack of EPO-receptor, the Jak/Tet Cross was very similar to the Tet2 HET model both in localisation and magnitude of the abundance changes, suggesting that outside of the erythroid trajectory, heterozygous loss of Tet2 is not impacted by the Jak2 V617F mutation. Given these findings, it was surprising to observe that the molecular response of the Jak/Tet Cross in the 'Erythroid-Middle/Late' cluster matched closely with the Jak2 model (Figure 3.14D). In particular, the Jak/Tet Cross appeared to downregulate the cluster of pro-myeloid transcription factors, whilst upregulating the clusters of pro-erythroid genes that were found to correlate with abundance changes previously.

To assess the the Jak/Tet Cross model over the entire erythroid trajectory, it was integrated into the supervised dDEG analysis performed in section 3.5. In the cluster of pro-myeloid genes, the Jak/Tet Cross behaves more like the Jak2 model than the Tet2 HET model in all genes apart from *Irf8* (Figure 3.15A). Nonetheless the magnitude of downregulation in *Mpo*, *Elane* and *Ctsg* is reduced slightly from that observed in the Jak2 model alone. Similarly for the cluster of interferon-related genes, there is greater similarity between the Jak/Tet Cross and the Jak2 models, but there is also a hint of dysregulation in the Jak/Tet Cross in genes such as *Stat1* and *Igtp* (Figure 3.15B).

Strikingly, the Jak/Tet Cross is clearly acting similarly to the Jak2 model alone in both clusters of pro-erythroid genes that were identified (Figure 3.16A,B). Hence this analysis strongly suggests that whilst the heterozygous loss of Tet2 is altering the composition of the Jak2 V617F erythroid trajectory at the cellular level, it remains largely unchanged at the molecular level, raising questions about how such a disconnect between molecular and cellular scales could occur. It is possible that the heterozygous loss of Tet2 impedes the proliferative nature of the Jak2 V617F erythroid cells - therefore altering cellular abundances - without affecting the signalling pathways that lead to the dysregulation of genes such as those in Figure 3.16. Another plausible possibility is that any pro-erythroid HSC skewing in the Jak2 model is dominated by the loss of Tet2, such that the HSC compartment is myeloid skewed and therefore the remaining erythroid cells cannot reach the abundance levels seen in the Jak2 model alone, despite normal proliferative behaviour. It is likely that both of these ideas are true to some extent, and both will be investigated further in Chapter 4. Nevertheless it is clear that the interaction between the Jak2 and Tet2 HET models is highly complex. dDEG analysis of the myeloid trajectories reveals that the Jak/Tet Cross behaves similarly to the Tet2 HET model at the molecular level. Hence the perturbation interaction is trajectory-specific; in the erythroid trajectory the loss of Tet2 negates some of the cellular effects of the Jak2 model but leaves the molecular response largely unchanged, whereas in the myeloid trajectory the loss of Tet2 dominates any effects that the Jak2 may be causing at both cellular and molecular scales. This ties in with the knowledge that different types

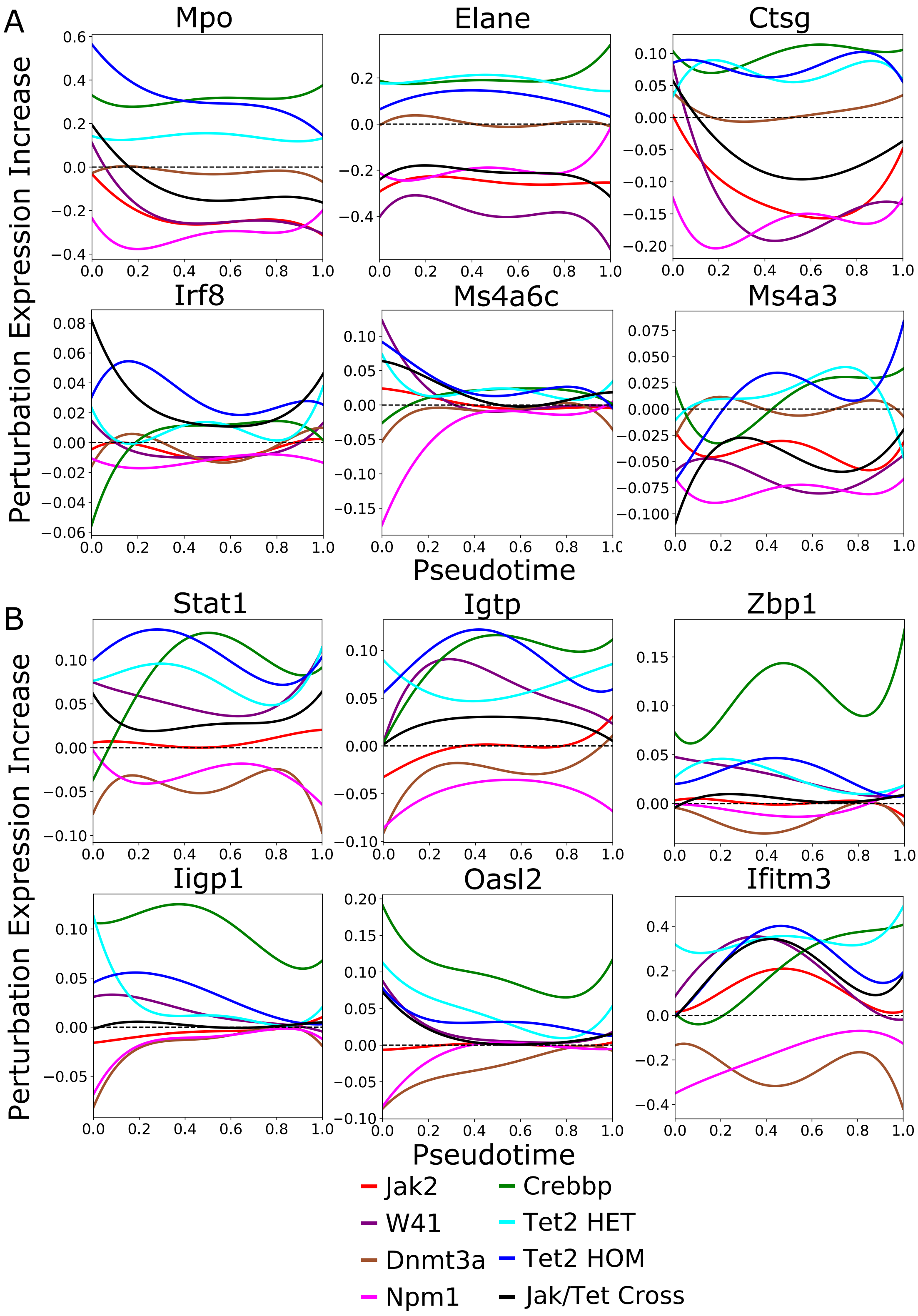


Fig. 3.15 The Jak2/Tet2 Cross model behaves like the Jak2 model at the molecular level in pro-myeloid and interferon-related genes. (A) Expression difference between perturbed and WT cells as a function of pseudotime from 0 (stem cells) to 1 (end of erythroid trajectory) for 6 pro-myeloid transcription factors. The Jak2/Tet2 Cross curve is in black. (B) Expression difference as a function of pseudotime along the erythroid trajectory for 6 interferon-related genes.

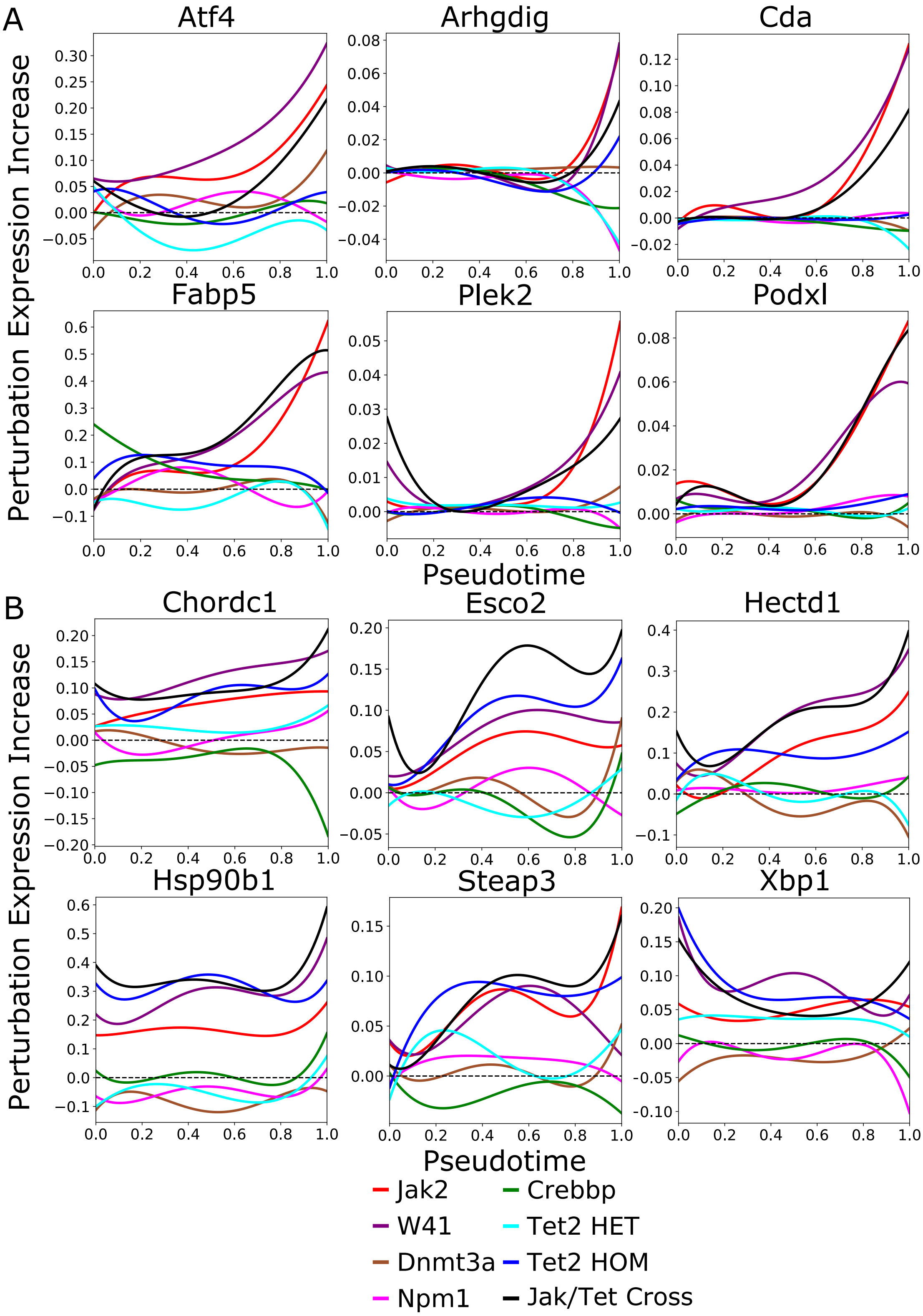


Fig. 3.16 The Jak2/Tet2 Cross model associates with the Jak2 model at the molecular level in pro-erythroid genes. (A) Expression difference between perturbed and WT cells as a function of pseudotime from 0 (stem cells) to 1 (end of erythroid trajectory) for 6 genes affecting the Jak2 and W41 models. (B) Expression difference as a function of pseudotime along the erythroid trajectory for 6 genes affecting the Jak2, W41 and Tet2 models.

of leukaemia respond differentially to certain perturbation combinations, and highlights the difficulty of understanding and treating patients with many different co-mutations.

3.7 Conclusions

This chapter describes an integrative framework for the analysis and comparison of many different scRNA-seq perturbation experiments at the scale of the entire blood system. By leveraging a large and densely sequenced dataset of WT haematopoiesis as a reference, a number of computational methods and visualisation tools have been created to allow comparison across several different perturbation models totalling nearly 250,000 single cells. These include coarse- and fine-grained approaches to calculating differential cellular abundances, a pipeline for integrating molecular information within specific cell types, and both supervised and unsupervised methods for analysing molecular dysregulation across entire differentiation trajectories. These approaches extend naturally to the study of complex perturbation interactions and therefore have a high degree of clinical relevance to pre-leukaemic haematological malignancies seen in patients.

3.7.1 The technical challenges of integrating scRNA-seq datasets

Any analysis of a stand-alone scRNA-seq dataset must contend with an array of technical effects such as doublets, cell filtering and normalisation difficulties. When trying to combine many different scRNA-seq experiments these difficulties are compounded by ‘batch effects’ such as differential sequencing depths or cell numbers across experiments. In addition, the work in this chapter requires that all the perturbation experiments cover the same LK landscape as each other; however differences in the antibody staining and FACS sorting parameters when isolating LK cells from the bone marrow will be inevitable. Add in some biological noise, as well as the fact that droplet-based sequencing methods achieve only relatively shallow sequencing depths and it is perhaps startling that any kind of biological insights can be gleaned from them, given the various competing sources of noise within the data. It should be noted, however, that the 10X Genomics pipeline results in comparatively high transcript capture compared to other droplet-based scRNA-seq methods (Giladi et al., 2018; Paul et al., 2015; Tusi et al., 2018).

Nevertheless this work has highlighted many of the important features necessary for good perturbation dataset integration. The first is the production of a high-quality reference dataset. Performing cellular-level analyses by mapping each perturbation to a reference dictates that regardless of the characteristic of each perturbation experiment, the mapping quality depends on the sampling density of the reference. Since true LT-HSCs represent less than 0.1% of WT LK cells, having a reference enriched in LSK cells was invaluable in accurately mapping cells at the top of the haematopoietic hierarchy. Secondly, the importance of having matched WT samples for each perturbation experiment cannot be overstated. As they were some of the first experiments to be performed, the W41 and Tet2 HOM models did not have matching WT samples, and at every stage this caused difficulties in their integration. Having more perturbations actually reduces this problem, as a suitable WT sample can then be paired with the unmatched perturbed sample that a) mirrors it in terms of sequencing depth and b) displays 'average' WT cellular abundances. Nonetheless differential expression comparisons between unmatched samples will always produce many more significant genes than between matched samples, and hence this must be accounted for by intersecting molecular results across more than one WT comparison.

Thirdly, both coarse-grained cluster mappings and fine-grained pseudotime alignments are useful tools for perturbation analysis. Ostensibly they serve the same purpose - to create homogeneous regions of cells across samples that are 'the same' based on their transcriptome. However due to their differences they will not always agree; a percentage of cells with the same pseudotime score may map to different clusters in the reference and vice-versa. Connecting cells across samples is key to the integration of perturbations and therefore choosing the method of alignment is of central importance in extracting signal from the noise. Work in this chapter suggests that when performing analyses across entire trajectories, pseudotime is to be preferred as it provides a finer delineation of the gradual changes occurring along a differentiation pathway. However when performing comparisons within a (supposedly) homogeneous group of cells, using clusters is arguably a better approach. For example, it would be possible to map the perturbations to a pseudotime defined on the reference dataset and then perform a differential expression analysis on all cells falling within a specific pseudotime window. This choice would nonetheless be arbitrary, and using a cluster defined through a computational method (such as Louvain clustering) designed to make the disparity between clusters as great as possible is a more data-driven approach.

3.7.2 Insights from cellular analyses

Calculating changes in the proportions of specific progenitor populations does not preclude the absolute numbers of any population changing in a different manner; for example, if the total number of LK cells is reduced due to a perturbation, a proportional increase in erythroid cells may still represent a decreased number in absolute terms. Similarly if the number of erythroid cells expands, the proportion of neutrophils may decrease despite no change in absolute numbers. Unfortunately this information is not available. It must be remembered that the perturbations analysed in this chapter are relatively small in magnitude and allow a homeostasis to be maintained in the blood system. They are therefore unlikely to dramatically change the absolute numbers of blood progenitor cells.

Analysis of the 'Stem Cell' cluster will form the basis of Chapter 4. It is nevertheless interesting to interpret this cluster's response on a global scale. The proportion of cells in this cluster remains unchanged in most models, but increases slightly in the Crebbp and Tet2. Tet2 KO is known to confer a self-renewal advantage in LT-HSCs, suggesting this may be due to a lack of differentiation activity rather than caused by excess proliferation of the most immature haematopoietic cells. The three clusters marked as 'immature' do not show high levels of marker gene expression for any specific single lineage, but based on their transcriptome they are likely to represent early stages of myeloid or lymphoid differentiation as opposed to erythroid, since they express low levels of markers such as *Mpo* and *Dntt*, but show almost no expression of erythroid transcription factors. Again these clusters show minimal proportional changes compared to more mature populations. The most myeloid-looking of the immature clusters (the orange cluster in Figure 3.1) is reduced in abundance in the Jak2 model, potentially connected to the reduction in neutrophils seen in Figure 3.4.

One of the more perplexing results is that the proportion of MEPs is reduced in both the Jak2, W41 and Tet2 models, despite the opposite behaviours of the later erythroid and megakaryocyte populations. It is difficult to arrive at any conclusions about this from scRNA-seq snapshot data alone. Assuming the myeloid/erythroid fate decision is occurring upstream of the MEP population, then possibly the Jak2, W41 and Tet2 models are skewed against erythroid cells, with the Jak2/W41 only recovering to have an overabundance of later erythroid progenitors due to an increased proliferation program downstream, which may itself be the cause of a skewing towards erythroid cells in an erythroid/megakaryocyte fate decision. However since no myeloid expansion is seen in the Jak2/W41, this scenario seems likely to only apply to the Tet2 model. It is also plausible that there is pro-erythroid skewing of immature cells in these models, with a loss of MEPs caused by rapid differentiation through this state. The latter explanation would appear to be supported in the Jak2/W41 by the molecular results suggesting genes upregulated above WT levels in the MEPs are associated with increased metabolic activity. However an explanation must then be found for why the Crebbp model has an overabundance of MEPs despite otherwise mirroring the Tet2 model. Fitting all of the data into a single model will require a new wave of lineage tracing perturbation experiments (Pei et al., 2020; Weinreb et al., 2020).

3.7.3 Insights from molecular analyses

Integrating molecular signals across perturbations necessarily depends on some degree of overlap existing between experiments. The approach taken in section 3.4 discards genes that were only differentially expressed in one perturbation model. Hence to properly analyse a single model in detail, an approach such as the unsupervised dDEG method introduced in section 3.5 must be used. Due to the relatively shallow depth of 10X scRNA-seq (or any droplet-based method), performing differential expression across an entire trajectory has far more power to identify upregulated genes than downregulated genes. This can be seen clearly in Figure 3.12C, where only one of the eight centroids identified contains a majority of downregulated genes. Unsurprisingly those genes which are found to be downregulated are generally highly expressed across the trajectory. The decision to measure the degree of differential expression using absolute expression difference also favours the identification of highly expressed genes, which will naturally be able to vary more in absolute terms; using fold changes instead favours lowly expressed genes where small differences can lead to big fold changes, many of which will not be biologically relevant.

Permutation testing was performed to assess the probability of a gene passing the threshold for dynamic differential expression due to chance. Using the Jak2 experiments, the aligned erythroid trajectories containing both WT and perturbed cells had both their cell labels and their pseudotime values randomly permuted before the dDEG method was performed. Using the same threshold for dynamic differential expression as in section 3.5, on average only 2-5 genes were identified as significant for any given permutation. Permuting only the pseudotime values (and not the cell labels) still identifies many of the dDEGs since their genotype information has not changed, however almost always with different dynamic shapes. These results suggest that the dDEGs identified and their shapes are highly unlikely to occur by chance, even when considering only a single experiment. In order to then compare the dynamic behaviour of a gene(s) across perturbations, it is essential that the same pseudotime value also represents the same differentiation stage across different experiments. This may not be true if the c-Kit FACS threshold differs between experiments or if there is highly varying cell densities in different models along the trajectory in question. To assess this, the computational tool cellAlign was used to check whether the same pseudotime value corresponded to the same expression states across perturbations, and to correct the pseudotime values if necessary (Alpert et al., 2018). Minimal correction was needed for all experiments. The alignment of pseudotime values was checked visually by comparing the expression profiles of marker genes across the trajectory.

The behaviour of the pro-myeloid enzymes such as *Mpo* and *Elane* strongly suggests that they play a role in many different perturbation responses across the entire LK landscape. It is interesting to speculate that during an initial myeloid/erythroid fate decision - when a multipotent cell is stochastically deciding to follow either a myeloid or erythroid differentiation program - the pro-erythroid perturbations such as Jak2 may either reduce these genes' ability to be transcribed or make their transcription more variable, creating a higher probability that in their absence, pro-erythroid signals can take hold. Studies have recently suggested that this molecular restriction could be achieved at the epigenetic level, implicating Jak2, Dnmt3a and Tet2 in the process (Izzo et al., 2020). It may also be that dysregulation of these genes confers latent plasticity to cells that would otherwise be committed in WT haematopoiesis; for example, early Tet2 MEPs may still have myeloid potential. Some of these ideas will be considered further in Chapter 4.

3.7.4 Future Work

This chapter presents a wide range of novel biological hypotheses that could form the starting point of further work. Further experimental validation of all the molecular results in this chapter is required. For example, the role of interferon-related genes in erythroid haematopoiesis and why they are dysregulated in only the W41, Tet2 and Crebbp models is currently unknown. Similarly the stress/heat shock/unfolded protein response seen only in the Jak2, W41 and Tet2 models requires further investigation. The cellular and molecular behaviour of the most immature HSCs and MPPs will form the basis of Chapter 4.

For each individual model, the dDEG method reveals clusters of genes upregulated at different stages of a trajectory. A powerful implication of this is that genes from a cluster upregulated at an earlier stage may be mechanistically driving those upregulated at a more mature stage. Investigating these potentially causal correlations along a trajectory would allow gene regulatory networks to be built that evolve over the course of a trajectory. Therapeutic targets designed to disrupt a specific stage of differentiation could then be designed.

3.7.5 Summary

In summary, the work in this chapter has built a computational framework for the integration of many different scRNA-seq perturbation datasets. The behaviour of >10 different experiments has been analysed at the scale of the entire blood system, with a variety of intriguing biological results discovered using a spate of new computational methods and tools for perturbation analysis.

Chapter 4

Dissecting the Pre-Leukaemic Perturbation Response of the Most Immature Haematopoietic Progenitors

Experimental work for this project was carried out by Nicola Wilson (isolation of primary bone marrow cells, scRNA-seq profiling). Initial preprocessing of the resulting data was carried out by Rebecca Hannah (running 10X Genomics CellRanger pipeline). After production of the raw count matrices from the CellRanger pipeline, all computational work was carried out by Sam Watcham, except where explicitly stated in the text.

4.1 Background

Elucidating the behaviour of the most immature blood progenitors, including true LT-HSCs capable of reconstituting the entire blood system from a single cell, has been a central goal in haematopoietic research for many decades. Prior to the advent of scRNA-seq, the classical haematopoietic hierarchy had three multipotent subpopulations at its peak: LT-HSCs, ST-HSCs and MPPs, defined primarily through their decreasing repopulation potential and their ability to develop into any of the main haematopoietic lineages (Eaves, 2015; Oguro et al., 2013). Immunophenotyping hinted at functional heterogeneity within the MPP compartment (Wilson et al., 2008), before single-cell technologies helped to cement the model of a continuous transcriptional landscape within the HSC and MPP compartments (Kowalczyk et al., 2015; Nestorowa et al., 2016) that forms the basis for haematopoiesis today.

Single-cell studies have highlighted that even within the FACS populations thought to best isolate LT-HSCS, there is significant transcriptional heterogeneity (Cabezas-Wallscheid et al., 2014; Dykstra et al., 2007; Wilson et al., 2015). The largest contribution to this heterogeneity arises from cell-cycle effects, with the 'best' repopulating LT-HSCs existing in deeply quiescent state marked by low biosynthetic activity (Cabezas-Wallscheid et al., 2017; Wilson et al., 2008). Transcriptional signatures corresponding to the most LT-HSC like states have been identified (Wilson et al., 2015), allowing a proxy for 'stemness' to be calculated from single-cell transcriptomes (Hamey and Gottgens, 2019). The evidence for lineage output bias in functional HSCs is still unclear, but several studies have hinted at the existence of megakaryocyte-biased HSCs and suggested there may be a direct route of differentiation from HSCs to megakaryocytes that exists in addition to the classical route through a bipotent MEP state (Grover et al., 2016; Rodriguez-Fraticelli et al., 2018; Sanjuan-Pla et al., 2013). Nevertheless transcriptionally, LT-HSCs do not exhibit clear lineage priming in native haematopoiesis, despite evidence for HSC-skewing from transplantation experiments (Benz et al., 2012; Muller-Sieburg et al., 2004).

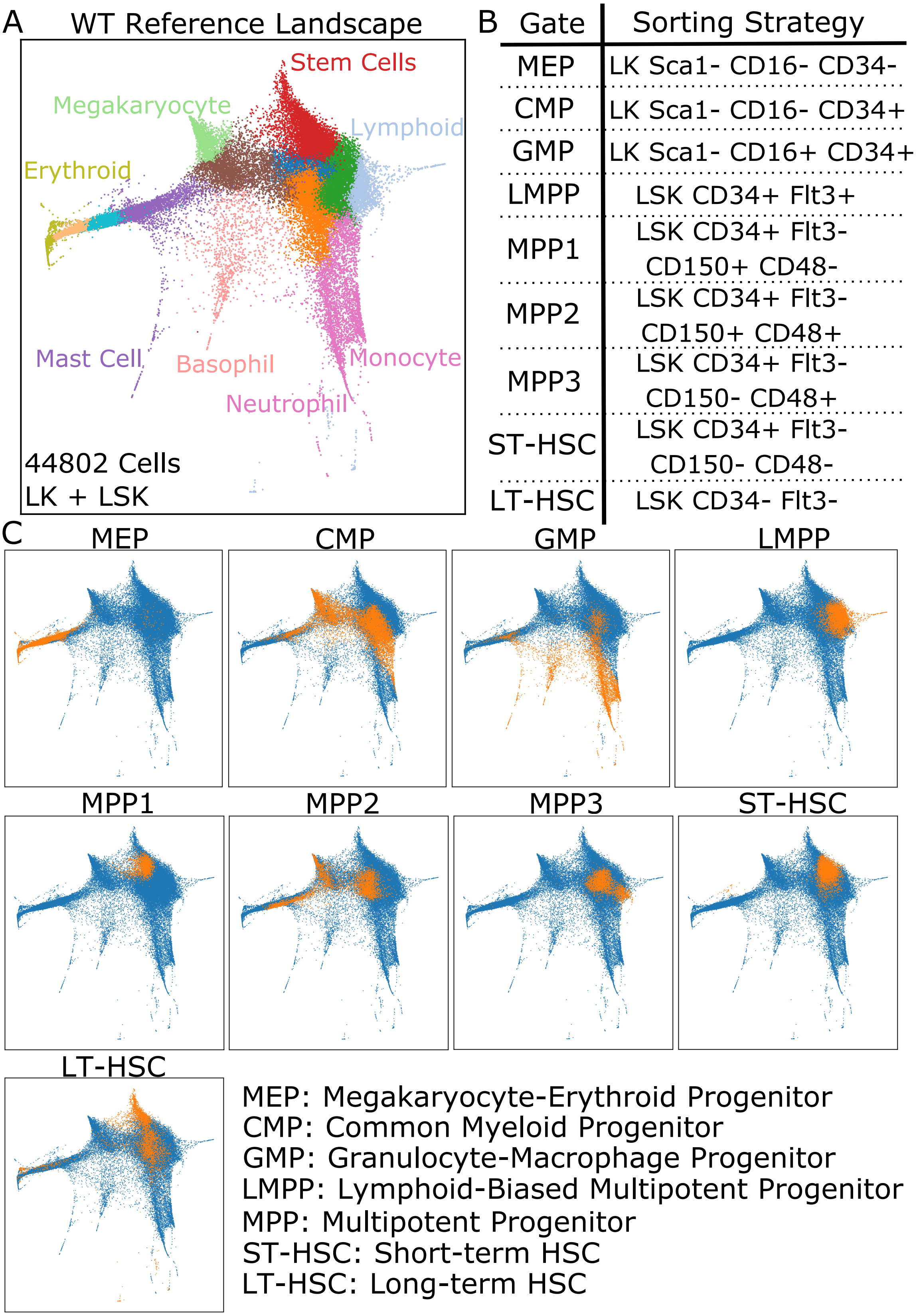
Within the classical ST-HSC/MPP compartments, scRNA-seq identified several subpopulations (known widely as MPP1-4) that exhibited clear transcriptional and functional biases, which have been confirmed *in vivo* using lineage barcoding (Rodriguez-Fraticelli et al., 2018). MPP2 is believed to be primed towards the erythroid/megakaryocytic fates, MPP3 towards the myeloid lineages, and MPP4 (also known as LMPP) towards a lymphoid fate (Pietras et al., 2015). This suggests that in truth, many key fate decisions are initiated upstream of these cells in either the MPP1 or LT-HSC states, although only cells in the latter state are potentially capable of long-term reconstitution of the blood system. However identifying these states on a transcriptional landscape is not trivial, especially since the acquisition of broad, unbiased droplet-based scRNA-seq datasets does not currently allow for the parallel capture of index sorting measurements. Building hypotheses that link transcriptomic states to functional output using scRNA-seq data alone is therefore very difficult, as it is impossible to know whether any observed lineage priming is actually the result of including already-committed cells in the analysis. Lineage tracing experiments are beginning to be able to answer these questions in native haematopoiesis (Weinreb et al., 2020). Furthermore, it has also been suggested that LT-HSCs do not contribute greatly to steady-state haematopoiesis, but rather that they preferentially contribute during injury response or transplantation (Busch et al., 2015; Rodriguez-Fraticelli et al., 2018). Instead it is the MPP1 compartment that take care of the majority of steady-state blood production.

Despite these technical limitations, scRNA-seq perturbation experiments provide an extremely valuable tool for discovering how specific mutations affect uncommitted HSC/MPP1 progenitors in a native setting (Izzo et al., 2020; Psaila et al., 2020). To date, minimal work has been performed in this setting that does not rely on prospectively sorting a specific group of cells (Ostrander et al., 2020; Shepherd et al., 2018). The aim of this work was to characterise the uncommitted progenitors from different pre-leukaemic perturbation models at both cellular and molecular levels, and to look for any evidence of lineage priming or other signals that correlated with the observed downstream progenitor abundance changes.

4.2 Cross-platform mapping delineates the HSC and MPP compartments in wild-type haematopoiesis

A major difficulty in building testable hypotheses from droplet-based scRNA-seq data stems from being unable to capture index-sorting measurements from each cell prior to sequencing. Recent multimodal techniques such as CITE-seq have begun to make this possible, but were not commercially available during this project (Stoeckius et al., 2017). Since index measurements are used to sort cells for functional assays, it is therefore not trivial to determine which cells or clusters in a scRNA-seq transcriptomic landscape correspond to classically defined populations. However, recent computational developments have led to a number of tools that can integrate single-cell datasets across different platforms, species and modalities. These include Seurat (Stuart et al., 2019), BB-KNN (Polanski et al., 2020) and MNN-Correct (Haghverdi et al., 2018) (see section 1.5.1). Tools such as these can potentially be used to map index-sorted cells from plate-based scRNA-seq platforms onto a droplet-based transcriptomic landscape, and hence provide an estimate of which classically defined FACS population each droplet-based cell would have belonged to.

Following this approach, a novel computational pipeline was built to map the Nestorowa et. al. dataset (Nestorowa et al., 2016), containing 1,654 index-sorted HSPCs sequenced using SmartSeq2, onto the WT reference landscape (Figure 4.1A). The HSPCs were collected from nine mutually exclusive FACS gates (Figure 4.1B). Whilst the exact gates used to define different MPP populations differ slightly across different research groups, these gates correspond almost exactly to the widely-accepted delineation of the MPP and HSC compartments used in Fraticelli et. al (Rodriguez-Fraticelli et al., 2018). The only differences are that 'LMPP' is used in Nestorowa et. al. instead of 'MPP4', and the 'MPP1' gate in Fraticelli et. al. is split into both 'MPP1' and 'ST-HSC' in Nestorowa et. al. The mapping was performed by integrating the datasets and performing MNN-Correct on the joint PCA



Gate	Sorting Strategy
MEP	LK Sca1- CD16- CD34-
CMP	LK Sca1- CD16- CD34+
GMP	LK Sca1- CD16+ CD34+
LMPP	LSK CD34+ Flt3+
MPP1	LSK CD34+ Flt3- CD150+ CD48-
MPP2	LSK CD34+ Flt3- CD150+ CD48+
MPP3	LSK CD34+ Flt3- CD150- CD48+
ST-HSC	LSK CD34+ Flt3- CD150- CD48-
LT-HSC	LSK CD34- Flt3-

MEP: Megakaryocyte-Erythroid Progenitor
CMP: Common Myeloid Progenitor
GMP: Granulocyte-Macrophage Progenitor
LMPP: Lymphoid-Biased Multipotent Progenitor
MPP: Multipotent Progenitor
ST-HSC: Short-term HSC
LT-HSC: Long-term HSC

Fig. 4.1 Mapping index-sorted SmartSeq2 HSPCs onto the 10X reference landscape. (A) The WT reference landscape, coloured by Louvain clusters. (B) The sorting strategies used in Nestorowa et. al. to collect murine HSPCs from classically-defined populations using SmartSeq2. (C) For each classical population, the 10X reference cells that map most closely to it are coloured in orange. LK: Lin- c-Kit+; LSK: Lin- Sca1+ c-Kit+.

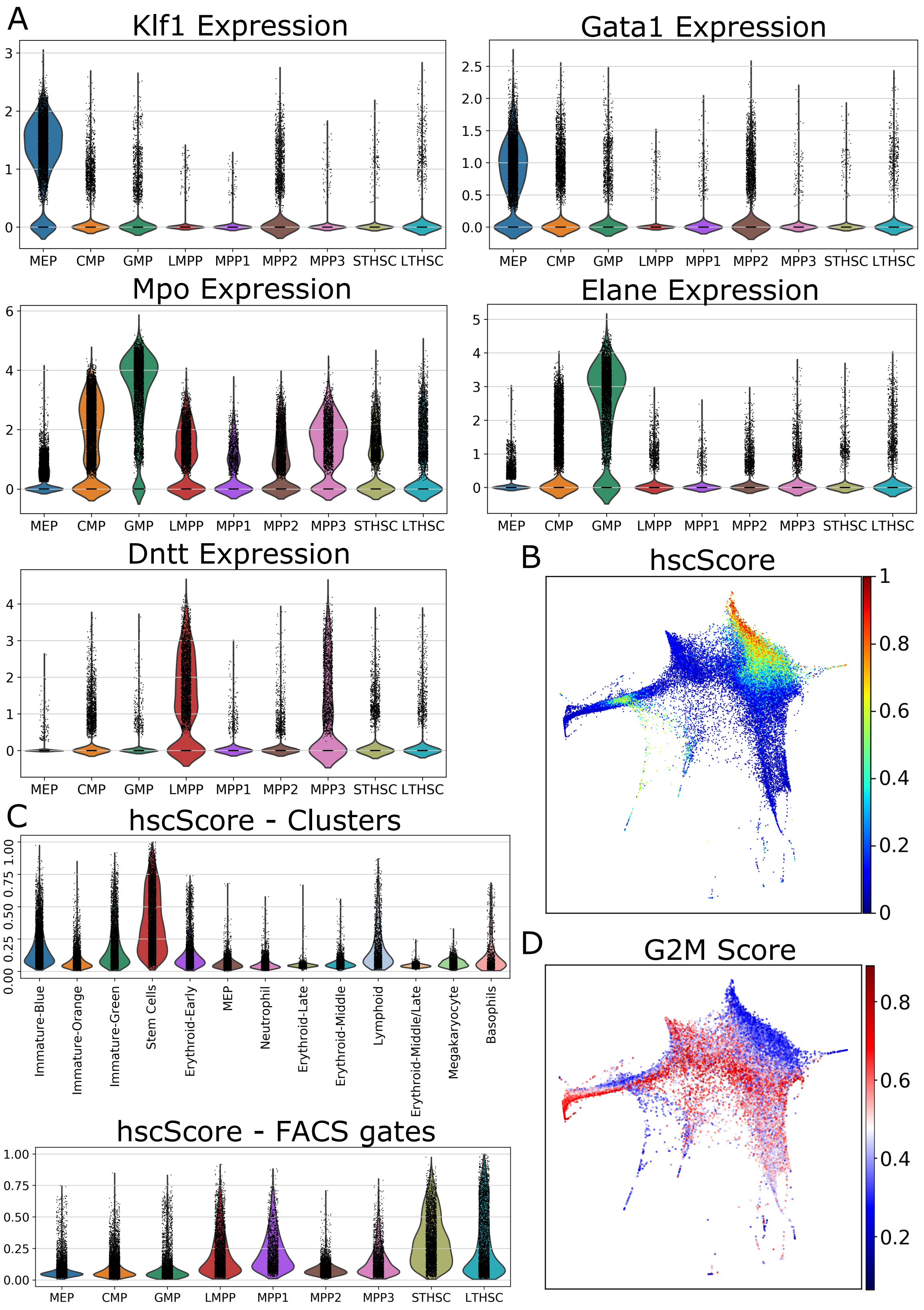


Fig. 4.2 Downstream MPP compartments are lineage primed in 10X WT data, while HSCs are quiescent. (A) Violin plots of lineage marker expression for each FACS gate mapped to the 10X reference landscape. (B) The hscScore, a measure of 'stemness', for each reference cell. (C) Violin plots of the hscScore split by Louvain clusters and mapped FACS gates. (D) The G2M score, a measure of cell cycle activity, for each reference cell.

space of the data (see methods). Each reference cell was then given an adjacency score for each FACS gate, weighted so that the total adjacency score for each gate across the entire reference dataset was the same. Each reference cell was then defined as belonging to the FACS gate with the highest adjacency score.

The result of this mapping appeared to be qualitatively sensible, with the most mature gates such as GMPs mapping in a high coordinated fashion to the expected region of the reference landscape that contained the neutrophil, monocyte and basophil progenitors (Figure 4.1C). In addition the delineation of the MPP and HSC compartments seemed sensible, with MPP2 located towards the erythroid trajectory, MPP3 towards the myeloid trajectories and LMPP (MPP4) towards the lymphoid trajectory, whilst MPP1 and ST-HSC were located nearer the top of the landscape containing the LT-HSCs. This behaviour directly matches with the functional and transcriptomic lineage priming of the MPP compartment expected from Fraticelli et. al (Rodriguez-Fraticelli et al., 2018). This was further confirmed by the expression of known erythroid, myeloid and lymphoid markers across the mapped FACS populations (Figure 4.2A). The erythroid transcription factors *Klf1* and *Gata1* were expressed in MEPs as expected, but also exhibit expression in MPP2, whilst having very little expression in any other MPP or HSC compartment. Similarly, the myeloid marker *Mpo* had greater expression in MPP3 than any other compartment (aside from GMP), and the lymphoid marker *Dntt* had the highest expression in LMPP (MPP4). Therefore the mapping appeared to have worked well, and provides confidence that the cells marked as LT-HSC, ST-HSC and MPP1 are likely to correspond to the functionally uncommitted and fully multipotent cells within the reference dataset.

However, the mapping did highlight some strange results, including that the MEP gate mapped to the later regions of the erythroid trajectory instead of the Louvain cluster annotated as MEP - which expressed both erythroid and megakaryocytic lineage markers. This cluster was instead mapped as belonging to the CMP gate. In addition, the MPP2 gate contained a number of cells that were located in the more mature regions of the erythroid and megakaryocyte trajectories. This again raises the question of whether the pro-erythroid lineage priming seen in the MPP2 compartment arises due to the presence of more mature, already committed cells still residing within this gate. The technical noise and limitations of the mapping process notwithstanding, it is unclear how to interpret this result in light of the published functional data on the MPP2 gate (Pietras et al., 2015; Rodriguez-Fraticelli et al., 2018).

As expected, many of the reference cells mapping to the LT-HSC gate were located at the top of the transcriptomic landscape, and were co-located with the expression of known

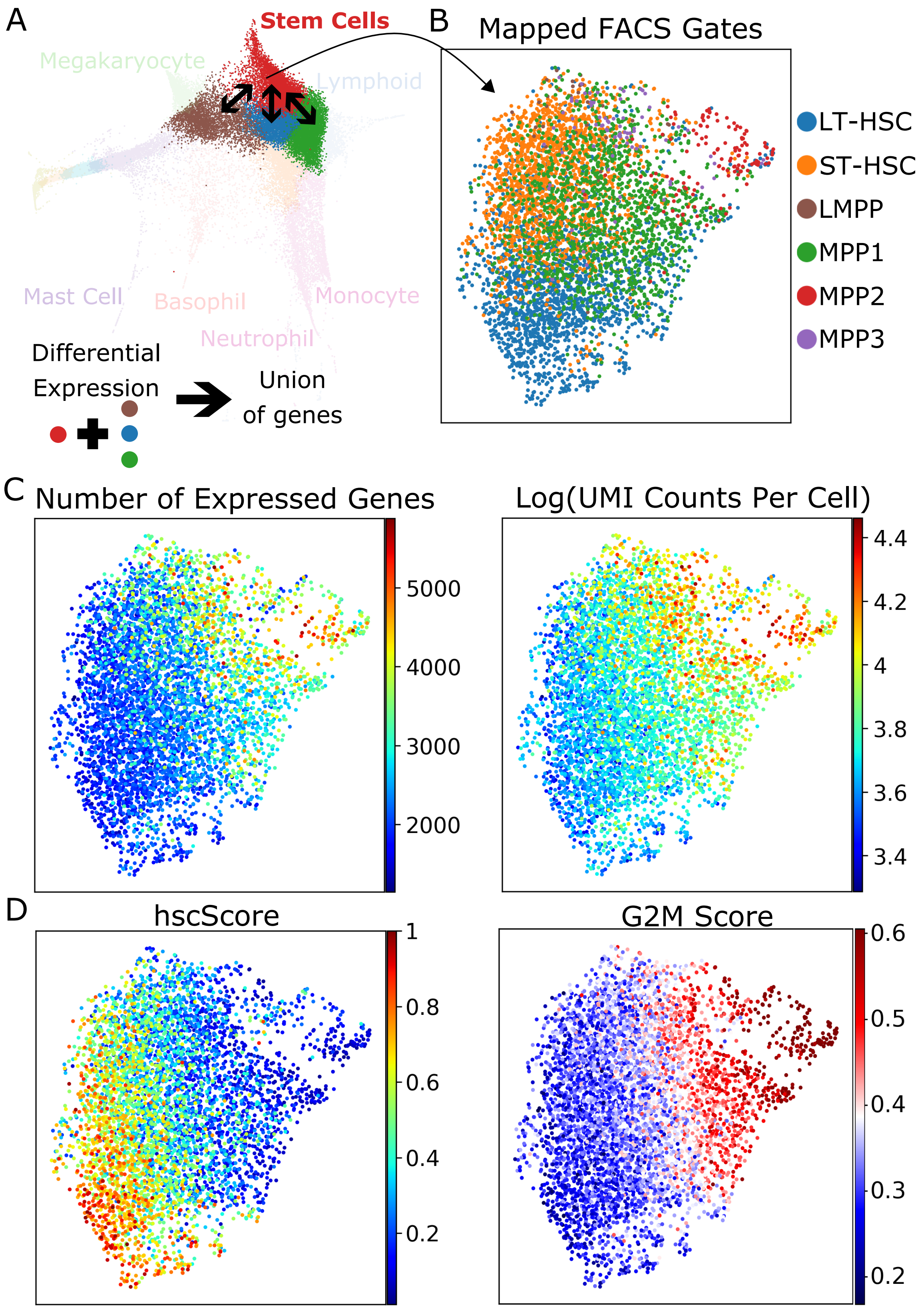


Fig. 4.3 Visualising the uncommitted progenitors within the WT reference data. (A) A geneset for visualisation of the Stem Cell cluster was created by taking the union of DEGs between the Stem Cell cluster and it's three neighbouring clusters. (B) The mapped FACS populations for the Stem Cell cluster. (C) The number of expressed genes and the $\log_{10}$(UMIs/cell) for each cell in the Stem Cell cluster. (D) The hscScore and G2M score for each cell.

HSC genes such as *Procr* (EPCR) and *Fgd5*. However, surprisingly some cells from the other immature clusters also mapped to the LT-HSC gate. To test where the best LT-HSCs reside within the landscape, the recently published hscScore method was applied (Hamey and Gottgens, 2019), which uses a machine learning model based on the expression of the set of MolO genes identified as marking functional murine LT-HSCs (Wilson et al., 2015). This provided a per-cell hscScore, with 1 corresponding to the 'best' HSC in the dataset. Plotting these scores on the reference revealed that the best HSCs were indeed located at the top of the landscape (Figure 4.2B), suggesting that the more mature cells identified as LT-HSC in Figure 4.1C were anomalous. Splitting the hscScore by Louvain cluster and mapped FACS gate revealed that, as expected, the 'Stem Cells' cluster and the ST-HSC/LT-HSC gates contain the best functional stem cells (Figure 4.2C).

HSCs with the greatest ability to reconstitute the blood system are associated with a quiescent signature and a lack of cell cycle activity (Cabezas-Wallscheid et al., 2017). To visualise this, a G2M score was designed to provide a proxy for cell cycle activity by calculating the geometric mean scaled expression of 198 genes known to correlate with the G2 and M phases of the cell cycle. Hence cells with a high G2M score are likely to be actively cycling. Plotting this score on the reference dataset clearly highlighted the correlation between HSC gene expression and quiescence (Figure 4.2D), providing further evidence for the identity of the true LT-HSCs within the dataset. MPP1 cells had noticeably higher G2M scores than either the LT-HSC or ST-HSC gates. The force-directed graph on which the G2M score is displayed was calculated after removal of cell-cycle correlated genes (see methods).

Considering these results, it was decided that the 'Stem Cell' Louvain cluster represented the best estimate for the group of uncommitted cells in the WT reference. It likely contains all the true LT-HSCs, as well as containing over 85% of the MPP1 cells and over 65% of the ST-HSCs identified from the cross-platform mapping. In order to extract as much structure from this cluster as possible, it was revisualised on a UMAP embedding using a geneset constructed from the union of all DEGs between the Stem Cell cluster and its three neighbouring clusters (Figure 4.3A). This pulled apart the different FACS populations (Figure 4.3B), and revealed how a small number of MPP2, MPP3 and LMPPs (MPP4) also reside in this cluster. These cells express minimal levels of all lineage marker genes. Visualising the number of expressed genes and UMI counts per cell alongside the hscScore and G2M score for each cell revealed a clear progression from LT-HSC to ST-HSC to MPP1 in terms of both stemness and quiescence (Figure 4.3C,D). The ST-HSC and MPP1 are known to be functionally very similar (Pietras et al., 2015), however there was a clear shift to higher levels of transcription, mRNA content and cell cycle activity between the two populations.

Whether this is a clear transition or simply a proportion of cells exiting quiescence is difficult to determine from snapshot data. Nevertheless it is clear that whist the overall structure of the uncommitted progenitors is rather flat and featureless, there is clear substructure driven largely by cell cycle activity and the downregulation of genes associated with a functional HSC state.

4.3 Pre-leukaemic perturbations push the uncommitted progenitor pool into a more active state

To assess the magnitude and direction of cellular abundance shifts in uncommitted progenitor cells due to the perturbations, the paired voting method introduced in Chapter 3 was applied using only the cells from each experiment that had mapped to the Stem Cell cluster. The results were then visualised on the UMAP embedding of the reference Stem Cell cluster (Figure 4.4A). Many models displayed a shift towards a greater abundance of cells in more active, cycling states at the expense of cells in a quiescent state. In addition, model-specific behaviour was seen. The Jak2 model displayed a clear abundance decrease in the LT-HSC region, consistent with the known phenotype of the model from functional data (Li et al., 2014; Shepherd et al., 2018). This was accompanied by a broad increase of MPP1 cells. By comparison the Tet2 HET and Tet2 HOM models displayed no reduction of LT-HSCs, and instead exhibited a shift from less active ST-HSC/MPP1 states to more active MPP1 states. The Jak/Tet cross displayed a combination of features from the Jak2 and Tet2 models; it was the only other model to show a clear reduction of LT-HSCs, but its shift from a less-active to more-active MPP1 states mirrored the Tet2 models rather than the Jak2 model. This suggests that both mutations are altering the transcriptomic landscape of these mice starting from the most immature regions of the haematopoietic hierarchy.

Both the Dnmt3a and W41 models also displayed a clear shift away from ST-HSC states towards more active MPP1 states. A hint of the same pattern is seen in the Crebbp model, albeit at a very low magnitude. Notably the Npm1 model and to some extent the p53 model displayed the opposite shifts, away from the most active MPP1 states. Overall, these abundance shifts were smaller in magnitude than those associated with the LK landscape as a whole (Figure 4.4B), but were still larger than the differences between any pair of WT samples.

Current models of haematopoiesis suggest that under steady-state conditions it is the MPP1 cells that replenish the majority of the blood system, with only occasional input

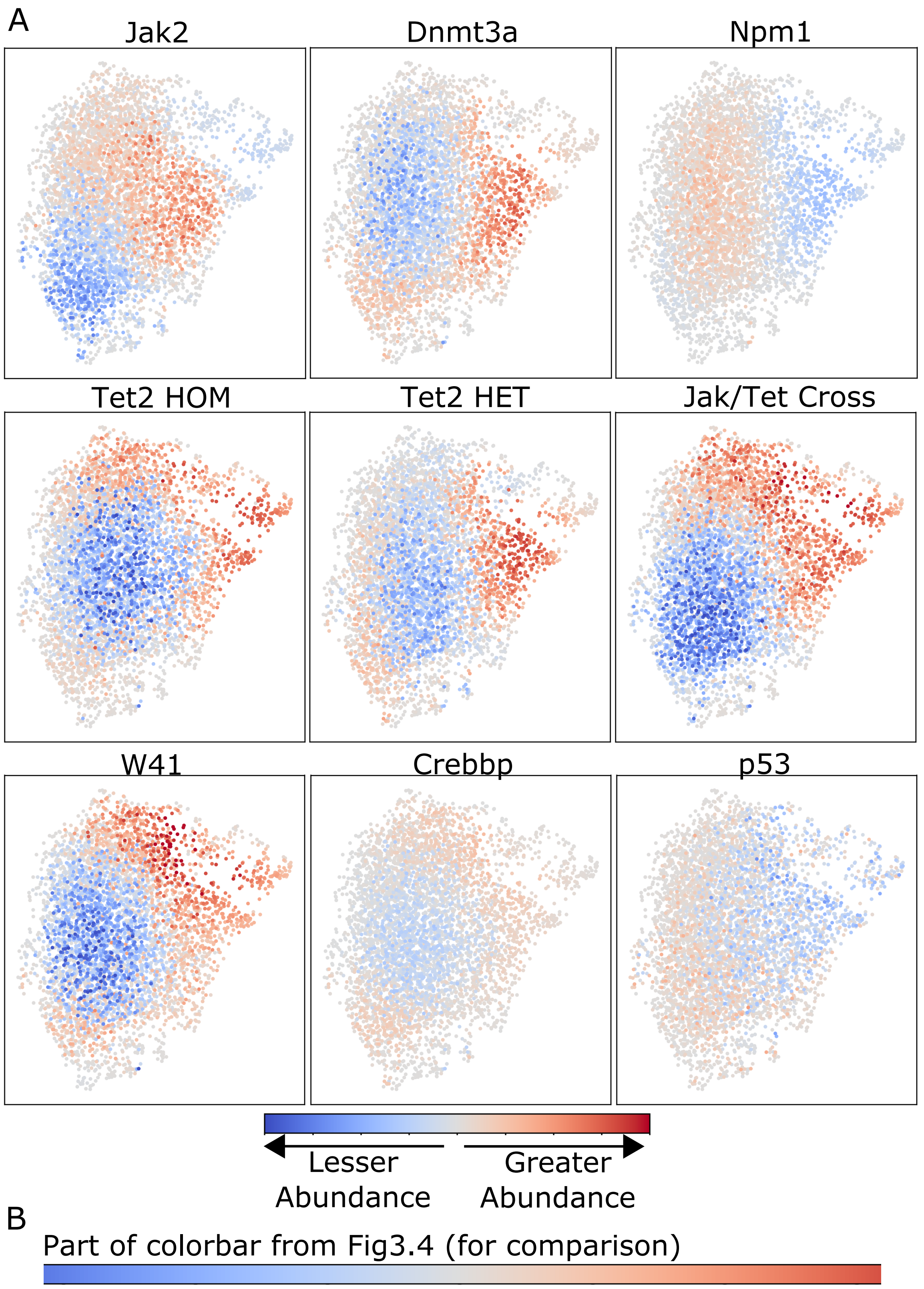


Fig. 4.4 Pre-leukaemic perturbations push uncommitted cells into a higher activity state. (A) Abundance changes within the 'Stem Cell' cluster visualised on the WT reference using the paired voting method described in Chapter 3. Red indicates greater abundance, blue lesser abundance. For each model, the average changes over biological and/or technical repeats is shown. (B) The colorbar from Figure 3.4 to scale with the colorbar from (A), indicating the relatively smaller abundance changes observed.

from LT-HSCs (Busch et al., 2015; Laurenti and Göttgens, 2018; Rodriguez-Fraticelli et al., 2018). An expansion of actively cycling MPP1 cells due to a perturbation may therefore allow mutated clones to out-compete WT clones when the mutation occurs in haematological malignancies. This would explain, for example, how the Jak2 mouse model displays a strong perturbation phenotype despite the fact that the Jak2 LT-HSCs do not exhibit any competitive advantage over WT LT-HSCs (Shepherd et al., 2018). Notably the abundance changes within the Stem Cell cluster did not correlate with downstream abundance changes in the erythroid or myeloid lineages, suggesting that an MPP1 expansion can occur regardless of the ultimate direction of the fate-skewing within the MPP1 compartment. Hence within the uncommitted progenitor pool, the main effect of cellular abundance shifts in response to perturbation is to drive the expansion of the perturbed blood system through a proportional increase in the most active MPP1 states, rather than to skew towards specific differentiation trajectories.

Despite this, molecular integration of cells mapping to the Stem Cell cluster revealed patterns of differential expression that appeared to correlate with observed differentiation skews. For each perturbation experiment, differential expression was performed between WT and perturbed samples in the Stem Cell cluster, and integrated by intersecting each pair of perturbations as in Chapter 3. When looking at the proportion of overlapping DEGs that are regulated concordantly, the three Jak2 experiments displayed a largely conserved molecular signature that also overlapped to a large degree with the W41 model (Figure 4.5A). Similarly, the Crebbp, Tet2 HET and Tet2 HOM models also share a molecular signature whilst being largely anti-correlated with the Jak2 model. Taken together, this hints at molecular changes which may be responsible for the observed downstream differentiation skews. However, the Dnmt3a and Npm1 models did not share a signature with each other or with the Jak2/W41 models, as might be expected given the similarities of these models in more mature populations. In addition, it is clear that the overall number of DEGs in the Stem Cell cluster is small, with the size of many overlaps between perturbations being less than 25 genes (Figure 4.5B). This lack of DEGs is not due to a lack of cells - cell numbers are comparable in magnitude to the erythroid cluster analysed in Chapter 3 - but it may be driven in part by smaller number of genes expressed in the most immature blood cells. Nevertheless it was clear that the transcriptional consequences of the perturbations were considerably smaller in magnitude than those seen in more mature populations, and Figure 4.5 should be interpreted in this context. Strikingly, the Jak/Tet Cross appeared to overlap strongly with both the Jak2 and the Tet2 models. It was also the model with the largest number of DEGs in the Stem Cell cluster, suggesting that the two driver mutations had combined to produce a larger degree of transcriptional perturbation than either of the mutations had managed alone.

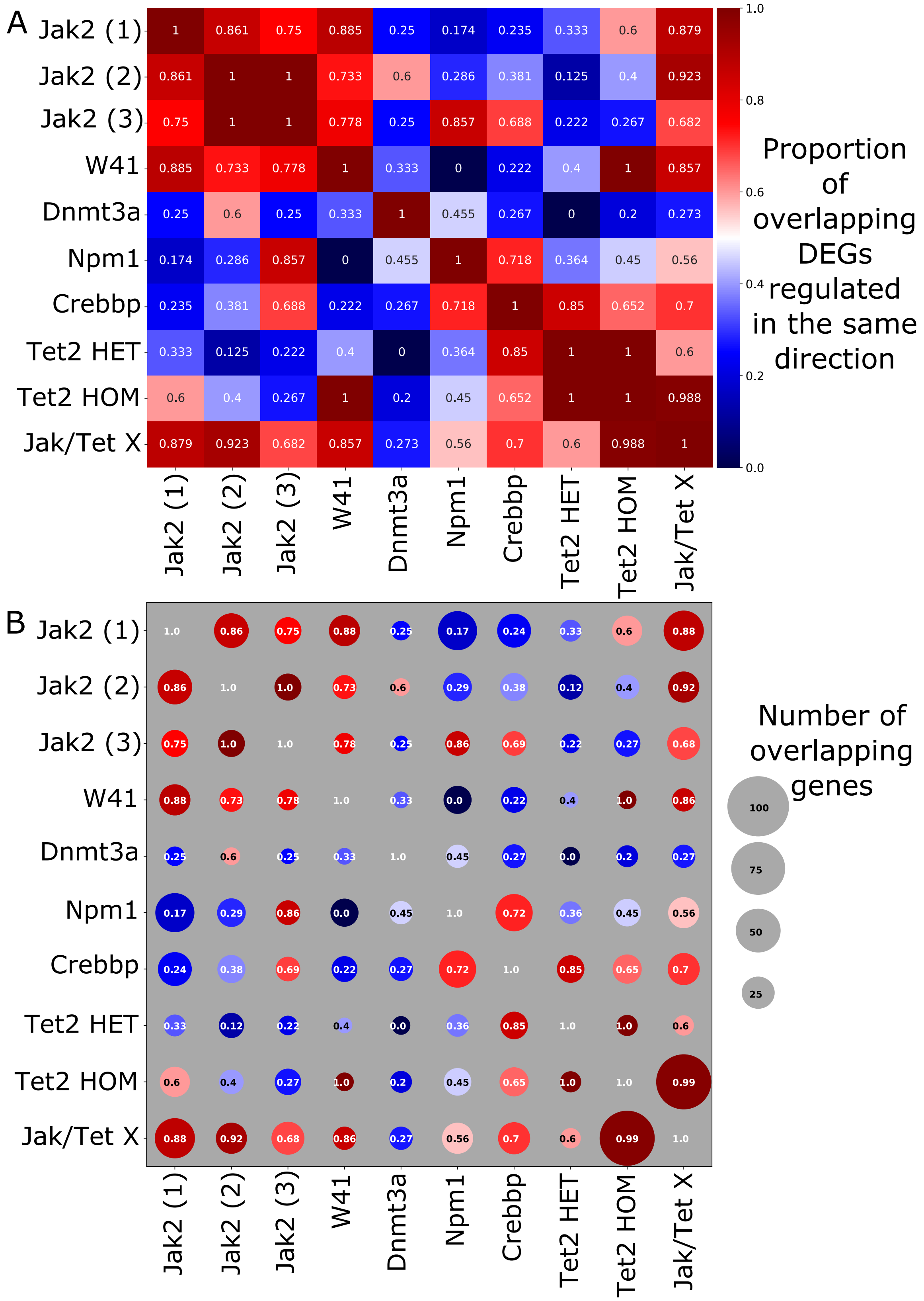


Fig. 4.5 Molecular changes of uncommitted progenitors loosely correlate with abundance outcomes (A) For each pair of perturbations, the heatmap is coloured by the proportion of overlapping DEGs in the Stem Cell cluster that are regulated in the same direction in both models. (B) The corresponding size of the overlap between each pair of perturbations.

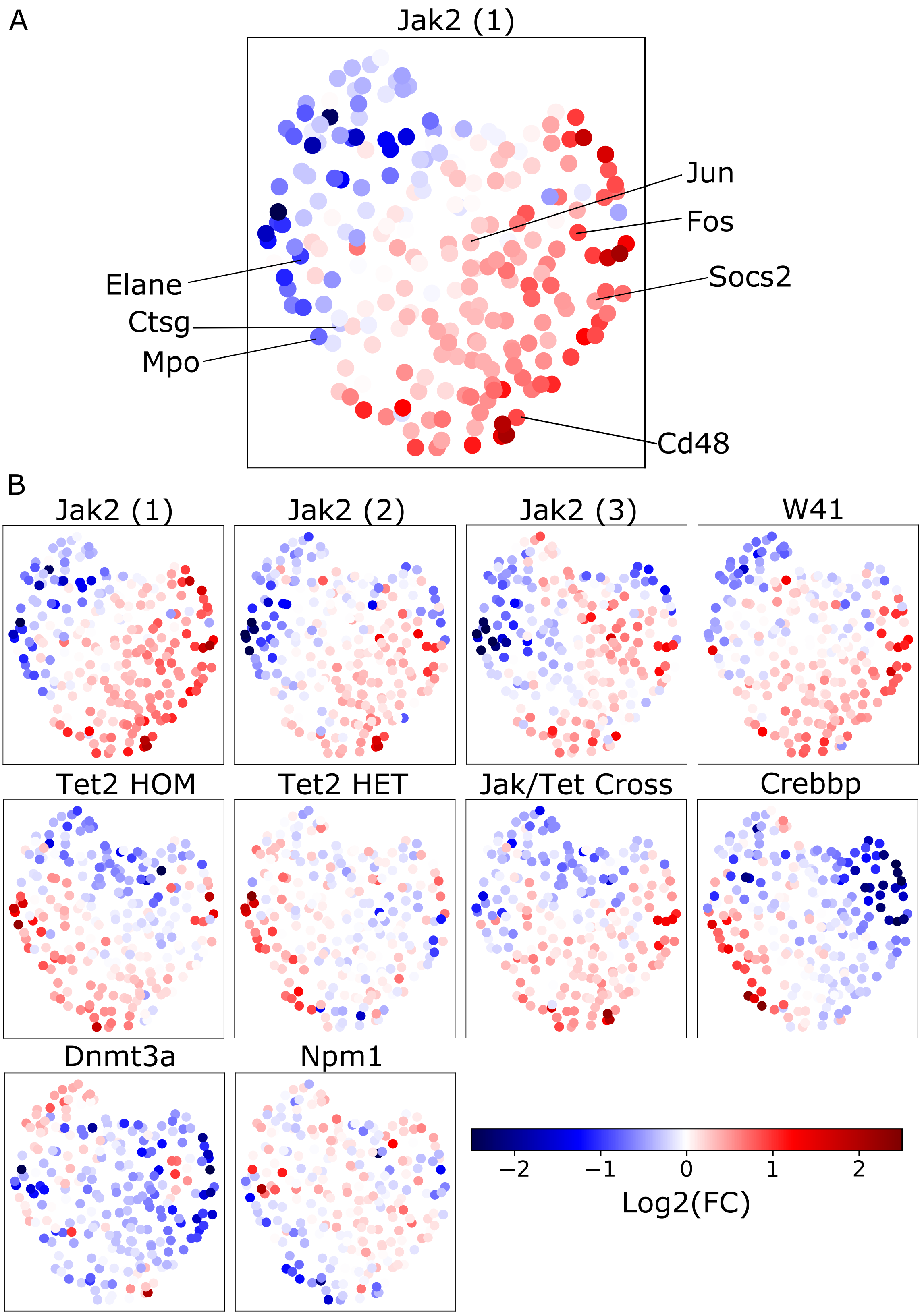


Fig. 4.6 Small-magnitude molecular changes have perturbation-specific features. (A) UMAP representation of the 231 genes identified as being dysregulated in two or more perturbations in the 'Stem Cell' cluster. Each gene is coloured by its fold-change in the first Jak2 experiment. (B) The same 231 genes coloured by their fold-change in each model.

Extracting all genes that were significantly dysregulated in at least two of the seven separate models (not including the Jak/Tet Cross) produced a list of 231 genes whose fold-changes across the perturbations were visualised on a UMAP of genes as in Chapter 3. Unlike for the erythroid cluster analysed previously, no clear clusters of coordinated genes were apparent (Figure 4.6A). Nonetheless clear large-scale differences in fold-changes between the erythroid-skewed models such as Jak2 and W41 and those skewed to myeloid differentiation were observed, as well as regions where the majority of genes were similarly up- or down-regulated across almost all of the different perturbations (Figure 4.6B). A majority of the genes upregulated across both pro-myeloid and pro-erythroid perturbations are cell-cycle related, lending evidence to the hypotheses presented above that a main feature of the immature cell perturbation response is to increase the proportion of ST-HSC/MPP1 cells in an active, cycling state. In addition there is a group of genes in the top left of the visualisation that are downregulated across all the models apart from the Dnmt3a model. These include *Erdr1*, *Mecom* and *Bcl11a*, whose upregulation has counter-intuitively been associated with increased leukaemic load (Tao et al., 2016). The significance of this group of genes remains unclear.

Manual inspection of the 231 genes revealed that once again a number pro-myeloid factors were strongly dysregulated (Figure 4.6A), including *Mpo*, *Ctsg* and *Elane*; this trio of genes was downregulated in the Jak2, Dnmt3a and Npm1 models whilst being upregulated in the Tet2, Crebbp and Jak/Tet Cross models. Whilst this is largely unsurprising given the results of Chapter 3, it is striking that these genes are dysregulated even amongst a cluster of supposedly uncommitted progenitor cells. To further confirm this, the expression of this trio across all cells mapping as MPP1 was visualised (Figure 4.7A). This confirmed that significant myeloid lineage priming (both pro- and anti-myeloid) was visible in cells identified as MPP1 across many different perturbations. No lineage priming was observed across known pro-erythroid transcriptions factors within the MPP1 compartment (Figure 4.7B). This may lend further support to a hypothesis that it is myeloid transcription factors which ultimately control the balance between myeloid and erythroid differentiation. It is potentially possible that some perturbed cells from certain models are already committed to a specific fate despite mapping to the MPP1 compartment. However, the complete lack of observed erythroid priming in any of the perturbation models does not support this idea.

In addition, a number of genes that have previously been implicated in the regulation of HSC behaviour were dysregulated. These included *Jun* and *Fos*, whose overexpression within HSCs in response to stress is known to be associated with increased differentiation and a decrease in self-renewal ability (Mallaney et al., 2019), as well as *Cd48*, whose expression

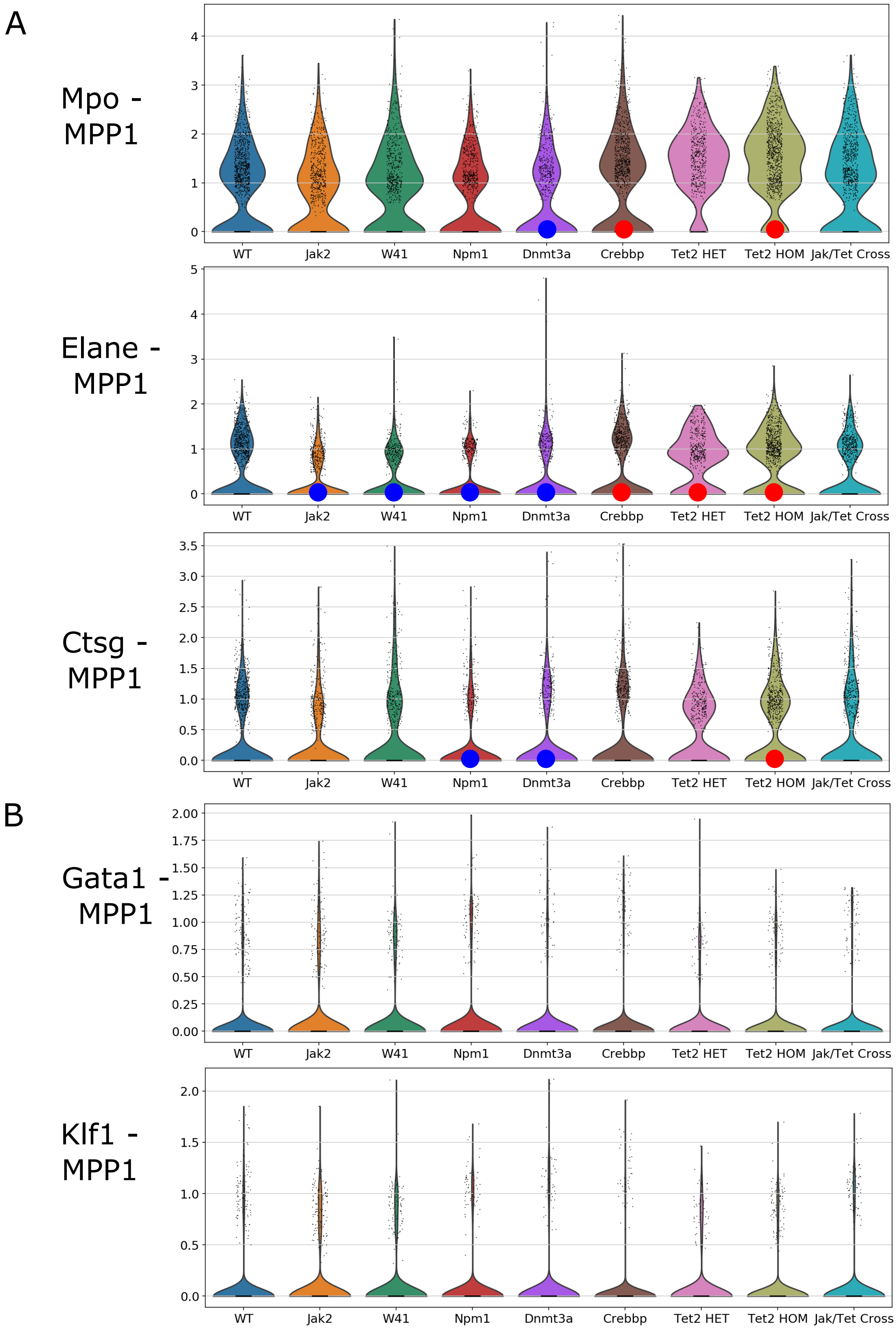


Fig. 4.7 Uncommitted progenitors exhibit myeloid-centric lineage priming. (A) Violin plots showing the expression of pro-myeloid transcription factors in cells identified as MPP1 within the Stem Cell cluster. Red/blue dots indicate these genes were significantly up/downregulated in the respective model. (B) Violin plots showing the expression of pro-erythroid transcription factors in cells identified as MPP1 within the Stem Cell cluster.

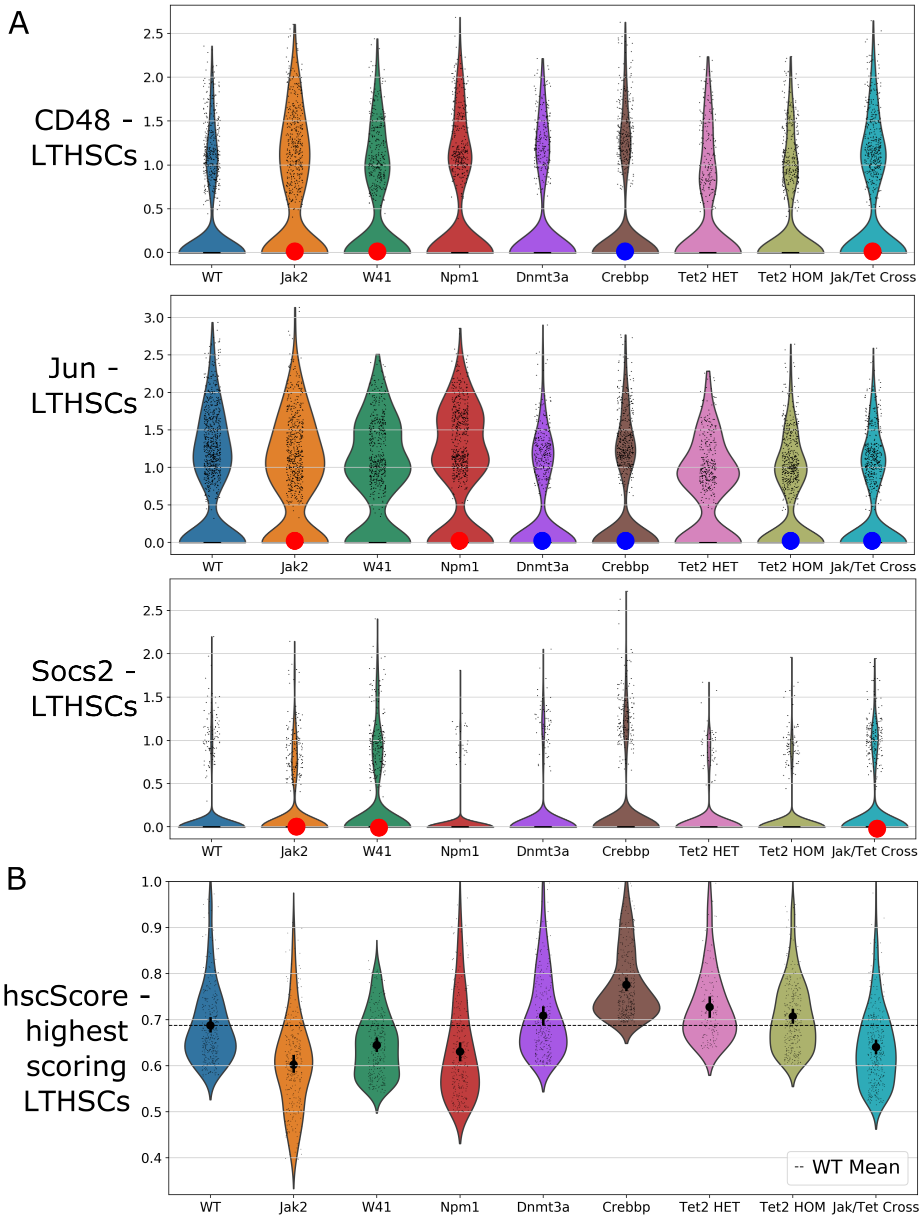


Fig. 4.8 Myeloid-skewing is correlated with increased HSC functionality. (A) Violin plots for HSC genes in cells identified as LTHSCs. Red/blue dots indicate significant up/downregulation in the respective model. (B) Violin plot of the hscScore for the top-scoring 10% of cells mapping to the Stem Cell cluster in each model. Black dots correspond to the sample mean, error bars to a 95% confidence interval for the mean as estimated through bootstrapping.

is known to anti-correlate with HSC quiescence and their functional abilities (Boles et al., 2011). A clear pattern emerged where the erythroid-skewed models had increased expression of these genes in cells mapping as LT-HSCs whilst the myeloid-skewed models display a reduction (Figure 4.8A). The exception to this is the Dnmt3a model, which had significant downregulation of *Jun*. Interestingly the Jak/Tet Cross had significant upregulation of *Cd48* in conjunction with significant downregulation of *Jun*, once again exhibiting the complex interplay between the two mutations. JAK/STAT pathway signalling targets such as *Socs2* and *Pim1* were also found to be significantly upregulated in the Jak2, W41 and Jak/Tet Cross models, indicating that the Jak2 V617F and c-Kit mutations already trigger these pathways in LT-HSCs (Figure 4.8A).

To assess and compare the quality of the best LT-HSCs in each model at a whole-transcriptome level, the hscScore algorithm was applied to each perturbation. The top 10% of highest scoring cells mapping to the Stem Cell cluster from each perturbation were then taken and their scores visualised (Figure 4.8B). All of these cells mapped to the LT-HSC gate, and can therefore sensibly compared as a result of the age and breeding-matched backgrounds of the models. HSCs from the Crebbp and Tet2 models displayed increased hscScores whilst the Jak2, W41 and Npm1 models displayed decreased scores. The Jak/Tet Cross also had decreased scores, suggesting that in the LT-HSCs, the Jak2 mutation ‘wins out’ over the Tet2 mutation and makes them less similar to the best LT-HSCs as defined by transplantation. This is consistent with functional work suggesting that Jak/Tet Cross HSCs lack the self-renewal advantage conferred upon Tet2 mutant HSCs (Shepherd et al., 2018). There is a clear correlation between overall myeloid skewing of the LK landscape and a shift towards ‘better’ LT-HSCs that more closely resemble true functional LT-HSCs at the transcriptomic level. This would support a hypothesis that the myeloid trajectory is the default differentiation pathway arising from uncommitted progenitor cells, whilst erythroid differentiation is promoted through HSCs responding to stressors which may be both cell-intrinsic and cell-extrinsic. This idea has been considered before (Laurenti and Göttgens, 2018; Pei et al., 2020), but further work is required to understand how a stressed HSC skews towards the erythroid lineage in the absence of any transcriptomic erythroid priming.

4.4 Increased single-cell variability is associated with erythroid differentiation skewing

Very recently it has been proposed epigenetic markers play a crucial role in explaining how an uncommitted blood cell may be mechanistically skewed towards specific differentiation pathways in the presence of pre-leukaemic mutations. For example, it is believed that hypermethylation of blood stem cells - potentially as a result of a functional loss of a demethylating enzyme such as Tet2 - preferentially affects pro-erythroid transcription factors due to the relatively high frequency of CpG methylation sites at erythroid promoters (Izzo et al., 2020). This could therefore transcriptionally silence erythroid factors to a greater extent than myeloid factors and could subsequently promote myeloid skewing (Baylin, 2005). Functional loss of a methylating enzyme such as Dnmt3a would therefore promote erythroid skewing in the same fashion. If Tet2 is a direct downstream target of Jak2 signalling as reported (Jeong et al., 2019), the Jak2 V617F mutation should also lead to hypomethylation and erythoid skewing. Hence this model potentially begins to explain the abundance shifts observed across different mutations.

Despite this regulation occurring at an epigenetic level, it was hypothesised that its effects may be visible at the transcriptional level in terms of the variability of single-cell expression, as has been previously suggested (Huh et al., 2013). Stochastic hypomethylation could potentially lead to cells within a population having highly variable abilities to transcribe a particular gene, leading to higher variability of that genes' expression across the population.

To test this, the VarID method was used as a starting point (Grün, 2020). Briefly, VarID constructs a k-nearest-neighbour graph from a transcriptomic landscape, and then calculates a variability score for each gene in each cell using the expression profiles of it's nearest neighbours. Hence the method produces another 'layer' of information that has the same dimensions as the original expression matrix, but instead provides a measure of how variable each gene is in the transcriptomic vicinity of each cell. So that these variabilities can be compared across genes, they are corrected for the differing expression levels of each gene. The global mean-variance relationship is modelled as quadratic in log-space (Kolodziejczyk et al., 2015), and this relationship is then corrected such that variabilities are comparable across genes regardless of their expression levels (Figure 4.9A, see methods). Comparing two lineage markers highlighted how for a given gene, variability was maximal in regions where expression levels were changing rapidly, and were minimal when expression levels were constant - regardless of the actual expression (Figure 4.9B). This can be seen further in Figure 4.9C, with the expression and variability of *Klf1* across the reference clusters shown. Variability was greatest in clusters 4 and 5 (corresponding to MEP/early-erythroid) despite expression being maximal in the latest erythroid clusters 10 and 7.

Unfortunately, the published version of VarID cannot be used to compare variabilities across samples. This is due to how the algorithm calculates the number of nearest neighbours over which it assesses a cell's variability; this number is highly dependent on both the sampling density and the sequencing saturation of each dataset. Taking the first Jak2 experiment as an example, VarID was performed on the WT and HOM samples, before

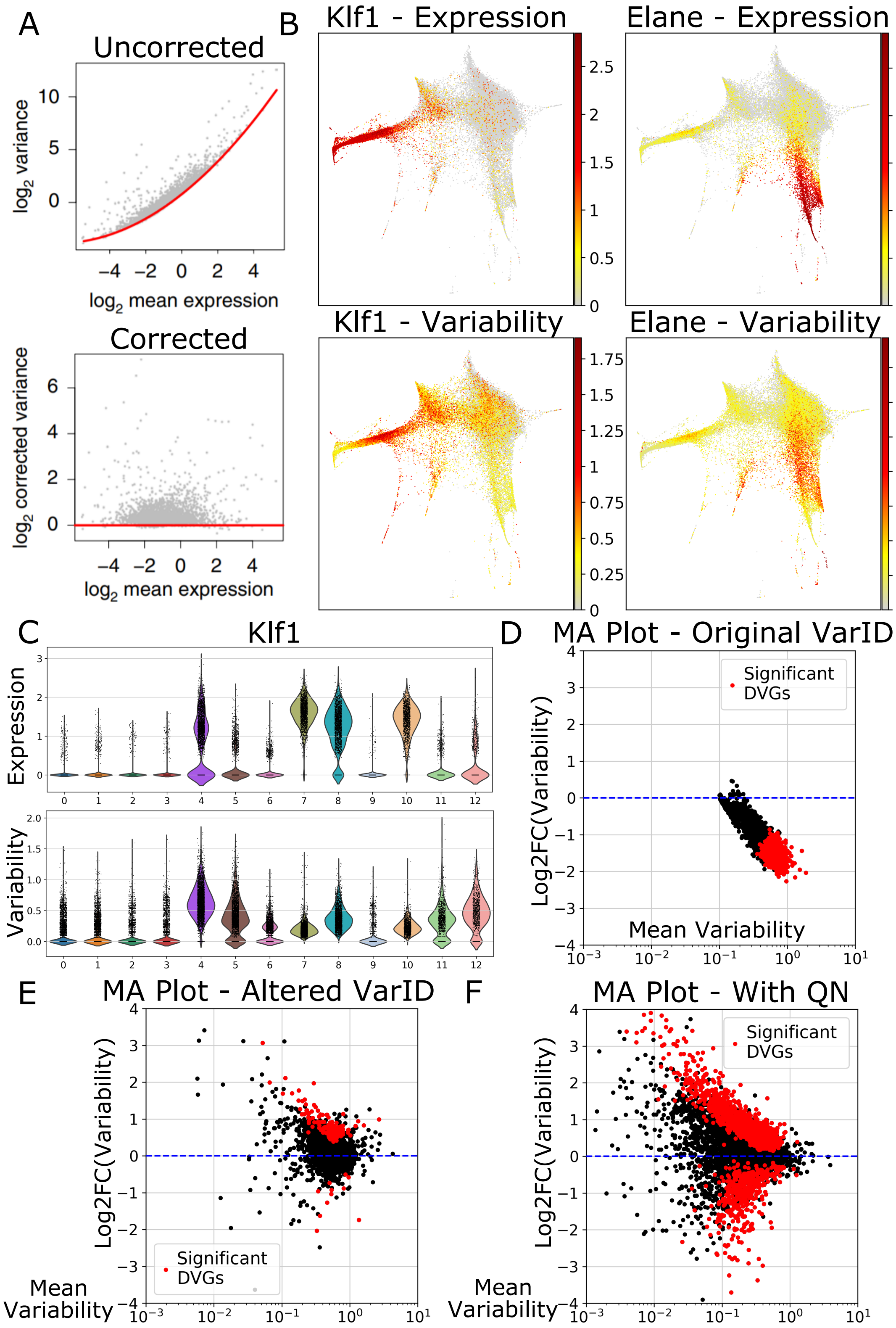


Fig. 4.9 An improved version of the VarID method reveals single-cell variabilities and allows cross-sample comparisons (A) To compare variabilities across genes, the mean-variance relationship is removed by modelling it as quadratic in log-space. (B) Comparison of expression and variability values for *Klf1* and *Elane*. (C) Violin plots comparing expression and variability of *Klf1*. (D) An MA plot of differential variability between two samples calculated using the original VarID method. (E) The same MA plot calculated using an altered version of VarID with constant-neighbour restrictions. (F) The same MA plot calculated with the additional inclusion of quantile normalisation.

differential variability testing was performed using the Wilcoxon rank-sum test in the same manner as for differential expression (see methods). An MA plot was calculated to visualise the results, with significantly differentially variable genes (DVGs) highlighted in red (Figure 4.9D). Due to the above factors, almost all genes displayed reduced variability in the HOM sample. To allow for cross-sample comparisons, VarID was rewritten from the ground up, incorporating a constant number of nearest neighbours for all variability calculations. In addition to greatly reducing computation time and memory load, this allowed sensible cross-sample comparisons of variability to be made (Figure 4.9E). However, sample-specific effects still remained, highlighted by the complete imbalance between the number of genes with significantly increased and decreased variability. To correct this further, quantile normalisation was performed on the variability values (see methods), in a similar manner to that used for microarray data (Rao et al., 2008). When combined, these changes resulted in improved sensitivity to detect DVGs and a sensibly balanced distribution of DVGs when analysing the perturbation experiments (Figure 4.9F). This improved version of VarID was therefore used in the analyses that follow.

Differential variability testing was performed on the Stem Cell cluster for all perturbation experiments. Due to the extreme variation in cell density and sequencing saturation between the WT and HOM samples in both the W41 and the third Jak2 experiment, the differential variability testing failed to return meaningful results for these experiments and they were excluded from further analysis. Subsampling was attempted to correct for this as in Chapter 3, but results were not improved. The concordance of the overlapping DVGs for each pair of perturbations was calculated as had been done previously for the expression data (Figure 4.10A). Strikingly, this revealed large similarities between the Jak2, Dnmt3a and Npm1 models that were not be observed when looking solely at the expression data in Section 4.3 (Figure 4.5). Combined with the anti-correlation of these models with the Crebbp and Tet2 models, this suggested that there was a conserved signal within the Stem Cell variability data priming these models for the pro-erythroid skewing they exhibit in more mature populations, and that this signal was largely invisible in the context of differential expression. In addition, the Dnmt3a and Npm1 models exhibited the largest number of DVGs, despite having the the lowest number of DEGs (excluding the Tet2 HET). Hence whilst these models may not appear to show a molecular response in the Stem Cell cluster, they are in fact significantly altered from a native state in terms of their molecular variability. Interestingly the Crebbp and Tet2 models were only correlated slightly more than would be expected by chance, potentially suggesting that whilst the differential expression response of these models within the Stem Cell cluster were similar, there may be different mechanisms driving these responses at a more basic level. The variability of the Jak/Tet cross displayed general agreement with both

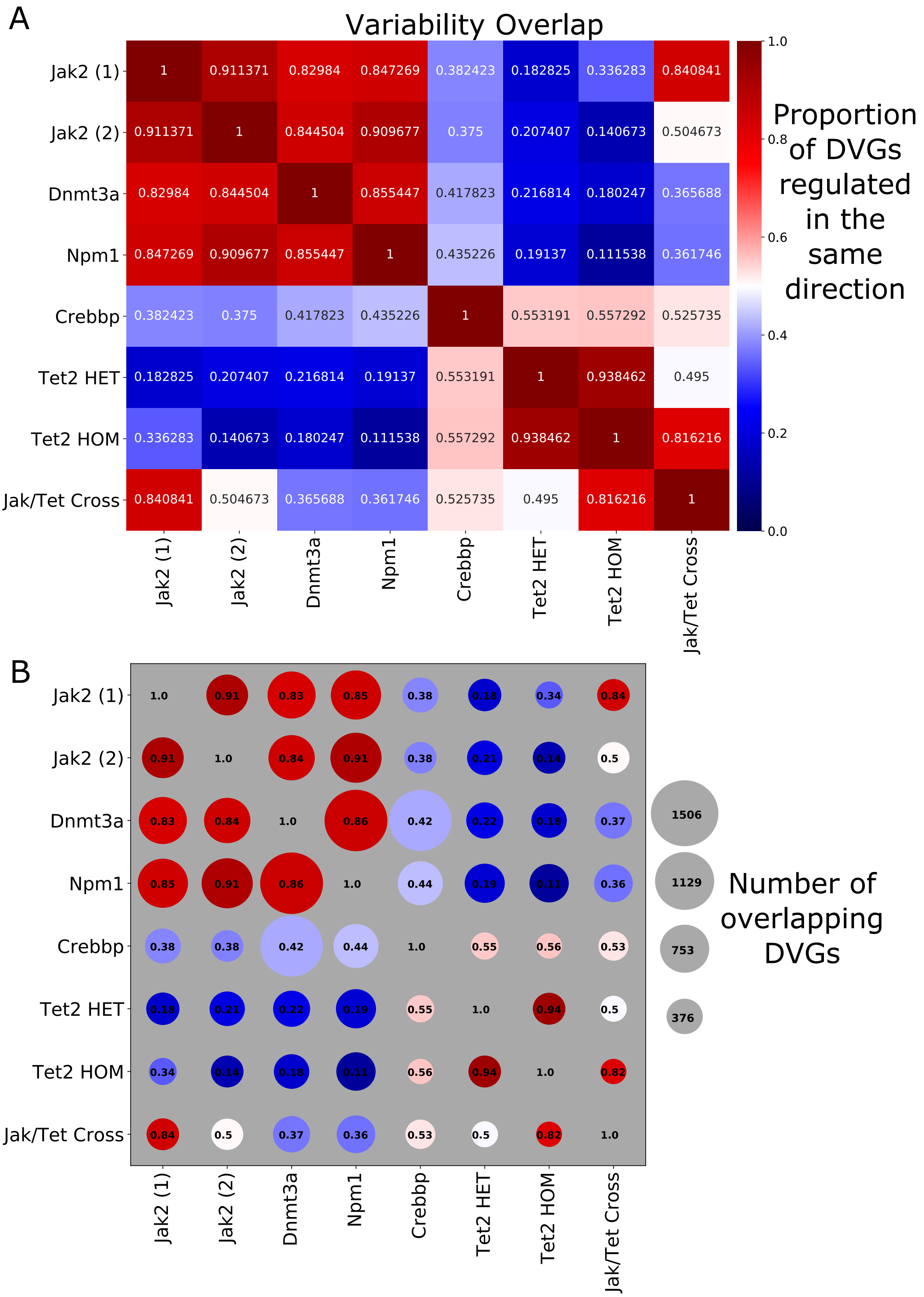


Fig. 4.10 Variability changes within uncommitted progenitors correlate with observed abundance skewing. (A) For each pair of perturbations, the heatmap is coloured by the proportion of overlapping DVGs that are differentially variable in the same direction. (B) For each pair of perturbations, the circle size corresponds to the number of overlapping DVGs.

the Jak2 and Tet2 HOM models, but strangely did not correlate strongly with the Tet2 HET model.

The number of significant DVGs was substantially larger than the corresponding numbers of DEGs in the Stem Cell cluster for each model, resulting in large overlaps of several hundred DVGs between each pair of models (Figure 4.10B). This provides a higher degree of confidence in the variability results compared to the expression results presented in Figure 4.5. A hypergeometric test was used to assess the overlap between DEGs and DVGs for each model; no significant overlap was found (Figure 4.11A). In addition, the overlap between DVGs in the Stem Cell cluster and DEGs in two later clusters (the 'Erythroid-Middle/Late cluster analysed in Chapter 3 and the 'Neutrophil/Monocyte' cluster) to see if differential variability was predictive of differential expression at later time points. However no significant overlap was found in either case.

To further decipher how variability changes were linked to abundance skewing, global variability changes in the Stem Cell cluster were visualised for each model (Figure 4.11B). This revealed a clear pattern in which the pro-erythroid models exhibited larger molecular variabilities globally than WT cells, whilst the pro-myeloid models exhibited reduced variabilities. This lends support to the hypothesis that hypomethylation occurring through a Jak2-activated increase in Tet2 or through functional loss of Dnmt3a leads to increased transcriptional variability, and vice versa in the case of functional Tet2 loss. The potential epigenetic links between Npm1, Crebbp and blood stem cells are not clear, although the effects of Npm1 loss on transcription have been shown to synergize with those occurring due to loss of Dnmt3a (Olausson et al., 2014), and Crebbp is a known histone acetyltransferase that interacts with chromatin-remodelling complexes (Chan et al., 2011; Ogryzko et al., 1996). Notably the Jak/Tet Cross displayed no overall increase or reduction in global variability, once again placing it as a middle ground between the Jak2 and Tet2 HET models. If constitutive Jak2 signalling led to increased variability solely through Tet2 activation, the loss of Tet2 should negate this, suggesting some other partially compensatory mechanism at work in the Jak2 model.

Gene ontology analysis was performed on the set of genes with significantly increased variability in the Jak2, Dnmt3a and Npm1 models whilst also having significantly decreased variability in the Crebbp, Tet2 HET and Tet2 HOM models (Figure 4.12A). This list of $\sim$100 genes was associated with both differentiation pathways (*regulation of haematopoietic stem cell differentiation, regulation of haematopoietic progenitor cell differentiation*) and RNA splicing pathways (*mRNA splicing via spliceosome, mRNA processing, RNA splicing*). Increased variability of these pathways makes sense in the context of the global variability

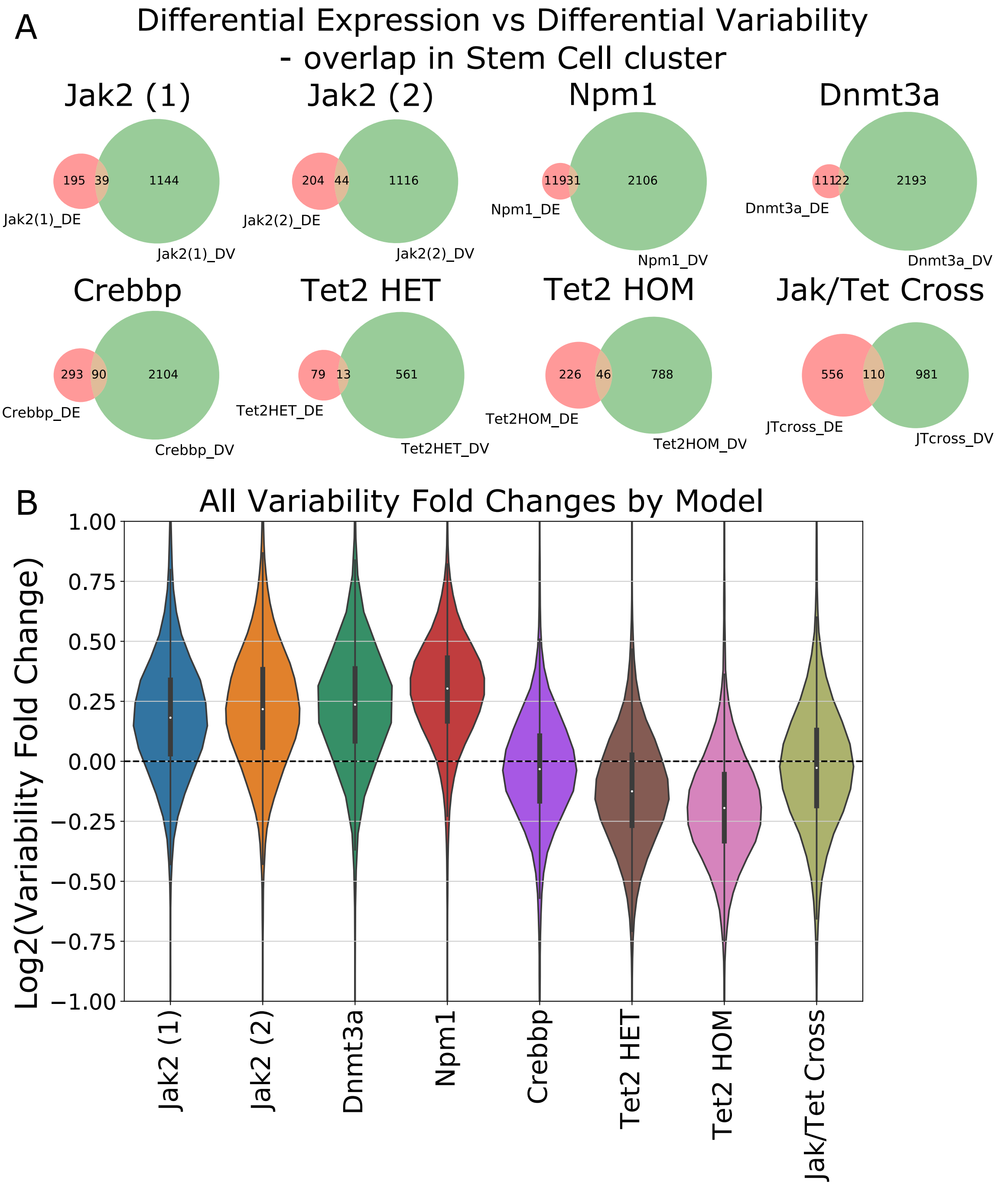
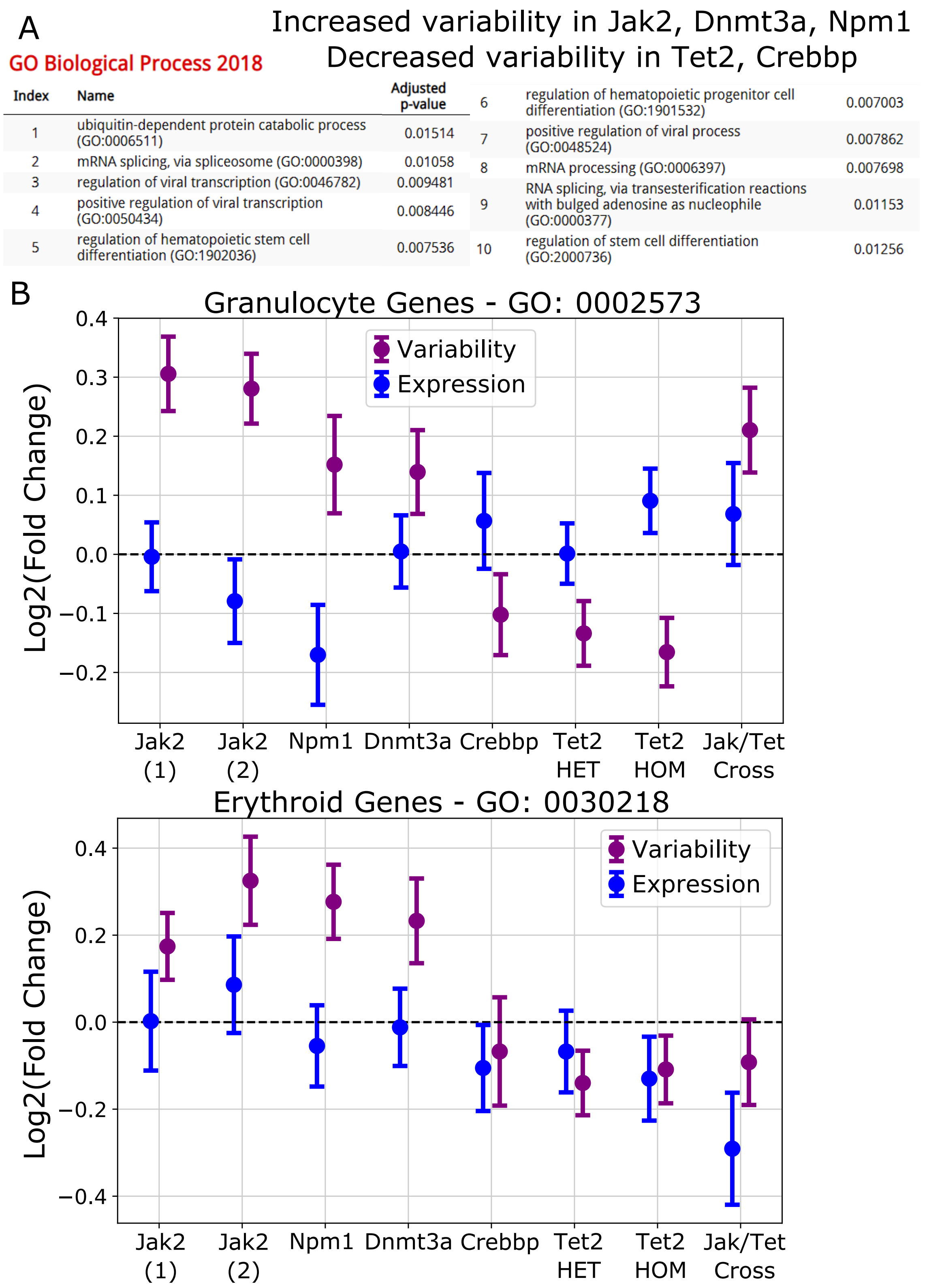


Fig. 4.11 Pre-leukaemic models with pro-erythroid skewing are associated with a global increase in molecular variability. (A) The overlap between differentially expressed genes (red) and differentially variable genes (green) for each model in the Stem Cell cluster. (B) Violin plot of all variability fold changes across the transcriptome for each model. Only genes whose variabilities had been assessed in all models were used, resulting in ∼8000 genes in total.



A

GO Biological Process 2018

Increased variability in Jak2, Dnmt3a, Npm1
Decreased variability in Tet2, Crebbp

Index	Name	Adjusted p-value
1	ubiquitin-dependent protein catabolic process (GO:0006511)	0.01514
2	mRNA splicing, via spliceosome (GO:0000398)	0.01058
3	regulation of viral transcription (GO:0046782)	0.009481
4	positive regulation of viral transcription (GO:0050434)	0.008446
5	regulation of hematopoietic stem cell differentiation (GO:1902036)	0.007536
6	regulation of hematopoietic progenitor cell differentiation (GO:1901532)	0.007003
7	positive regulation of viral process (GO:0048524)	0.007862
8	mRNA processing (GO:0006397)	0.007698
9	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	0.01153
10	regulation of stem cell differentiation (GO:2000736)	0.01256

Fig. 4.12 Erythroid and myeloid differentiation genesets display similar variability responses. (A) Gene ontology results for the set of genes with significantly increased variability in each of the Jak2, Dnmt3a and Npm1 experiments, which also had significantly decreased variability in each of the Crebbp, Tet2 HET and Tet2 HOM experiments. (B) For each experiment, the mean and standard deviation of expression/variability foldchanges was calculated for either a pro-erythroid geneset (GO term 0030218) or a pro-myeloid geneset (GO term 0002573).

increases seen in the pro-erythroid models. To assess whether disparate differentiation pathways responded differently in terms of molecular variability, the variabilities of two genesets corresponding to gene ontology terms *myeloid leukocyte differentiation* and *erythrocyte differentiation* respectively was calculated for each model, alongside the expression changes for the same genesets (Figure 4.12B). Both genesets displayed similar variability trends across models, with increases in the pro-erythroid model and decreases in the pro-myeloid models. The exception was the Jak/Tet Cross, which exhibited increased variability of myeloid genes (agreeing with Jak2) but a decreased variability of erythroid genes (agreeing with Tet2). It is not clear how to interpret this result, though it does suggest that in complex perturbations with multiple driver mutations, the molecular variability response may be trajectory-specific in the same way that the molecular expression response was found to be in Chapter 3.

Overall, this data suggests that global molecular variability changes within the uncommitted progenitor pool are strongly correlated with aberrant differentiation skews. It is plausible that these changes may be being driven by the epigenetic affects of the mutations, though further experimental work is needed to validate this hypothesis. The observation of mRNA splicing pathways being differentially variable suggests a potential positive feedback loop where variability of DNA transcription induced through epigenetic alterations triggers further variability in mature mRNA production if RNA splicing genes are compromised. Linking these results to Section 4.3, it is sensible to suggest that an increase in molecular variability is associated with a decrease in HSC ‘stemness’ and vice-versa. The stress response seen in the LT-HSCs from the pro-erythroid models may be connected to an increase in molecular variability, though which is driving the other remains difficult to say. It is certainly clear, however, that molecular variability is a useful layer of information that can be informative even when traditional differential expression does not reveal large changes across conditions, as is the case in the flat transcriptomic landscape of uncommitted blood progenitor cells.

4.5 Cell cycle perturbations in uncommitted progenitors predict whole-transcriptome cell cycle shifts in mature populations

The role of cell cycle in the WT HSC compartment has been reasonably well established, with HSCs believed to exist along a ‘dormancy axis’ where the most quiescent HSCs are associated with the best long-term engraftment and reconstituting ability (Cabezas-Wallscheid et al., 2017; Wilson et al., 2008). These quiescent LT-HSCs are thought to provide only a small contribution to steady-state haematopoiesis, with the majority of blood cells being maintained from the ST-HSC and MPP1 compartments ((Busch et al., 2015; Laurenti and Göttgens, 2018)). However the veracity of this result has been questioned (Sawai et al., 2016). Nonetheless there is a clear correlation between cell cycle activity and repopulating ability amongst undifferentiated blood cells. The cell cycle response of these compartments in response to severe stresses such as infection has been investigated, with greater proliferative behaviour observed in response to IFN-γ and IFN-α (Baldridge et al., 2010; Vainieri et al., 2016). More recently, the effect of leukaemic mutations on HSC and MPP proliferation have been described in an AML mouse-model using a dual pulse-chase labelling system (Akinduro et al., 2018). This revealed no evidence of significant changes in HSC or MPP proliferation over the course of AML progression. However these mice achieve a high leukaemic load within 3 weeks, and hence do not represent a homeostatic system such as those observed in the pre-leukaemic models described in this thesis.

The G2M score introduced in Section 4.2 is a proxy for cell cycle activity that is calculated from the transcriptome using a set of 198 genes annotated as being upregulated during the G2/M phases of cell division. Whilst only a proxy, applying this score to the reference dataset clearly demarcates between highly quiescent LT-HSCs, relatively quiescent MPP1 and more proliferative fate-restricted progenitors (Figure 4.3D, 4.2D). To test whether there was transcriptional evidence for cell-cycle changes across the pre-leukaemic models, the G2M score was calculated for each experiment. Overall, the G2M scores in pre-leukaemic cells mapping as LT-HSCs were roughly consistent with those observed across the WT samples (Figure 4.13A). The exceptions were the Jak2 model, which had a statistically significant increase in LT-HSC G2M scores compared to WT, and the Crebbp model, which displayed reduced LT-HSC G2M scores. In the MPP1 compartment, all models apart from the Npm1 and Crebbp models had higher mean G2M scores than WT, consistent with the abundance shifts to more active MPP1 states shown in Figure 4.4. However only the Jak2, W41 and Crebbp models returned statistically significant changes in the G2M score. The Jak2 results corroborate functional results from the same mouse model, which suggested that Jak2 V617F LT-HSCs have shorter cell cycle transit time and exit quiescence quicker than WT when stimulated (Shepherd et al., 2018). The same study also reported that Tet2 homozygous KO did not alter the divisional kinetics of LT-HSCs, which appears to also be the case transcriptionally. However the divisional kinetics of Jak2 V617F mice crossed with Tet2 HOM KO were found to be more similar to the Jak2 single-mutant mice. However, the Jak/Tet Cross model investigated here (which is HET for Tet2 KO) did not display significant changes in G2M score compared to WT, and appeared to be more similar to the Tet2 HET and Tet2 HOM models. Functional cell cycle analyses of LSK cells from the Crebbp model

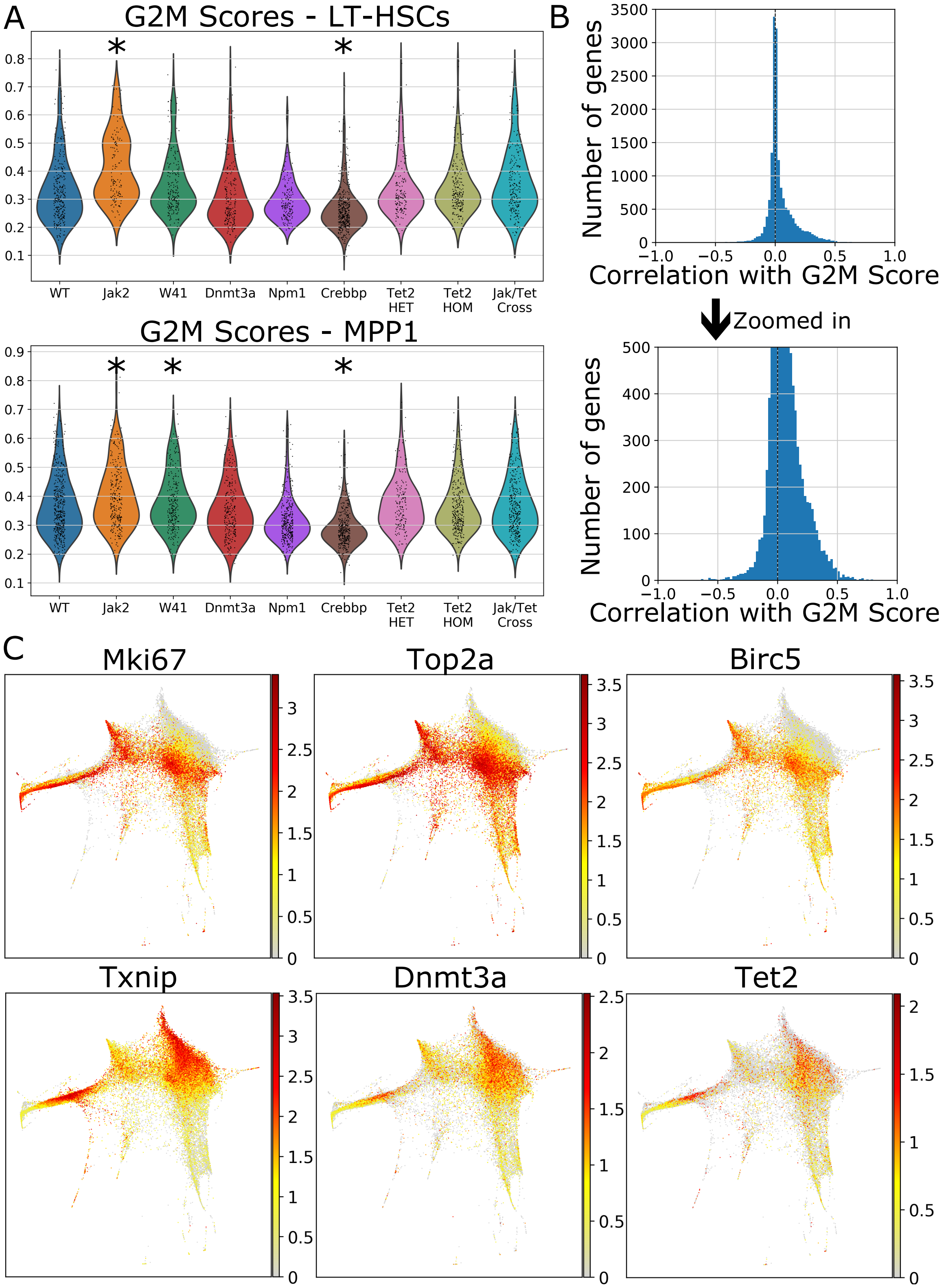


Fig. 4.13 Jak2 and Crebbp LT-HSCs/MPP1s display small changes in cell cycle activity compared to WT. (A) Violin plots of the G2M score in cells mapping as LT-HSCs or MPP1 across the pre-leukaemic models. Asterisks denote significant differences from WT using a two-sided t-test with a p-value cutoff of 0.05. (B) Histogram of gene correlations with the G2M score across the entire transcriptome. Bottom panel is a magnified version of the top panel. (C) Force-directed graphs showing the expression of some genes highly correlated with the G2M score.

have previously suggested an increase of quiescent cells in G0 compared to WT controls, again lining up well with the transcriptomic data (Chan et al., 2011).

It was therefore hypothesised that the cell cycle changes observed in the Jak2 and Crebbp uncommitted progenitors may lead to larger changes in downstream progenitor populations that would not be present in other models. It is important to note that within the 10X drop-seq datasets analysed in this thesis, cell cycle is an extremely large confounding factor that contributes greatly to the the expression profile of any individual cell. Despite only 198 genes being used to calculate the G2M score, more than 2100 genes have an absolute spearman correlation of > 0.2 with the G2M score. Around 75% of these are positively correlated and around 25% are negatively correlated (Figure 4.13B). These negatively correlated genes are therefore associated with G0 and/or G1 cell cycle states. The expression of some highly correlated genes is shown in Figure 4.13C. This highlights how cell-cycle effects are pertinent across the entire transcriptome; for example, *Mki67* is used to calculate the G2M score, but *Top2a* and *Birc5* are not, despite their high correlation. It is notable that several epigenetic enzymes including *Dnmt3a* and *Tet2* were amongst those genes anti-correlated with the G2M score across the LK landscape. Plotting an expression heatmap of the top 2000 correlated and anti-correlated genes further highlights the global reach of cell cycle effects, and how genes with strong positive correlations with the G2M score are more prevalent than those which are anti-correlated (Figure 4.14A).

Visual inspection of the G2M score (Figure 4.2D) suggests that within the LK landscape, the erythroid trajectory displayed a wide range of G2M values. The force-directed graph layout in Figure 4.2D is calculated after removal of 380 genes known to associate with the cell cycle (see methods), and hence cell cycle effects in the visualisation are diminished. Visualisations without these genes removed revealed two clearly separated ‘highways’ of erythroid differentiation, split by their G2M score and hence by their apparent cell cycle state. To investigate this further, the WT reference was subset to just the four clusters annotated as erythroid (Figure 4.14B). A histogram of the G2M scores from these erythroid cells immediately revealed a bimodal distribution, indicating that cells are generally found in either a G2M-Low (G2M score<0.5) or a G2M-High (G2M score>0.5) state when sequenced (Figure 4.14C, top-left). However this histogram represents an average over all the erythroid differentiation stages captured in the LK gate. To visualise the shifting nature of the G2M-Low and G2M-High populations along differentiation, pseudotime was again used to give each cell a score from 0 (most immature erythroid cell) to 1 (most mature erythroid cell). A sliding window was then used to observe the distribution of G2M scores present at any specific point along the erythroid trajectory (Figure 4.14C, top-right). There was a clear

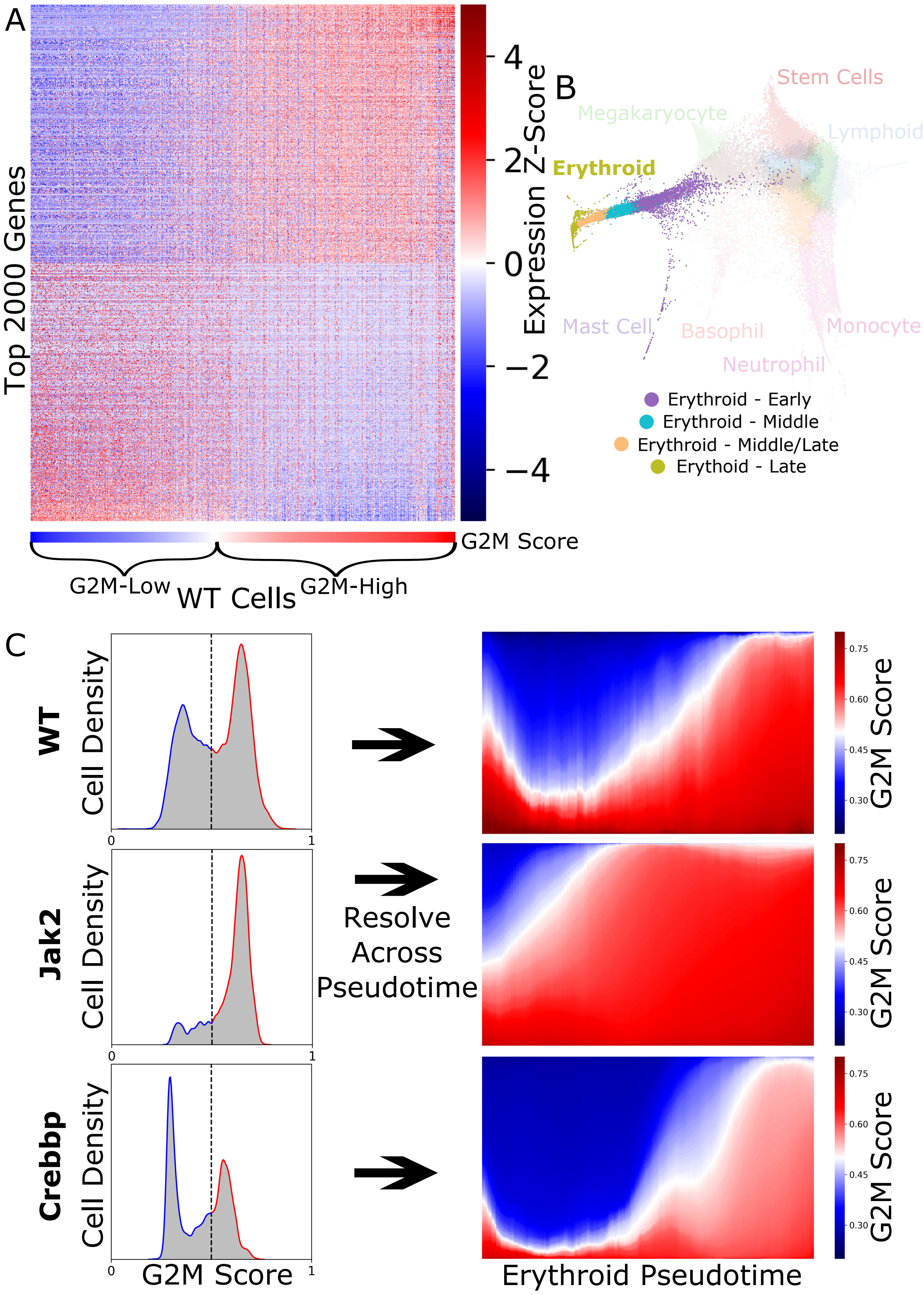


Fig. 4.14 Erythroid progenitors exhibit a bimodal distribution of G2M scores which are severely altered in the Jak2 and Crebbp models. (A) Heatmap showing scaled gene expression of the top 2000 genes positively- and negatively-correlated with the G2M score. (B) The four erythroid clusters used in the downstream cell cycle analysis are highlighted. (C) Left Column: histograms of G2M scores across all four erythroid clusters for the WT reference, Jak2 (1) and Crebbp experiments. Right Column: the histograms expanded across erythroid maturation using pseudotime (see methods).

trend towards the G2M-high states as differentiation progressed, and by the time cells exit the LK gate they are almost all found to be in a G2M-High state. It seems plausible that any given erythroid progenitor is likely to be cycling between G2M-Low and G2M-High states, and therefore the shifting distributions along pseudotime observed here are due to a larger proportion of cells being in a G2M-High state at any given time point. This result appears to contradict previous work in foetal liver that suggests tight synchronisation between cell cycle state and erythroid differentiation landmarks (Pop et al., 2010), however such work was not performed with single-cell resolution and focused on later stages of the erythroid trajectory.

The same analysis was then performed for each perturbation experiment. Strikingly, both the Jak2 experimental repeats and the Crebbp experiment showed significantly altered erythroid cell cycle behaviour. In both models, the proportion of cells in a G2M-High state at the start of the trajectory were roughly similar. However in the Jak2 model this proportion increased quickly, such that almost all cells were found in a G2M-High state at around the same pseudotime coordinate when only half of the WT cells were (Figure 4.14C, middle row). In the Crebbp model, almost no cells were found in a G2M-High state during the first half of the trajectory. Eventually the distribution did shift such that the majority of cells were in a G2M-High state by the time they exited the LK gate, however even at this stage their G2M scores were noticeably lower than the WT control (Figure 4.14C, bottom row). The results of this analysis for the other perturbation experiments is summarised in Figure 4.15A, which visualises the percentage of erythroid cells in a G2M-High state as a function of pseudotime (e.g. it plots the 'white line' from Figure 4.14C for each experiment). The average trend across all WT samples (including the reference) and its standard deviation is shown in black and shaded grey respectively. The Dnmt3a, Npm1, Tet2 HET and Tet2 HOM experiments all fall close to the WT trend, whilst the three repeats of the Jak2 experiment and the Crebbp experiments fall clearly outside of it on either side. Both the W41 and the Jak/Tet Cross experiments appear to differ from WT to some extent in the same fashion as the Jak2 model. In the case of the Jak/Tet Cross this is significant, since it appears that whilst the introduction of heterozygous Tet2 KO strongly reduces the hyper-proliferative phenotype observed along the Jak2 erythroid trajectory, it does not abrogate it completely. Hence Jak2-mediated cell-cycle abnormalities may be acting partially - but not completely - through interaction with Tet2 or a downstream target of Tet2. For comparison, the results for both the (younger) p53 WT mouse and the mutant p53 mouse are shown. The aged WT exhibits a shift towards more G2M-Low cells compared to the other WTs. It is also notable that the mutant p53 mouse increases this lower-proliferation phenotype in the same manner as Crebbp; p53 has shown to be a downstream target of Crebbp acetylation and requires this acetylation to activate (Tang et al., 2008).

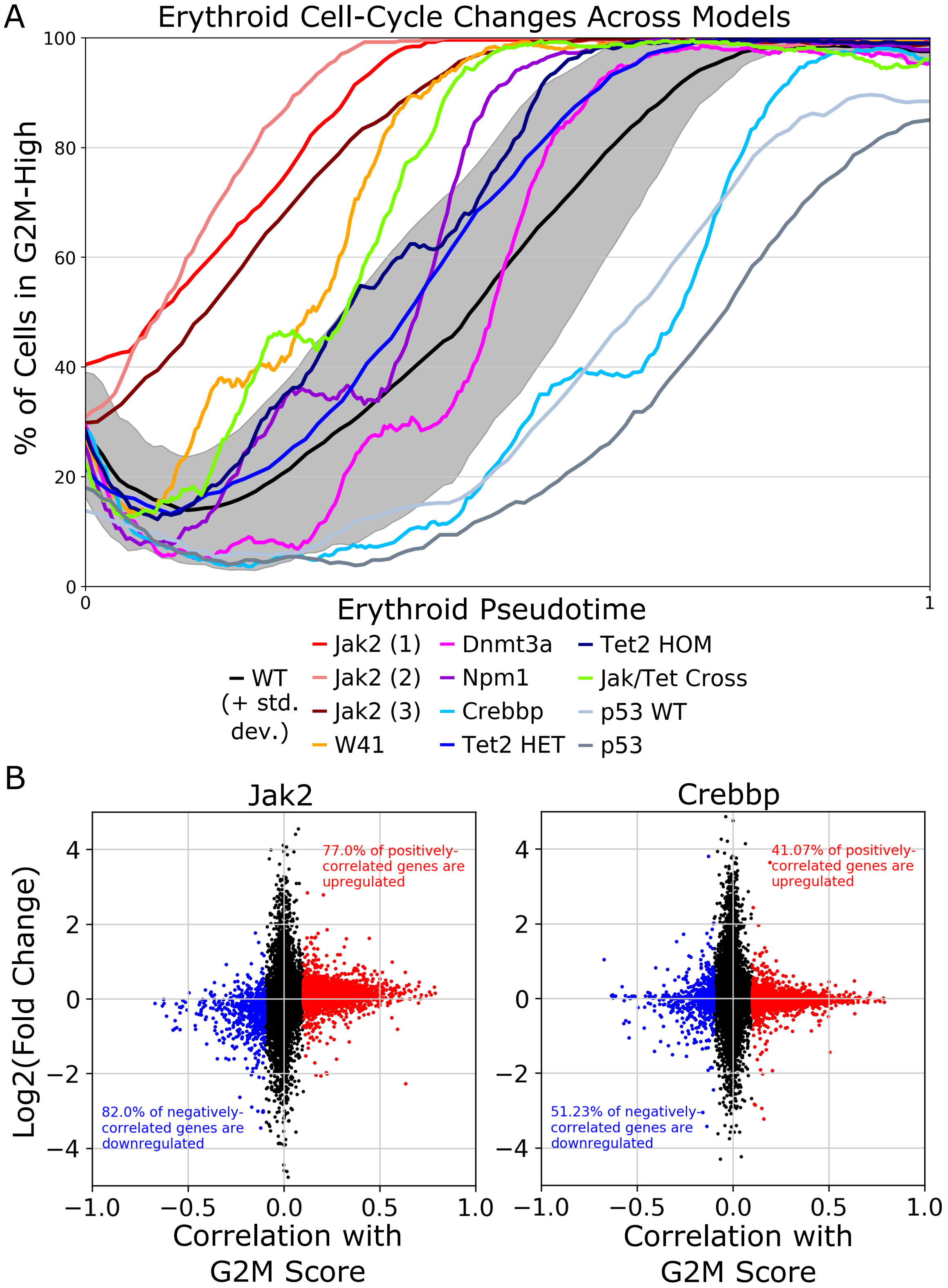


Fig. 4.15 The Jak2 and Crebbp models exhibit whole-transcriptome cell cycle shifts. (A) The percentage of cells in a G2M-High state (G2M score>0.5) as a function of pseudotime across the erythroid clusters, coloured by each model. The shaded area represents the standard deviation of the WT results across all WT samples (including the reference but not including the p53 WT). (B) For each gene, its log2(Fold Change) in the 'erythroid-middle/late' cluster is plotted against it's correlation with the G2M score in WT cells. Genes with a pearson correlation greater/less than 0.2/-0.2 are plotted in red/blue respectively.

Finally, the extent to which the Jak2 and Crebbp models alter expression of cell cycle correlated genes globally was assessed. Differential expression was performed in cells mapping to the 'erythroid-middle/late' cluster. However to account for different numbers of G2M-Low and G2M-High cells in this cluster between WT and perturbed samples (and the subsequent impact on DE this would have), the cells used for the DE analysis were subset such that each sample had identical distributions of G2M scores. DE was then performed and the fold-changes of each gene in these cells were plotted against its correlation with the G2M score (Figure 4.15B). In the Jak2 model, 77% of genes that were positively correlated (r>0.2) with the G2M score were upregulated and 82% of negatively correlated genes were downregulated. Since the overall shift to more G2M-High cells had already been accounted for, this revealed that Jak2 erythroid cells are altering their entire transcriptome towards a more proliferative state. Conversely in the Crebbp model, only 41% of positively correlated genes were upregulated, suggesting an opposite transcriptional shift that is smaller in magnitude than that affecting the Jak2 cells, but which nonetheless induces large scale alterations to the Crebbp KO transcriptome. No other model displayed significant changes across the transcriptome when analysed in this manner.

Taken together, this data would suggest that aberrant expression of cell cycle genes within the uncommitted progenitor compartment leads to profound changes in the cell cycle state and expression of mature progenitors along the erythroid trajectory. The bimodal nature of the G2M score along the erythroid trajectory is somewhat surprising, and is suggestive of a transcriptional switch between dividing and non-dividing red blood cell progenitors. Whilst it is potentially plausible that cells on the erythroid trajectory would have some 'memory' of the fact that the MPP cells they differentiated from were more/less quiescent that in WT - via epigenetic or other means - this is difficult to investigate using snapshot data. However the finding that the entire cell cycle transcriptome is shifted in the Jak2 and Crebbp models may support this idea, since it is difficult to imagine how a single driver mutation would affect so many thousands of genes purely through its action within the erythroid trajectory.

In terms of correlation to abundance skewing, it is clear that hyper-proliferation in the Jak2 and W41 models during the early stages of the erythroid trajectory may be contributing to the observed abundance increases of later erythroid clusters. Results in Chapter 3 suggested that whilst the Crebbp and Tet2 models looked similar in terms of their differential abundances, their molecular response along the erythroid trajectory differed in several respects. The cell cycle data further suggests that the mechanisms behind their lack of erythroid cells may be different, as both Tet2 HET and Tet2 HOM models have a non-significant yet small increase in the proportion of G2M-High cells across erythroid pseudotime, whilst the Crebbp model has a significant reduction. It is possible that whilst both of these models undergo pro-myeloid skewing from uncommitted progenitors, only the Crebbp model has this phenotype further exacerbated by cell-cycle defects.

4.6 Conclusions

Work performed in this chapter presents a wide-ranging analysis of the uncommitted haematopoietic progenitor landscape across pre-leukaemic perturbation models using scRNA-seq. Cross-platform mapping was used to retrospectively assign classical FACS labels to an unbiased WT reference of haematopoiesis, allowing the identification of LT-HSC, ST-HSC and MPP1 states. The cellular and molecular shifts within these compartments driven by the perturbations were characterised. An improved tool for calculating and comparing single-cell variability revealed further insights that were not possible using the expression counts themselves. Finally the aberrant cell cycle characteristics of the perturbations were analysed and found to exist across the LK transcriptomic landscape.

4.6.1 Strengths and weaknesses of cross-platform mapping

Droplet-based scRNA sequencing methods have dramatically reduced the per-cell cost of procuring single cell transcriptomes and hence have allowed the creation of large, unbiased datasets. Since by definition, an unbiased approach uses only a small number of surface markers to identify cells, detailed index sorting measurements of relevant surface markers are not available for these cells. New multimodal techniques such as CITE-seq for measuring the transcriptome and protein level of single-cells in a high-throughput manner are becoming increasingly available, but remain in their infancy in terms of analytical techniques (Mimitou et al., 2019; Stoeckius et al., 2017). Hence cross-platform mapping of index sorted populations onto drop-seq landscapes is currently a powerful tool to link transcriptomic state with functional potential.

In this chapter the mapping appeared to perform excellently and largely reconstructed the expected relationships between different haematopoietic FACS gates within the WT reference dataset. However it remains very difficult to independently assess and quantify the quality of these mappings. Using the hscScore method to corroborate the location of the best LT-HSCs is useful, but no such signatures exist for other populations. Applying the mapping procedure to data where a ‘ground truth’ is known *a priori* and assessing its accuracy is appealing, but arguably native haematopoiesis is already the model where this ground truth is known with the highest accuracy. Here the mapping was used to identify the set of uncommitted progenitor cells within the WT reference, in conjunction with the abscence of lineage marker expression. However even within this group of cells, the expression of known pro-myeloid transcription factors such as *Mpo* was reasonably high. It remains to be shown with greater confidence whether the cells annotated as uncommitted in this analysis are truly multipotent.

4.6.2 Marked similarities and marked differences between perturbations with opposite effects

Many of the perturbation models display a shift towards more active/cycling MPP1 states within the ST-HSC/MPP1 compartments, regardless of the ultimate direction of their differentiation skewing. This can be interpreted as a mechanism by which mutant clones could outcompete WT clones in a competitive setting through increased proliferation. It could also be interpreted as highlighting an inability of mutant cells to remain in a self-renewing, quiescent state once differentiation programs start to push them out of a LT-HSC state. Jak2 V617F HSCs have been shown to produce differentiated cells quicker than their WT counterparts, although no significant differences were seen in Tet2 HOM cells (Shepherd et al., 2018). Yet despite the similarities at a cellular level, the perturbation models exhibit differences at a molecular level which largely align with their downstream cellular and molecular shifts. Clearly a leukaemic perturbation needs to do two things to be ‘successful’; it needs to amplify and take over the bone marrow, and it needs to drive some kind of skewed differentiation that ultimately results in leukaemic symptoms (be that anaemia, neutropenia, myelofibrosis etc). Whilst there are many ways to do the latter, it appears there is a conserved method of doing the former amongst many of the perturbations analysed here.

The molecular analysis of expression values suggests myeloid - but not erythroid - priming of MPP1 cells is predictive of downstream differentiation skews. This is perhaps unsurprising given that almost no expression of pro-erythroid genes is observed in WT MPP1 cells. However it once again raises the question of whether perturbed cells mapping to MPP1 actually remain functionally uncommitted or not. Nevertheless the evidence of increased stress response and transcriptionally inferior LT-HSCs in the erythroid-skewed perturbations is very suggestive that there is a coordinated response propogating from HSCs into the MPP compartment and driving the observed lineage priming.

4.6.3 Single-cell variability is a useful and informative new tool for scRNA-seq analysis

In addition to the insights gained from the expression values, a great deal of insight was gained from transforming the data into single cell variability data. This modality has largely been overlooked by the single-cell community but has clear potential within the setting of perturbation analysis. The finding that most perturbation models had far more differentially variable genes than differentially expressed genes is difficult to interpret. On one hand, this could be considered simply as an artefact of the numerical transformation; the variability data has no zero-dropouts but necessarily still contains all the technical noise associated with them (as they were part of the expression data used to create the variability data). It could be argued that this noise is mistakenly being interpreted as true signal in the variability data. On the other hand, it could be argued that the apparent increase in statistical power is essentially the result of combining the expression values with information on the connectivity of different cells through the nearest neighbour graph. The fact that differential variability actually correlates better with downstream abundance skews than the differential expression data certainly highlights how there is more information to be gleaned from scRNA-seq datasets by transforming the gene expression values in innovative ways.

A notable example of this is exhibited by the Dnmt3a and Npm1 models, which have by far the largest ratio between their number of DVGs and DEGs. There is accumulating evidence that whilst the loss of Dnmt3a predisposes to haematological malignancies, it does not by itself drive transformation into acute disease (Buscarlet et al., 2017; Chaudry and Chevassut, 2017; Ostrander et al., 2020). Of the mouse models here, the Dnmt3a model is the only one in which the mice do not develop leukaemia unless a second mutational hit occurs (data unpublished). Similarly, evidence suggests that the Npm1 model requires cooperating mutations to drive AML (Vassiliou et al., 2011). This may suggest that these mutations place the blood system in a ‘poised’ state with high single cell variability, that can easily be driven to further transformation (Ortmann et al., 2015). This is supported by the fact that whilst the variability within the HSC/MPP compartment is very similar to models such as Jak2, this similarity does not carry over to the differential expression results.

Overall, the uncommitted progenitor landscapes of the perturbations exhibit a clear trend transcriptionally impaired LT-HSCs, increased expression of stress markers and increased single cell variability levels are associated with differentiation skewing towards the erythroid lineage and vice versa, which first manifests itself in anti-myeloid lineage priming within the MPP1 compartment. The hypothesis that this is at least partially driven by epigenetic alterations in which DNA hypomethylation is associated with increased expression variability (and vice-versa) - and that this leads to a stress response and an increased proportion of cells not entering the ‘standard’ route of myeloid differentiation - is attractive, but much work is needed to validate it. Nevertheless results from this chapter have proved that despite the large amounts of noise inherent to droplet-based scRNA-seq data and the largely featureless transcriptomic landscape displayed by mouse HSC/MPPs, it is still possible to extract a range of coherent signals from the data through innovative analyses.

4.6.4 Future Work

There are a myriad of directions in which the work in this chapter could be continued. Work is under way to experimentally validate both the myeloid lineage priming and the shift to higher activity states observed in the MPP1 compartment of the perturbation models. Work to further understand the potential interplay between epigenetic state and transcriptional variability will focus on single-cell DNA methylation/ATAC-seq assays, potentially parallel to transcriptome sequencing through multimodal techniques such as scNMT-seq (Argelaguet et al., 2019). If further evidence can be found for causal links between differential transcription factor methylation and HSC fate bias, this will greatly aid our mechanistic understanding of these perturbations (Izzo et al., 2020).

Computationally, an exciting direction is to build this perturbation data into time-resolved flux models of haematopoiesis. Modelling how a shift in the MPP1 state is likely to impact the size and kinetics of downstream progenitor compartments could prove invaluable in taking single-cell perturbation analysis out of the ‘snapshot’ era and into the realm of real-time experiments, such as pulse-chase labelling systems capable of measuring flux through haematopoietic compartments (Akinduro et al., 2018).

4.6.5 Summary

In summary, this chapter has discussed a comprehensive analysis of the uncommitted progenitor compartment across different pre-leukaemic perturbation models. Insights have been gained at both the cellular and molecular scales - including single cell variability and cell cycle analyses - that provide an integrated view of how these perturbations are driving aberrant haematopoiesis.

Chapter 5

Dissecting Haematopoietic Responses in a Range of Other Perturbation Models

Parts of this chapter have been modified from Haltalli et al. (2020), Dingler et al. (2020), and Prins et al. (2020) on which Sam Watcham carried out bioinformatic analysis of single-cell data. In Haltalli et. al. (2020), single-cell experimental work was carried out by Myriam Haltalli and Nicola Wilson. In Dingler et. al. (2020), single-cell experimental work was carried out by Felix Dingler, Meng Wang and Nicola Wilson. In Prins et. al. (2020), single-cell experimental work was carried out by Daniel Prins, June Park, Juan Li and Nicola Wilson. In each case, initial preprocessing of the data was carried out by Rebecca Hannah (running 10X Genomics CellRanger pipeline). Subsequent to this, all analyses were performed and all figures were produced by Sam Watcham, except where explicitly stated in the text.

5.1 Background

This chapter is comprised of three separate computational analyses of haematopoietic perturbations, each placed within the context of a larger body of experimental work. Relevant background information is located at the start of each section.

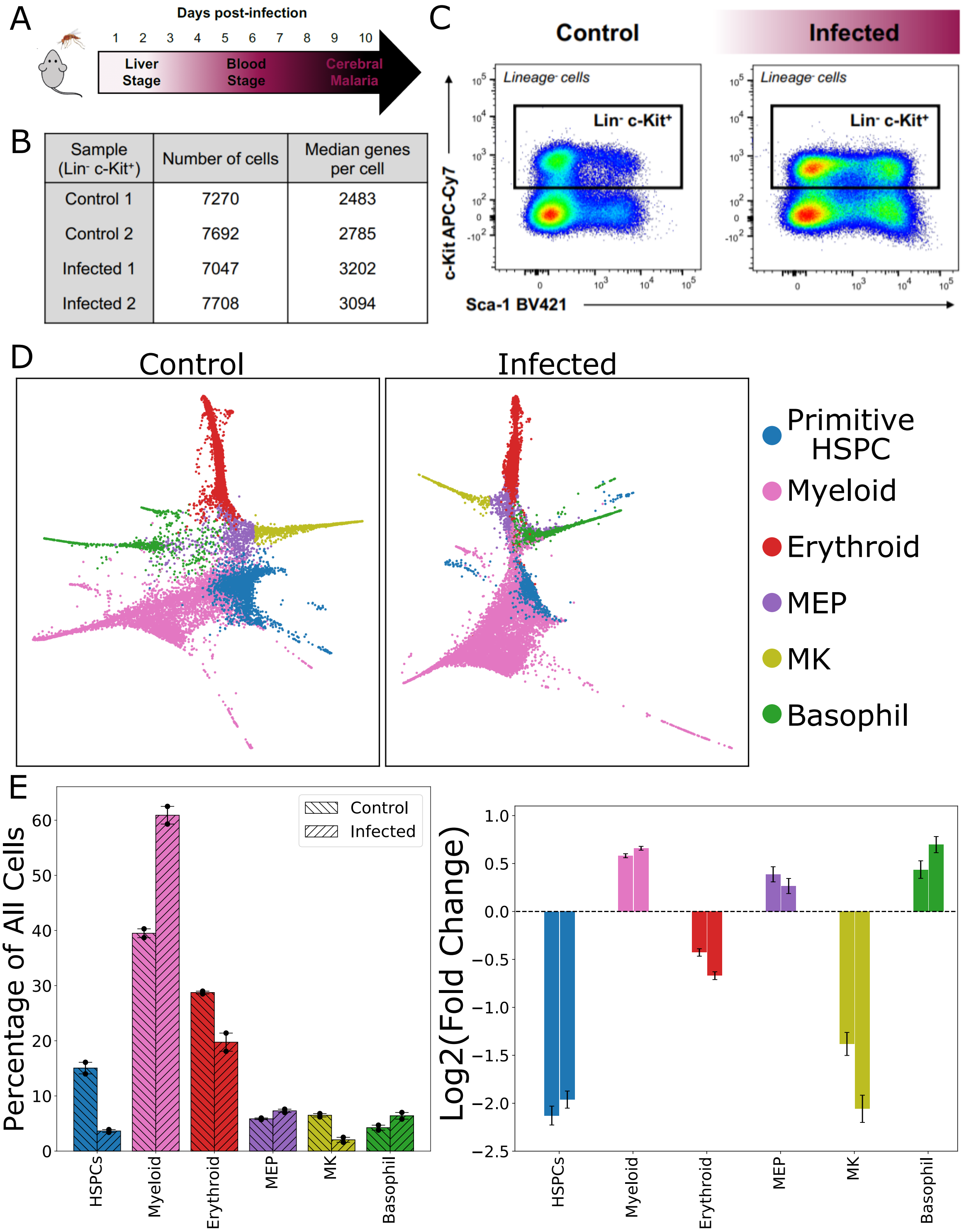
5.2 Revealing the extent of haematopoietic reprogramming during *P. berghei* infection

The response of the haematopoietic system to acute infection has been well studied. The destruction of mature cells through both the innate and adaptive immune response is known to induce skewed differentiation in immature HSPCs (Esplin et al., 2011; King and Goodell, 2011). Furthermore the infection response has been shown to specifically impact HSC dynamics, self renewal and their interaction with the bone marrow microenvironment (Baldridge et al., 2010; Essers et al., 2009; Takizawa et al., 2011). For example, HSCs can be driven out of quiescence and into a highly mobilised, cycling state through the presence of inflammatory cytokines, potentially leading to stem cell exhaustion (Takizawa et al., 2012). This further manifests itself as a loss of HSC engraftment within a transplantation setting, once again highlighting the link between increased cell cycle activity and loss of HSC functionality.

Malaria is a severe and widespread infectious disease that accounts for at least 400,000 deaths a year. Is is caused by pathogenic eukaryotes of genus *Plasmodium*, and is transmitted through the bites of *Plasmodium*-infected female mosquitoes. The infection starts in the liver before leading to large-scale perturbation of the haematopoietic system and the specific targeting of red blood cells. It is often fatal in humans and other species after causing anaemia, organ failure and cerebral malaria through obstruction of blood capillaries by the parasite. *Plasmodium berghei* has been successfully used as a model for investigating human malaria, due to its similarities to human *Plasmodium* species and its ability to infect rodents. In this model, mice are infected with malaria through mosquito bites, therefore simulating all stages of the disease (Vainieri et al., 2016).

In this work, a *P. berghei*-infected mouse model was used to assess the effects of malaria infection on the blood progenitor landscape as a whole. Parallel to looking experimentally at the proliferation dynamics of infected HSCs, scRNA-seq was applied to gain a global overview of the haematopoietic shifts driven by the infection, and to gain a fine-resolution dataset of the transcriptional impact on specific haematopoietic cell types including HSCs. Despite the severity of malaria in the human population, relatively little work has looked specifically at its impact on immature blood cells at a single cell level. Additionally the specific mechanisms through which HSPCs are impacted during malarial infection have not been identified.

The Lin- c-Kit+ (LK) progenitor compartments of two infected and two control (mice bitten by mosquitoes not carrying *P. berghei*) mice were sequenced 7 days after infection, during the height of the ‘blood stage’ of infection progression, but before the onset of cerebral malaria and death (Figure 5.1A). Similar cell numbers were retrieved for each mouse, with the median number of genes detected per cell being higher in the infected mice (Figure 5.1B). FACS sorting highlighted a large upregulation of Sca-1 surface protein in the infected mice (Figure 5.1C). This has been observed previously (Vainieri et al., 2016), where it was unclear if this expansion of the Sca-1+ compartment had arisen from mature progenitors



Sample (Lin- c-Kit+)	Number of cells	Median genes per cell
Control 1	7270	2483
Control 2	7692	2785
Infected 1	7047	3202
Infected 2	7708	3094

Fig. 5.1 Malaria infection skews haematopoietic differentiation towards the myeloid lineages. (A) Schematic of disease progression in infected mice. (B) Summary of sequencing results for the two control and two infected mice. (C) Representative sorting gate used to identify LK cells from control and infected mice at day 7 after infection. (D) Force-directed graphs for control and infected cells coloured by cluster identity. (E) Quantification of changes in cluster proportions between control and infected mice. Panels A, B and C are taken from Haltalli et al. (2020). MK: Megakaryocyte, HSPC: Haematopoietic Stem and Progenitor Cell.

re-expressing Sca-1 or from a significant increase in the proportion of the most immature cells.

After quality control, force-directed graphs were calculated for the control and infected cells separately. The control cells were partitioned using Louvain clustering into six clusters. Each infected cell was then mapped to its nearest control cluster using the same method introduced in Chapter 3 (Figure 5.1D), and the changes in cluster proportions between control and infected mice were quantified (Figure 5.1E). This revealed that 7 days after the onset of infection, the haematopoietic system of infected mice had shifted towards the myeloid and basophil trajectories at the expense of the erythrocyte and megakaryocyte trajectories. It is notable that whilst the infection targets mature red blood cells, it causes differentiation skewing away from the erythroid trajectory in the bone marrow. Additionally, the number of cells mapping to the 'Primitive HSPC' cluster - containing the most immature cells as identified by marker genes - was significantly reduced in the infected mice. To reconcile this with the increase in the Sca-1+ compartment seen by FACS, *Sca1* expression was visualised (Figure 5.2A). This clearly showed that the infection results in expression of *Sca1* by all cells in the LK gate at the transcriptomic level, despite many of them clearly mapping to regions that are *Sca1*- in the control mice. Hence the proportion of the most immature haematopoietic progenitors actually decreases due to the infection.

To further assess the effects of the infection on the most immature progenitors, the hscScore algorithm (Hamey and Gottgens, 2019) introduced in Chapter 4 was applied to both control and infected cells (Figure 5.2B). This revealed that the highest-scoring infected cells displayed greatly reduced transcriptional similarities to functional HSCs when compared to control. Quantification of the hscScores from the top 2% of highest-scoring cells from both control and infected mice confirmed this (Figure 5.2C). Experimental work performed by Myriam Haltalli corroborated these findings, with FACS-sorted HSCs from infected mice unable to reconstitute irradiated mice in primary transplantation assays (Haltalli et al., 2020) (Figure 5.2D). In addition, the absolute number of ESLAM HSCs (CD48- CD150+ CD45+ EPCR+) and CD48neg HSCs (LSK CD150+ CD48neg, (Akinduro et al., 2018)) was significantly reduced in infected mice (Figure 5.2D). Hence the transcriptional data supports the hypothesis that HSCs from infected mice are reduced in number and are functionally impaired. This occurs despite an apparent increase in the proportion of LSK cells due to the erroneous upregulation of *Sca1* by infected cells. Notably, when HSCs from primary recipients were sorted and a secondary transplantation performed, engraftment levels showed no difference between infected and control mice. This suggests that a significantly reduced but still measurable number of HSCs from infected mice maintain their reconstituting ability

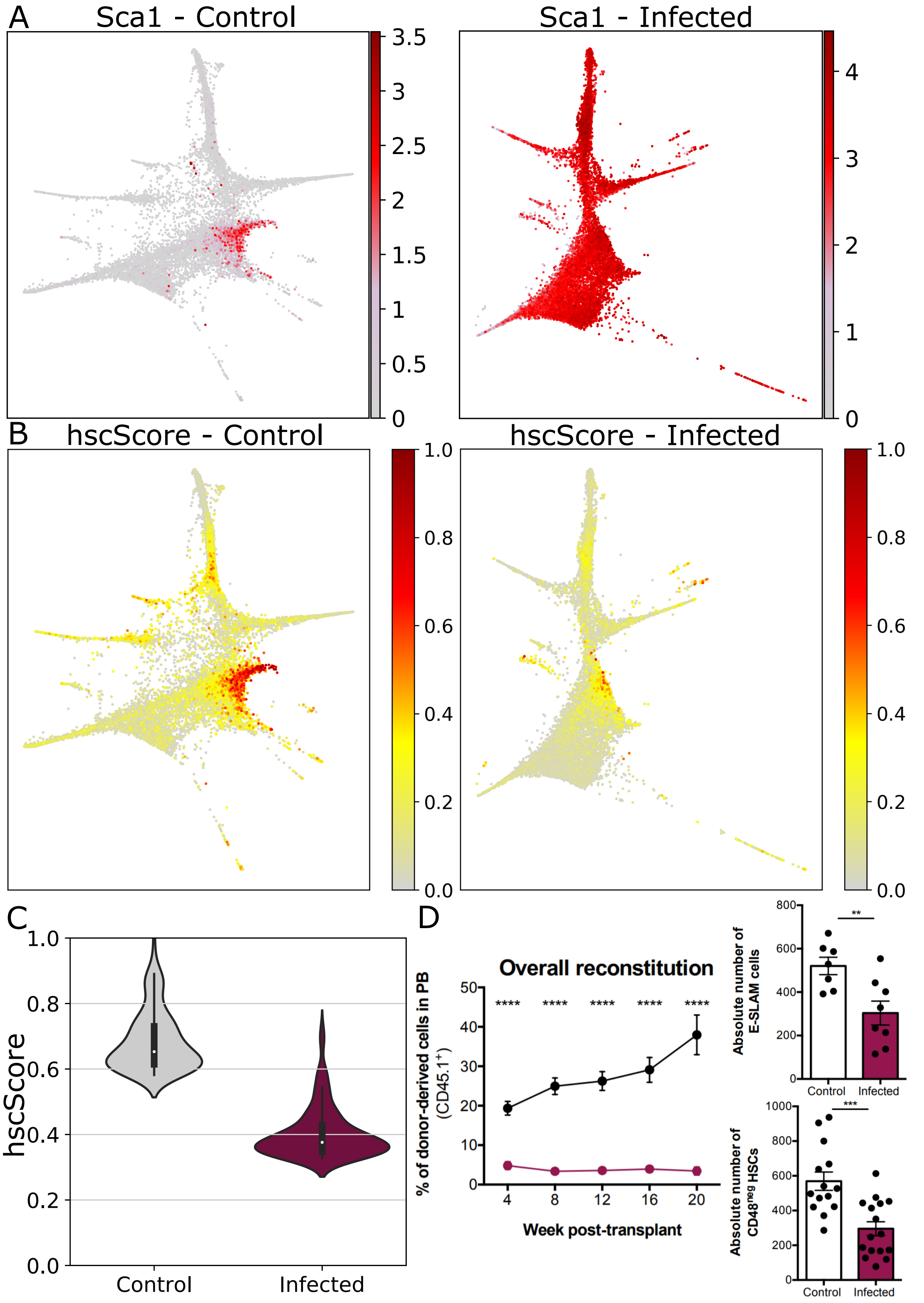


Fig. 5.2 Infected HSCs are functionally impaired compared to control. (A) Expression of *Sca1* in control and infected cells. (B) hscScores for control and infected cells. (C) The top 2% of cells with the highest hscScores from control and infected cells respectively. (D) 200 HSCs (LSK SLAM) were transplanted into lethally irradiated primary recipients with 300,000 whole bone marrow (WBM) support. WBM was pooled after 20 weeks and total blood reconstitution of engrafted HSCs was calculated. The absolute number of ESLAM HSCs or CD48neg HSCs were also calculated. Error bars represent the mean ± s.e.m. Statistical differences were determined using Student's t-tests with Bonferroni correction (** represents $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$).

across serial transplantation, potentially indicating that whilst the majority of HSCs from infected mice exhaust in response to the infection, the most quiescent can avoid any functional impairment.

Differential expression was performed between control and infected cells for each cluster. It was immediately clear that the principal pathway being highlighted across all clusters was interferon (IFN) cytokine signalling, and in particular IFN-γ signalling. To highlight this, a list of 20 genes that had at least four-fold upregulation in all six clusters were identified. These included *Igtp* (interferon gamma-induced GTPase), *Irf7* (interferon regulatory factor 7) and *Gbp2* (interferon-induced guanylate-binding protein 2) amongst many others related to IFN signalling (Figure 5.3A). Using these 20 genes as a base set, the transcriptome was searched for genes whose expression was highly correlated with the base set (see methods). This identified 89 further genes relating to IFN signalling for a total of 109 ‘driving genes’. The importance of these genes was observed when calculating force-directed graphs for the combined control and infected cells, with and without the driving genes (Figure 5.3B). Removing the driving genes also removed much of the distance between control and infected cells in the visualisation, suggesting IFN signalling is responsible for the majority of the differences between conditions. Even after removing these genes, the most significant DEGs that remained were also almost exclusively related to IFN signalling. It was notable that almost no cluster-specific DEGs could be identified in the infected cells, such was the dominance of the interferon response. This was highlighted in the MA plots for each cluster, where in each case the most significant DEGs were the 109 driving genes that could be identified by simply performing differential expression between all cells at once (Figure 5.3C,D). This remained true even in the ‘Primitive HSPC’ cluster, indicating that IFN upregulation may be responsible for the impairment of HSC functionality (Baldridge et al., 2010).

Overall, this work identified how *P. berghei* infection led to pro-myeloid differentiation skewing within the LK progenitor compartment, accompanied by a clear proportional loss of the most immature blood cells. Transcriptionally, immature infected cells lose their HSC signature suggesting impaired functionality. These changes appear to be driven largely by IFN signalling pathways driven by high levels of IFN cytokines in the bone marrow. It is interesting to note that in this infection model, a loss of ‘stemness’ in HSCs as measured by the hscScore occurs alongside pro-myeloid differentiation skewing. This is in contradiction to the results for the pre-leukaemic perturbation models analysed in Chapter 4. However emergency myelopoiesis has been observed in infection models previously (Boettcher et al., 2012; Schurch et al., 2014), and it is likely that the bone marrow cytokine signalling induced

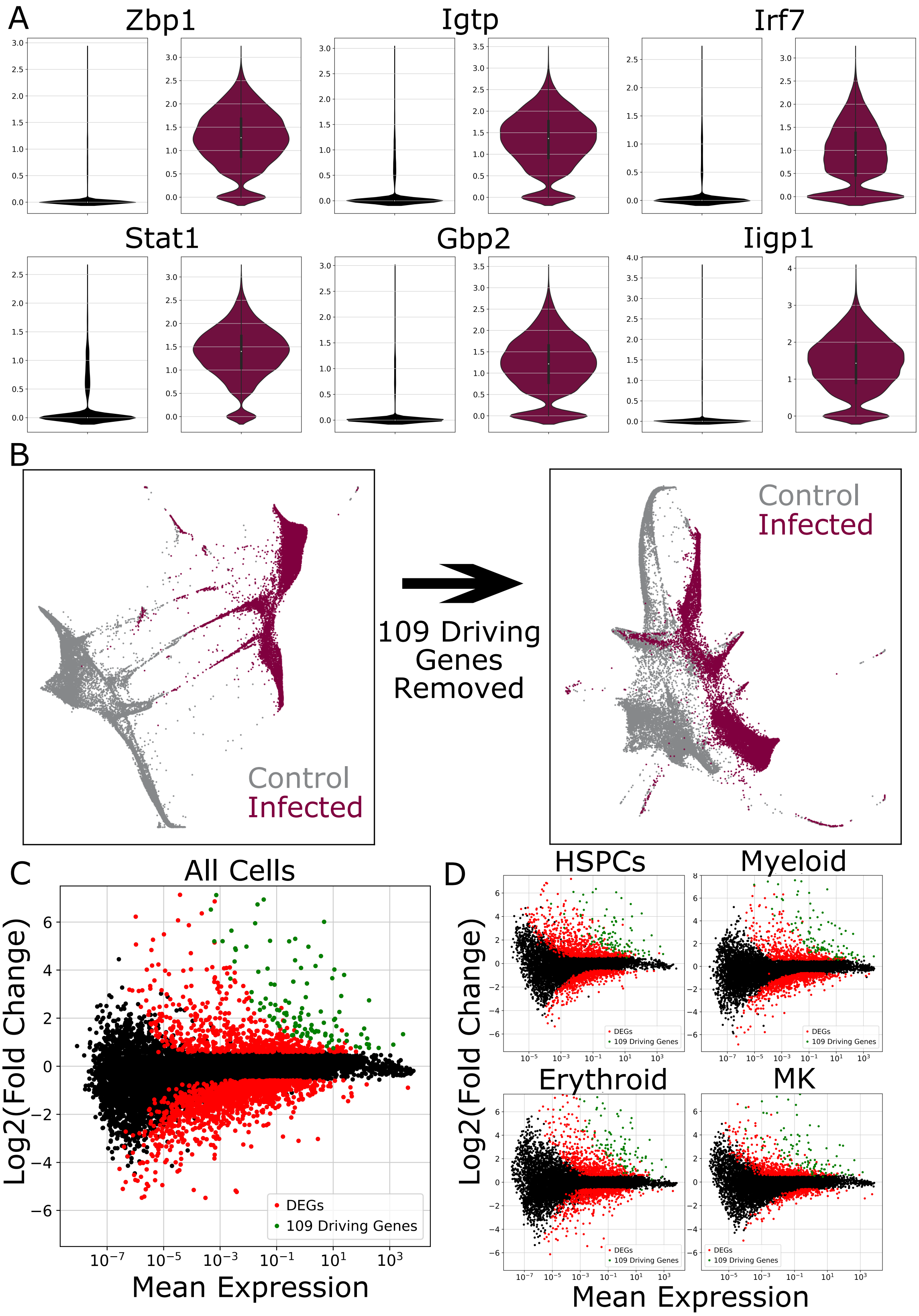


Fig. 5.3 The molecular response to malaria infection is driven by interferon signalling. (A) Force-directed graphs of the combined control and infected cells with all genes (left) and with the 109 IFN driving genes removed (right). (B) Violin plots of representative IFN gene expression in control and infected cells. (C) MA plot comparing all infected cells to all control cells. The 109 driving genes are highlighted in green. (D) Cluster-specific MA plots with the 109 driving genes highlighted in green.

by the infection overrides all other signals to such an extent that it is not reasonable to compare the malaria model to any others. Indeed, the cellular abundance changes observed in the malaria model are far greater than in any of the pre-leukaemic perturbations. Interestingly, when the improved varID method was applied to the malaria experiment and the results for the ‘Primitive HSPC’ cluster were analysed, no overall shift to greater or lesser variabilities was observed across the transcriptome, such as those in Figure 4.11. There is little evidence to suggest that *P. berghei* infection leads to epigenetic changes such as DNA methylation in HSCs. Hence this may be indirect evidence that large transcriptome-wide single-cell variability changes occur as a result of epigenetic alterations.

5.3 Aldehyde accumulation disrupts haematopoietic development

Endogenous aldehydes are generated by numerous cellular processes such as lipid peroxidation, glycation and response to oxidative stress (Voulgaridou et al., 2011). Many of them are highly genotoxic and can react directly with DNA, leading to DNA damage such as interstrand crosslinks (Hodskinson et al., 2020). Their carcinogenic and mutagenic effects have been implicated in the pathology of various diseases ranging from neurodegenerative disorders to heart disease. In blood, the DNA damage caused by endogenous aldehydes has been shown to limit the function of HSCs and impair the immature haematopoietic system through double stranded DNA breaks and chromosome rearrangements, leading to loss of reconstitution, diminished self renewal and increased apoptosis (Garaycoechea et al., 2018; Rossi et al., 2007).

Across the human genome there are at least 19 aldehyde-dehydrogenase (ALDH-) genes responsible for detoxifying aldehydes (Jackson et al., 2011). ALDH2 in particular is responsible for oxidising acetaldehyde to inert acetate. Acetaldehyde is a direct product of ethanol oxidation and likely confers the majority of alcohol’s toxic effects in humans. ADH5 is another enzyme responsible for aldehyde detoxification; however instead of acting on free aldehydes, it oxidises glutathione, a conjugate of formaldehyde. In both cases, the result of ALDH2 and ADH5 activity is the production of inert metabolites from damaging aldehydes.

It has been shown that crossing either *Aldh2-/-* or *Adh5-/-* mice with mice that exhibit loss-of-function of the DNA crosslink repair gene *Fancd2* leads to mice that rapidly develop leukaemia and complete haematopoietic failure (Garaycoechea et al., 2018; Langevin et al., 2011). However a strong haematopoietic phenotype is not seen in the *Aldh2-/-* or *Adh5-/-* mice themselves. These findings have suggested that even if a single aldehyde detoxifying gene is compromised, DNA repair genes are able to compensate for this loss. However a subsequent loss of the DNA repair pathway leaves the haematopoietic system defenceless. Hence there appears to be two 'tiers' of protection against the effects of aldehydes. Nevertheless it remains unclear whether different aldehyde detoxifying enzymes are functionally related and/or redundant, or whether specific aldehydes have greater physiological importance than others.

In this study, scRNA-seq was performed on WT, *Aldh2-/-*, *Adh5-/-* and double knockout (DKO, *Aldh2-/-Adh5-/-*) mice as part of a larger body of experimental work that aimed to

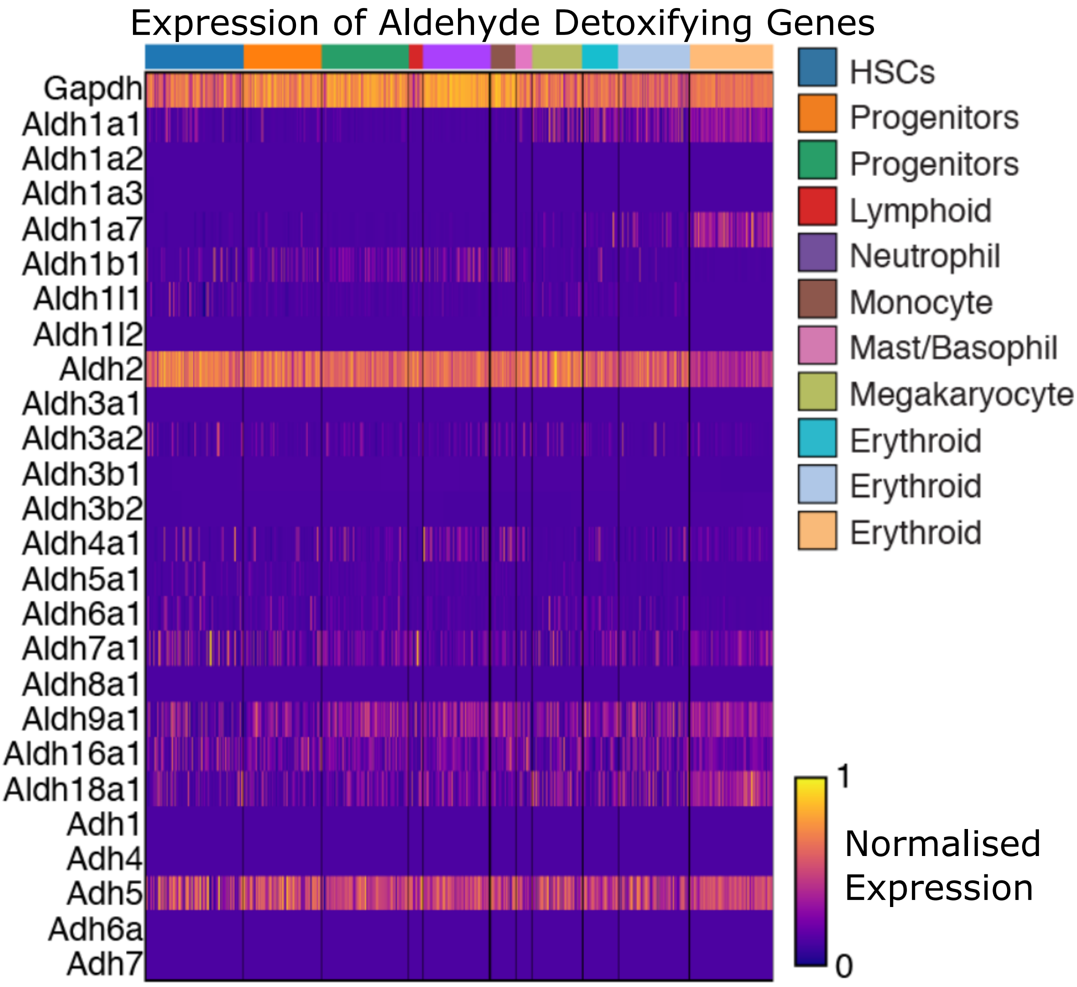


Fig. 5.4 ***Aldh2* and *Adh5* are the two most important aldehyde detoxifying genes in the blood system.** Heatmap of aldehyde detoxifying gene expression in WT haematopoiesis. Cells are coloured by cluster annotation introduced in Figure 5.5. Expression scores are normalised between zero and one for each gene. *Gapdh* is included as a housekeeping gene.

understand the effects of losing these two key detoxifying enzymes on the blood system. Initial analysis of WT LK cells revealed that of the many aldehyde detoxifying genes in the transcriptome, *Aldh2* and *Adh5* are the two genes with the highest and most consistent expression across all haematopoietic lineages (Figure 5.4). This highlights the importance of these enzymes within the blood system. As mentioned above, the single knockout *Aldh2-/-* and *Adh5-/-* mice show only mild haematopoietic phenotypes. However the DKO *Aldh2-/- Adh5-/-* mice display significantly retarded growth compared to the single knockouts and are predisposed to leukaemia and early death (Dingler et al., 2020). In addition they present with mild anaemia and depressed lymphocyte counts. Therefore there is clear evidence of genetic redundancy between ALDH2 and ADH5, as inactivation of both genes simultaneously leads to growth failure, malignancy and early death.

The LK compartment was sequenced for each of the four genotypes, resulting in over 33,000 transcriptomes after quality control (Figure 5.5A). Unsupervised Louvain clustering resulted in 12 clusters that were manually annotated based upon marker gene expression (Figure 5.5B). Whilst the WT and single knockout mice appeared to be well mixed on the UMAP representation, the DKO cells displayed clear changes compared to other genotypes. The most striking of these were in the immature 'HSC' cluster and the erythroid trajectory (Figure 5.5C). Visualising the expression of *Aldh2* and *Adh5* across the LK landscape revealed that whilst the knockouts had worked effectively, there was no evidence of compensatory expression changes of *Adh5* in *Aldh2-/-* mice and vice versa (Figure 5.5D). This supports a hypothesis in which ALDH2 and ADH5 are genetically redundant.

To begin quantifying the cellular and molecular changes occurring in the perturbed mice, the proportion of each cluster belonging to each genotype was calculated (Figure 5.6A). This confirmed a significant under-representation of DKO cells in the 'HSC' cluster, alongside a clear overabundance of cells identified as lymphoid. However whilst the end of the DKO erythroid trajectory is separate from the other genotypes in the UMAP, the abundance of erythroid cells is not significantly altered in the erythroid trajectory overall. To assess the corresponding molecular changes, differential expression was performed for each genotype across all clusters. The number of significant DEGs (both upregulated and downregulated) was greatly increased in the DKO compared to the single knockouts in all cell types (Figure 5.6B). This suggests that the entire LK landscape is responding transcriptionally to the combined inactivation of ALDH2 and ADH5.

Given the clear changes occurring in DKO cells within the 'HSC' cluster at both the cellular and molecular levels, further investigation of this cluster was undertaken. Cells belonging to this cluster were revisualised and reclustered, generating seven fine clusters

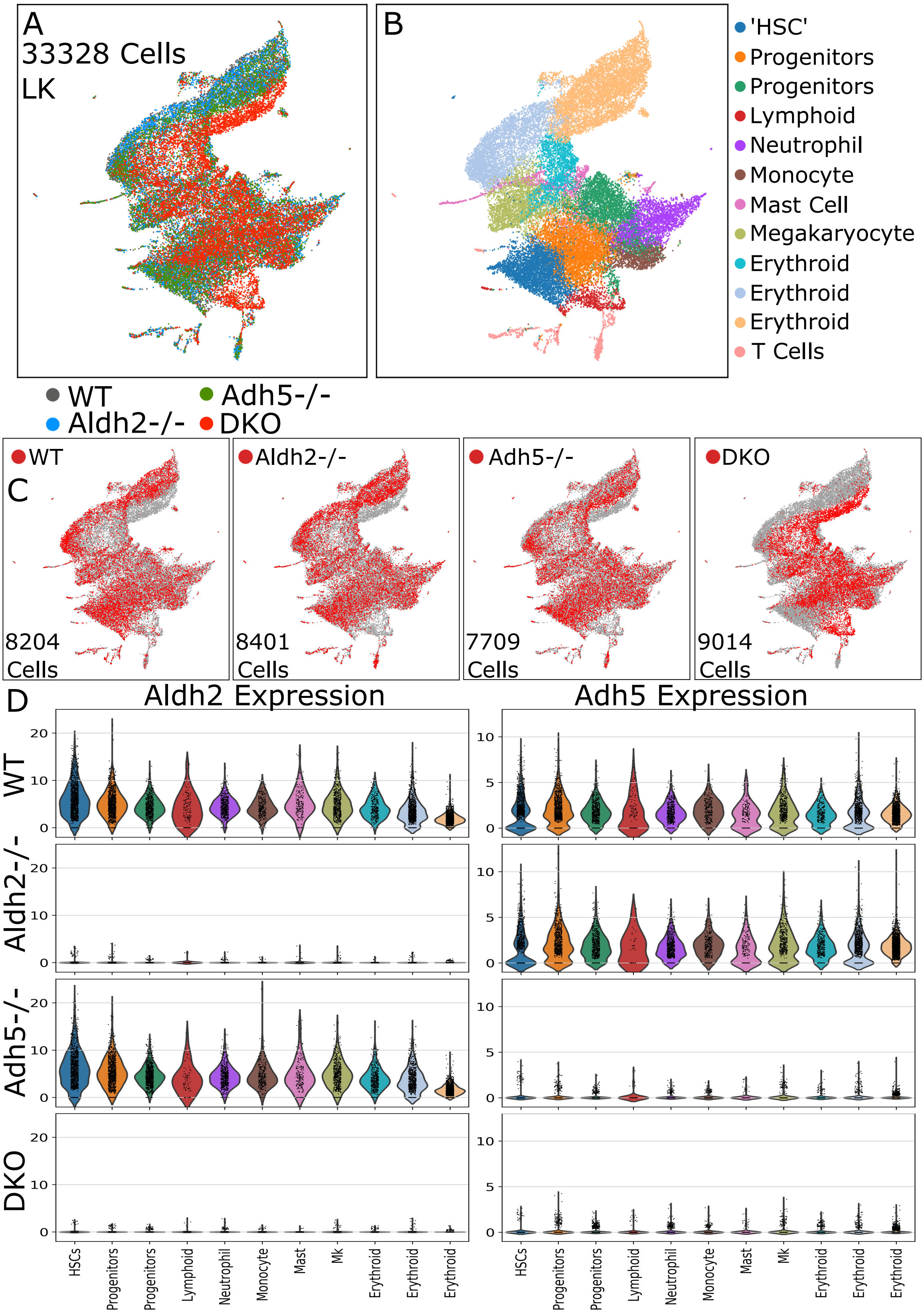


Fig. 5.5 The LK compartment of *Aldh2-/-Adh5-/-* mice displays altered distributions of cellular populations compared to single knockouts. (A) UMAP representation of the LK landscape, coloured by genotype. (B) Louvain clustering of the data, annotated according to marker gene expression. (C) Cells specific to each genotype are highlighted on the UMAP in red. (D) Violin plots showing the expression of *Aldh2* and *Adh5* across genotype and across clusters.

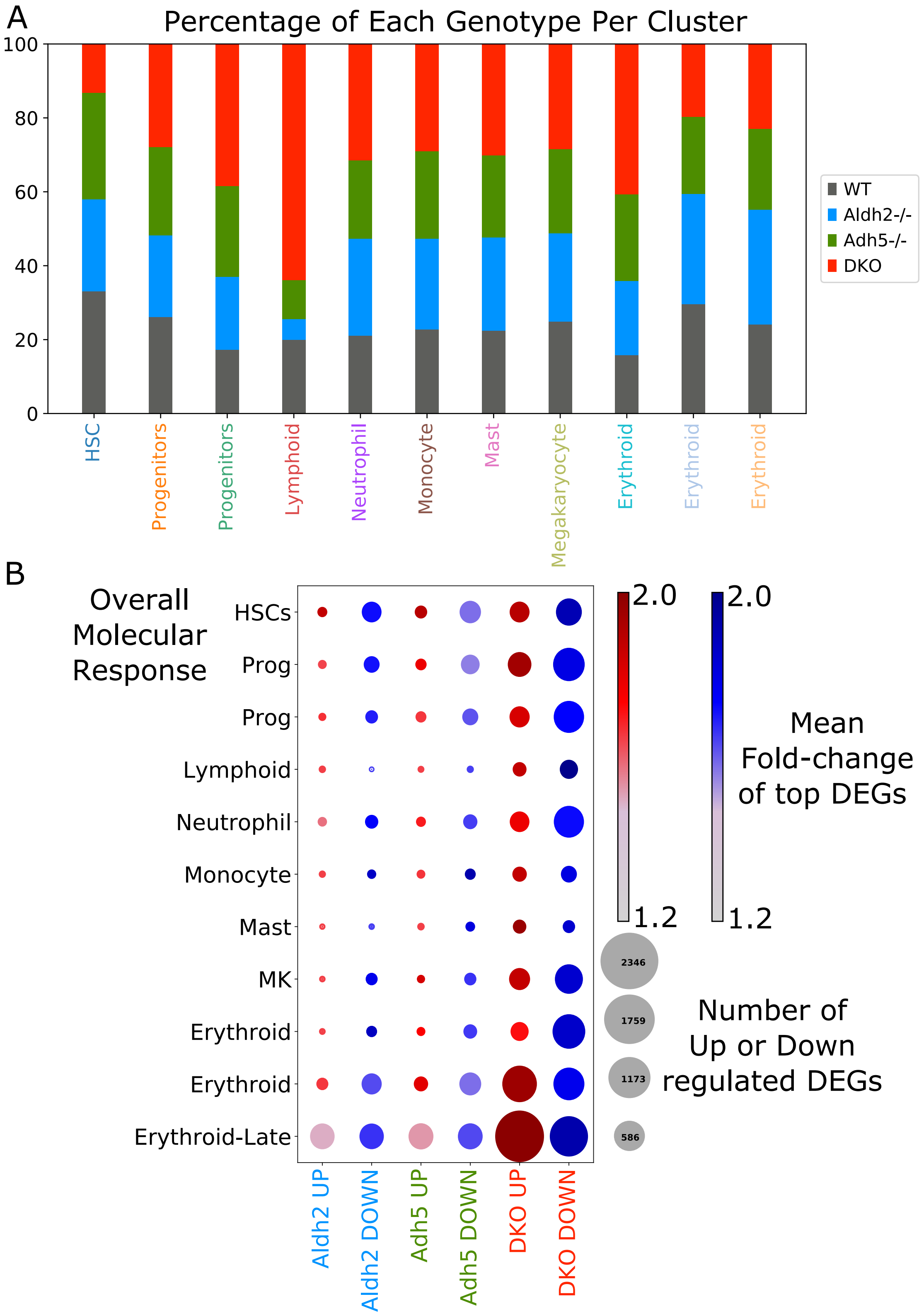


Fig. 5.6 ***Aldh2-/-Adh5-/-* mice display dramatic transcriptional changes.** (A) Bar chart showing the proportion of each cluster arising from each genotype. (B) For each genotype, DEGs were calculated in each cluster using cutoffs of adj. p-value<0.05 and fold-change>1.2. Circle size represents the number of DEGs either up- or down-regulated in a specific cluster. Circle colour represents the mean fold-change of the top 50 up- or down-regulated DEGs (measured by adj. p-value) in each cluster, using red for upregulated genes and blue for downregulated genes.

(Figure 5.7A). These were annotated as either HSC, lymphoid-biased or myeloid-biased based upon the onset of known marker genes such as *Procr*, *Flt3* and *Mpo* (Figure 5.7B). DKO cells were clearly depleted in abundance within the HSC and lymphoid-biased clusters compared to the other genotypes, suggesting a loss of the most functional HSCs (Figure 5.7C). Applying the hscScore algorithm (Hamey and Gottgens, 2019) to the data further highlighted a clear loss of HSC transcriptional identity within the DKO mice (Figure 5.7D). Even when considering the LK compartment as a whole, the 'HSC' cluster was significantly depleted compared to controls (Figure 5.7E). Overall, it is clear that the DKO mice have substantial cellular and transcriptional defects within the HSC compartment. Further investigation of the functional potential of DKO HSCs may help to explain the severe haematopoietic phenotype observed in these mice, and will help to understand the links to similar mouse models such as *Aldh2-/-Fancd2-/-* (Garaycoechea et al., 2018).

Since aldehyde accumulation is known to lead to DNA damage, it was important to address whether the mutant mice display transcriptional evidence of a DNA repair response. Furthermore, the tumour suppressor p53 is responsible for regulating the cellular response to DNA damage (Lakin and Jackson, 1999), and aldehyde accumulation has been shown to increase p53 levels in blood cells (Garaycoechea et al., 2018). Hence it is interesting to ask whether p53 target genes are also transcriptionally affected at the single-cell level. Firstly, differential expression was performed between genotypes using all LK cells to capture as much signal as possible. Intersecting these results with the gene ontology term for 'DNA Repair' (GO:0006281) immediately highlighted a large DNA repair response in the DKO that is absent from the single knockouts (Figure 5.8A). This is clear evidence that across the LK landscape as a whole, DNA damage and hence upregulation of DNA repair genes is only required when both ALDH2 and ADH5 enzymes are absent. Significant hits include key DNA repair genes such as *Neil3* and *Pclaf* (Li et al., 2016; Semlow et al., 2016), as well as well-known p53 targets *Cdkn1a* (p21) and *Pcbp4* (Ceccaldi et al., 2012; Scoumanne et al., 2011) (Figure 5.8B). Further upregulated hits include the Fanconi-anaemia genes *Fancd2*, *Fanci*, *Brca1* and *Brca2*, all of which are known to be key players in DNA repair and targets of p53.

Further quantification of the DNA repair and p53 target response was observed by calculating a per-cell score for each pathway based on gene expression, in an identical fashion to the G2M score introduced in Chapter 4. Genes belonging to the DNA Repair GO term were used to calculate the DNA repair module score (in total, 474 genes), whilst a validated list of p53 target genes were used to calculate the 'p53 Targets' module score (in total, 147 genes). Eleven genes were present in both lists; these were removed from the DNA

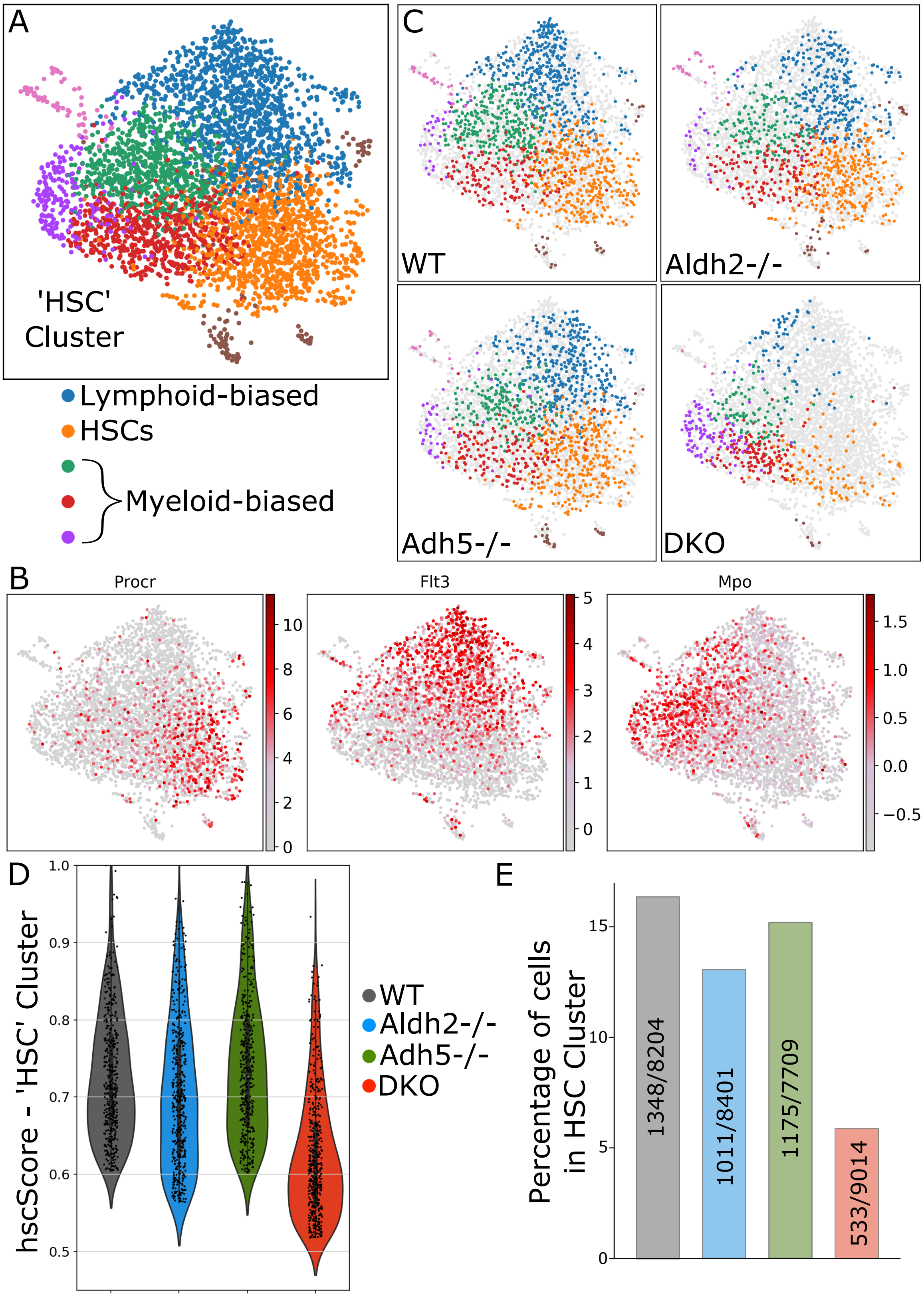


Fig. 5.7 The HSC compartment of *Aldh2-/-Adh5-/-* mice is defective both transcriptionally and in terms of abundance (A) UMAP of the 'HSC' cluster, reclustered using Louvain clustering. (B) Example marker gene expression used to annotate the new cluster in panel A. (C) Cells from each genotype are highlighted, with colours corresponding to the new louvain clusters. (D) Violin plot of hscScores for all cells belonging to the 'HSC' cluster. (E) The proportion of LK cells within the 'HSC' cluster for each genotype.

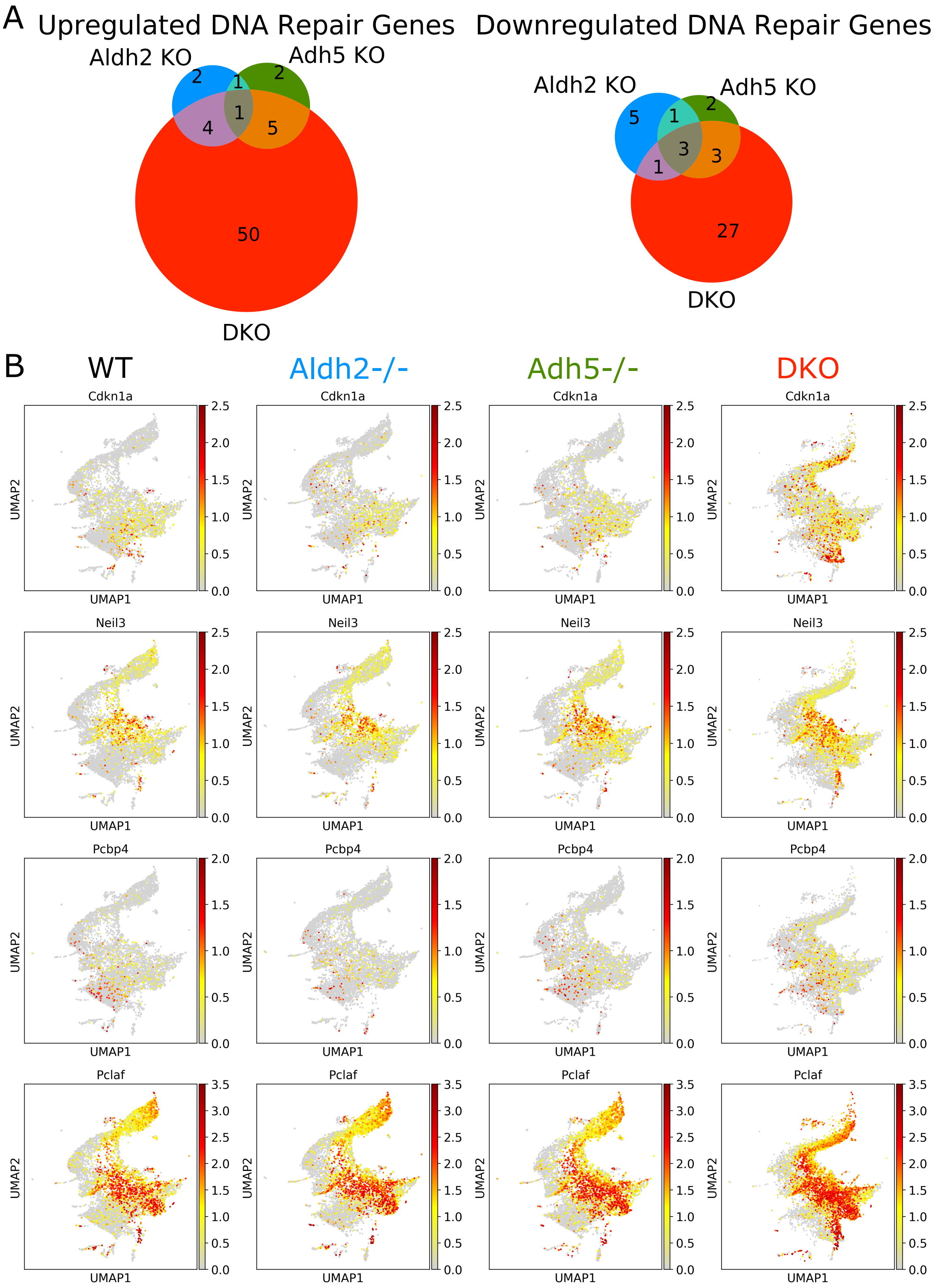


Fig. 5.8 A large DNA repair response occurs in *Aldh2-/-Adh5-/-* mice. (A) Venn diagrams showing the number of genes belonging to the DNA Repair GO term which are differentially regulated across the three mutant genotypes. (B) UMAPs of gene expression for four DNA repair genes, split by genotype.

repair genelist to avoid duplication and create independent scores. Plotting these scores for each genotype across the LK landscape revealed a clear increase in both pathways within the DKO cells (Figure 5.9A). To test whether this response was being driven by specific cell types, both genelists were intersected with the differential expression results for each cluster, and the overall molecular response of each list across each cluster was visualised (Figure 5.9B). Once again, both pathways showed clear activation in the DKO that was not present in the single knockouts. Notably, the effects were clearly visible in all clusters including HSCs and progenitors. This is clear evidence that a system-wide transcriptional upregulation of DNA repair and p53 targets occurs as a result of combined ALDH2 and ADH5 inactivation that does not occur when only one of these enzymes is lost.

The above analyses suggest that whilst there are clear transcriptional changes occurring across the entire progenitor compartment of the DKO mice, the largest changes appear to be occurring within the erythroid trajectory. This is backed up by the UMAP representation of the data, where the DKO erythroid trajectory is pulled apart from the other genotypes. It was hypothesised that changes in cell cycle or apoptosis may be driving this separation. Differential expression of the three erythroid clusters combined revealed over 1700 DEGs in the DKO, compared to less than 300 in each of the single knockouts (Figure 5.10A). Intriguingly, many of the top upregulated DEGs occurring only in the DKO were the same genes upregulated in the erythroid trajectory of the Jak2 and W41 perturbation models (Figures 3.7, 3.11A). These included *Fabp5*, *Cda*, *Podxl* and *Atf4*. This may suggest that these genes are upregulated as part of a generic 'stressed erythropoiesis' response. It could also suggest that DNA damage plays a role in the Jak2 and W41 models, although further evidence of this was not observed. In the Jak2 and W41 models these changes were associated with a large increase in cellular abundance of erythroid cells; however, this is not seen in the DKO model. Cell cycle analysis of the erythroid cells suggested that there were a greater proportion of cells in G2/M and S phase in the DKO compared to control (Figure 5.10B, see methods). This is concordant with the Jak2 and W41 erythroid cells where a clear increase in cycling activity was observed (Figure 4.15A). There was also transcriptional evidence of increased erythroid apoptosis in all three genotypes compared to WT, as quantified by a module score created using the 'Intrinsic Apoptotic Signalling Pathway' GO term (Figure 5.10C). The strongest effect was observed in the DKO cells. Large changes in cell cycle or apoptosis scores were not observed for any other cell type in the DKO cells. Hence the evidence would suggest that on top of the cell-type independent DNA repair and p53 response occuring in the DKO, there are also erythroid-specific aberrations including increased cell cycling and cell death that may be contributing to the separation observed on the UMAP.

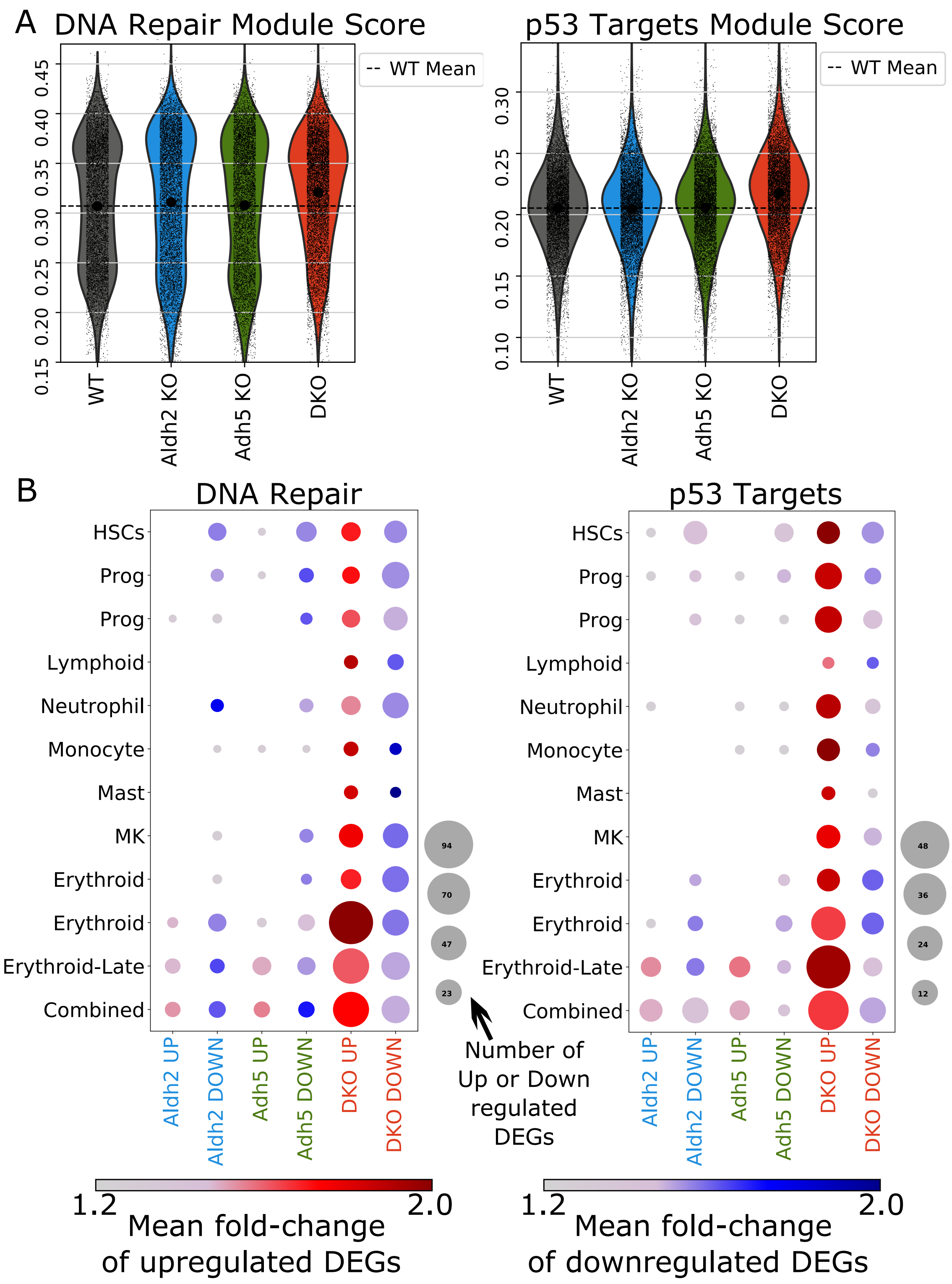


Fig. 5.9 DNA Repair and p53 target pathways are upregulated across all cell types in *Aldh2-/-Adh5-/-* mice. (A) Module scores for the 'DNA Repair' and 'p53 Targets' genelists. For each genotype the violin plot contains all LK cells. (B) For each genotype, the DNA repair and p53 target genelists were intersected with the DEGs for each cluster. Circle size represents the number of DNA Repair/p53 target genes either up- or down-regulated in a specific cluster. Circle colour represents the mean fold-change of the up- or down-regulated DEGs in each cluster, using red for upregulated genes and blue for downregulated genes.

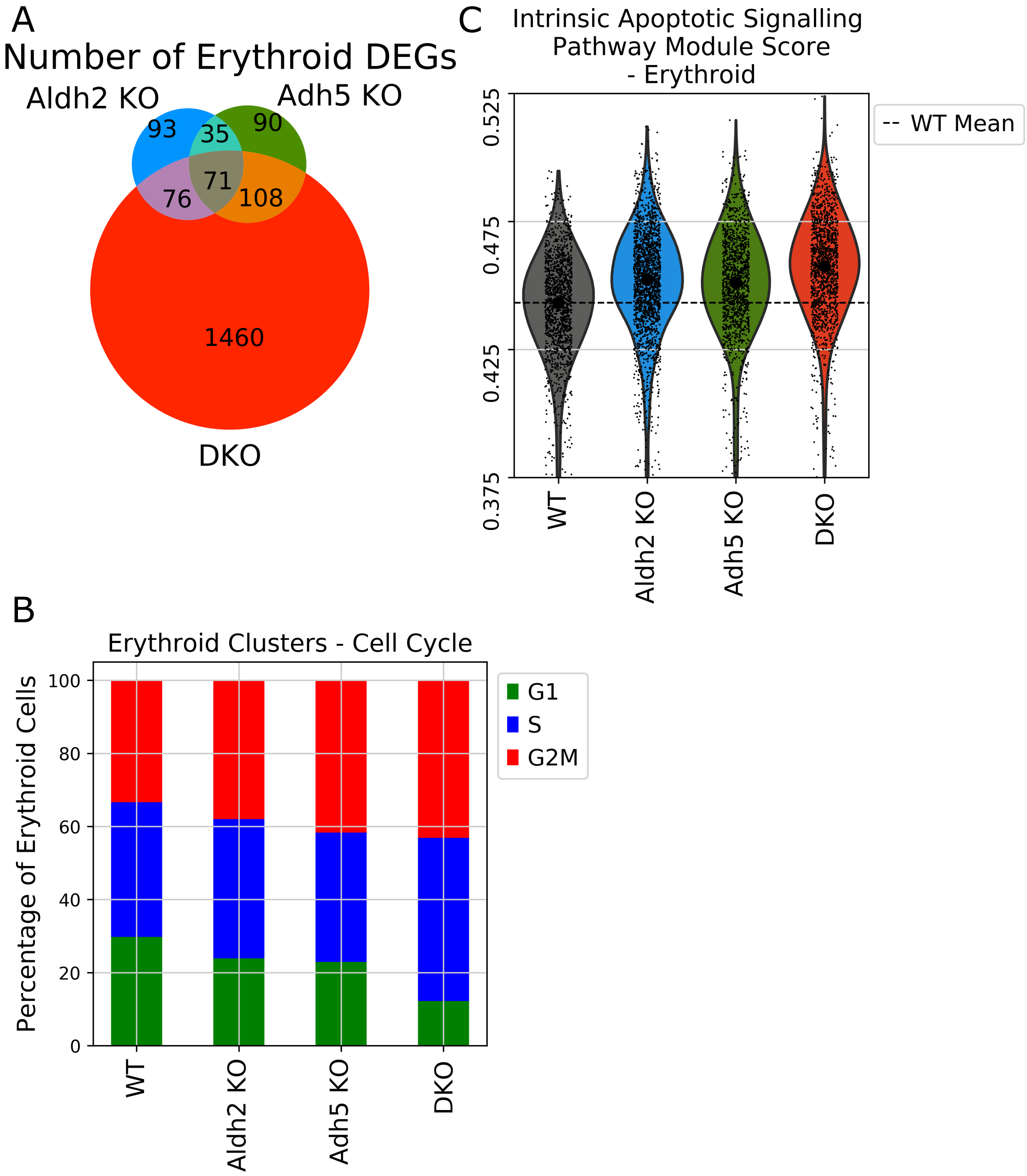


Fig. 5.10 Erythroid DKO cells display specific cell cycle and apoptosis alterations (A) Number of DEGs found for each genotype when combining the three erythroid clusters from Figure 5.4. (B) Each erythroid cell was assigned a cell cycle phase based on their transcriptome, and plotted across genotypes. (C) Module score for the 'Intrinsic Apoptotic Signalling Pathway' GO term for each genotype.

Overall, this single-cell data has revealed that whilst single knockout *Aldh2-/-* or *Adh5-/-* mice show only minor transcriptional changes compared to WT, DKO mice lacking both these detoxifying genes are severely compromised at a transcriptional level. Large scale molecular changes are observed across the LK landscape, including sizeable contributions from both DNA repair and p53 target pathways. Experimental work from this project has confirmed that endogenous formaldehyde accumulates in DKO cells as a result of the perturbation, and that this is responsible for DNA damage. Ongoing work is attempting to determine whether the haematopoietic phenotypes of the DKO perturbation are solely due to the effects of aldehydes on DNA, or whether certain aldehydes can also cause modifications at the RNA or protein levels. What is certainly clear is that loss of both ALDH2 and ADH5 has a profound effect on HSCs, greatly reducing their abundance and similarity to functional HSCs in terms of their transcriptome. Further work will be needed to assess the true functional impairment of these perturbed HSCs.

5.4 A mouse model of essential thrombocythaemia reshapes the haematopoietic progenitor landscape

In Chapters 3 and 4, a number of pre-leukaemic perturbation models were analysed. Many of these are associated with either clonal haematopoiesis (Tet2 and Dnmt3a) or more specifically with myeloproliferative neoplasms (MPNs), such as the Jak2 V617F model. As discussed in Chapter 1, the family of disorders labelled as MPNs includes polycythaemia vera (PV, characterised by an overproduction of red blood cells), myelofibrosis (MF, characterised by bone marrow fibrosis) and essential thrombocythaemia (ET, characterised by an overproduction of platelets). Indeed, mutations in Jak2 are found in >99% of PV patients and >50% of MF and ET patients (Grinfeld et al., 2017). The second most common mutation in MPNs occurs in the gene encoding for calreticulin (*Calr*), found in 25-40% of ET and MF patients (Klampfl et al., 2013; Nangalia et al., 2013). Frameshift mutations can cause mutant CALR to act as a rogue ligand for the thrombopoietin receptor MPL (TPOR), the major regulator of megakaryocyte and platelet formation. This leads to overactivation of numerous signalling pathways through MPL including JAK/STAT signalling (Elf et al., 2016; Nivarthi et al., 2016).

Recently, a mouse model of ET has been generated that displays the key phenotypes of the human disease, including a large overproduction of platelets (Li et al., 2018). In this project, scRNA-seq was performed using this mutant CALR model as part of a larger body of experimental work aiming to characterise the model at the single-cell level. Whilst mutant CALR is known to act through MPL, the specific differentiation stages, mechanisms and consequences of this interaction remain unknown. Further to this, there is increasing evidence that there may be multiple differentiation trajectories from HSCs to megakaryocytes, potentially including a direct trajectory from the earliest HSCs as a result of intrinsically MK-biased HSCs (Rodriguez-Fraticelli et al., 2018; Sanjuan-Pla et al., 2013; Sawai et al., 2016). The impact of perturbations on these potential trajectories remains largely unexplored, and may help to better understand megakaryocytic differentiation in wild-type haematopoiesis.

10X scRNA-seq was performed on LK and LSK samples from two pairs of WT and CALR homozygous mutant mice (henceforth called CALR DEL). The first pair were sequenced at 3 months of age, the second pair at 6 months of age. This occurred because the sequencing results for the first pair of mice returned far fewer WT cells than expected (972 WT LSK, 2479 WT LK, 4548 CALR DEL LSK, 7824 CALR DEL LK). Hence the sequencing was repeated; however age-matched mice were not available. The second pair returned better cell numbers (5959 WT LSK, 5139 WT LK, 7732 CALR DEL LSK, 7815 CALR DEL LK). This did however allow the affects of age on the CALR DEL model to be investigated. Both LK and LSK samples were collected because there was particular interest in the HSC region, which is enriched in the LSK cells.

Starting from the WT reference landscape introduced in Chapter 3 (Figure 5.11A), a fine-resolution Louvain clustering of the data was performed, to better delineate potential stages along the megakaryocytic trajectory. This results in 60 fine clusters, each of which belonged to one of the original clusters from Chapter 3. Partition-based Graph Abstraction (PAGA) was then performed in order to generate a graph that visualised the potential connections between these fine clusters (Figure 5.11B) (Wolf et al., 2019). Experimental work (performed by Daniel Prins and June Park) had located a novel WT cell population representing an intermediate state along the megakaryocyte trajectory that was highly MK-biased but retained proliferative behaviour compared to classical MK progenitors sorted using CD150 and CD41. This population was defined to be Lin-EPCR-CD48-CD150+CD45+, and termed ‘proliferative MK-progenitors’ or pMKPs. Analysis of gene expression within the WT reference suggested that this population should be associated with a single fine cluster that showed no expression of *Procr* or *Cd48* but clear expression of *Slamf1* (CD150) and *Ptprc* (CD45) (Figure 5.11C, pMKPs marked by arrow).

In CALR DEL mice, the pMKP population appeared to be greatly expanded when sorted using FACS (Figure 5.11D,E). To put this in the context of abundance changes over the entire LK landscape, the voting method for visualising differential abundance was applied to both pairs of sequenced mice (Figure 5.12A). For comparison, the scalebar in Figure 5.12A is identical to that used in Figure 3.4 for the other pre-leukaemic models. At 3 months, there was a clear expansion of the HSC and MK regions of the landscape (it is worth remembering

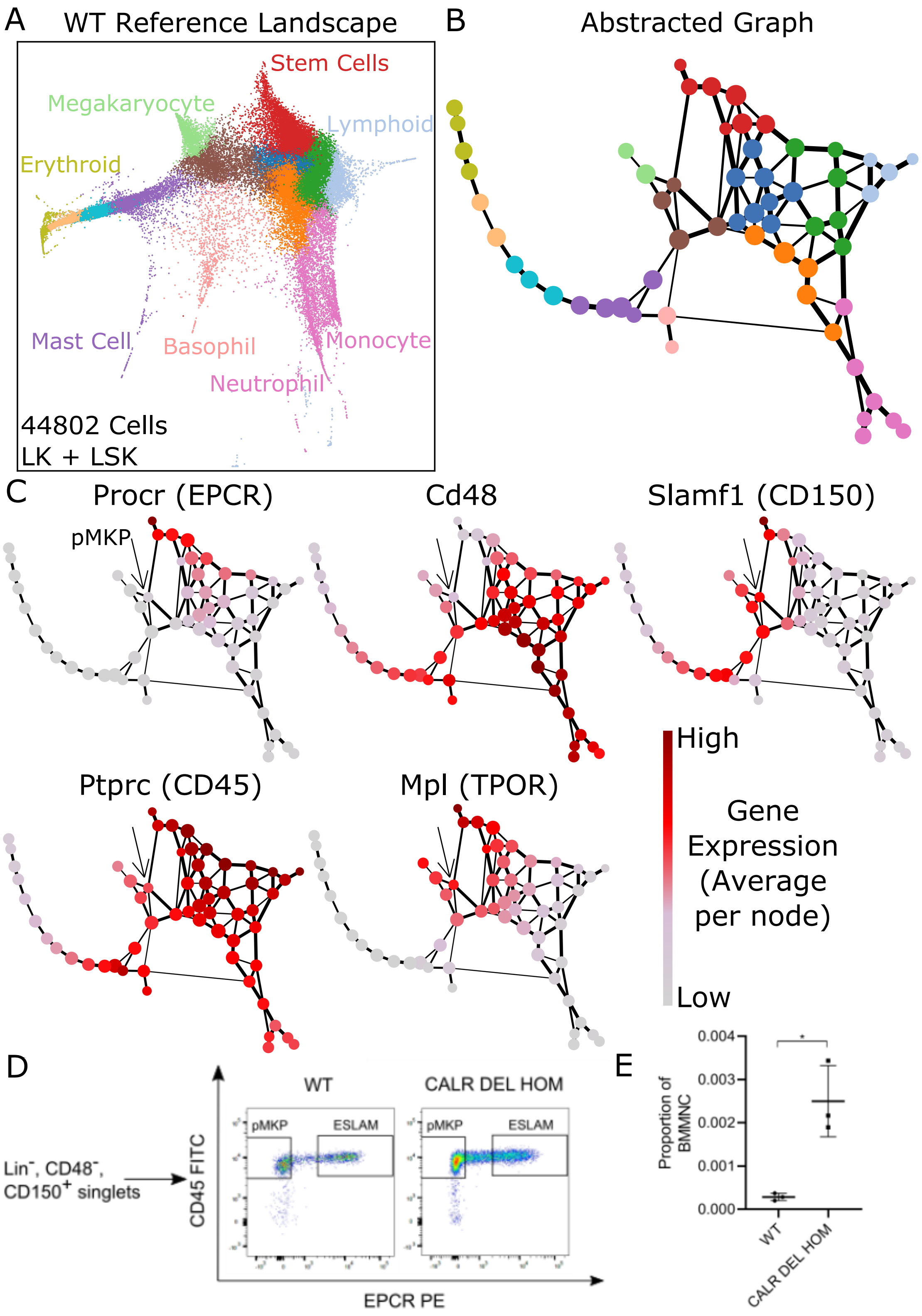


Fig. 5.11 pMKPs represent an intermediate megakaryocyte population and can be located on the WT reference landscape. (A) The WT 'reference' transcriptomic landscape. (B) Fine clustering of the reference resulted in 60 clusters. PAGA was used to measure and visualise the connectivities of the fine clusters. Only connections with weights passing a certain threshold are shown. Each fine cluster is a node, coloured as in panel A. (C) The mean gene expression per fine cluster is plotted. Each gene has a different scalebar. (D) FACS gating of the pMKP population in WT and CALR DEL mice. (E) The proportion of bone marrow mononuclear cells (BMMNC) within the pMKP gate in WT and CALR DEL.

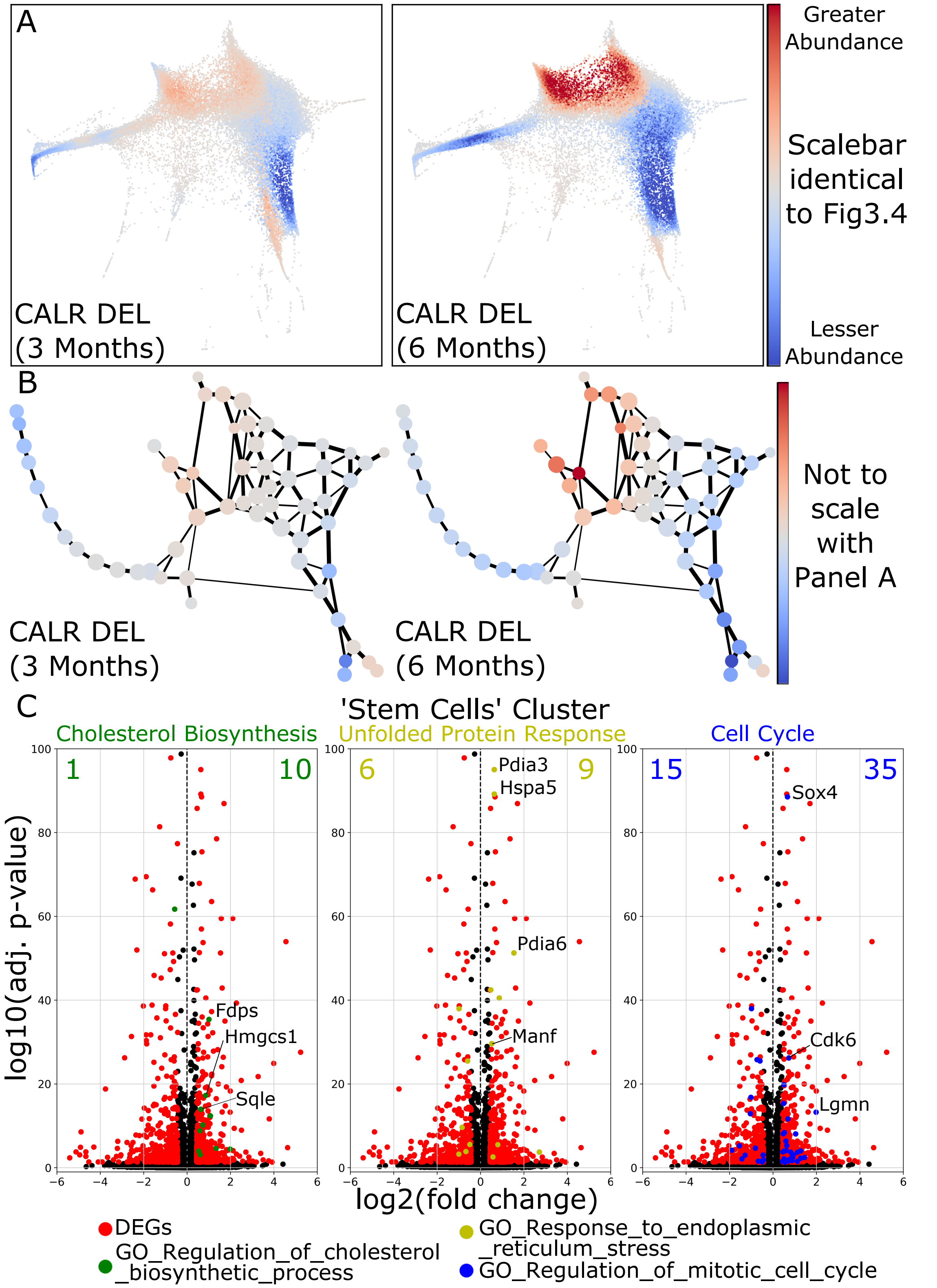


Fig. 5.12 The pMKP population is preferentially overabundant in CALR DEL mice. (A) Differential abundances for the two pairs of sequenced mice displayed using the voting method from Chapter 3. The scalebar is identical to that from Figure 3.4. (B) For each node, the average abundance change is displayed. The scalebar is consistent within the panel, but is not identical to the scalebar in panel A. (C) Volcano plots of DE results from the 'Stem Cells' cluster in the older mice. Genes belonging to three GO terms associated with cholesterol biosynthesis, unfolded protein response and cell cycle are shown. Numbers represent the number of genes belonging to these GO terms that are up/downregulated.

that all the other pre-leukaemic models were also sequenced at 3 months). This was matched by a decrease in erythroid cells and a hitherto unobserved decrease in the monocyte region and increase at the end of the neutrophil trajectory. Strikingly, at 6 months these abundance changes were highly strengthened, with the expansion of the HSC and MK regions having a magnitude that was far greater than any changes observed in any other pre-leukaemic model. The decrease in erythroid and myeloid cells was similarly increased in magnitude to compensate, though the very end of the neutrophil trajectory still displayed an overabundance. It is therefore clear the ET phenotype of the CALR DEL mouse increases strongly with age.

The abundance changes of both pairs of mice were visualised on the PAGA graph by averaging votes over each fine cluster (Figure 5.12B). Comparing the expression of *Mpl* - the gene encoding for the TPO receptor - from Figure 5.11C suggested that the expanded region in CALR DEL mice aligns almost exactly with cells that clearly express *Mpl*. Additionally, in both pairs of mice the fine cluster with the largest overabundance due to the perturbation was the cluster associated with the pMKPs. Experimental work has found evidence that pMKPs are produced from HSCs in an MPP2-independent manner; in a mouse model allowing inducible depletion of HSCs (but not MPPs or later progenitors (Schoedel et al., 2016)), a 75% reduction in HSCs was measured alongside a 68% reduction in pMKPs, but no reduction in MPP2s or any other MPP population. Thus it is possible that the pMKP population is mainly produced on a differentiation trajectory directly from HSCs, and that this trajectory is being greatly stimulated in the CALR DEL model. Notably, the PAGA analysis displayed a direct connection between the HSC and pMKP nodes. On its own this is very weak evidence of a direct trajectory between these populations, but combined with the experimental work this is very suggestive that the CALR mutation is directly influencing the very earliest HSCs and promoting a ‘direct’ megakaryocyte differentiation trajectory.

To characterise the molecular impacts within *Mpl*-expressing cells, differential expression was performed in the older mice across the ‘Stem Cells’ and megakaryocyte clusters. Unfortunately there were too few cells within the specific pMKP node for any meaningful differential expression results to be obtained. Analysis of the DEGs from the ‘Stem Cells’ cluster revealed the upregulation of three major biological pathways; cholesterol biosynthesis, unfolded protein response and cell cycle (Figure 5.12C). These pathways were all also observed in the megakaryocyte cluster (Figure 5.13A). Ingenuity Pathway Analysis (IPA) was performed by Michele Vacca to visualise these findings (Figure 5.13B). The cell cycle signature is perhaps unsurprising given the abundance changes observed within these cells. The unfolded protein response is also expected, given CALR’s role as a protein chaperone (Nam et al., 2019). The upregulation of cholesterol biosynthesis was unexpected; however,

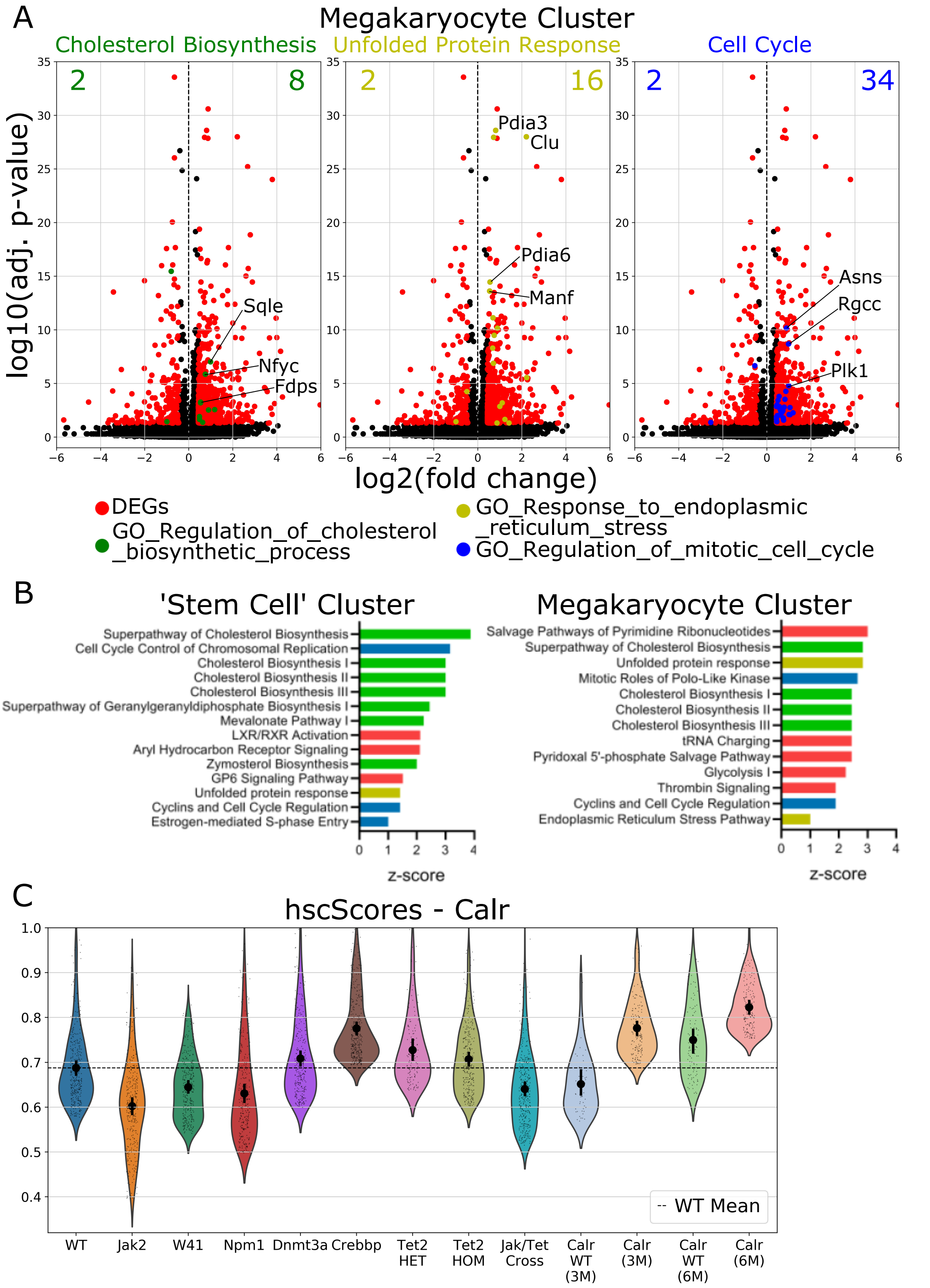


Fig. 5.13 Mutant CALR HSCs display upregulation of cholesterol biosynthesis, unfolded protein response and cell cycle pathways. (A) Volcano plots of DE results from the Megakaryocyte cluster in the older mice. Genes belonging to three GO terms associated with cholesterol biosynthesis, unfolded protein response and cell cycle are shown. Numbers represent the number of genes belonging to these GO terms that are up/downregulated. (B) The results of an IPA analysis using the DE results. Z-scores of upregulated pathways are shown, in colours corresponding to panel A. (C) Violin plots of hscScores across the pre-leukaemic perturbations and CALR experiments (an extension of Figure 4.8B).

it is biologically plausible. It has been shown that mice with impaired cholesterol efflux have more proliferative HSCs and megakaryocytes, as observed in the CALR DEL model (Murphy et al., 2013; Yvan-charvet et al., 2010). The link between cholesterol metabolism and CALR mutations will form the basis of further work.

Finally, hscScores were calculated for the two pairs of mice. Since the breeding background of the WT and CALR DEL mice used in this project are very similar to the mouse models analysed in Chapters 3 and 4, it is reasonable to compare them. Notably, in the 3 month old mice (age matched with the other models), the CALR DEL HSCs displayed significantly higher scores than their WT or the other WTs (Figure 5.13C). The same shift was observed in the 6-month old mice between the CALR DEL HSCs and their WT counterparts, which themselves displayed much higher scores than any other wild-type sample. It is difficult to assess the extent to which this latter comparison is valid, however the overall trend in the CALR DEL mice raises intriguing questions. In Chapter 4 it was suggested that increases in proliferation of HSCs may lead to lower hscScores, however the opposite effect is observed here. The CALR DEL HSCs do not show either an advantage or disadvantage compared to WT HSCs in competitive transplantation assays (Li et al., 2018), but it is possible that overactivation of MPL improves their functional potential in a native setting. This could be linked to a shift towards an MK-biased HSC state associated with the production of pMKPs, despite no clear molecular priming of CALR DEL HSCs towards the megakaryocyte lineage being observed.

Overall, this work suggests that in a disease model of essential thrombocythaemia, all *Mpl*-expressing LK progenitors are affected by mutations in CALR and display large overabundances compared to WT haematopoiesis. This is particularly prevalent in the newly defined pMKP population, which represents an intermediate megakaryocytic state that may be related to a differentiation trajectory originating directly from HSCs in an MPP-independent manner. The perturbation induces clear transcriptomic alterations in *Mpl*-expressing cells including novel upregulation of cholesterol biosynthesis genes, which may be a promising avenue for the identification of new therapeutic strategies.

5.5 Conclusions

Work presented in this chapter has formed part of three collaborations with the Lo Celso (Imperial), Patel (Cambridge) and Green (Cambridge) laboratories. In each case, the effect of different haematopoietic perturbations were analysed through scRNA-seq, and placed in the context of ongoing experimental work. Many analyses were project-specific, and this necessitated novel analysis and visualisation methods; this was particularly necessary to detect molecular signals of specific biological pathways that had been found experimentally. Where suitable, these perturbations were compared to the pre-leukaemic models analysed in Chapters 3 and 4.

5.5.1 The suitability of droplet-based scRNA-seq as a hypothesis-testing tool

The shallow but broad nature of droplet-based scRNA-seq approaches such as 10X lends itself towards being a hypothesis-generating tool; it is extremely powerful at illuminating the key features of a perturbation experiment at a large scale, and suggesting avenues down which targeted experimental and computational work needs to head. This was the case in the Malaria and CALR projects from sections 5.2 and 5.4, with the data generating hypotheses about interferon response and cholesterol biosynthesis respectively that were tested experimentally. Conversely, in the aldehyde project in section 5.3 the scRNA-seq data was largely used to confirm prior hypotheses concerning DNA damage and p53 response in haematopoietic progenitors, by specifically looking for those pathways within the data. In section 5.3 this approach was extremely successful, but this belies the fact that often, trying to confirm hypotheses about specific genes using droplet-based scRNA-seq data did not work. This necessitated the unsupervised ‘ground-up’ approaches introduced in Chapters 3 and 4 to analyse the pre-leukaemic models, which did produce a variety of interesting and novel hypotheses. In general, the supervised approach worked poorly due to many genes of potential interest being lowly- or not-expressed within the drop-seq data, possibly as a result of the polyA bias inherent to these sequencing techniques.

5.5.2 Comparing perturbations across extremely different biological settings

The extent to which the three haematopoietic perturbations discussed in this chapter can be integrated into the perturbation framework from Chapters 3 and 4 largely depends on the similarity of the WT mice used to the WT mice from earlier chapters. In all cases, integrating cellular abundance information was more straightforward, since all samples originated from the same LK gate. The abundance perturbation response in the malaria model, for example, was of greater magnitude than any of the pre-leukaemic models. The abundance changes in the DKO aldehyde model and the CALR model (at 3 months) were of similar magnitude to the pre-leukaemic models.

Molecularly, the CALR model was able to integrate seamlessly into the perturbation framework from Chapter 3 (as evidenced in Figure 5.13C). This was unsurprising given the similarities of its breeding background to the other pre-leukaemic models. Both the malaria and aldehyde models used mice originating from different facilities with disparate breeding backgrounds, and this was evident in that their WT samples looked molecularly very different compared to the pre-leukaemic WT samples. This in turn led to differences in DEG results between these models and the pre-leukaemic models that were confounded by the large differences in the WT animals. These differences were likely caused by a combination of genetic and environmental factors. In such situations, where there is a large number of models from one facility and a single model from another facility, it should be possible to correct this by using one of the numerous batch correction tools available. This is an important next step in building large perturbation comparison frameworks that are usable out-of-the-box by anyone anywhere in the world.

5.5.3 Graph-abstraction as a link between discrete and continuous approaches

The graph-abstraction (PAGA) approach used in section 5.4 proved to be a powerful tool for connecting specific FACS based cell populations to the transcriptomic landscape. In addition, combining the PAGA algorithm with a very fine-grained clustering allowed the relationship between the original clusters from the ‘reference’ landscape to be viewed in a new light. This technique sits as an intermediate between coarse grained differential expression of large clusters and the pseudotime techniques for identifying molecular drivers introduced in Chapter 3. Whilst performing differential expression on fine clusters is difficult due to reduced statistical power, it may be worth it if the exact regions where abundances are most affected (such as the pMKP node) can be located. In addition, although the ‘connectivity’ between different nodes is hard to interpret in the context of continuous differentiation trajectories, it nonetheless provides important information about which clusters have the most similar transcriptomes, even if they are not located next to each other on a 2D visualisation of the data. For this reason alone it should be considered as part of any single-cell analysis pipeline.

5.5.4 Further work

Further integration of the perturbations analysed in this chapter with the pre-leukaemic models discussed previously is possible, and may be highly informative. In particular, further comparisons of the HSC region are warranted since all three perturbations display clear phenotypes in the most immature cells, over and above what is observed across the pre-leukaemic models. Mapping the malaria and aldehyde models to the ‘reference’ clusters and interrogating their abundance, molecular and variability changes with the Stem Cell cluster should be relatively straightforward. A tougher challenge will be to assess whether any sensible comparisons can be made that will strengthen the hypotheses stated in Chapter 4. It is not immediately clear how the abundance shifts in any of these three perturbations would be driven by epigenetic changes. The CALR model undoubtedly belongs in the framework set out in Chapter 3, and the striking increase in hscScores amongst the mutant CALR mice are worthy of careful investigation. The CALR model appears to produce a very different set of cellular/molecular shifts than any other pre-leukaemic model, but its effects within the erythroid trajectory should be further compared to the other models that show a proportional loss of erythrocytes, such as Tet2. Of course given the enormous overproduction of megakaryocyte progenitors observed, it may be that in absolute terms, erythrocyte numbers are not decreased and align more close with models like the Jak2.

5.5.5 Summary

In summary, the work in this chapter analysed transcriptomic data from three diverse haematopoietic perturbations to identify the pathways driving the observed cellular and molecular shifts. Many of the generated hypotheses have been validated experimentally and have in all cases revealed large impacts upon the HSC transcriptomic states.

Chapter 6

The Trajectory Inference Method TITANS Reveals Complex Cellular Trajectories across Different Species and Systems

The computational method described in this chapter was created and implemented by Sam Watcham. It was tested on single-cell datasets from Dahlin et. al. (2018) and Pijuan-Sala et. al. (2019), using data generated by a number of people including Nicola Wilson, Blanca Pijuan-Sala and Carolina Guibentif.

6.1 Background

The introduction of scRNA-seq in recent years has provided a huge amount of insight into the heterogeneous nature of differentiating cell populations. Results from single-cell studies have led classical models of cell differentiation containing discrete transitions to be reworked (Nestorowa et al., 2016). Instead, the notion of a continuous transcriptional landscape has been formulated, in which a cell differentiates probabilistically through a transcriptomic landscape in a manner defined by its current and previous cell states (Laurenti and Göttgens, 2018; Nimmo et al., 2015).

Accurately reconstructing these differentiation trajectories from scRNA-seq data may be possible provided certain assumptions are satisfied, including the key assumption that within a given cell population, differentiation is asynchronous (Weinreb et al., 2018b). This means that the ‘snapshot’ of different cell states captured in a scRNA-seq experiment can potentially act as a proxy for the differentiation path of an individual cell over time. A number of algorithms have been developed within this framework to reconstruct differentiation trajectories from primitive to more mature cells (Herman et al., 2018; Schiebinger et al., 2019; Setty et al., 2018; Street et al., 2018; Weinreb et al., 2018b; Welch et al., 2016). These methods and others like them have been thoroughly reviewed and quantified in Saelens et. al. (Saelens et al., 2019).

However, increases in experimental throughput have led to an explosion of very large (>100,000 cells) and very sparse scRNA-seq datasets that sample from broad and unbiased cell populations (generated by droplet methods such as 10X and inDrops) without time-course information (Tusi et al., 2018; Wagner et al., 2018). Reconstructing accurate differentiation trajectories from such data remains a significant challenge, as the transcriptomic changes related to differentiation are extremely subtle and can be heavily obscured by both technical and biological noise (Grün et al., 2014). In addition, many current methods do not scale well to large cell numbers, making trajectory inference in such datasets infeasible (Svensson et al., 2018).

To facilitate more accurate trajectory inference in this breed of datasets, TITANS (Trajectory Inference Through Iterative Ancestral Search) was developed specifically to work with large, sparse single-cell datasets taken from extremely broad cell populations. The key advantages of TITANS over published methods are that 1) it does not rely on or compute a graph-based representation of the dataset, meaning the calculated trajectories are more closely related to the underlying expression data and can capture the complexities present in the high-dimensional space, 2) it is efficient and scalable, allowing a trajectory containing hundreds of thousands of cells to be calculated in minutes and 3) its iterative procedure for adding cells to a specific trajectory allows very subtle transcriptomic changes to be perceived even in very sparsely sequenced data, without requiring any user-defined temporal information. This chapter describes how TITANS can accurately reconstruct cellular trajectories and determine cellular hierarchy across different species and systems including adult haematopoiesis and early mammalian development. Furthermore, TITANS was compared and contrasted with several current state-of-the-art published methods for single-cell trajectory inference.

6.2 TITANS: Trajectory Inference Through Iterative Ancestral Search

The aim of the TITANS algorithm is to infer the most likely trajectory through a transcriptomic landscape between the most primitive, immature cells in the data and an identified

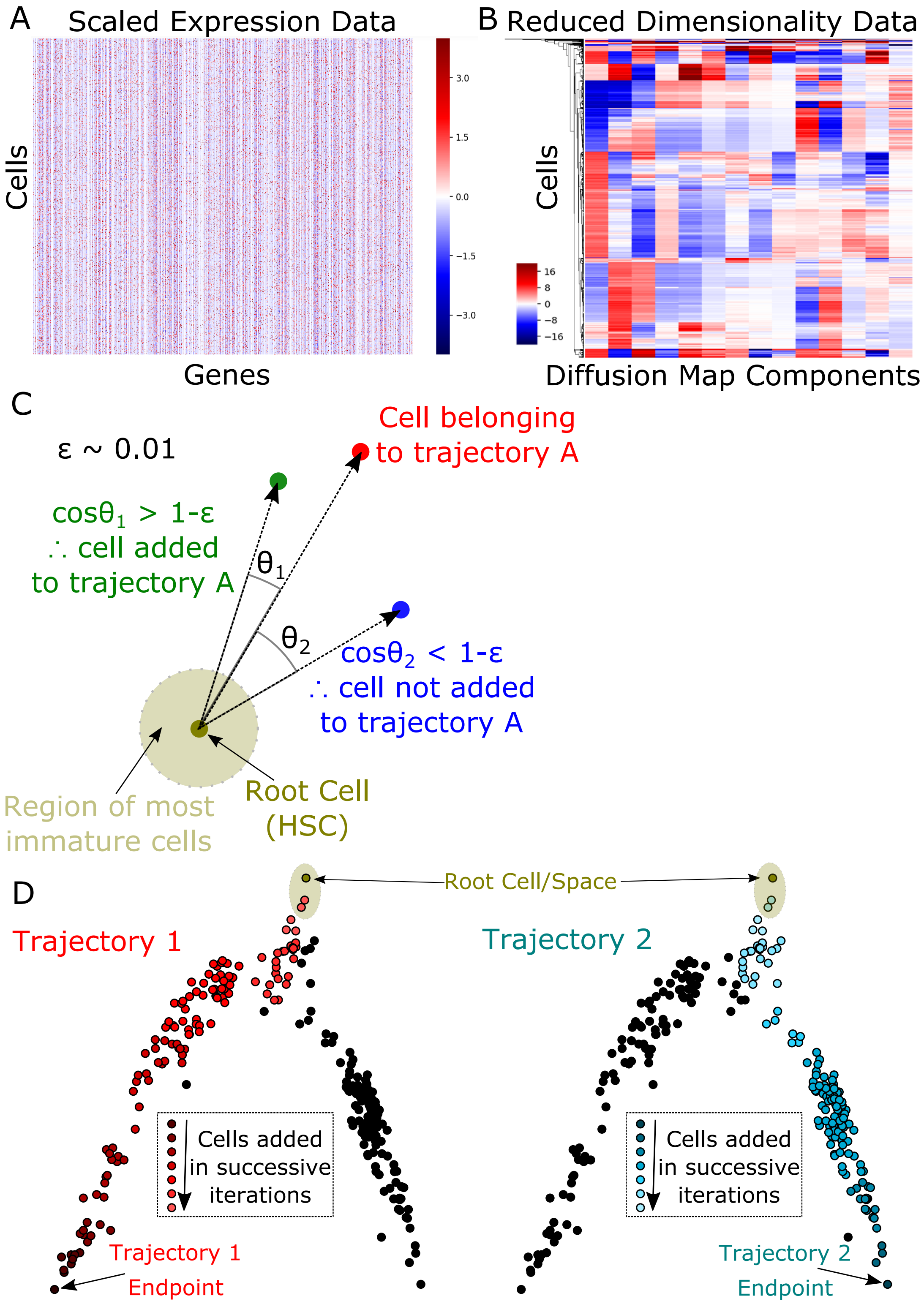


Fig. 6.1 Visual overview of the TITANS algorithm. (A) Representative heatmap of scaled expression data generated in a droplet-based sequencing experiment. 5000 cells and genes were randomly selected from the LK WT reference dataset. (B) The data from panel A, now embedded within a 15-dimensional diffusion-map space. (C) For a given endpoint and a given iteration, all cells added in the previous iteration (example in red) are tested to find other cells that are highly aligned with the root space in the diffusion-map space (example in green) according to the geometric threshold that $cos\theta > 1 - \varepsilon$, where epsilon is a small, user-defined threshold parameter. (D) Schematic of the iterative procedure from two different endpoints back to the root space within a diffusion-map embedding.

trajectory endpoint. For a full description of the TITANS method, see section 2.20. Briefly, the dataset being analysed is first embedded in a d-dimensional diffusion-map space where $g >> d > m$, with g the number of genes in the original dataset and m the expected number of trajectories within the data. Diffusion maps have been widely used in single-cell genomics to perform a similar function to principal component analysis, and have been shown to capture greater variance with a lower number of dimensions (Haghverdi et al., 2016). The only input required to the algorithm is information about the most primitive cells in the data (either a list of genes known to associate with them, or a list of cell identities from a clustering analysis). From this, the 'root space' within the diffusion-map space is identified. Using a pseudotime analysis, n potential trajectory endpoints within the data are identified. For each of these n endpoints, a trajectory is constructed by iteratively tracing back towards the root space, adding in only the cells that are highly geometrically aligned with cells added in the previous iteration. Once the trajectory reaches the most primitive cells within the root space, trajectory inference is stopped. This process allows the complex winding and branching behaviour of trajectories within a complex dataset to be captured without requiring prior biological knowledge.

This process is shown visually in Figure 6.1. Droplet-based single-cell data is typically extremely sparse and noisy, with the scaled expression data containing a lot of uninformative genes that make it hard to observe structure (Figure 6.1A). Performing dimensionality reduction on the scaled expression data to a d-dimensional diffusion-map embedding allows this structure to be seen with greater clarity (Figure 6.1B). Once the root space and putative endpoints within a dataset have been identified, iterative trajectory reconstruction begins from each endpoint. During each iteration, cells are only added to a trajectory if they are highly aligned with the cells added in the previous iteration (Figure 6.1C). After many iterations, these trajectories will have traced back all the way to the most primitive cells (Figure 6.1D). Importantly, the trajectories identified from each endpoint are independent; each cell can belong to none, one or more trajectories (Figure 6.1D). TITANS can then construct the most likely cellular hierarchy across all of the endpoints identified within a dataset.

To test whether TITANS was able to provide sensible trajectory inference in a well characterised system, the algorithm was applied to the LK cells from the WT reference dataset, published in Dahlin et. al. (Dahlin et al., 2018). To recap, cells were sorted from the Lin- c-Kit+ (LK) gate which broadly captures haematopoietic stem and progenitor cells from the top of the blood hierarchy, including the true long term haematopoietic stem cells (HSCs) and a range of progenitor populations such as the classically defined MPPs, GMPs and MEPs (Adolfsson et al., 2005; Pietras et al., 2015). Droplet based scRNA-seq was then

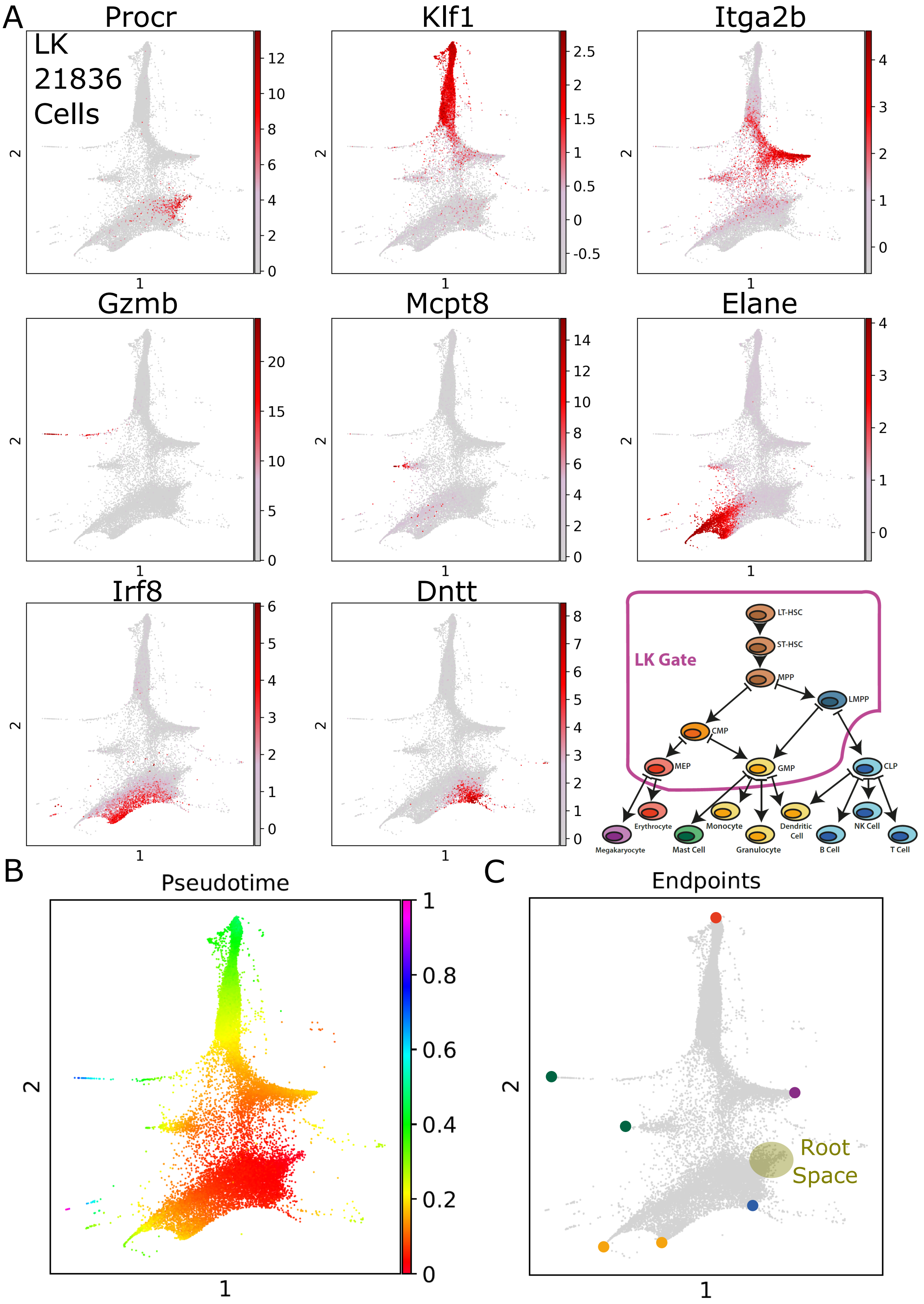


Fig. 6.2 Application to the LK reference dataset (A) Marker gene expression on the force-directed graph representation of the 21,836 LK cells from the WT reference dataset. A schematic of the haematopoietic tree in mice in also shown, displaying the populations residing within the LK gate. (B) The pseudotime score given to each cell, using a root cell as identified by the TITANS algorithm. (C) The location of seven endpoints used as starting points for trajectory inference, as located by TITANS in an unsupervised manner. Endpoint colour corresponds to cell types from panel A.

performed using 10X Chromium, with 21,836 transcriptional profiles being retained after quality control (see methods). The median number of genes detected per cell was 2842. This is an example of a very sparse dataset; 90% of the raw expression matrix contains zeros, and the majority of nonzero entries have a UMI count of one, making the bulk of the dataset binary in nature.

The transcriptional landscape of the LK cells was visualised in two dimensions using a force-directed graph, with unsurprising similarities to the WT reference dataset as a whole. Whilst such two-dimensional representations (alternatives include UMAP, tSNE, PCA) are necessary to visualise cellular trajectories, is it important to remember that the visualisations themselves cannot be used for trajectory inference as their structure is often misleading (Watcham et al., 2019). The force-directed graph of the murine LK cells revealed the existence of several branches which were shown to correspond to different haematopoietic lineages, identified by the expression of known marker genes (Figure 6.2A). As described above, the only input the TITANS algorithm requires is a list of genes associated with the most primitive cells in the dataset. Here we used a list of the MolO genes, identified by Wilson et. al. as being enriched in the most primitive HSCs (Wilson et al., 2015). From these, TITANS identified the root space within the data, calculated a $d = 15$ dimensional diffusion-map embedding of the data and constructed a pseudotime value for each cell (Figure 6.2B). Next, a list of seven possible endpoints was calculated and these were visualised on the force-directed graph (Figure 6.2C). It is important to note that these endpoints were located in a completely unsupervised manner, and no other endpoints were located. These seven endpoints were identified as being the most mature cells within the dataset corresponding to erythroid, megakaryocyte, mast cell, basophil, neutrophil, monocyte and lymphoid trajectories respectively.

Starting from each of these endpoints, the TITANS algorithm iteratively added cells to each trajectory based on their alignment in the high-dimensional diffusion-map representation of the data. Inference was stopped once each trajectory had traced back close enough to the root space and no new cells could be added. Each trajectory could then be separately visualised in the force-directed graph, with colours corresponding to the cells added in each iteration (Figure 6.3). In an attempt to reconstruct the putative cellular hierarchy within the data, all seven trajectories were combined into a single visualisation (Figure 6.4A). By starting with the most primitive cells and moving along the trajectories, a cellular hierarchy was revealed in which the earliest branching point is to either move in an erythroid/megakaryocytic direction or a mast/basophil/neutrophil/monocyte/lymphoid direction (Figure 6.4B). The former then split into individual erythroid and megakaryocytic trajectories. The latter

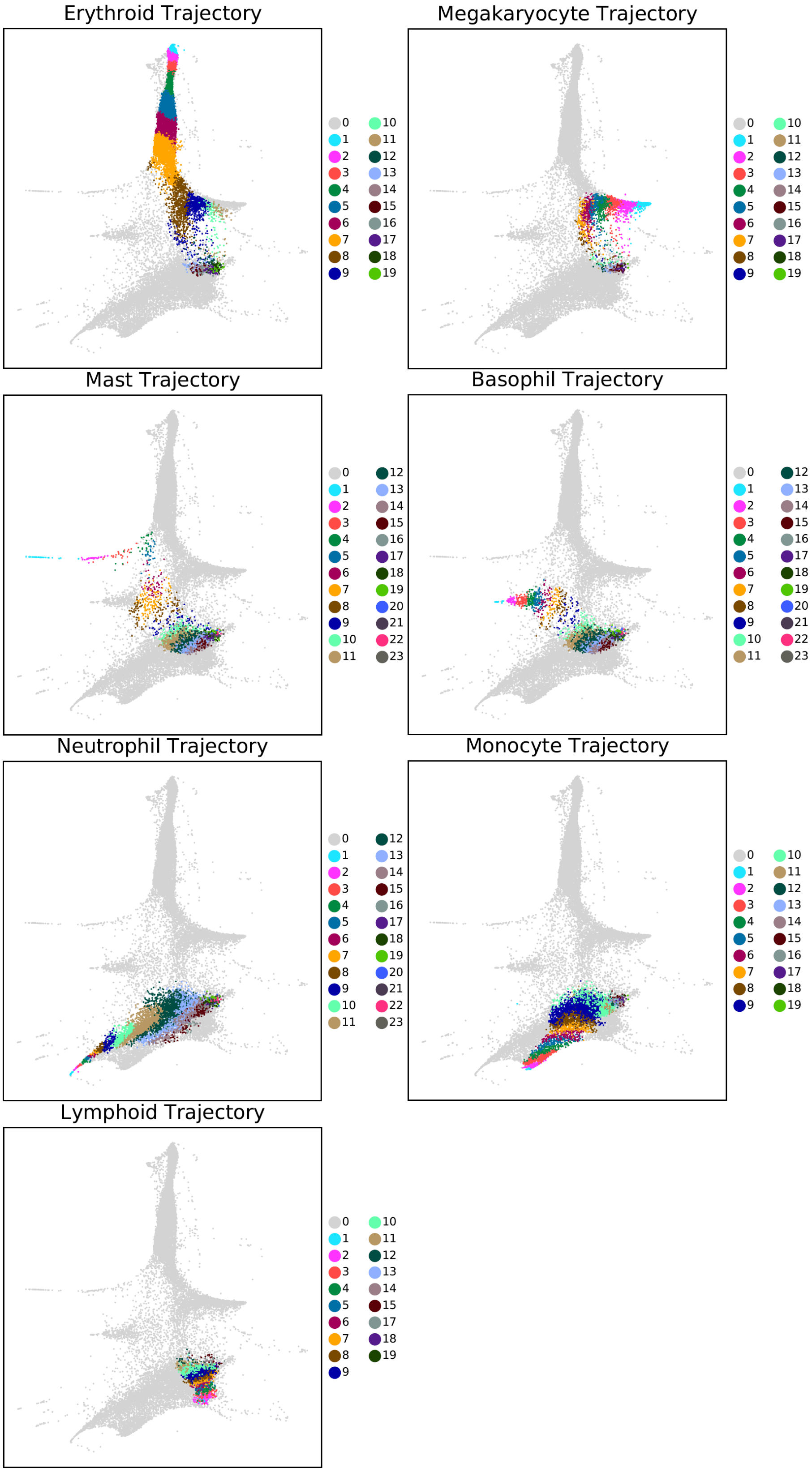


Fig. 6.3 Analysis of calculated haematopoietic trajectories reveals distinct progenitor populations. For each identified endpoint, the resulting annotated trajectory between the root space and the endpoint is shown. Cells added in each iteration are separated by colour, allowing the path of the trajectory to be traced.

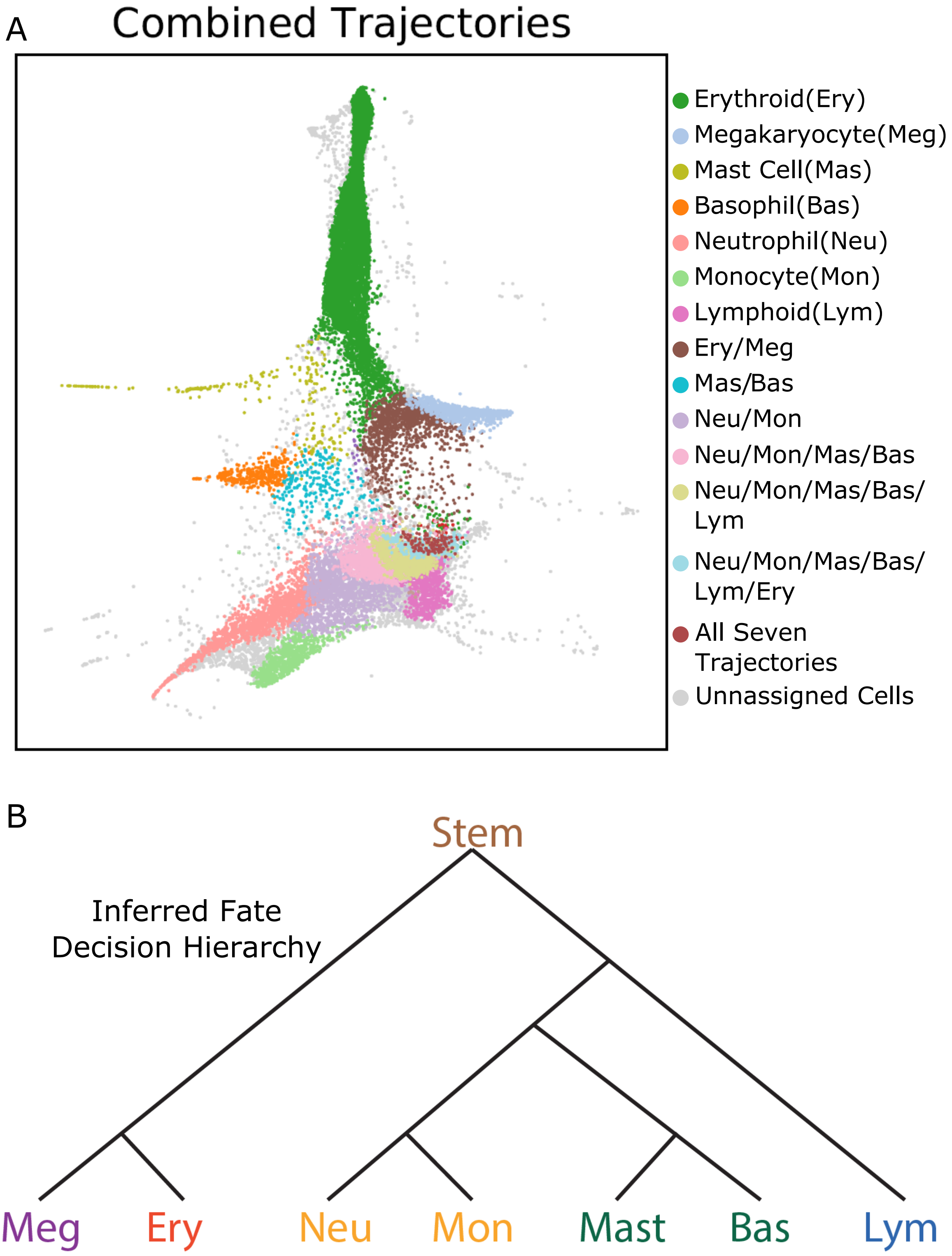


Fig. 6.4 TITANS infers a cellular hierarchy of murine haematopoiesis. (A) For each LK cell, a vector of cell assignments was calculated (1 if it belonged to a trajectory, 0 if it did not). Each unique assignment vector was turned into a distinct cluster, before being annotated according to the trajectories that cells of that cluster belonged to. Hence green cells were assigned only to the erythroid trajectory, brown cells to only the erythroid and megakaryocyte trajectories etc. (B) The hierarchy of fate decisions in murine haematopoiesis as inferred from the TITANS results in panel A.

direction then sees the lymphoid trajectory branch off, before branching into two bipotent populations containing the mast/basophil and neutrophil/monocyte trajectories respectively. These populations then branch into their individual trajectories.

There is currently much debate about the specific hierarchy of murine haematopoiesis, for example as to whether mast cells and/or basophils have their origin in erythroid or myeloid progenitors (Grootens et al., 2018; Tusi et al., 2018; Wolf et al., 2019). The TITANS analysis presented here supports the latter conclusion, and suggests there is a shared mast/basophil progenitor population that branches off from a population which also contains the neutrophil and monocyte trajectories. Additionally, TITANS also defines a small population of the most primitive cells, which are found in all seven of the calculated trajectories. This population overlaps highly with known markers of HSCs such as *Procr* and provides evidence that TITANS is capable of capturing complete trajectories through the data. The earliest branching points in murine haematopoiesis are also hotly debated, with TITANS suggesting that a branching to erythroid/megakaryocyte trajectories is one of the first to occur, which is supported by recent single-cell lineage tracing experiments (Carrelha et al., 2018; Rodriguez-Fraticelli et al., 2018; Sanjuan-Pla et al., 2013). In total, 86% of the cells in the dataset were attributed to at least one trajectory. The remaining cells may have not been found because they are outliers, or belong to trajectories not identified in the data. Additionally, they may have been excluded due to transcriptional changes caused by cell intrinsic processes such as cell division.

The inferred hierarchy presented here is not intended to be taken as a ground truth concerning murine haematopoiesis. Rather, it is intended to describe the most accurate hierarchy that can be inferred given the raw expression data that was collected. Calculating each of the seven trajectories for this dataset (the largest containing 11,000 cells) takes less than thirty seconds on a standard laptop. Despite the sparsity of the data, these complex trajectories can be captured thanks to the large cell numbers and the broad sorting gate used. Under these conditions TITANS performs accurately and efficiently, paving its way for use in less well characterised systems.

6.3 Comparison to published methods

To facilitate a better understanding of the strengths and weaknesses of the TITANS algorithm, three state-of-the-art published methods for trajectory inference - FateID, Population Balance Analysis and Slingshot - were applied to the same dataset of murine haematopoietic cells. Of the three, FateID is the most similar to TITANS in design, as it also works iteratively back from endpoints to define trajectories using a Random Forest classifier (Herman et al., 2018). This results in each cell having an associated vector of probabilities for each trajectory. However FateID classifies cells from all trajectories at once and does not take into account any information about the location of the most immature cells in the dataset. This results in the algorithm ‘overextending’ some trajectories in unlikely directions away from the most primitive cells. Additionally the fact that FateID synchronously calculates all trajectories means that some cells which would be classified into one trajectory have already been ‘used up’ by a different trajectory, unduly biasing the final calculated trajectories.

Applied to the murine haematopoietic data, these issues are clear (Figure 6.5A). In particular those cells with the highest probability assigned to the mast cell trajectory are found well into the erythroid region of the landscape, despite the expression of well described erythroid markers such as Klf1 and Gata1. Similarly, almost all of the most immature cells are assigned with the highest probability to the lymphoid trajectory, a result that makes very little biological sense given that the dataset is sure to contain a significant number of multipotent cells. These results are independent of the specific parameters used to run the FateID algorithm. Whilst FateID does successfully capture aspects of the landscape - such as the separation of the most mature neutrophil and monocyte cells – it is unable to fully elucidate the overall transcriptional landscape due to its size and complexity.

Population Balance Analysis (PBA) approaches the problem of trajectory inference in a physically-motivated way (Weinreb et al., 2018b). It attempts to reconstruct the dynamics of a single cell transcriptomic landscape by asymptotically solving a drift-diffusion equation to estimate a potential function which underlies the landscape. From this cell transition probabilities and a fate potential vector can be calculated for each cell. Whilst the rigorous nature of this approach is to be greatly commended, the PBA algorithm requires as input a vector of cell sink rates from each of the identified terminal states. This information is not known *a priori* and very small changes in this input can lead to very large changes in the calculated fate probabilities (Figure 6.5B), greatly limiting the ability of the algorithm to accurately calculate the correct dynamics from the transcriptomic data. Applied to the murine haematopoietic data, sensible results could only be achieved through manual fine tuning of this sink rate vector. Currently there is very little published data suggesting how sensible values of this input can be calculated (Busch et al., 2015). Therefore performing accurate trajectory inference in a complex dataset with several terminal fates is not currently suited to the PBA approach.

Finally, the Slingshot algorithm calculates a minimum spanning tree through the data using user-defined clusters, before defining a pseudotime and fitting principal curves through

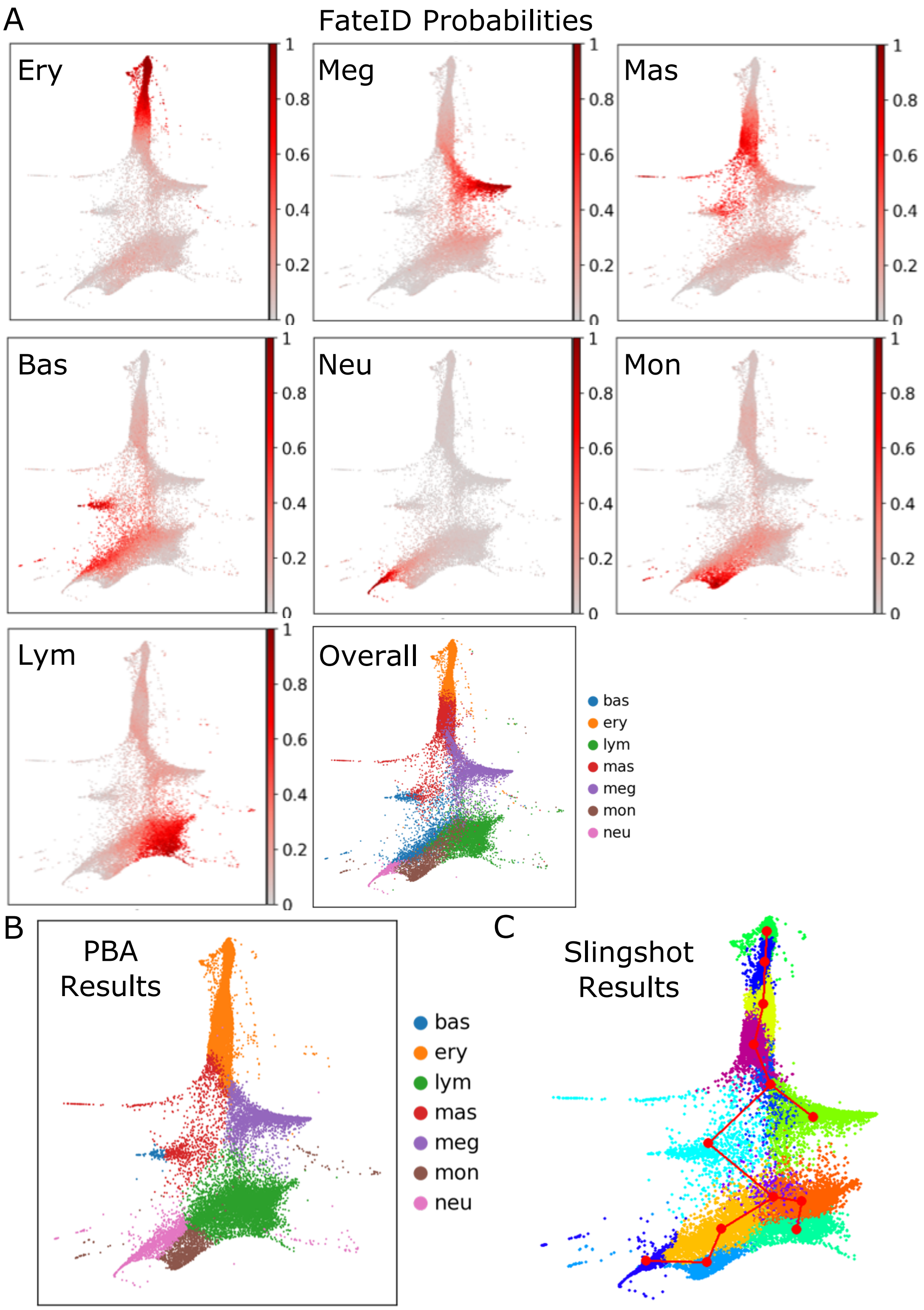


Fig. 6.5 Comparison between TITANS and three published methods. (A) FateID produced a vector of seven fate probabilities for each cell in the landscape. The individual probabilities for each fate are shown, as well as each cell coloured by the fate with the highest probability. (B) Population Balance Analysis was used to calculate a vector of seven fate probabilities for each cell, given a user-defined vector of sink rates from each fate. Cells are coloured by the fate with the highest probability. (C) Louvain clustering was used to cluster the landscape into 13 populations. Given the most immature cells (dark orange) as a starting cluster, Slingshot was used to infer the possible trajectories through the data, given by the overlaid red dots and paths.

the inferred structure (Street et al., 2018). This is the latest of a number of algorithms based on a similar approach, and has been extended to deal with branching topologies (Shin et al., 2015; Trapnell et al., 2014). The drawbacks of algorithms such as these are that they are heavily reliant on the quality of the supplied clusters, especially when applied to large, continuous transcriptomic landscapes. Here Slingshot was applied using clusters generated by a standard Louvain clustering of the LK data, and using a diffusion-map representation of the data rather than simply the expression values as suggested by the authors. The cluster containing the most immature cells was annotated to be the starting cluster for the algorithm. Slingshot calculated a branching topology that did not split the neutrophil and monocyte populations into different trajectories, despite there being clearly separated clusters for each of these lineages (Figure 6.5C). Instead the monocyte-like cluster was an intermediate step on the way to the neutrophil-like cluster. Additionally, the trajectory from the most immature cells to the erythroid lineage inaccurately passed through a basophil/mast cell-like cluster. Whilst this may be in part due to the particular clustering supplied, this indicates how Slingshot cannot be relied upon to calculate trajectories in the setting of a continuous landscape.

Overall, FateID performs best of the three tested algorithms on the murine haematopoietic data, but it is nevertheless unable to capture the complex branching topology of the transcriptional landscape to the same extent as TITANS. Whilst FateID has been extensively validated to perform well on continuous datasets, it is likely that the sparsity of drop-seq data diminishes its ability to accurately classify cells into their associated trajectories. On the other hand PBA attempts to explicitly solve the dynamics of the system, but is greatly hindered by the inputs required for such a calculation. Slingshot and other inference tools which rely on a clustering of the data are typically unsuitable for the type of large, continuous datasets that TITANS is designed to work with. In such a setting, trajectory inference using TITANS outperforms these state-of-the-art methods and is able to capture branching processes in a clear and accurate manner.

6.4 Application to a more complex transcriptional landscape

To comprehensively test TITANS in an extremely complex system, the algorithm was applied to a recently published gastrulation atlas of 116,312 mouse embryo cells collected at 9 sequential time points between 6.5 and 8.5 days post fertilisation (Pijuan-Sala et al., 2019). Cells were sequenced using the 10X Chromium platform, with a median of 3,436 genes detected per cell. Previous annotation of this dataset described 37 clusters, starting from pluripotent epiblast cells and diversifying into progenitors from each of the three main germ layers, shown here using a 2D UMAP representation (Figure 6.6A). Using a set of known epiblast markers as input, TITANS located the root space in the dataset, constructed a pseudotime and identified a number of possible endpoints as before (Figure 6.6B). Of the 19 identified endpoints, six were chosen to display the ability of TITANS to uncover developmental trajectories in the dataset (Figure 6.6C). Not all endpoints could be easily annotated as belonging to a specific developmental trajectory; these trajectories may nonetheless offer new insight into the data when combined with specialised biological knowledge.

The trajectories calculated by TITANS reveal a rich complexity within the gastrulation dataset (Figure 6.6D). Strikingly, the trajectory towards the blood progenitor cells moves from the early epiblast/primitive streak cells to the nascent and mixed mesoderm populations, before entering a thin ‘highway’ through the UMAP representation that takes the trajectory through the haematoendothelial progenitor populations to the blood progenitors and finally the erythroid progenitor cells. The endothelial trajectory follows the same path before branching off the highway in the haematoendothelial progenitor population. The ability of TITANS to capture the known relationship between these two populations - despite the strongly varying cell types along the course of the trajectories – suggests TITANS is well suited for trajectory inference even at the scale of whole organisms.

The calculated cardiomyocyte trajectory also moves from the primitive streak into the nascent mesoderm, but then deviates from the haematoendothelial trajectories. Instead it moves into the intermediate and somatic mesoderm populations, before moving to the pharyngeal mesoderm and finally into the cardiomyocyte population. Comparatively, the calculated allantois trajectory moves directly from the nascent mesoderm to a small section of the intermediate mesoderm and then into the allantois population directly. This suggests that the fate decision towards the allantois occurs somewhere within this intermediate mesoderm population. It is clear that there is a large degree of complexity within the various annotated mesoderm populations, which TITANS is nevertheless able to dissect.

Similarly, the calculated neural crest trajectory moves from the primitive streak into the rostral neurectoderm, before entering the brain progenitors (forebrain/midbrain/hindbrain) and then the neural crest population. This simple summary belies the complexity of the trajectory, which largely avoids the surface ectoderm population, despite them looking extremely intertwined in the UMAP representation. Finally the notochord trajectory is also notable, since it passes from the primitive streak to the notochord population through a very

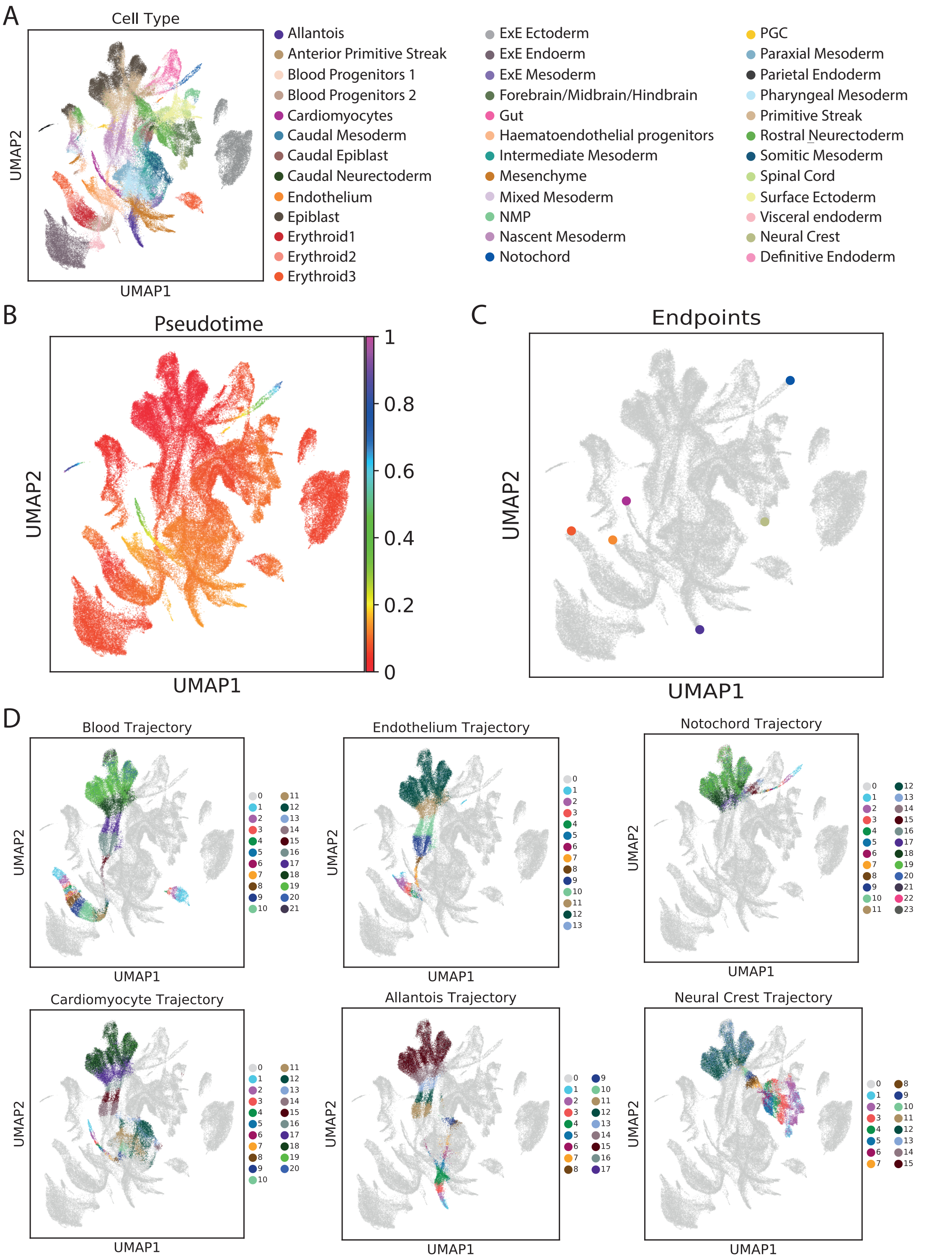


Fig. 6.6 TITANS reveal the location of potential fate decisions during mouse gastrulation. (A) A UMAP representation of the 116,312 cells in the mouse gastrulation dataset. Data was labelled according to previous annotation and clustering of the data. (B) The pseudotime score given to each cell, using a root cell identified by TITANS as being closest to the root space within the data. (C) The location of six endpoints identified by TITANS, with colours corresponding to those in panel A. (D) The calculated trajectories for each of the six identified endpoints. Colours correspond to the iteration at which that cell was added to the trajectory.

small subpopulation of the definitive endoderm, whilst most of the definitive endoderm is not contained within the trajectory. Overall, TITANS provides new insights into the developmental paths underlying mouse gastrulation, and highlights how the TITANS algorithm is capable of capturing complex cellular trajectories regardless of the heterogeneity, sparsity or size of the underlying dataset.

6.5 Conclusions

TITANS has been specifically designed to infer cellular trajectories in large, sparse scRNAseq datasets. The algorithm infers cellular trajectories by first identifying the root space (the location of the most primitive cells) in a high-dimensional diffusion-map representation of the dataset, before identifying potential endpoints in an unsupervised manner. From each of these endpoints, cells are iteratively added to the trajectory if they are highly aligned with the cells added in the previous iteration, until the trajectory has been traced back to the root space. This procedure allows the complex topology of the high-dimensional space to be deconstructed and easily understood, facilitating trajectory inference in the variety of large scRNAseq datasets/atlases currently being produced (Benoist et al., 2017; Pijuan-Sala et al., 2019; Tusi et al., 2018; Wagner et al., 2018).

When applied to mouse haematopoietic cells, TITANS infers the presence of recently characterised progenitor populations such as mast/basophil and neutrophil/monocyte populations in mouse haematopoietic cells (Dahlin et al., 2018; Görgens et al., 2013; Villani et al., 2017; Yáñez et al., 2017). Furthermore, the method is able to segregate increasingly immature cell populations and construct a putative hierarchy within murine haematopoiesis. Importantly, this analysis further infers the presence a small group of cells as belonging to all seven calculated trajectories. These cells are located in the correct part of the transcriptomic landscape, as evidenced by the expression of known mouse HSC markers such as *Procr*. Downstream analysis of gene expression differences between cell populations which TITANS highlights as having different lineage potentials has the potential to implicate new genes as drivers of haematopoietic fate decisions.

More complex systems such as murine gastrulation can also be efficiently analysed using TITANS, and reveal the nature of a number of developmental pathways which have not previously been observed in scRNAseq data. Our analysis reveals the location of several putative cell fate decisions in the mesodermal and ectodermal lineages, and provides a framework for integrating a myriad of different cell types into continuous cellular trajectories. It is foreseen that the application of TITANS to further atlases of development - such as those that have been published for Xenopus, Zebrafish and Planaria (Briggs et al., 2018; Fincher et al., 2018; Wagner et al., 2018) - should be similarly enlightening.

Compared to some of the best performing published trajectory inference methods as quantified by Saelens et. al., TITANS excels at analysing large, complex and sparse datasets of the kind which the field of single-cell genomics is fast moving towards (Saelens et al., 2019). By not relying on any kind of graph- or cluster-based approximation to the transcriptomic landscape, complex differentiation trajectories are captured with greater detail than other methods. Additionally the largely unsupervised approach removes the bias inherent to manually choosing trajectory endpoints. By treating trajectories independently and combining then only after the cell assignments are complete, TITANS maintains the speed and scalability of cluster-based approaches such as PAGA (Wolf et al., 2019) whilst not being limited by their drawbacks, and is capable of analysing datasets containing hundreds of thousands of cells in a short time on standard hardware.

The main limitation of TITANS is the need for the user-defined alignment threshold parameter to be manually chosen (see Figure 6.1). Whilst the results presented in this chapter are robust over a large range of this parameter, setting the threshold too low will nevertheless lead to artificial connections between trajectories. This could occur due to inherent biological and technical noise causing cells from two distinct trajectories to be reasonably close together in a diffusion map representation of the data. Alternatively, setting the threshold too high will cause the majority of cells to be unclassified. Due to TITANS' speed, it is quick and easy to test different threshold parameters and arrive at a reasonable value. Often the parameter can be calibrated using prior biological knowledge, such as the proportion of multipotent stem cells expected within the data; however in many cases this is not known. In addition, the magnitude of the 'correct' threshold parameter depends slightly on the size of the dataset, as this affects the scale of the diffusion map representation. Work is currently ongoing to to make this parameter scale-independent.

Chapter 7

Discussion

This chapter will review the overarching conclusions, themes and challenges of the work presented in this thesis. It will also discuss potential future directions for both this work and the field of single-cell biology at large. Detailed conclusions concerning specific results from this thesis are presented at the end of Chapters 3-6.

7.1 scRNA-seq perturbation experiments as a tool for studying pre-leukaemic states

7.1.1 Perturbed transcriptional landscapes

The concept of a transcriptional landscape has been applied to the blood system with great effect, with multiple studies attempting to reconstruct haematopoietic differentiation in the context of single-cell datasets (for example see Dahlin et al., 2018; Tusi et al., 2018). Work in this thesis has showcased how the same paradigm can be equally useful in understanding haematopoietic perturbations. By exploiting computational techniques in innovative ways, single-cell perturbation experiments can provide novel biological insights at both cellular and molecular scales, using the underlying perturbed transcriptional landscape as the link through which aberrant behaviour can be understood. For example, clustering and mapping techniques were used in Chapter 3 to quantify abundance shifts across a landscape with extremely fine resolution, and targeted differential expression testing combined with dimensionality reduction revealed potential molecular drivers of this dysregulation. Incorporating pseudotime methods with a sliding window approach further extended the molecular analysis of perturbations across entire differentiation trajectories, and provides a basis upon which regulatory networks of perturbation response can be built. Taken together, the framework for perturbation analysis introduced in Chapter 3 is widely applicable to other perturbation settings and can be used to study multiple differentiation trajectories in both supervised and unsupervised manners.

7.1.2 The need for integration across multiple conditions

A central theme to this work has been the increase in signal-to-noise ratio achieved by the integration of many different perturbations. Many of the hypotheses introduced in Chapters 3 and 4 would not have been observed if the pre-leukaemic perturbations had been analysed individually. Work in Chapter 3 strongly suggests that given the technical and biological noise inherent to droplet-based scRNA-seq data, the significance of observed abundance shifts or differentially regulated genes due to a perturbation can *only* be sensibly assessed in the context of multiple different perturbations. For example, the seemingly crucial role of pro-myeloid factors in driving aberrant haematopoietic fate decisions is only observable as a result of multiple mutations displaying correlated molecular and cellular shifts - both towards and away from - the myeloid terminal fate. A corollary to this is the clear advantages gained from using a suitably large and densely sampled 'reference' dataset to anchor multiple perturbations. Not only does this approach allow for easier comprehension of the overall biological signals exhibited by a group of perturbations, it also provides greater confidence that these signals are not an artefact of poor sampling within an individual single-cell experiment. With the current increase in the number and size of cell-type atlases being produced for different organisms, this approach is likely to remain viable in the future. The integrative framework from Chapter 3 further allowed for simple comparison of combinatorial perturbations such as the Jak/Tet Cross with the Jak2 and Tet2 single mutations respectively. From this work it was clear that the perturbations induced by such models are trajectory-specific, and similarities to their component perturbations depend on the scale at which they are observed. As an example, the erythroid trajectory of the Jak/Tet Cross looked very similar to the Jak2 model at the molecular level, but arguably more similar to the Tet2 model at the cellular level. Again, such observations would not be possible without an integrative framework for the inclusion of different perturbations.

7.1.3 Incorporating new layers of information from both external and internal sources

Extracting true signal from sparse droplet-based scRNA-seq data is challenging, and the returned count matrices often need to be transformed or supplemented in innovative ways to access this information. Work in Chapter 4 employed a variety of methods to extract signals from the ‘flat’ transcriptomic landscape containing uncommitted haematopoietic progenitors. By using external, index sorted blood progenitors to assign cells sequenced using droplet-based methods to classical progenitor populations, tentative links between transcriptional state and functional potential were achieved. This was used to highlight the cellular similarities (many perturbations pushing their MPP1 cells towards more active states) and molecular differences (perturbations up- or down-regulating pro-myeloid and HSC stress genes) between models that were correlated with their abundance changes in more mature progenitor states. Internal information from the perturbation datasets themselves was also combined across genes as part of both existing and new metrics such as the hscScore and G2M score, which were able to extract otherwise obscured information about the effects of pre-leukaemic states on HSCs. Excitingly, per-gene single-cell variability was also introduced as a new, internal source of information that may uncover otherwise hidden signals within single-cell data. For example, it appears that pre-leukaemic perturbations inducing greater single-cell variability are associated with downstream pro-erythroid skewing and vice versa. Further investigation of variability as a new, easy-to-calculate layer of information within scRNA-seq data is certainly warranted.

7.1.4 Fate decisions and differentiation trajectories in the era of high-throughput transcriptomics

The mechanistic underpinnings of haematopoietic fate decisions remain incompletely understood, but recent work has implicated epigenetic marks such as DNA methylation as being one route through which genetic mutations (in genes such as Tet2 and Dnmt3a) can drive aberrant fate decisions (Izzo et al., 2020). Work in Chapter 4 provided evidence that supposedly uncommitted MPP1 cells display pro- or anti-myeloid priming as a result of such mutations. Furthermore, it is conceivable that single-cell variability may provide a link between these epigenetic and transcriptional alterations, with greater variability potentially induced in response to a lack of global DNA methylation and vice versa.

New methods of trajectory inference are needed to computationally test whether these (perturbation-driven) aberrant fate decisions are therefore occurring in different regions of the transcriptomic landscape compared to wild-type. Additionally, the rise of microfluidic-based protocols for scRNA-seq has meant that datasets containing entire densely-sampled differentiation trajectories can now be produced with relative ease. Work in Chapter 6 aimed at creating a trajectory inference algorithm that was specifically designed to work with sparse droplet-based scRNA-seq datasets. Future work applying this algorithm to perturbed datasets may help to provide further hypotheses concerning the fate potential of perturbed MPPs and HSCs.

7.2 Limitations of scRNA-seq perturbation datasets

7.2.1 Technical variation

Mitigating the impact of technical noise is a key consideration for any scRNA-seq analysis involving multiple experimental conditions. As discussed in Chapter 1, numerous methods for single-cell normalisation and batch correction have been proposed (Polanski et al., 2020; Stuart et al., 2019). Nonetheless, analysis of the pre-leukaemic states in Chapters 3 and 4 was greatly impacted by the lack of matched WT samples for a subset of perturbations (W41, Tet2 HOM and Jak/Tet Cross). This flaw in the experimental design meant that the molecular effects of these perturbations were strongly confounded by the increased technical variation between the perturbed sample and any WT sample it was compared to. These effects were minimised by averaging over comparisons with multiple WT samples, but could not be eliminated. It is therefore advisable to produce matched control samples for all scRNA-seq perturbation experiments, even if a suitable reference dataset exists for the system of interest. In general, technical variation between perturbation experiments that both have matched WT samples is less of a problem, as integrative frameworks such as the one introduced in Chapter 3 compare each experiment to a reference and therefore will not require direct comparisons between two perturbed conditions.

Intra-experiment technical noise can also cause problems, as was the case for the third Jak2 experiment analysed in this thesis, where the perturbed sample was sequenced to a far higher saturation than the matched WT sample (see Figure 3.5). This resulted in many of the same difficulties as if the WT and perturbed sample were not matched. In this case it was found that cluster-based downsampling of the over-sequenced sample was able to mitigate much of this technical variation, but any cluster-based downsampling procedure will inevitably introduce bias, due to cells being downsampled based on a target cell-size distribution that is not completely accurate for the cell in question. It remains an open question as to whether all scRNA-seq experiments with multiple conditions should use downsampling as a standard preprocessing step; however ‘throwing away’ data by downsampling is probably inadvisable unless proved necessary by the failure of downstream analyses. It should be noted that attempts were made to correct for the lack of matched WT samples in the W41 and Tet2 HOM experiments by downsampling these perturbed samples to match an alternative WT sample. However this approach failed to produce sensible differential expression results.

7.2.2 Sparsity of droplet-based approaches

The high throughput attained by droplet-based scRNA-seq approaches is balanced by a smaller number of reads obtained per cell as compared with plate-based protocols. This trade-off - combined with the fact that only a fraction of the mRNA molecules within a cell are transcribed for sequencing - results in large, sparse datasets such as those analysed in this thesis. This sparsity leads to two important effects. Firstly, the expression of many genes is extremely noisy across different cells, and secondly, only highly expressed genes exhibit a large dynamic range of expression values across a dataset, with lowly expressed genes typically limited to either having one UMI count (‘on’) or zero UMI counts (‘off’). Hence analysis techniques that search for dynamic molecular changes across a trajectory - such as the dDEG method introduced in Chapter 3 - are heavily biased towards highly expressed genes. Furthermore, metrics used to quantify differential expression from such sparse data must be carefully chosen; lowly expressed genes can exhibit large fold changes due to biological noise, whilst small fold changes in highly expressed genes can be of important significance. Using absolute expression differences between conditions is also possible, but this again suffers from the fact that many potentially important genes do not have the dynamic range to achieve large absolute changes. A number of imputation methods for scRNA-seq have recently been published, with the aim of reducing biological noise by imputing expression values that are corrected for low sequencing depths (Eraslan et al., 2019; Lopez et al., 2018). This could help to identify new molecular drivers of perturbation response that are currently hidden by noise, but the accuracy of imputation methods in complex transcriptional landscape is difficult to assess and they are currently treated with a healthy amount of scepticism. Overall, the conclusions presented in this thesis concerning molecular drivers of perturbation response must be considered alongside the limitations of droplet-based scRNA-seq technology, as well as the advantages and drawbacks of the metrics used to assess significance.

7.2.3 Loss of temporal information

Pseudotime algorithms have been widely integrated into computational analyses of scRNA-seq data, and form the backbone of several new approaches to analysing perturbation experiments in this thesis. However there will not necessarily be any correlation between differences in pseudotime and differences in real time over the course of a differentiation trajectory. Parametrising one in terms of the other may be possible by using cell labelling techniques to estimate the real time rate of cell flux between haematopoietic compartments, but this remains a challenge (Barile et al., 2020; Busch et al., 2015; Sawai et al., 2016). It is therefore difficult to sensibly assess whether genes upregulated at one point along a trajectory are causally linked to genes upregulated at earlier or later points. Similarly, reconstructing lineage hierarchies from scRNA-seq data alone will always rely on a level of inference about the relevance of cell-to-cell distances within a transcriptomic landscape; real-time information about HSC dynamics obtained through lineage tracing or pulse-chase experiments is therefore required to fully elucidate the topology of haematopoietic fate decisions.

7.3 New directions for single-cell perturbation models

7.3.1 Towards complete cell-state information

The single-cell perturbation landscapes analysed in this thesis were constructed using information from a subset of mRNA molecules present in the original cells. As a result, potential cellular heterogeneity present at the epigenetic, protein or spatial levels will remain hidden. A complete understanding of how driver mutations impact haematopoietic differentiation will rely on single-cell methods capable of capturing a more complete picture of an individual cell's state, and associated methods for integrating this information across modalities (Kucinski and Gottgens, 2020). Examples of recently published multiomics approaches include methods for simultaneous detection of RNAs with individual proteins (Mimitou et al., 2019; Stoeckius et al., 2017), chromatin accessibility (Buenrostro et al., 2018; Cao et al., 2018), DNA methylation (Clark et al., 2018) and protein-chromatin binding (Moudgil et al., 2020; Xie et al., 2017). The combination of whole-transcriptome and whole-proteome single-cell information is particularly exciting, given that mRNA abundance explains only a small fraction of protein abundance in single cells (Gong et al., 2017; Yang et al., 2020). Furthermore, 2D spatial information can now be recovered at a whole-transcriptome scale using methods such as seqFISH+ (Eng et al., 2019), and even 3D information can now be captured for hundreds of genes using STARMAP (Wang et al., 2018). These will be of direct importance for investigating perturbed HSC niches and their interaction with the bone-marrow microenvironment. To date, very little work has been performed using these multiomics approaches to look at haematopoietic perturbations, aside from the use of TARGET-seq to profile a cell's mutational status alongside its transcriptome (Rodriguez-Meira et al., 2019). Epigenetic information in particular will be vitally important, and may provide evidence for or against many of the hypotheses concerning HSCs presented in Chapter 4. Linking mRNA levels to chromatin accessibility and methylation status will facilitate the inference of links between regulatory elements and their gene targets, as well as highlighting transcription factors responsible for driving the observed expression. These multiomics methods should see improved sensitivity and throughput over the next few years, and likely will replace single modality datasets as a matter of course.

7.3.2 Increasing perturbation throughput using CRISPR

Mouse models of pre-leukaemic driver mutations - such as those analysed in this thesis - can be difficult and time consuming to construct and functionally characterise. Given the large diversity of somatic mutations observed in patients with haematological malignancies, alternative approaches that allow a large number of perturbations to be assessed in the same experiment have arisen. The Perturb-Seq and CRISP-seq methods (Dixit et al., 2016; Jaitin et al., 2016) were the first to combine scRNA-seq with multiplexed CRISPR (clustered regularly interspaced short palindromic repeats) screening, where a library of barcoded guide RNAs target and knockout one of a pre-selected group of genes. By identifying which cells are associated with which guide RNA, the transcriptional effects of knocking out specific genes can be observed in a high throughout manner. Subsequent advances using this idea have allowed specific pairs of genetic perturbations to be studied at a whole-transcriptome scale (Boettcher et al., 2018; Datlinger et al., 2017; Giladi et al., 2018). High-throughput analysis of combinatorial perturbations should provide clear insights into why certain mutational combinations lead to leukaemia progression in patients, whilst others do not (Ortmann et al., 2015). Recovering molecular regulatory networks from such data remains computationally challenging, but has the potential to transform our understanding of perturbed haematopoiesis. Unpublished data from the Göttgens group has perturbed over 30 haematopoietic transcription factors - mutations in many of which are associated with leukaemic states - in a myelo-lymphoid progenitor cell line and attempted to decode the underlying regulatory networks from the perturbed transcriptomes, uncovering previously unknown molecular interactions.

7.3.3 Linking perturbed states to perturbed fates

Ultimately, answering the most fundamental questions about the topology of the haematopoietic differentiation hierarchy cannot be done using snapshot data alone. Instead, information is needed that can directly link a cell's current state (via its transcriptome, epigenome, proteome etc) to the future fate of the exact same cell. There is a currently a revolution of methods that combine single-cell snapshot measurements (primarily still trascriptomics) with single-cell lineage tracing, whereby a cell is labelled at a starting time point, and its clonal progeny tracked at a later time point using the inherited label (Wagner and Klein, 2020). This simultaneous measurement of state and fate potentially allows for global models of differentiation dynamics across an entire cellular system. These methods use DNA sequence barcodes as labels for lineage tracing, which can then be retrospectively recorded as part of high-throughput sequencing. As long as the barcodes have high enough complexity, thousands of distinctly labelled clones can be investigated from a single sequencing experiment. Furthermore, continuous labelling results in cumulative barcoding, whereby cells obtain successively more labels over time. This greatly expands the possibilities for lineage reconstruction from the sequenced barcodes.

The first wave of methods, including scGESTALT, ScarTrace and LINNAEUS (Alemany et al., 2018; Raj et al., 2018; Spanjaard et al., 2018), used CRISPR-cas9 activity to generate random insertions and deletions that could be used as barcodes. Subsequent approaches have alternatively used genetic recombination or retroviral integration techniques, such as in the PolyLox and LARRY methods (Pei et al., 2019; Weinreb et al., 2020). The latter method was applied to wild-type haematopoietic progenitor cells to link early biases in gene expression to a cell's fate potential, focusing on the neutrophil/monocyte fate decision. Furthermore, Weinreb et al. suggested that transcriptionally very similar MPPs exhibited clear bias towards different fate choices, highlighting the limitations of standalone transcriptional landscapes. Computational methods to make sense of clonal barcodes from single cells remain in their infancy, with maximum-parsimony and maximum-likelihood models having been suggested by various groups (Feng et al., 2019; Salvador-Martinez et al., 2019). However a number of technical issues including barcode dropouts, non-uniform likelihood of specific barcodes occurring and lack of barcode diversity currently hamper efforts to produce a gold-standard method for lineage reconstruction from single-cells.

Excitingly, a study from the Camargo group recently presented an inducible CRISPR lineage tracing mouse model that will be able to simultaneously interrogate lineage and transcriptomic information *in vivo* (Bowling et al., 2020). Using this model, intrinsic fate biases in fetal liver HSCs were observed, and it will surely play a role in supporting the next generation of studies into native haematopoiesis (Lareau et al., 2020; Morcos et al., 2020). However as of yet, no studies have focused on using this technology to look at differences in lineage trees between steady-state and perturbed haematopoiesis. When this data is available, it will be hugely exciting to try and link differences in fate biases to the underlying transcriptomes of perturbed blood progenitors, therefore identifying the true drivers of aberrant haematopoietic fate decisions.

7.4 Concluding Remarks

The past decade has seen single-cell transcriptomics grow from a revolutionary idea to a mature field in its own right. The study of haematopoiesis has greatly advanced as a result, with new insights into cellular heterogeneity, differentiation pathways and fate decisions. In the future, scRNA-seq will be combined with other single-cell modalities and time-resolved measurements of differentiation to further advance our understanding of the blood system as a whole.

This PhD thesis focused on the analysis of leukaemic perturbations to the haematopoietic system through scRNA-seq. Ultimately, an improved understanding of how these perturbations alter haematopoiesis will form the basis for new therapies aimed at treating patients with blood disorders. To this end, a integrative computational framework for the analysis of single-cell perturbation experiments was built, revealing a range of significant cellular abundance shifts in response to specific leukaemic driver mutations. Molecular integration of gene expression changes highlighted coordinated gene modules associated with perturbation response at different stages of differentiation. Furthermore, analysis of a combinatorial perturbation showcased the complexity that arises across the blood progenitor landscape as a result of successively accumulated mutations, with direct relevance to the onset of leukaemia in human patients. Since many key fate decisions occur in the HSC and MPP compartments, a detailed analysis of perturbation response in these cells revealed changes in both gene expression and variability that correlated with differentiation skewing in downstream populations. A conserved cellular shift towards more active MPP1 states was also observed, alongside cell-cycle alterations that correlated with large cell-cycle transformations along the erythroid trajectory.

Overall, this thesis has introduced a range of new computational techniques for perturbation analysis and trajectory inference from scRNA-seq data. These have led to a number of novel biological hypotheses concerning pre-leukaemic states of haematopoiesis. It is hoped that these insights will pave the way for new experiments that further elucidate our understanding of leukaemia progression.

References

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C. T., Bryder, D., Yang, L., Borge, O. J., Thoren, L. A., Anderson, K., Sitnicka, E., Sasaki, Y., Sigvardsson, M., and Jacobsen, S. E. W. (2005). Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: A revised road map for adult blood lineage commitment. *Cell*, 121(2):295–306.

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I. L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(March):193–197.

Akinduro, O., Weber, T. S., Ang, H., Haltalli, M. L. R., Ruivo, N., Duarte, D., Rashidi, N. M., Hawkins, E. D., Duffy, K. R., and Celso, C. L. (2018). Proliferation dynamics of acute myeloid leukaemia and haematopoietic progenitors competing for bone marrow space. *Nature Communications*, 9(159):1–12.

Alemany, A., Florescu, M., Baron, C. S., Peterson-maduro, J., and Oudenaarden, A. V. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(April):108–124.

Alpert, A., Moore, L. S., Dubovik, T., and Shen-orr, S. S. (2018). Alignment of single-cell trajectories to compare cellular expression dynamics. *Nature Methods*, 15(4):267–270.

Anand, S., Stedham, F., Beer, P., Gudgin, E., Ortmann, C. A., Bench, A., Erber, W., Green, A. R., and Huntly, B. J. P. (2011). Effects of the JAK2 mutation on the hematopoietic stem and progenitor compartment in human myeloproliferative neoplasms. *Blood*, 118(1):177–181.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:1–12.

Andrews, T. S. and Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59:114–122.

Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-a., Imaz-rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., and Ibarra-soria, X. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(October 2018).

Asada, N., Kunisaki, Y., Pierce, H., Wang, Z., Fernandez, N. F., Birbrair, A., Ma, A., and Frenette, P. S. (2017). Differential cytokine contributions of perivascular haematopoietic stem cell niches. *Nature Cell Biology*, 19(3):214–225.

Balazs, A. B., Fabian, A. J., Esmon, C. T., and Mulligan, R. C. (2006). Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow. *Blood*, 107(6):2317–2321.

Baldridge, M. T., King, K. Y., Boles, N. C., Weksberg, D. C., and Goodell, M. A. (2010). Quiescent haematopoietic stem cells are activated by IFN-gamma in response to chronic infection. *Nature*, 465(June):793–798.

Barile, M., Busch, K., Fanti, A.-k., Greco, A., Wang, X., Oguro, H., Zhang, Q., Morrison, S. J., Rodewald, H.-r., and Hofer, T. (2020). Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. *bioRxiv*, doi.org/10.

Baumeister, J., Chatain, N., Hubrich, A., Maié, T., Costa, I. G., Denecke, B., Han, L., Küstermann, C., Sontag, S., Seré, K., Strathmann, K., Zenke, M., Schuppert, A., Brümmendorf, T. H., Kranc, K. R., Koschmieder, S., and Gezer, D. (2020). Hypoxia-inducible factor 1 (HIF-1) is a new therapeutic target in JAK2 V617F-positive myeloproliferative neoplasms. *Leukemia*, 34:1062–1074.

Baxter, J., Scott, L. M., Campbell, P. J., East, C., Fourouclas, N., Swanton, S., Vassiliou, G. S., Bench, A. J., Boyd, E. M., Curtin, N., Scott, M. A., Erber, W. N., and Green, A. R. (2005). Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet*, 365:1054–1061.

Baylin, S. B. (2005). DNA methylation and gene silencing in cancer. *Nature Clinical Practice*, 2(December):4–11.

Bendall, S. C., Davis, K. L., Amir, E. A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'Er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725.

Bennett, L. F., Liao, C., Quickel, M. D., Yeoh, S., Vijay-kumar, M., and Hankey-giblin, P. (2019). Inflammation induces stress erythropoiesis through heme-dependent activation of SPI-C. *Science Signaling*, 12(September):1–14.

Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I. A. N., Eberwine, J., Eils, R., Enard, W., Kriegstein, A., Lein, E. D., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Reik, W., Rozenblatt-rosen, O., Sanes, J., Satija, R., Ton, N., Theis, F. J., Uhlen, M., Oudenaarden, A. V. A. N., Wagner, A., Watt, F., Weissman, J., and Wold, B. (2017). The Human Cell Atlas. *eLIFE*, 6:1–30.

Benz, C., Copley, M. R., Kent, D. G., Wohrer, S., Cortes, A., Aghaeepour, N., Ma, E., Mader, H., Rowe, K., Day, C., Treloar, D., Brinkman, R. R., and Eaves, C. J. (2012). Hematopoietic Stem Cell Subtypes Expand Differentially during Development and Display Distinct Lymphopoietic Programs. *Stem Cell*, 10(3):273–283.

Berndt, D. J. and Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. *AAAI*, pages 359–370.

Bincoletto, C., Saad, S., Soares, E., and Queiroz, M. L. S. (1999). Haematopoietic response and bcl-2 expression in patients with acute myeloid leukaemia. *European Journal of Haematology*, 62:38–42.

Blondel, V. D., Guillaume, J.-l., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *arXiv*, pages 1–12.

Boettcher, M., Tian, R., Blau, J. A., Markegard, E., Wagner, R. T., Wu, D., Mo, X., Biton, A., Zaitlen, N., Fu, H., Mccormick, F., Kampmann, M., and Mcmanus, M. T. (2018). Dual gene activation and knockout screen reveals directional dependencies in genetic networks. *Nature Biotechnology*, 36(2):170–178.

Boettcher, S., Ziegler, P., Schmid, M. A., Takizawa, H., Rooijen, N. V., Kopf, M., Heikenwalder, M., and Manz, M. G. (2012). Cutting Edge: LPS-Induced Emergency Myelopoiesis Depends on TLR4-Expressing Nonhematopoietic Cells. *The Journal of Immunology*, 188:5824–5828.

Boisset, J.-c. and Robin, C. (2012). On the origin of hematopoietic stem cells : Progress and controversy. *Stem Cell Research*, 8(1):1–13.

Boles, N. C., Lin, K. K., Lukov, G. L., Bowman, T. V., Baldridge, M. T., and Goodell, M. A. (2011). CD48 on hematopoietic progenitors regulates stem cells and suppresses tumor formation. *Blood*, 118(1):80–87.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Bonnet, D. and Dick, J. E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature*, 3(7):730–745.

Bose, P. and Verstovsek, S. (2017). Perspective JAK2 inhibitors for myeloproliferative neoplasms: what is next? *Blood*, 130(2):36–39.

Bowling, S., Sritharan, D., Osorio, F. G., Orkin, S. H., Hormoz, S., and Camargo, F. D. (2020). An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell*, 181(6):1410–1422.

Bowman, R. L., Busque, L., and Levine, R. L. (2018). Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. *Cell Stem Cell*, 22(2):157–170.

Bowman, R. L. and Levine, R. L. (2017). TET2 in Normal and Malignant Hematopoiesis. *Cold Spring Harbor Perspectives in Medicine*, pages 1–12.

Brady, G., Barbara, M., and Iscove, N. N. (1990). Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies. *Methods in Molecular and Cellular Biology*, 25:17–25.

Brennan, M. S. (2019). The Role of HECTD1 and MCL-1 in the Regulation of Normal and Malignant Haematopoiesis. *PhD Thesis*.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1102.

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W., and Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(980):1–9.

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., Majeti, R., Chang, H. Y., and Greenleaf, W. J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*, 173(6):1535–1548.e16.

Burda, P., Laslo, P., and Stopka, T. (2010). The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia*, 24:1249–1257.

Buscarlet, M., Provost, S., Zada, Y. F., Barhdadi, A., Bourgoin, V., Guylaine, L., and Busque, L. (2017). DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood*, 130(6):753–762.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S. M., Reth, M., Höfer, T., and Rodewald, H. R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542–546.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, 16(January):43–49.

Cabezas-Wallscheid, N., Buettner, F., Sommerkamp, P., Klimmeck, D., Ladel, L., Thalheimer, F. B., Pastor-Flores, D., Roma, L. P., Renders, S., Zeisberger, P., Przybylla, A., Schönberger, K., Scognamiglio, R., Altamura, S., Florian, C. M., Fawaz, M., Vonficht, D., Tesio, M., Collier, P., Pavlinic, D., Geiger, H., Schroeder, T., Benes, V., Dick, T. P., Rieger, M. A., Stegle, O., and Trumpp, A. (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell*, 169(5):807–823.e19.

Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D. B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., Von Paleske, L., Renders, S., Wünsche, P., Zeisberger, P., Brocks, D., Gu, L., Herrmann, C., Haas, S., Essers, M. A., Brors, B., Eils, R., Huber, W., Milsom, M. D., Plass, C., Krijgsveld, J., and Trumpp, A. (2014). Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell*, 15(4):507–522.

Campbell, P. J. and Green, A. R. (2006). The Myeloproliferative Disorders. *The New England Journal of Medicine*, 355(23):2452–2466.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., Mcfaline-figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., and Adey, A. C. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(September):1380–1385.

Carrelha, J., Meng, Y., Kettyle, L. M., Luis, T. C., Norfo, R., Alcolea, V., Boukarabila, H., Grasso, F., Gambardella, A., Grover, A., Högstrand, K., Lord, A. M., Sanjuan-Pla, A., Woll, P. S., Nerlov, C., and Jacobsen, S. E. W. (2018). Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature*, 554(7690):106–111.

Ceccaldi, R., Parmar, K., Mouly, E., Delord, M., Kim, J. M., and Regairaz, M. (2012). Bone Marrow Failure in Fanconi Anemia Is Triggered by an Exacerbated p53 / p21 DNA Damage Response that Impairs Hematopoietic Stem and Progenitor Cells. *Cell Stem Cell*, 11:36–49.

Chachoua, I., Pecquet, C., El-khoury, M., Nivarthi, H., Albu, R.-i., Marty, C., Gryshkova, V., Defour, J.-p., Ngo, A., Koay, A., Raslova, H., Courtoy, P. J., Choong, M. L., Plo, I., Vainchenker, W., Kralovics, R., and Constantinescu, S. N. (2016). Thrombopoietin receptor activation by myeloproliferative neoplasm associated calreticulin mutants. *Blood*, 127(10):1325–1335.

Challen, G. A., Boles, N., Lin, K. K.-y., and Goodell, M. A. (2009). Mouse Hematopoietic Stem Cell Identification and Analysis. *Cytometry Part A*, 75:14–24.

Challen, G. A., Sun, D., Mayle, A., Jeong, M., Luo, M., Rodriguez, B., Mallaney, C., Celik, H., Yang, L., Xia, Z., Cullen, S., Berg, J., Zheng, Y., Darlington, G. J., Li, W., and Goodell, M. A. (2014). Dnmt3a and Dnmt3b Have Overlapping and Distinct Functions in Hematopoietic Stem Cells. *Cell Stem Cell*, 15:350–364.

Chan, W.-i., Hannah, R. L., Dawson, M. A., Pridans, C., Foster, D., Joshi, A., Go, B., Deursen, J. M. V., and Huntly, B. J. P. (2011). The Transcriptional Coactivator Cbp Regulates Self-Renewal and Differentiation in Adult Hematopoietic Stem Cells. *Molecular and Cellular Biology*, 31(24):5046–5060.

Chatterjee, A., Ghosh, J., and Kapur, R. (2015). Mastocytosis- a mutated KIT receptor induced myeloproliferative disorder. *Oncotarget*, 6(21):18250–18269.

Chaudry, S. F. and Chevassut, T. J. T. (2017). Epigenetic Guardian : A Review of the DNA Methyltransferase DNMT3A in Acute Myeloid Leukaemia and Clonal Haematopoiesis. *BioMed Research International*, 2017:1–13.

Chen, E., Schneider, R. K., Breyfogle, L. J., Rosen, E. A., Poveromo, L., Elf, S., Ko, A., Brumme, K., Levine, R., Ebert, B. L., and Mullally, A. (2015). Distinct effects of concomitant Jak2V617F expression and Tet2 loss in mice promote disease progression in myeloproliferative neoplasms. *Blood*, 125(2):327–335.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *Bioinformatics*, 14(128):1–14.

Chen, G., Bradford, W. D., Seidel, C. W., and Li, R. (2012). Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. *Nature*, 482(7384):246–250.

Chopra, M. and Bohlander, S. K. (2019). The cell of origin and the leukemia stem cell in acute myeloid leukemia Martin Chopra. *Genes, Chromosomes & Cancer*, (April):850–858.

Clark, S. J., Argelaguet, R., Stubbs, C.-a. K. T. M., Lee, H. J., Alda-catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9:781–790.

Corces-Zimmerman, M. R. and Majeti, R. (2014). Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia*, 28(12):2276–2282.

Coue, J.-p. L., Lacout, C., Staerk, J., Raslova, H., Berger, R., Bennaceur-griscelli, A., Villeval, J. L., Constantinescu, S. N., Casadevall, N., and Vainchenker, W. (2005). A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature*, 434(April):2–6.

Dahlin, J. S., Hamey, F. K., Pijuan-Sala, B., Shepherd, M., Lau, W. W., Nestorowa, S., Weinreb, C., Wolock, S., Hannah, R., Diamanti, E., Kent, D. G., Göttgens, B., and Wilson, N. K. (2018). A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood*, 131(21):e1–e11.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–307.

Ding, L., Saunders, T. L., Enikolopov, G., and Morrison, S. J. (2012). Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature*, 481:457–468.

Dingler, F. A., Wang, M., Mu, A., Millington, C. L., Oberbeck, N., Watcham, S., Pontel, L. B., Kamimae-lanning, A. N., Langevin, F., Nadler, C., Cordell, R. L., Monks, P. S., Yu, R., Wilson, N. K., Gottgens, B., Hodskinson, M. R., and Patel, K. J. (2020). Two aldehyde clearance systems are essential to prevent lethal formaldehyde accumulation in mice and humans. *Molecular Cell*, In Press.

Dixit, A., Parnas, O., Li, B., Weissman, J. S., Friedman, N., Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-arnon, L., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. (2016). Perturb-Seq : Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Resource Perturb-Seq : Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1857.e17.

Dore, L. C. and Crispino, J. D. (2011). Transcription factor networks in erythroid cell and megakaryocyte development. *Blood*, 118(2):231–239.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J. E. (2012). Hematopoiesis: A Human Perspective. *Stem Cell*, 10(2):120–136.

Drissen, R., Buza-Vidas, N., Woll, P., Thongjuea, S., Gambardella, A., Giustacchini, A., Mancini, E., Zriwil, A., Lutteropp, M., Grover, A., Mead, A., Sitnicka, E., Jacobsen, S. E. W., and Nerlov, C. (2016). Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nature Immunology*, 17(6):666–676.

Dwyer, D. F., Barrett, N. A., Austen, K. F., and The Immunological Genome Project Consortium (2016). Expression profiling of constitutive mast cells reveals a unique identity within the immune system. *Nature Immunology*, 17(7):878–887.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S. J., Brinkman, R., and Eaves, C. (2007). Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo. *Cell Stem Cell*, 1(2):218–229.

Dzierzak, E. and Philipsen, S. (2013). Erythropoiesis : Development and Differentiation. *Cold Spring Harbor Perspectives in Biology*, pages 1–16.

Eaves, C. (2015). Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood*, 125(17):2605–2614.

Elf, S., Abdelfattah, N. S., Chen, E., Emily, A., Ko, A., Peisker, F., Florescu, N., Giannini, S., Wolach, O., Morgan, E. A., Tothova, Z., Losman, J.-a., Schneider, R. K., Al-shahrour, F., and Mullally, A. (2016). Mutant calreticulin requires both its mutant C-terminus and the thrombopoietin receptor for oncogenic transformation. *Cancer Discovery*, 6(4):368–381.

Eng, C.-h. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-c., and Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(April):235–250.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(390):1–14.

Esplin, B. L., Shimazu, T., Welner, R. S., Garrett, K. P., Nie, L., Zhang, Q., Beth, M., Yang, Q., Borghesi, L. A., and Kincade, P. W. (2011). Chronic Exposure to a TLR Ligand Injures Hematopoietic Stem Cells. *The Journal of Immunology*, 186(5367):5375.

Essers, M. A. G., Offner, S., Blanco-bose, W. E., Waibler, Z., Kalinke, U., Duchosal, M. A., and Trumpp, A. (2009). IFN-alpha activates dormant haematopoietic stem cells in vivo. *Nature*, 458(April):904–914.

Feng, J., Iii, W. S. D., Mckenna, A., Simon, N., Willis, A., and V, F. A. M. I. (2019). Estimation of cell lineage trees by maximum-likelihood phylogenetics. *bioRxiv*, doi.org/10.

Fialkow, P. J. (1974). The origin and development of human tumors studied with cell markers. *The New England Journal of Medicine*, 291(1):26–35.

Figueroa, M. E., Abdel-wahab, O., Lu, C., Ward, P. S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H. F., Tallman, M. S., Sun, Z., Wolniak, K., Peeters, J. K., Liu, W., Choe, S. E., Fantin, V. R., Paietta, E., Lo, B., Licht, J. D., Godley, L. A., Delwel, R., Valk, P. J. M., Thompson, C. B., and Levine, R. L. (2010). Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype , Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, 18:553–567.

Fincher, C. T., Fincher, C. T., Wurtzel, O., Hoog, T. D., Kravarik, K. M., and Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science*, 1736(April):1–20.

Fu, G. K., Hu, J., Wang, P.-h., and Fodor, S. P. A. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences*, 108(22):9026–9031.

Fujino, T. and Kitamura, T. (2020). ASXL1 mutation in clonal hematopoiesis. *Experimental Hematology*, 83:74–84.

Gambacorti-passerini, C., Antolini, L., Mahon, F.-x., Guilhot, F., Deininger, M., Fava, C., Nagler, A., Maria, C., Casa, D., Morra, E., Abruzzese, E., Emilio, A. D., Stagno, F., Coutre, P., Santini, V., Martino, B., Pane, F., Piccin, A., Giraldo, P., Assouline, S., Durosinmi, M. A., Leeksma, O., Pogliani, E. M., Puttini, M., Jang, E., Reiffers, J., Valsecchi, M. G., and Kim, D.-w. (2011). Multicenter Independent Assessment of Outcomes in Chronic Myeloid Leukemia Patients Treated With Imatinib. *Journal of the National Cancer Institute*, pages 553–561.

Gano, J. J. and Simon, J. A. (2010). A Proteomic Investigation of Ligand-dependent HSP90 Complexes Reveals CHORDC1 as a Novel ADP-dependent HSP90-interacting Protein. *Molecular & Cellular Proteomics*, pages 255–270.

Garaycoechea, J. I., Crossan, G. P., Langevin, F., Mulderrig, L., Louzada, S., Yang, F., Guilbaud, G., Park, N., Roerink, S., Nik-Zainal, S., Stratton, M. R., and Patel, K. J. (2018). Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature*, 553(January):171–180.

Genovese, G., Lindberg, J., Rose, S. A., Bakhoum, S. F., Chambert, K., Mick, E., Neale, B. M., Fromer, M., Purcell, S. M., Svantesson, O., Sullivan, P. F., Sklar, P., Grönberg, H., Hultman, C. M., and Mccarroll, S. A. (2014). Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *The New England Journal of Medicine*, 371(26):2477–2487.

Gezer, D., Vukovic, M., Soga, T., Pollard, P. J., and Kranc, K. R. (2014). Concise Review : Genetic Dissection of Hypoxia Signaling Pathways in Normal and Leukemic Stem Cells. *Stem Cells*, 32:1390–1397.

Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-wallscheid, N., Dress, R., Ginhoux, F., Trumpp, A., Tanay, A., and Amit, I. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nature Cell Biology*, 20(July):836–846.

Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P. S., Povinelli, B. J., Booth, C. A., Sopp, P., Norfo, R., Rodriguez-Meira, A., Ashley, N., Jamieson, L., Vyas, P., Anderson, K., Segerstolpe, Å., Qian, H., Olsson-Strömberg, U., Mustjoki, S., Sandberg, R., Jacobsen, S. E. W., and Mead, A. J. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nature Medicine*, 23(6):692–702.

Godfrey, A. L., Chen, E., Pagano, F., Silber, Y., Campbell, P. J., and Green, A. R. (2013). Clonal analyses reveal associations of JAK2V617F homozygosity with hematologic features , age and gender in polycythemia vera and essential thrombocythemia. *Haematologica*, 98(5):718–722.

Gong, H., Wang, X., Liu, B., Boutet, S., Holcomb, I., Ooi, A., Sanada, C., Sun, G., and Ramakrishnan, R. (2017). Single-cell protein-mRNA correlation analysis enabled by multiplexed dual-analyte co- detection. *Scientific Reports*, 7(2776):1–8.

Görgens, A., Radtke, S., Möllmann, M., Cross, M., Dürig, J., Horn, P. A., and Giebel, B. (2013). Revision of the Human Hematopoietic Tree: Granulocyte Subtypes Derive from Distinct Hematopoietic Lineages. *Cell Reports*, 3(5):1539–1552.

Grinfeld, J., Nangalia, J., and Green, A. R. (2017). Molecular determinants of pathogenesis and clinical phenotype in myeloproliferative neoplasms. *Haematologica*, 102(1):7–17.

Grootens, J., Ungerstedt, J. S., Ekoff, M., Rönnberg, E., Klimkowska, M., Amini, R.-m., Arock, M., Söderlund, S., Mattsson, M., Nilsson, G., Dahlin, J. S., and Dv, K. I. T. (2019). Single-cell analysis reveals the KIT D816V mutation in haematopoietic stem and progenitor cells in systemic mastocytosis. *EBioMedicine*, 43:150–158.

Grootens, J., Ungerstedt, J. S., Nilsson, G., and Dahlin, J. S. (2018). Deciphering the differentiation trajectory from hematopoietic stem cells to mast cells. *Blood Advances*, 2(17):2273–2281.

Grover, A., Sanjuan-Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., Macaulay, I., Mancini, E., Luis, T. C., Mead, A., Jacobsen, S. E. W., and Nerlov, C. (2016). Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nature Communications*, 7.

Grün, D. (2020). Revealing dynamics of gene expression variability in cell state space. *Nature Methods*, 17(January):45–49.

Grün, D., Kester, L., and Oudenaarden, A. V. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–641.

Grün, D., Muraro, M. J., Boisset, J. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., de Koning, E. J., and van Oudenaarden, A. (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, 19(2):266–277.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D., and Robson, P. (2010). Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell*, 18(4):675–685.

Guo, G., Luc, S., Marco, E., Lin, T.-w., Peng, C., Kerenyi, M. A., Beyaz, S., Kim, W., Xu, J., Das, P. P., Neff, T., Zou, K., Yuan, G.-c., and Orkin, S. H. (2013). Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire. *Stem Cell*, 13(4):492–505.

Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(296):1–15.

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., and Alpermann, T. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, (October 2013):241–247.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Gene expression Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

Haltalli, M. L. R., Watcham, S., Wilson, N. K., Eilers, K., Ang, H., Birch, F., Anton, S. G., Pirillo, C., Ruivo, N., Vainieri, M. L., Pospori, C., Sinden, R. E., Luis, T. C., Duffy, K. R., Gottgens, B., Blagborough, A. M., and Lo Celso, C. (2020). Manipulating niche composition limits damage to haematopoietic stem cells during Plasmodium infection. *Nature Cell Biology*, In Press.

Hamey, F. K. and Göttgens, B. (2018). Sorting apples from oranges in single-cell expression comparisons. *Nature Methods*, 15(5):321–322.

Hamey, F. K. and Gottgens, B. (2019). Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. *Experimental Hematology*, 78:11–20.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., Leeuw, Y. D., Anavy, L., Gennert, D., Li, S., Livak, K. J., Rozenblatt-rosen, O., Dor, Y., and Regev, A. (2016). CEL-Seq2 : sensitive highly-multiplexed. *Genome Biology*, pages 1–7.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673.

Hehlmann, R. and Hochhaus, A. (2007). Chronic myeloid leukaemia. *Lancet*, 370:342–359.

Herman, J. S., Sagar, and Grün, D. (2018). FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(June).

Hodskinson, M. R., Bolner, A., Sato, K., Kamimae-Lanning, A. N., Rooijers, K., Witte, M., Mahesh, M., Silhan, J., Petek, M., Williams, D. M., Kind, J., Chin, J. W., Patel, K. J., and Knipscheer, P. (2020). Alcohol-derived DNA crosslinks are repaired by two distinct mechanisms. *Nature*, 579(March):603–610.

Hoppe, P. S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K. D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Gambardella, A., Ahmed, N., Etzrodt, M., Coutu, D. L., Rieger, M. A., Marr, C., Strasser, M. K., Schauberger, B., Burtscher, I., Ermakova, O., Bürger, A., Lickert, H., Nerlov, C., Theis, F. J., and Schroeder, T. (2016). Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature*.

Howie, H. L., Hay, A. M., Wolski, K. D., Waterman, H., Lebedev, J., Fu, X., Culp-hill, R., Alessandro, A. D., Gorham, J. D., Ranson, M. S., Roback, J. D., Thomson, P. C., and Zimring, J. C. (2019). Differences in Steap3 expression are a mechanism of genetic variation of RBC storage and oxidative damage in mice. *Blood Advances*, 3(15):2272–2285.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., and Heyworth, C. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. *Genes & Development*, 11:774–785.

Huh, I., Zeng, J., Park, T., and Yi, S. V. (2013). DNA methylation and transcriptional noise. *Epigenetics & Chromatin*, 6(9):1–10.

Ibarra-Soria, X., Jawaid, W., Pijuan-Sala, B., Ladopoulos, V., Scialdone, A., Jörg, D. J., Tyser, R. C., Calero-Nieto, F. J., Mulas, C., Nichols, J., Vallier, L., Srinivas, S., Simons, B. D., Göttgens, B., and Marioni, J. C. (2018). Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nature Cell Biology*, 20(2):127–134.

Ingram, B. D. A., Yang, F.-c., Travers, J. B., Wenning, M. J., Hiatt, K., New, S., Hood, A., Shannon, K., Williams, D. A., and Clapp, D. W. (2000). Genetic and Biochemical Evidence that Haploinsufficiency of the Nf1 Tumor Suppressor Gene Modulates Melanocyte and Mast Cell Fates In Vivo. *Journal of Experimental Medicine*, 191(1):181–187.

Inoue, D., Bradley, R. K., and Abdel-wahab, O. (2016). Spliceosomal gene mutations in myelodysplasia : molecular links to clonal abnormalities of hematopoiesis. *Genes & Development*, 30:989–1001.

Islam, S., Kja, U., Moliner, A., Zajac, P., Fan, J.-b., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21:1160–1167.

Islam, S., Zeisel, A., Joost, S., Manno, G. L., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(1):163–170.

Isobe, T., Baba, E., Arita, S., Komoda, M., Tamura, S., Shirakawa, T., Ariyama, H., Takaishi, S., Kusaba, H., Ueki, T., and Akashi, K. (2011). Human STEAP3 maintains tumor growth under hypoferric condition. *Experimental Cell Research*, 317(18):2582–2591.

Izzo, F., Lee, S. C., Poran, A., Chaligne, R., Gaiti, F., Gross, B., Murali, R. R., Deochand, S. D., Ang, C., Jones, P. W., Nam, A. S., Kim, K.-t., Kothen-hill, S., Schulman, R. C., Ki, M., Lhoumaud, P., Skok, J. A., Viny, A. D., Levine, R. L., Kenigsberg, E., Abdel-wahab, O., and Landau, D. A. (2020). DNA methylation disruption reshapes the hematopoietic differentiation landscape. *Nature Genetics*, 52(April):378–387.

Jackson, B., Brocker, C., Thompson, D. C., Black, W., Vasiliou, K., Nebert, D. W., and Vasiliou, V. (2011). Update on the aldehyde dehydrogenase gene (ALDH) superfamily. *Human Genomics*, 5(4):283–303.

Jacobson, L., Simmons, E. L., Marks, E. K., and Eldredge, J. H. (1951). Recovery from Radiation Injury. 113:510–512.

Jaiswal, S., Chavez, A., Higgins, J. M., Moltchanov, V., Kuo, F. C., Kluk, M. J., Henderson, B., Sc, L. K. M., Koistinen, H. A., Ladenvall, C., Haiman, C., Atzmon, G., Wilson, J. G., Neuberg, D., Altshuler, D., and Ebert, B. L. (2014). Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *The New England Journal of Medicine*, 371(26):2488–2497.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(February):776–786.

Jaitin, D. A., Weiner, A., Yofe, I., Tanay, A., Oudenaarden, A. V., Amit, I., Jaitin, D. A., Weiner, A., Yofe, I., Lara-astiaso, D., Keren-shaul, H., and David, E. (2016). Dissecting Immune Circuits by Linking CRISPR- Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7):1883–1888.e15.

James, C., Mazurier, F., Dupont, S., Chaligne, R., Lamrissi-garcia, I., Tulliez, M., Lippert, E., Pasquet, J.-m., and Etienne, G. (2008). The hematopoietic stem cell compartment of JAK2V617F-positive myeloproliferative disorders is a reflection of disease heterogeneity. *Blood*, 112(6):1–3.

Jamieson, C. H. M., Gotlib, J., Durocher, J. A., Chao, M. P., Mariappan, M. R., Lay, M., Jones, C., Zehnder, J. L., Lilleberg, S. L., and Weissman, I. L. (2006). The JAK2 V617F mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proceedings of the National Academy of Sciences*, 103(16):6224–6229.

Jan, M., Snyder, T. M., Corces-zimmerman, M. R., Vyas, P., Weissman, I. L., Quake, S. R., and Majeti, R. (2012). Clonal Evolution of Preleukemic Hematopoietic Stem Cells Precedes Human Acute Myeloid Leukemia. *Leukemia*, 4(149).

Jeong, J. J., Gu, X., Nie, J., Sundaravel, S., Liu, H., Kuo, W.-L., Bhagat, T. D., Pradhan, K., Cao, J., Nischal, S., Mcgraw, K. L., Bhattacharyya, S., Bishop, M. R., Artz, A., Thirman, M. J., Moliterno, A., Ji, P., Levine, R. L., Godley, L. A., Steidl, U., Bieker, J., List, A. F., Saunthararajah, Y., He, C., Verma, A., and Wickrema, A. (2019). Cytokine regulated phosphorylation and activation of TET2 by JAK2 in hematopoiesis. *Cancer Discovery*, 10.1158/21:1–19.

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., Kim, S.-b., Yang, L., Ko, M., Chen, R., Göttgens, B., Lee, J.-s., Gunaratne, P., Godley, L. A., Darlington, G. J., Rao, A., Li, W., and Goodell, M. A. (2013). Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature Genetics*, 46(1):17–23.

Johnson, W. E. and Li, C. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

Julius, M. H., Masuda, T., and Herzenberg, L. A. (1972). Demonstration That Antigen-Binding Cells Are Precursors of Antibody- Producing Cells After Purification with a Fluorescence-Activated Cell Sorter. *Proceedings of the National Academy of Sciences*, 69(7):1934–1938.

Jung, M. K., Park, Y., Song, S. B., Cheon, S. Y., Park, S., Houh, Y., Ha, S., Kim, H. J., Park, J. M., Kim, T. S., Lee, W. J., Cho, B. J., Bang, S. I., Park, H., and Cho, D. (2011). Erythroid Differentiation Regulator 1, an Interleukin 18-Regulated Gene, Acts as a Metastasis Suppressor in Melanoma. *Journal of Investigative Dermatology*, 131(10):2096–2104.

Kang, Z. J., Liu, Y. F., Xu, L. Z., Long, Z. J., Huang, D., Yang, Y., Liu, B., and Feng, J. X. (2016). The Philadelphia chromosome in leukemogenesis. *Chinese Journal of Cancer*, 35(48):1–15.

Kent, D. G., Li, J., Tanna, H., Fink, J., Kirschner, K., Pask, D. C., Silber, Y., Hamilton, T. L., Sneade, R., Simons, B. D., and Green, A. R. (2013). Self-Renewal of Single Mouse Hematopoietic Stem Cells Is Reduced by JAK2V617F Without Compromising Progenitor Cell Expansion. *PLOS Biology*, 11(6).

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):18–22.

Kiel, M. J., Yilmaz, Ö. H., Iwashita, T., Yilmaz, O. H., Terhorst, C., Morrison, S. J., and Arbor, A. (2005). Hematopoietic Stem and Progenitor Cells and Reveal Endothelial Niches for Stem Cells. *Cell*, 121:1109–1121.

King, K. Y. and Goodell, M. A. (2011). Inflammatory modulation of HSCs: viewing the HSC as a foundation for the immune response. *Nature Reviews Immunology*, 11:685–692.

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap : projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5):359–365.

Kisseleva, T., Bhattacharya, S., Braunstein, J., and Schindler, C. W. (2002). Signaling through the JAK / STAT pathway , recent advances and future challenges. *Gene*, 285:1–24.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):3–7.

Klampfl, T., Casetti, I. C., Antonio, E. S., Ferretti, V., Elena, C., Schischlik, F., Cleary, C., Six, M., Schalling, M., Schönegger, A., Bock, C., Malcovati, L., Pascutto, C., Superti-furga, G., Cazzola, M., and Kralovics, R. (2013). Somatic Mutations of Calreticulin in Myeloproliferative Neoplasms. *The New England Journal of Medidine*, 369(25):2379–2390.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Ko, M., Bandukwala, H. S., An, J., Lamperti, E. D., Thompson, E. C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S. B., and Rao, A. (2011). Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proceedings of the National Academy of Sciences*, 108(35):14566–14571.

Kohler, G. and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256:495–500.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4):610–620.

Kondo, M., Weissman, I. L., and Akashi, K. (1997). Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow. *Cell*, 91:661–672.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., and Loh, P.-r. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(December):1289–1300.

Kouchkovsky, I. D. and Abdul-Hay, M. (2016). 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood Cancer Journal*, 6(April):1–12.

Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research*, 25(12):1860–1872.

Kucinski, I. and Gottgens, B. (2020). Advancing Stem Cell Research through Multimodal Single-Cell Analysis. *Cold Spring Harbor Perspectives in Biology*.

Kurimoto, K., Yabuta, Y., Ohinata, Y., Ono, Y., Uno, K. D., Yamada, R. G., Ueda, H. R., and Saitou, M. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic acids research*, 34(5):1–17.

Lakin, N. D. and Jackson, S. P. (1999). Regulation of p53 in response to DNA damage. *Oncogene*, 18:7644–7655.

Lambe, T., Simpson, R. J., Dawson, S., Bouriez-jones, T., Crockford, T. L., Lepherd, M., Latunde-dada, G. O., Robinson, H., Raja, K. B., Campagna, D. R., Jr, G. V., Clive, J., Goodnow, C. C., Fleming, M. D., Mckie, A. T., and Cornall, R. J. (2009). Identification of a Steap3 endosomal targeting motif essential for normal iron metabolism. *Blood*, 113(8):1805–1808.

Langevin, F., Crossan, G. P., Rosado, I. V., Arends, M. J., and Patel, K. J. (2011). Fancd2 counteracts the toxic effects of naturally produced aldehydes in mice. *Nature*, 475(July):53–60.

Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R., and Buenrostro, J. D. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, 37(August):916–929.

Lareau, C. A., Ludwig, L. S., Muus, C., Gohil, S. H., Zhao, T., Chiang, Z., Pelka, K., Verboon, J. M., Luo, W., Christian, E., Rosebrock, D., Getz, G., Boland, G. M., Chen, F., Buenrostro, J. D., Hacohen, N., Wu, C. J., and Aryee, M. J. (2020). Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nature Biotechnology*, doi.org/10.

Laurenti, E. and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689):418–426.

Letellier, E. and Haan, S. (2015). SOCS2 : physiological and pathological functions. *Frontiers in Bioscience*, 8:189–204.

Levine, J. H., Simonds, E. F., Bendall, S. C., Downing, J. R., Pe, D., Nolan, G. P., Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Downing, J. R., Pe, D., and Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197.

Levy, E. and Slavov, N. (2018). Single cell protein analysis for systems biology. *Essays in Biochemistry*, 62(July):595–605.

Li, G., Luna, C., and Gonzalez, P. (2016). miR-183 Inhibits UV-Induced DNA Damage Repair in Human Trabecular Meshwork Cells by Targeting of KIAA0101. *Investigative Opthalmology*, 57:2178–2186.

Li, J., Kent, D. G., Godfrey, A. L., Manning, H., Nangalia, J., Aziz, A., Chen, E., Saeb-parsy, K., Fink, J., Sneade, R., Hamilton, T. L., Pask, D. C., Silber, Y., Zhao, X., Ghevaert, C., Liu, P., and Green, A. R. (2014). JAK2V617F homozygosity drives a phenotypic switch in myeloproliferative neoplasms , but is insufficient to sustain disease. *Blood*, 123(20):3139–3151.

Li, J., Prins, D., Park, H. J., Grinfeld, J., Gonzalez-arias, C., Loughran, S., Dovey, O. M., Klamp, T., Bennett, C., Hamilton, T. L., Pask, D. C., Sneade, R., Williams, M., Aungier, J., Ghevaert, C., Vassiliou, G. S., Kent, D. G., and Green, A. R. (2018). Mutant calreticulin knockin mice develop thrombocytosis and myelofibrosis without a stem cell self-renewal advantage. *Blood*, 131(6):649–660.

Liang, S., Wang, F., Han, J., and Chen, K. (2020). Latent periodic process inference from single-cell RNA-seq data. *Nature Communications*, 11:1–8.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2016). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425.

Liu, L., Zhao, M., Jin, X., Ney, G., Yang, K. B., Peng, F., Cao, J., Iwawaki, T., Valle, J. D., Chen, X., and Li, Q. (2019). Adaptive endoplasmic reticulum stress signalling via IRE1α–XBP1 preserves self-renewal of haematopoietic and pre-leukaemic stem cells. *Nature Cell Biology*, 21(March):328–337.

Livesey, F. J. (2003). Strategies for microarray analysis of limiting amounts of RNA. *Briefings in Functional Genomics and Proteomics*, 2(1):31–36.

Lopes, M. R., Kleber, J., Pereira, N., and Campos, P. D. M. (2017). De novo AML exhibits greater microenvironment dysregulation compared to AML with myelodysplasia-related changes. *Scientific Reports*, 7:1–12.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(December):1053–1058.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15:1–23.

Luis, T. C., Lawson, H., and Kranc, K. R. (2020). Divide and Rule: Mitochondrial Fission Regulates Quiescence in Hematopoietic Stem Cells. *Stem Cell*, 26(3):299–301.

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(75):1–14.

Maaten, L. V. D. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Mallaney, C., Ostrander, E. L., Celik, H., Kramer, A. C., Martens, A., Kothari, A., Koh, K. W., Haussler, E., Iwamori, N., Gontarz, P., Zhang, B., and Challen, G. A. (2019). Kdm6b regulates context-dependent hematopoietic stem cell self-renewal and leukemogenesis. *Leukemia*, 33:2506–2521.

Mango, R. L., Wu, Q. P., West, M., Mccook, E. C., Serody, J. S., and Deventer, H. W. V. (2014). C-C Chemokine Receptor 5 on Pulmonary Mesenchymal Cells Promotes Experimental Metastasis via the Induction of Erythroid Differentiation Regulator 1. *Molecular Cancer Research*, 12(February):274–283.

Mazumdar, C., Shen, Y., Hong, W.-j., Howard, Y., Mazumdar, C., Shen, Y., Xavy, S., Zhao, F., Reinisch, A., Li, R., Corces, M. R., and Flynn, R. A. (2015). Leukemia-Associated Cohesin Mutants Dominantly Enforce Stem Cell Programs and Impair Human Hematopoietic Progenitor Differentiation. *Cell Stem Cell*, 17:675–688.

Mcinnes, L. and Healy, J. (2018). UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, pages 1–18.

Meer, L. T. V. D., Jansen, J. H., and Reijden, B. A. V. D. (2010). Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia*, 24:1834–1843.

Menendez-gonzalez, J. B., Vukovic, M., Abdelfattah, A., Saleh, L., Almotiri, A., Azevedo, A., Menezes, A. C., Tornillo, G., Thomas, L.-a., Edkins, S., Kong, K., Giles, P., Anjos-afonso, F., Tonks, A., Boyd, A. S., Kranc, K. R., and Rodrigues, N. P. (2019). Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia. *Stem Cell Reports*, 13:291–306.

Millot, S., Letteron, P., Lyoumi, S., Hurtado-nedelec, M., Karim, Z., Thibaudeau, O., Bennada, S., Charrier, J.-l., Lasocki, S., and Beaumont, C. (2010). Erythropoietin stimulates spleen BMP4-dependent stress erythropoiesis and partially corrects anemia in a mouse model of generalized inflammation. *Blood*, 116(26):6072–6079.

Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., Satija, R., Sanjana, N. E., Koralov, S. B., and Smibert, P. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(May).

Moignard, V., Woodhouse, S., Fisher, J., and Göttgens, B. (2013). Blood Cells , Molecules and Diseases Transcriptional hierarchies regulating early blood cell development. *Blood Cells, Molecules, and Diseases*, 51(4):239–247.

Morales-mantilla, D. E. and King, K. Y. (2018). The Role of Interferon-Gamma in Hematopoietic Stem Cell Development, Homeostasis, and Disease. *Current Stem Cell Reports*, 4:264–271.

Moran-crusio, K., Reavie, L., Shih, A., Abdel-wahab, O., Ndiaye-lobry, D., Lobry, C., Figueroa, M. E., Vasanthakumar, A., Patel, J., Zhao, X., Perna, F., Pandey, S., Madzo, J., Song, C., Dai, Q., He, C., Ibrahim, S., Beran, M., Zavadil, J., Nimer, S. D., Melnick, A., Godley, L. A., Aifantis, I., and Levine, R. L. (2011). Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. *Cancer Cell*, 20(1):11–24.

Morcos, M. N. F., Li, C., Munz, C. M., Greco, A., Dressel, N., Reinhardt, S., Dahl, A., Becker, N. B., Roers, A., Höfer, T., and Gerbaulet, A. (2020). Hematopoietic lineages diverge within the stem cell compartment. *bioRxiv*, doi.org/10.

Moudgil, A., Wilkinson, M. N., Chen, X., Morris, S. A., Dougherty, J. D., Mitra, R. D., Moudgil, A., Wilkinson, M. N., Chen, X., He, J., Cammack, A. J., and Vasek, M. J. (2020). Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell*, 182:992–1008.

Mullally, A., Lane, S. W., Ball, B., Megerdichian, C., Okabe, R., Al-shahrour, F., Paktinat, M., Haydu, J. E., Housman, E., Lord, A. M., Wernig, G., Kharas, M. G., Mercher, T., Kutok, J. L., Gilliland, D. G., and Ebert, B. L. (2010). Physiological Jak2V617F Expression Causes a Lethal Myeloproliferative Neoplasm with Differential Effects on Hematopoietic Stem and Progenitor Cells. *Cancer Cell*, 17(6):584–596.

Muller-Sieburg, C. E., Cho, R. H., Karlsson, L., Huang, J. F., and Sieburg, H. B. (2004). Myeloid-biased hematopoietic stem cells have extensive self-renewal capacity but generate diminished lymphoid progeny with impaired IL-7 responsiveness. *Blood*, 103(11):4111–4118.

Munugalavadla, V. and Kapur, R. (2005). Role of c-Kit and erythropoietin receptor in erythropoiesis. *Critical Reviews in Oncology and Hematology*, 54:63–75.

Murphy, A. J., Bijl, N., Yvan-charvet, L., Welch, C. B., Bhagwat, N., Reheman, A., Wang, Y., Shaw, J. A., Levine, R. L., Ni, H., Tall, A. R., and Wang, N. (2013). Cholesterol efflux in megakaryocyte progenitors suppresses platelet production and thrombocytosis. *Nature Medicine*, 19(5):586–600.

Naik, S. H., Perié, L., Swart, E., Gerlach, C., Van Rooij, N., De Boer, R. J., and Schumacher, T. N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*, 496(7444):229–232.

Nam, A. S., Kim, K.-t., Chaligne, R., Izzo, F., Ang, C., Taylor, J., Myers, R. M., Abu-zeinah, G., Brand, R., Omans, N. D., Alonso, A., Sheridan, C., Mariani, M., Dai, X., Harrington, E., Pastore, A., Cubillos-ruiz, J. R., Tam, W., Hoffman, R., Rabadan, R., Abdel-Wahab, O., Smilbert, P., and Landau, D. A. (2019). Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*, 571:355–366.

Nangalia, J., Massie, C. E., Baxter, E. J., Nice, F. L., Gundem, G., Wedge, D. C., Avezov, E., Li, J., Kollmann, K., Kent, D. G., Aziz, A., Godfrey, A. L., Hinton, J., Martincorena, I., Loo, P. V., Jones, A. V., Guglielmelli, P., Tarpey, P., Harding, H. P., Fitzpatrick, J. D., Teague, J. W., Meara, S. O., Mclaren, S., Bianchi, M., Silber, Y., Dimitropoulou, D., Bloxham, D., Mudie, L., Maddison, M., Robinson, B., Keohane, C., Maclean, C., Hill, K., Orchard, K., Tauro, S., Du, M., Greaves, M., Bowen, D., Huntly, B. J. P., Campbell, P. J., and Green, A. R. (2013). Somatic CALR Mutations in Myeloproliferative Neoplasms with Nonmutated JAK2. *The New England Journal of Medicine*, 369(25):2391–2405.

Nestorowa, S., Hamey, F. K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N. K., Kent, D. G., and Göttgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8):e20–e31.

Network, T. C. G. A. R. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *The New England Journal of Medicine*, 368(22):2059–2074.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69:1–15.

Nimmo, R. A., May, G. E., and Enver, T. (2015). Primed and ready: Understanding lineage commitment through single cell analysis. *Trends in Cell Biology*, 25(8):459–467.

Nivarthi, N., Chen, D., Cleary, C., Kubesova, B., Jager, R., Bogner, E., Marty, C., Pecquet, C., Vainchenker, W, Constantinescu, S., and Kralovics, R. (2016). Thrombopoietin receptor is required for the oncogenic function of CALR mutants. *Leukaemia*, 30(February):1759–1763.

Nocka, K., Tan, J. C., Chiu, E., Chu, T. Y., Ray, P., and Traktman, P. (1990). Molecular bases of dominant negative and loss of function mutations at the murine c-kit / white spotting locus: W37, WV, W41 and W. *The EMBO Journal*, 9(6):1805–1813.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O. I., Wilson, G., Kaufmann, K. B., McLeod, J., Laurenti, E., Dunant, C. F., McPherson, J. D., Stein, L. D., Dror, Y., and Dick, J. E. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*, 351(6269).

Ocasio, J., Babcock, B., Malawsky, D., Weir, S. J., Loo, L., Simon, J. M., Zylka, M. J., Hwang, D., Dismuke, T., Sokolsky, M., Rosen, E. P., Vibhakar, R., Zhang, J., Saulnier, O., Vladoiu, M., El-hamamy, I., Stein, L. D., Taylor, M. D., Smith, K. S., and Northcott, P. A. (2019). scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy. *Nature Communications*, 10:1–21.

Ochiai, H., Hayashi, T., Umeda, M., Yoshimura, M., Harada, A., Shimizu, Y., Nakano, K., Saitoh, N., Liu, Z., Yamamoto, T., Okamura, T., Ohkawa, Y., Kimura, H., and Nikaido, I. (2020). Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells. *Science Advances*, 6:1–19.

Ocone, A., Haghverdi, L., Mueller, N. S., and Theis, F. J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96.

Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H., and Nakatani, Y. (1996). The Transcriptional Coactivators p300 and CBP Are Histone Acetyltransferases. *Cell*, 87:953–959.

Oguro, H., Ding, L., and Morrison, S. J. (2013). SLAM Family Markers Resolve Functionally Distinct Subpopulations of Hematopoietic Stem Cells and Multipotent Progenitors. *Stem Cell*, 13(1):102–116.

Olausson, K. H., Nistér, M., and Lindström, M. S. (2014). Loss of Nucleolar Histone Chaperone NPM1 Triggers Rearrangement of Heterochromatin and Synergizes with a Deficiency in DNA Methyltransferase DNMT3A to Drive Ribosomal DNA Transcription. *The Journal of Biological Chemistry*, 289(50):34601–34619.

Olsson, A., Venkatasubramanian, M., Chaudhri, V. K., Aronow, B. J., Salomonis, N., Singh, H., and Grimes, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, 537(7622):698–702.

Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell*, 132:631–644.

Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., Baxter, E. J., Massie, C. E., Harrison, C. N., Vassiliou, G. S., and Vannucchi, A. (2015). Effect of Mutation Order on Myeloproliferative Neoplasms. *The New England Journal of Medicine*, 372(7):601–612.

Osborne, G. W. (2010). A Method of Quantifying Cell Sorting Yield in "Real Time". *Cytometry Part A*, 77A:983–989.

Ostrander, E. L., Kramer, A. C., Mallaney, C., Celik, H., Koh, W. K., Fairchild, J., Haussler, E., Zhang, C. R. C., and Challen, G. A. (2020). Divergent Effects of Dnmt3a and Tet2 Mutations on Hematopoietic Progenitor Cell Fitness. *Stem Cell Reports*, 14(4):551–560.

Panigrahi, A. K. and Pati, D. (2012). Higher-order orchestration of hematopoiesis: Is cohesin a new player? *Experimental Hematology*, 40(12):967–973.

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Loo, P. V., Yoon, C. J., Ellis, P., Wedge, D. C., Pellagatti, A., Shlien, A., Groves, M. J., Forbes, S. A., Raine, K., Hinton, J., Mudie, L. J., Mclaren, S., Hardy, C., Latimer, C., Porta, M. G. D., Meara, S. O., Ambaglio, I., Galli, A., Butler, A. P., Walldin, G., Teague, J. W., Quek, L., Sternberg, A., Gambacorti-passerini, C., Cross, N. C. P., Green, A. R., Boultwood, J., Vyas, P., Hellstrom-lindberg, E., Bowen, D., Cazzola, M., Stratton, M. R., and Campbell, P. J. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627.

Passegué, E., Jamieson, C. H. M., Ailles, L. E., and Weissman, I. L. (2003). Normal and leukemic hematopoiesis: Are leukemias a stem cell disorder or a reacquisition of stem cell characteristics ? *Proceedings of the National Academy of Sciences*, 100:11842–11849.

Patel, J. P., Racevskis, J., Vlierberghe, P. V., Dolgalev, I., Thomas, S., Aminova, O., Huberman, K., Cheng, J., Viale, A., Socci, N. D., Heguy, A., Cherry, A., Vance, G., Higgins, R. R., Ketterling, R. P., Gallagher, R. E., Litzow, M., Brink, M. R. M. V. D., Lazarus, H. M., Rowe, J. M., Luger, S., Ferrando, A., Paietta, E., Tallman, M. S., Melnick, A., Abdel-Wahab, O., and Levine, R. L. (2012). Prognostic Relevance of Integrated Genetic Profiling in Acute Myeloid Leukemia. *The New England Journal of Medicine*, 366(12):1079–1089.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., and Amit, I. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, 163(7):1663–1677.

Pei, W., Feyerabend, T. B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., Chen, W., Sauer, S., Wolf, S., Höfer, T., and Rodewald, H.-r. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature Publishing Group*, 548(7668):456–460.

Pei, W., Shang, F., Wang, X., Fanti, A.-k., Greco, A., Busch, K., Klapproth, K., Zhang, Q., Quedenau, C., Sauer, S., Feyerabend, T. B., Hofer, T., and Rodewald, H.-R. (2020). Resolving fate and transcriptome of hematopoietic stem cell clones. *bioRxiv*, doi.org/10:1–16.

Pei, W., Wang, X., Rössler, J., Feyerabend, T. B., and Rodewald, H.-r. (2019). Using Cre-recombinase-driven Polylox barcoding for in vivo fate mapping in mice. *Nature Protocols*, 14(June):1820–1840.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181.

Pietras, E. M., Reynaud, D., Kang, Y. A., Carlin, D., Calero-Nieto, F. J., Leavitt, A. D., Stuart, J. A., Göttgens, B., and Passegué, E. (2015). Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell*, 17(1):35–46.

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-nieto, F. J., Mulas, C., Ibarra-soria, X., Tyser, R. C. V., Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C., and Gottgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566:490–495.

Pinho, S. and Frenette, P. S. (2019). Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology*, 20(May):303–320.

Pinho, S., Marchand, T., Yang, E., Wei, Q., Nerlov, C., Frenette, P. S., Pinho, S., Marchand, T., Yang, E., Wei, Q., Nerlov, C., and Frenette, P. S. (2018). Lineage-Biased Hematopoietic Stem Cells Are Regulated by Distinct Niches. *Developmental Cell*, 44(5):634–641.e4.

Polanski, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A., and Park, J.-e. (2020). BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(August 2019):964–965.

Pop, R., Shearstone, J. R., Shen, Q., Liu, Y., Hallstrom, K., Koulnis, M., Gribnau, J., and Socolovsky, M. (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):1–18.

Povinelli, B. J., Rodriguez-Meira, A., and Mead, A. J. (2018). Single cell analysis of normal and leukemic hematopoiesis. *Molecular Aspects of Medicine*, 59:85–94.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K., and Ren, B. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience*, 21(March):432–450.

Prick, J., Haan, G. D., Green, A. R., and Kent, D. G. (2014). Clonal heterogeneity as a driver of disease variability in the evolution of myeloproliferative neoplasms. *Experimental Hematology*, 42(10):841–851.

Psaila, B., Wang, G., Rodriguez-Meira, A., Li, R., Heuston, E. F., Murphy, L., Yee, D., Hitchcok, I. S., Sousos, N., O'Sullivan, J., Anderson, S., Senis, Y. A., Weinberg, O. K., Calicchio, M. L., Iskander, D., Royston, D., Milojkovic, D., Roberts, I., Bodine, D. M., Thongjuea, S., and Mead, A. J. (2020). Single-Cell Analyses Reveal Megakaryocyte-Biased Hematopoiesis in Myelofibrosis and Identify Mutant Clone-Specific Targets. *Molecular Cell*, 78:477–492.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–988.

Raaijmakers, M. H. G. P., Mukherjee, S., Guo, S., Zhang, S., Kobayashi, T., Schoonmaker, J. A., Ebert, B. L., Al-shahrour, F., Hasserjian, R. P., Scadden, E. O., Aung, Z., Matza, M., Merkenschlager, M., Lin, C., Rommens, J. M., and Scadden, D. T. (2010). Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature*, 464(April):852–865.

Radovanovic, M., Nanopoulos, A., and Ivanovic, M. (2010). Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 11:2487–2531.

Raj, B., Wagner, D. E., Mckenna, A., Pandey, S., Klein, A. M., and Shendure, J. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36(5):442–450.

Rampal, R., Al-shahrour, F., Abdel-wahab, O., Patel, J. P., Brunel, J.-p., Mermel, C. H., Bass, A. J., Pretz, J., Ahn, J., Hricik, T., Kilpivaara, O., Wadleigh, M., Busque, L., Gilliland, D. G., Golub, T. R., Ebert, B. L., and Levine, R. L. (2014). Integrated genomic analysis illustrates the central role of JAK-STAT pathway activation in myeloproliferative neoplasm pathogenesis. *Blood*, 123(22):e123–133.

Ramsköld, D., Luo, S., Wang, Y.-c., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–788.

Rao, Y., Ruppert, A. S., Rao, Y., Lee, Y., Jarjoura, D., Ruppert, A. S., Liu, C.-g., and Hagan, J. P. (2008). A Comparison of Normalization Techniques for MicroRNA Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-p. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(284):1–17.

Rocca, S., Carrà, G., Poggio, P., Morotti, A., and Brancaccio, M. (2018). Targeting few to help hundreds: JAK , MAPK and ROCK pathways as druggable targets in atypical chronic myeloid leukemia. *Molecular Cancer*, 17(40):1–12.

Rodriguez-Fraticelli, A. E., Wolock, S. L., Weinreb, C. S., Panero, R., Patel, S. H., Jankovic, M., Sun, J., Calogero, R. A., Klein, A. M., and Camargo, F. D. (2018). Clonal analysis of lineage fate in native haematopoiesis. *Nature*, 553(7687):212–216.

Rodriguez-Meira, A., Buck, G., Clark, S.-a., Jacobsen, S. E. W., Thongjuea, S., Mead, A. J., Rodriguez-meira, A., Buck, G., Clark, S.-a., Povinelli, B. J., Alcolea, V., and Louka, E. (2019). Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular Cell*, 73(6):1292–1305.

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., and Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 182(April):176–182.

Rossi, D. J., Bryder, D., Seita, J., Nussenzweig, A., Hoeijmakers, J., and Weissman, I. L. (2007). Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature*, 447(June):725–730.

Rothenberg, E. V. (2014). Transcriptional Control of Early T and B Cell Developmental Choices. *Annual Review of Immunology*, 32:283–321.

Rowley, J. W., Oler, A. J., Tolley, N. D., Hunter, B. N., Low, E. N., Nix, D. A., Yost, C. C., Zimmerman, G. A., and Weyrich, A. S. (2011). Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. *Blood*, 118(14):101–111.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(May):547–558.

Salvador-Martinez, I., Grillo, M., Averof, M., and Telford, M. J. (2019). Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLIFE*, 8:1–23.

Sanjuan-Pla, A., Macaulay, I. C., Jensen, C. T., Woll, P. S., Luis, T. C., Mead, A., Moore, S., Carella, C., Matsuoka, S., Jones, T. B., Chowdhury, O., Stenson, L., Lutteropp, M., Green, J. C., Facchini, R., Boukarabila, H., Grover, A., Gambardella, A., Thongjuea, S., Carrelha, J., Tarrant, P., Atkinson, D., Clark, S. A., Nerlov, C., and Jacobsen, S. E. W. (2013). Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*, 502(7470):232–236.

Sarkar, A. A. and Zohn, I. E. (2012). Hectd1 regulates intracellular localization and secretion of Hsp90 to control cellular behavior of the cranial mesenchyme. *Journal of Cell Biology*, 196(6):789–800.

Sasca, D., Yun, H., Giotopoulos, G., Szybinski, J., Evan, T., Wilson, N. K., Gerstung, M., Gallipoli, P., Green, A. R., Hills, R., Russell, N., Osborne, C. S., Papaemmanuil, E., Berthold, G., Campbell, P., and Huntly, B. J. P. (2019). Cohesin-dependent regulation of gene expression during differentiation is lost in cohesin-mutated myeloid malignancies. *Blood*, 134(24):2195–2208.

Savino, A. D., Panuzzo, C., Rocca, S., Familiari, U., Piazza, R., Crivellaro, S., Carr, G., Ferretti, R., Fusella, F., Giugliano, E., Camporeale, A., Franco, I., Miniscalco, B., Cutrin, J. C., Turco, E., Silengo, L., Hirsch, E., Rege-cambrin, G., Gambacorti-passerini, C., Pandolfi, P. P., Papotti, M., Saglio, G., Tarone, G., Morotti, A., and Brancaccio, M. (2015). Morgana acts as an oncosuppressor in chronic myeloid leukemia. *Blood*, 125(14):2245–2253.

Sawai, C. M., Babovic, S., Upadhaya, S., Knapp, D. J. H. F., Lavin, Y., Lau, C. M., Merad, M., Eaves, C. J., and Reizis, B. (2016). Hematopoietic Stem Cells Are the Major Source of Multilineage Hematopoiesis in Adult Animals. *Immunity*, 45(3):597–609.

Schepers, K., Pietras, E. M., Reynaud, D., Flach, J., Binnewies, M., Garg, T., and Wagers, A. J. (2013). Myeloproliferative Neoplasia Remodels the Endosteal Bone Marrow Niche into a Self-Reinforcing Leukemic Niche. *Cell*, 13:285–299.

Schiebinger, G., Shu, J., Tabaka, M., Jaenisch, R., Regev, A., and Lander, E. S. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943.

Schoedel, K. B., Morcos, M. N. F., Zerjatke, T., Roeder, I., Grinenko, T., Voehringer, D., and Joachim, R. G. (2016). The bulk of the hematopoietic stem cell population is dispensable for murine steady-state and stress hematopoiesis. *Blood*, 128(19):2285–2296.

Schulte, R., Wilson, N. K., Prick, J. C. M., Cossetti, C., Maj, M. K., Gottgens, B., and Kent, D. G. (2015). Index sorting resolves heterogeneous murine hematopoietic stem cell populations. *Experimental Hematology*, 43(9):803–811.

Schurch, C. M., Riether, C., and Ochsenbein, A. F. (2014). Cytotoxic CD8 + T Cells Stimulate Hematopoietic Progenitors by Promoting Cytokine Release from Bone Marrow Mesenchymal Stromal Cells. *Cell Stem Cell*, 14:460–472.

Schwemmers, S., Will, B., Waller, C. F., Abdulkarim, K., Johansson, P., Andreasson, B., and Pahl, H. L. (2007). JAK2 V617F-negative ET patients do not display constitutively active JAK/STAT signaling. *Experimental Hematology*, 35:1695–1703.

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61.

Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535(7611):289–293.

Scoumanne, A., Cho, S. J., Zhang, J., and Chen, X. (2011). The cyclin-dependent kinase inhibitor p21 is regulated by RNA-binding protein PCBP4 via mRNA stability. *Nucleic acids research*, 39(1):213–224.

Seita, J. and Weissman, I. L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. *Systems Biology & Medicine*, 2(November/December):640–665.

Semlow, D. R., Zhang, J., Budzowska, M., Drohat, A. C., Walter, J. C., Semlow, D. R., Zhang, J., Budzowska, M., Drohat, A. C., and Walter, J. C. (2016). Replication-Dependent Unhooking of DNA Interstrand Cross-Links by the NEIL3 Glycosylase. *Cell*, 167(2):498–511.e14.

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2018). Palantir characterizes cell fate continuities in human hematopoiesis. *bioRxiv*.

Setty, M., Tadmor, M. D., Reich-zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(6):614–637.

Shah, S., Lubeck, E., Schwarzkopf, M., He, T.-f., Greenbaum, A., and Sohn, C. H. (2016). Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. *Development*, 143:2862–2867.

Shepherd, M. S. and Kent, D. G. (2019). Emerging single-cell tools are primed to reveal functional and molecular heterogeneity in malignant hematopoietic stem cells. *Current Opinion in Hematology*, 26(4):214–221.

Shepherd, M. S., Li, J., Wilson, N. K., Oedekoven, C. A., Li, J., Belmonte, M., Fink, J., Prick, J. C. M., Dean, C., Hamilton, T. L., Löffler, D., Rao, A., Schroeder, T., Green, A. R., and Kent, D. G. (2018). Single cell approaches identify the molecular network driving malignant hematopoietic stem cell self-renewal. *Blood*, 132(8):1–16.

Shin, J., Berg, D. A., Christian, K. M., Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., and Bonaguidi, M. A. (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell*, 17(3):360–372.

Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., Kennedy, J. A., Schimmer, A. D., Schuh, A. C., Yee, K. W., Mcleod, J. L., Doedens, M., Medeiros, J. J. F., Marke, R., Kim, H. J., Lee, K., Mcpherson, J. D., Hudson, T. J., Halt, T., Panel, P.-l. G., Brown, A. M. K., Yousif, F., Trinh, Q. M., and Stein, L. D. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*.

Shokouhian, M., Bagheri, M., Poopak, B., Chegeni, R., Davari, N., and Saki, N. (2020). Altering chromatin methylation patterns and the transcriptional network involved in regulation of hematopoietic stem cell fate. *Journal of Cellular Physiology*, (January):1–20.

Siegel, R. L., Miller, K. D., and Jemal, A. (2015). Cancer Statistics, 2015. *CA: A Cancer Journal for Clinicians*, 65(1):5–29.

Silvennoinen, O. and Hubbard, S. R. (2015). Molecular insights into regulation of JAK2 in myeloproliferative neoplasms. *Blood*, 125(22):3388–3392.

Simons, B. D. and Clevers, H. (2011). Strategies for Homeostatic Stem Cell Self-Renewal in Adult Tissues. *Cell*, 145(6):851–862.

Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(March):204–220.

Soneson, C. and Robinson, M. D. (2018). Bias , robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261.

Spanjaard, B., Hu, B., Mitic, N., Olivares-chauvet, P., Janjuha, S., Ninov, N., and Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR – Cas9-induced genetic scars. *Nature Biotechnology*, 36(5).

Sperling, A. S., Gibson, C. J., and Ebert, B. L. (2017). The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nature Reviews Cancer*, 17:5–19.

Ståhlberg, A. and Bengtsson, M. (2010). Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *Methods*, 50(4):282–288.

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., Xu, Y., and Fertig, E. J. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends in Genetics*, 34(10):790–805.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(477):1–16.

Stuart, T., Butler, A., Hoffman, P., Stoeckius, M., Smibert, P., Satija, R., Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Iii, W. M. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21.

Sun, J., Ramos, A., Chapman, B., Johnnidis, J. B., Le, L., Ho, Y.-j., Klein, A., Hofmann, O., and Camargo, F. D. (2014). Clonal dynamics of native haematopoiesis. *Nature*, 514(7522):322–327.

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(February):147–150.

Svensson, V., Vento-tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604.

Takizawa, H., Boettcher, S., and Manz, M. G. (2012). Demand-adapted regulation of early hematopoiesis in infection and inflammation. *Blood*, 119(13):2991–3002.

Takizawa, H., Regoes, R. R., Boddupalli, C. S., Bonhoeffer, S., and Manz, M. G. (2011). Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. *Journal of Experimental Medicine*, 208(2):273–284.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–384.

Tang, Y., Zhao, W., Chen, Y., Zhao, Y., and Gu, W. (2008). Acetylation Is Indispensable for p53 Activation. *Cell*, 133:612–626.

Tanikawa, C., Zhang, Y.-z., Yamamoto, R., Tsuda, Y., Tanaka, M., and Funauchi, Y. (2017). The Transcriptional Landscape of p53 Signalling Pathway. *EBioMedicine*, 20:109–119.

Tao, H., Ma, X., Su, G., Yin, J., Xie, X., Hu, C., Chen, Z., Tan, D., Xu, Z., Zheng, Y., Liu, H., He, C., Jenny, Z., Yin, H., Wang, Z., Chang, W., Peter, R., Chen, Z., Wu, D., and Yin, B. (2016). BCL11A expression in acute myeloid leukemia. *Leukemia Research*, 41:71–75.

Taylor, J., Xiao, W., and Abdel-wahab, O. (2017). Diagnosis and classification of hematologic malignancies on the basis of genetics. *Blood*, 130(4):410–423.

Tekippe, M., Harrison, D. E., and Chen, J. (2003). Expansion of hematopoietic stem cell phenotype and activity in Trp53 -null mice. *Experimental Hematology*, 31:521–527.

Thota, S., Viny, A. D., Makishima, H., Spitzer, B., Radivoyevitch, T., Przychodzen, B., Sekeres, M. A., Levine, R. L., and Maciejewski, J. P. (2014). Genetic alterations of the cohesin complex genes in myeloid malignancies. *Blood*, 124(11):1790–1798.

Tiedt, R., Hao-shen, H., Sobas, M. A., Looser, R., Dirnhofer, S., and Skoda, R. C. (2008). Ratio of mutant JAK2 -V617F to wild-type Jak2 determines the MPD phenotypes in transgenic mice. *Blood*, 111(8):3931–3940.

Tirosh, I., Izar, B., Prakadan, S. M., Ii, M. H. W., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-sichani, M., Dutton-regester, K., Lin, J.-r., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., and Kolb, K. E. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–203.

Traag, V. A., Waltman, L., and Eck, N. J. V. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:1–12.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Tsuyuzaki, K., Sato, H., and Sato, K. (2020). Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology*, 21(9):1–17.

Tusi, B. K., Wolock, S. L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J. R., Klein, A. M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, 555(7694):54–60.

Ugajin, T., Kojima, T., Mukai, K., Obata, K., and Kawano, Y. (2009). Basophils preferentially express mouse mast cell protease 11 among the mast cell tryptase family in contrast to mast cells. *Journal of Leukocyte Biology*, 86(December):1417–1425.

Ugo, V., Marzac, C., Debili, N., Vainchenker, W., and Casadevall, N. (2004). Multiple signaling pathways are involved in erythropoietin-independent differentiation of erythroid progenitors in polycythemia vera. *Experimental Hematology*, 32:179–187.

Urso, P. and Congdon, C. C. (1956). The Effect of the Amount of Isologous Bone Marrow Injected on the Recovery of Hematopoietic Organs, Survival and Body Weight after Lethal Irradiation Injury in Mice. *Blood*, 12(3):251–260.

Vainieri, M. L., Blagborough, A. M., Maclean, A. L., Haltalli, M. L. R., Ruivo, N., Fletcher, H. A., Stumpf, M. P. H., Sinden, R. E., and Celso, C. L. (2016). Systematic tracking of altered haematopoiesis during sporozoite- mediated malaria development reveals multiple response points. *Open Biology*, 6.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS : Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Computational Biology*, DOI:10.137:1–18.

Valletta, S., Thomas, A., Meng, Y., Ren, X., Drissen, R., Genua, C. D., and Nerlov, C. (2020). Micro-environmental sensing by bone marrow stroma identifies IL-6 and TGFB1 as regulators of hematopoietic ageing. *Nature Communications*, 11(4075):1–13.

Vardiman, J. W., Arber, D. A., Brunning, R. D., Borowitz, M. J., Porwit, A., Harris, N. L., Beau, M. M. L., and Hellstro, E. (2009). The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*, 114(5):937–951.

Vassiliou, G. S., Cooper, J. L., Rad, R., Li, J., Rice, S., Uren, A., Rad, L., Ellis, P., Andrews, R., Banerjee, R., Grove, C., Wang, W., Liu, P., Wright, P., Arends, M., and Bradley, A. (2011). Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice. *Nature Genetics*, 43(5):470–475.

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W. K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A., and Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19(4):271–281.

Verstovsek, S. (2012). Advanced systemic mastocytosis: the impact of KIT mutations in diagnosis, treatment, and progression. *European Journal of Haematology*, 90(7):89–98.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P. L., Rozenblatt-Rosen, O., Lane, A. A., Haniffa, M., Regev, A., and Hacohen, N. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573.

Voulgaridou, G.-p., Anestopoulos, I., Franco, R., Panayiotidis, M. I., and Pappa, A. (2011). DNA damage induced by endogenous aldehydes : Current state of knowledge. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 711(1-2):13–27.

Wagner, D. E. and Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(July):410–25.

Wagner, D. E., Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 4362(April):1–13.

Walkley, C. R., Olsen, G. H., Dworkin, S., Fabb, S. A., Swann, J., Mcarthur, G. A., Westmoreland, S. V., Chambon, P., Scadden, D. T., and Purton, L. E. (2007). A Microenvironment-Induced Myeloproliferative Syndrome Caused by Retinoic Acid Receptor gamma Deficiency. *Cell*, 129:1097–1110.

Wang, X., Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F.-a., and Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 5691(June):1–18.

Watcham, S., Kucinski, I., and Gottgens, B. (2019). New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. *Blood*, 133(13):1415–1426.

Weinreb, C., Rodriguez-Fraticelli, A. E., Camargo, F. D., and Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(755):1–9.

Weinreb, C., Wolock, S., and Klein, A. M. (2018a). Gene expression SPRING : a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(August):1246–1248.

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., and Klein, A. M. (2018b). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, page 201714723.

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016). SLICER : inferring branched , nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, pages 1–15.

Welch, J. S., Ley, T. J., Link, D. C., Miller, C. A., Larson, D. E., Koboldt, D. C., Wartman, L. D., Lamprecht, T. L., Liu, F., Xia, J., Kandoth, C., Fulton, R. S., Mclellan, M. D., Dooling, D. J., Wallis, J. W., Chen, K., Harris, C. C., Schmidt, H. K., Kalicki-veizer, J. M., Lu, C., Zhang, Q., Lin, L., Laughlin, M. D. O., Mcmichael, J. F., Delehaunty, K. D., Fulton, L. A., Magrini, V. J., Mcgrath, S. D., Demeter, R. T., Vickery, T. L., Hundal, J., Cook, L. L., Swift, G. W., Reed, J. P., Alldredge, P. A., Wylie, T. N., Walker, J. R., Watson, M. A., Heath, S. E., Shannon, W. D., Varghese, N., Nagarajan, R., Payton, J. E., Baty, J. D., Kulkarni, S., Klco, J. M., Tomasson, M. H., Westervelt, P., Walter, M. J., Graubert, T. A., Dipersio, J. F., Ding, L., Mardis, E. R., and Wilson, R. K. (2012). The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell*, 150:264–278.

Wilson, A., Laurenti, E., Oser, G., Wath, R. C. V. D., Blanco-bose, W., Jaworski, M., Macdonald, H. R., Offner, S., Dunant, C. F., Eshkind, L., Bockamp, E., and Lio, P. (2008). Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell*, 135:1118–1129.

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schu, J., Kaimakis, P., Chilarska, P. M., Kinston, S., Ouwehand, W. H., Dzierzak, E., and Pimanda, J. E. (2010). Combinatorial Transcriptional Control In Blood Stem/Progenitor Cells: Genome-wide Analysis of Ten Major Transcriptional Regulators. *Cell Stem Cell*, 7:532–544.

Wilson, N. K., Kent, D. G., Buettner, F., Shehata, M., Macaulay, I. C., Calero-Nieto, F. J., Sánchez Castillo, M., Oedekoven, C. A., Diamanti, E., Schulte, R., Ponting, C. P., Voet, T., Caldas, C., Stingl, J., Green, A. R., Theis, F. J., and Göttgens, B. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell*, 16(6):712–724.

Witthuhn, B. A., Queile, F. W., Yi, T., Tang, B., Miura, O., and Ihle, J. N. (1993). JAK2 Associates with the Erythropoietin Receptor and Is Tyrosine Phosphorylated and Activated following Stimulation with Erythropoietin. *Cell*, 74:227–236.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(15):1–5.

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(59):1–9.

Wolock, S. L., Lopez, R., and Klein, A. M. (2018). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *bioRxiv*, pages 1–18.

Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881.

Xie, S., Duan, J., Li, B., Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G. C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, 66(2):285–299.e5.

Yáñez, A., Coetzee, S. G., Olsson, A., Muench, D. E., Berman, B. P., Hazelett, D. J., Salomonis, N., Grimes, H. L., and Goodridge, H. S. (2017). Granulocyte-Monocyte Progenitors and Monocyte-Dendritic Cell Progenitors Independently Produce Functionally Distinct Monocytes. *Immunity*, 47(5):890–902.e4.

Yang, L., George, J., and Wang, J. (2020). Deep Profiling of Cellular Heterogeneity by Emerging Single-Cell Proteomic Technologies. *Proteomics*, 20:1–12.

Yvan-charvet, L., Pagler, T., Gautier, E. L., Avagyan, S., Siry, R. L., Han, S., Welch, C. L., Wang, N., Randolph, G. J., Snoeck, H. W., and Tall, A. R. (2010). ATP-Binding Cassette Transporters and HDL Suppress Hematopoietic Stem Cell Proliferation. *Science*, 328(June):1689–1694.

Zhang, J., Niu, C., Ye, L., Huang, H., He, X., Harris, S., Wiedemann, L. M., Mishina, Y., and Li, L. (2003). Identification of the haematopoietic stem cell niche and control of the niche size Jiwang. *Nature*, 425(October):836–841.

Zhang, X., Su, J., Jeong, M., Ko, M., Huang, Y., Park, H. J., Guzman, A., Lei, Y., Huang, Y.-h., Rao, A., Li, W., and Goodell, M. A. (2016). DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nature Genetics*, 48(9):1014–1021.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:1–12.

Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell multimodal omics: the power of many. *Nature Methods*, 17(January):11–14.