Charting the single-cell transcriptional landscape of haematopoiesis

Fiona Kathryn Hamey

Department of Haematology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

King’s College

August 2018

Charting the single-cell transcriptional landscape of haematopoiesis

Fiona Kathryn Hamey

High turnover in the haematopoietic system is sustained by stem and progenitor cells, which divide and mature to produce the range of cell types present in the blood. This complex system has long served as a model of differentiation in adult stem cell systems and its study has important clinical relevance. Maintaining a healthy blood system requires regulation of haematopoietic cell fate decisions, with severe dysregulation of these fate choices observed in diseases such as leukaemia. As transcriptional regulation is known to play a role in this regulation, the gene expression of many haematopoietic progenitors has been measured. However, many of the classic populations are actually extremely heterogeneous in both expression and function, highlighting the need for characterising the haematopoietic progenitor compartment at the level of individual cells.

The first aim of this work was to chart the single-cell transcriptional landscape of the haematopoietic stem and progenitor cell (HSPC) compartment. To build a comprehensive map of this landscape, 1,654 HSPCs from mouse bone marrow were profiled using single-cell RNA-sequencing. Analysis of these data generated a useful resource, and reconstructed changes in gene expression, cell cycle and RNA content along differentiation trajectories to three blood lineages.

To investigate how single-cell gene expression can be used to learn about regulatory relationships, data measuring the expression of 41 genes (including 31 transcription factors) in 2,167 stem and progenitor cells were used to construct Boolean gene regulatory network models describing the regulation of differentiation from stem cells to two different progenitor populations. The inferred relationships revealed positive regulation of Nfe2 and Cbfa2t3h by Gata2 that was unique to differentiation towards megakaryocyte-erythroid progenitors, which was subsequently experimentally validated.

The next study focused on investigating the link between transcriptional and functional heterogeneity within blood progenitor populations. Single-cell profiles of human cord blood progenitors revealed a continuum of lympho-myeloid gene expression. Culture assays performed to assess the functional output of single cells found both unilineage and bilineage output and, by investigating the link between surface marker expression and function, a new sorting strategy was devised that was able to enrich for function within conventional lympho-myeloid progenitor sorting gates.

The final project aimed to study changes to the HSPC compartment in a perturbed state. A droplet-based single-cell RNA-sequencing dataset of 44,802 cells was analysed to identify entry points to eight blood lineages and to characterise gene expression changes in this transcriptional landscape. Mapping single-cell data from W41/W41 *Kit* mutant mice highlighted quantitative shifts in progenitor populations such as a reduction in mast cell progenitors and an increase towards more mature progenitors along the erythroid trajectory. Differential gene expression identified upregulation of stress response and a reduction of apoptosis during erythropoiesis as potential compensatory mechanisms in the *Kit* mutant progenitors.

Together this body of work characterises the HSPC compartment at single-cell level and provides methods for how single-cell data can be used to discover regulatory relationships, link expression heterogeneity to function, and investigate changes in the transcriptional landscape in a perturbed environment.

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. Specifc details of work done in collaboration are given at the start of relevant chapters. The contents of 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. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. The total length of the main body of this dissertation including figure legends is 42,629 words and therefore does not exceed the limit of 60,000 words for such a dissertation.

Fiona Kathryn Hamey
August 2018

Acknowledgements

Firstly, I would like to thank my supervisor, Bertie Göttgens, for providing fantastic mentorship and support over the last four years. I am also indebted to the entire Göttgens lab, as I am honoured to have worked alongside this group of kind, funny and helpful people. Particularly I would like to acknowledge Sonia Nestorowa, Joakim Dahlin and Nicola Wilson, who all provided me with a great deal of support and were a pleasure to work with. Special thanks goes to the members of the bioinformatics office: Lila, Beccy, Xiaonan, Sam, Ivan, Chee, Wajid (and occasional guest Blanca) for being such great office mates in such a small space. I have enjoyed our laughs, rants and (too much of) our cake over the years.

I am lucky to have had some brilliant collaborators and would like to thank Alex Wolf, Laleh Haghverdi and Fabian Theis for inviting me to visit Munich, and Dimitris Karamitros, Bilyana Stoilova, and Paresh Vyas for involving me in their work. I would also like to express my graditude to Alfonso Martinez Arias for our many interesting discussions. My thanks also goes to Brian Hendrich and the other members of the Stem Cell Institute who decided to give me a place on their PhD programme, and to the Medical Research Council for funding my PhD. I would like to thank Jo Jack and Martin Dawes who always went above and beyond answering any of my questions. And thanks also to Anastasiya, Caroline, Livvi and Lucia for the tea and cake meet-ups and all of our WhatsApp conversations.

Thank you to all of my maths friends who let me invade your happy hour, and in particular to the MMT crew for our ambitious cooking adventures—especially Benjamin Barrett and James Munro who have been my housemates this past year. Finishing a PhD would have been nowhere near as fun without all of our brilliant but time-consuming parties, home brewing and foraging quests. I would like to thank my parents and my sister for their love and support, including the countless car journeys to ferry my belongings to and from Cambridge. And lastly, I would like to thank James for being by my side throughout this PhD, always helping me to calm down when I got stressed, bringing so much great wave artwork into my life and for putting up with all of the houseplants and owls.

Table of contents

List of figures

List of tables

List of abbreviations

AML	Acute myeloid leukaemia
ANOVA	Analysis of variance test
cDNA	Complementary DNA
ChIP-seq	Chromatin immunoprecipitation sequencing
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
CRISPR	Clustered regularly interspaced short palindromic repeats
Ct	Cycle-threshold
ΔCt	Change in cycle-threshold
CyTOF	Cytometry by time of flight
DPT	Diffusion pseudotime
E	Erythroid lineage
ERCC	External RNA controls consortium
Ery	Erythroid
FACS	Fluorescence-activated cell sorting
FSC-H	Forward-scattered light-height

GM	Granulocyte-macrophage lineage
GMP	Granulocyte-macrophage progenitor
HSC	Haematopoietic stem cell
HSPC	Haematopoietic stem/progenitor cell
ICA	Independent component analysis
IVT	In vitro transcription
L	Lymphoid lineage
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
MEP	Megakaryocyte-erythroid progenitor
Mk	Megakaryocyte
MLP	Multi-lymphoid progenitor
MolO	Molecular overlapping population of HSCs defined by Wilson et al. (2015)
mRNA	Messenger RNA
PBA	Population balance analysis
PC	Principal component
PCA	Principal component analysis
preMegE	Pre-megakaryocyte-erythroid progenitor

Prog	Progenitor (Lin$^-$ c-Kit$^+$ Sca1$^-$) sorting gate
qRT-PCR	Quantitative reverse-transcription polymerase chain reaction
RNA-seq	Sequencing of RNA
scATAC-seq	Single-cell assay for transposase-accessible chromatin using sequencing
scRNA-seq	Single-cell RNA-sequencing
SPADE	Spanning-tree progression analysis of density-normalised events
SSC	Side-scatter
ST-HSC	Short-term haematopoietic stem cell
t-SNE	t-distributed stochastic neighbour embedding
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
W^{41}/W^{41}	Mouse model with V831M mutation in the *Kit* gene
WT	Wild-type

Papers arising from this PhD

* represents equal contribution of both authors to paper

1. **A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation.** Nestorowa S*, Hamey FK*, Pijuan-Sala B, Diamanti E, Shepherd M, Laurenti E, Wilson NK, Kent DG and Göttgens B. (2016). *Blood.*
2. **Reconstructing blood stem cell regulatory network models from single-cell molecular profiles.** Hamey FK*, Nestorowa S*, Kinston SJ, Kent DG, Wilson NK and Göttgens B. (2017). *Proceedings of the National Academy of Sciences.*
3. **A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice.** Dahlin JS*, Hamey FK*, Pijuan-Sala B, Shepherd M, Lau WWY, Nestorowa S, Weinreb C, Wolock S, Hannah R, Diamanti E, Kent DG, Göttgens B and Wilson NK. (2018). *Blood.*
4. **Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells.** Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L and Theis FJ. (2017). *BioRxiv.*
5. **Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells.** Karamitros D, Stoilova B, Aboukhalil Z, Hamey FK, Reinisch A, Samitsch M, Quek L, Otto G, Repapi E, Doondeea J, Usukhbayar B, Calvo J, Taylor S, Goardon N, Six E, Pflumio F, Porcher C, Majeti R, Göttgens B and Vyas P. (2018). *Nature Immunology.*
6. **Mbd3/NuRD controls lymphoid cell fate and inhibits tumorigenesis by repressing a B cell transcriptional program.** Loughran SJ, Comoglio F, Hamey FK, Giustacchini A, Errami Y, Earp E, Göttgens B, Jacobsen SEW, Mead AJ, Hendrich B and Green AR. (2017). *Journal of Experimental Medicine.*

7. **Advancing haematopoietic stem and progenitor cell biology through single-cell profiling.** Hamey FK*, Nestorowa S*, Wilson NK and Göttgens B. (2016). *FEBS Letters*.

8. **Demystifying blood stem cell fates.** Hamey FK and Göttgens B. (2017). *Nature Cell Biology*.

9. **Sorting apples from oranges in single-cell expression comparisons.** Hamey FK and Göttgens B. (2018). *Nature Methods*.

Chapter 1

Introduction

Parts of this sections have been adapted from the news and views articles and the review by F. Hamey, which were written as part of this PhD (Hamey and Göttgens, 2017, 2018; Hamey et al., 2016).

1.1 Haematopoiesis

Blood contains a complex mixture of specialised cell types, ranging from the erythrocytes which transport oxygen around the body, to the lymphocytes such as T cells that make up a vital part of the immune system. It has been estimated that an adult human has a turnover of around 10^{12} blood cells per day (Ogawa, 1993). To sustain itself, in order to continue performing all of its vital functions, the haematopoietic system requires constant production of new cells. This process is known as haematopoiesis.

Remarkably, the whole complex mixture of blood cell types can be generated from a single cell: the haematopoietic stem cell (HSC) (Bryder et al., 2006). Experiments in the 1960s from Till and McCulloch demonstrated that there are cells in the bone marrow with the abilities of self-renewal, where a cell can give rise to more cells of the same type, and differentiation towards multiple blood lineages (Till and McCulloch, 1961). Extensive research over the past 60 years has refined our knowledge of haematopoiesis, and led to the discovery of a number of haematopoietic stem and progenitor cell (HSPC) populations. The existence of cells such as common myeloid progenitors (CMPs), capable of producing both erythroid and myeloid colonies (Akashi et al., 2000), and common lymphoid progenitors (CLPs), restricted to only lymphoid output upon differentiation (Kondo et al., 1997), supported the model of a haematopoietic tree, where cells undergo stepwise differentiation along branches towards different fates, gradually becoming more specialised (Fig. 1.1). Multipotent progenitors at the top of the hierarchy were further characterised to identify the true stem cells capable of indefinite self-renewal (long-term HSCs or LT-HSCs), those cells with limited self-renewal capacity (short-term HSCs or ST-HSCs) and cells that had lost self-renewal but retained the ability to differentiate towards all of the blood lineages (multipotent progenitors or MPPs). Haematopoiesis is a widely-used model in stem cell biology due to the accessibility of material, the clinical relevance of the blood system, and the existence of sophisticated assays that are able to test for stem and progenitor function, which are lacking in many other adult stem cells systems.

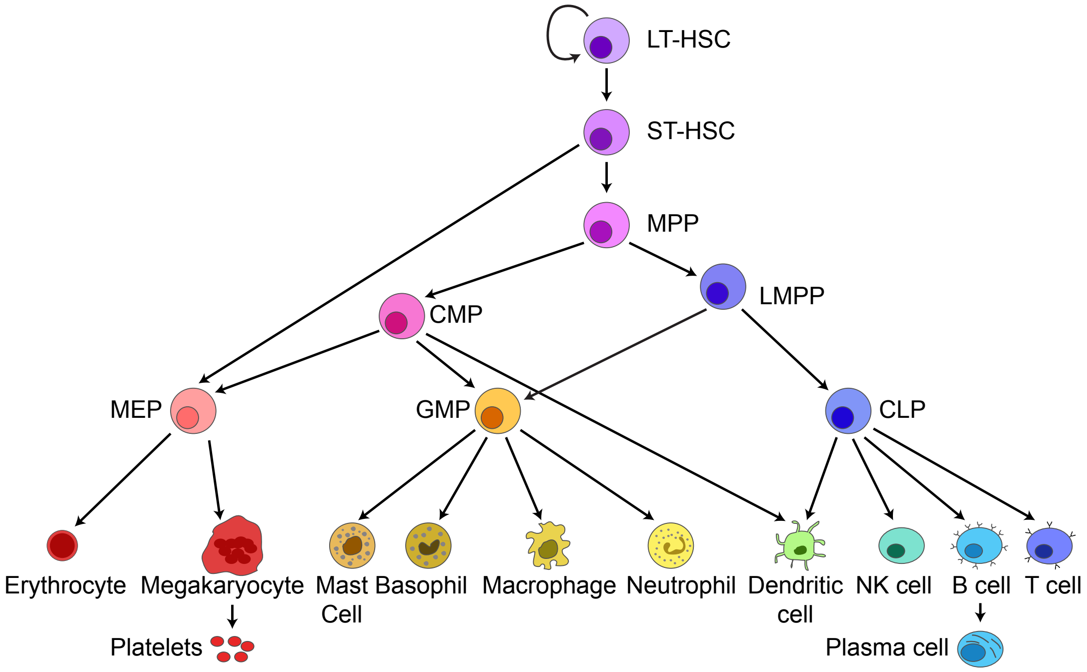


Fig. 1.1. The haematopoietic tree. Classic view of the haematopoietic hierarchy based on Moignard et al. (2013) and Wilson et al. (2015). LT-HSC, long-term haematopoietic stem cell; ST-HSC, short-term haematopoietic stem cell; MPP, multipotent progenitor; MEP, megakaryocyte-erythroid progenitor; GMP, granulocyte-macrophage progenitor; CMP, common myeloid progenitor; LMPP, lymphoid-primed multipotent progenitor; CLP, common lymphoid progenitor; NK cell, natural killer cell.

1.1.1 Cell fate decisions in haematopoiesis

Maintaining the correct numbers of mature cell types in this complex system requires careful balancing of cell fate decisions. Many serious blood disorders, including several leukaemia, show evidence of severely disrupted haematopoietic decision-making (Tenen, 2003), highlighting the need to understand the processes regulating HSPC differentiation. There are a wide range of factors regulating haematopoiesis, such as extrinsic signalling and epigenetic regulation. Amongst these lies transcriptional control, with a number of transcription factors known to play key roles in blood cell differentiation (Shivdasani and Orkin, 1996). For example, the transcription factor protein Scl (encoded by the gene Tal1) has been demonstrated to have a key role in haematopoiesis, with disruption of Tal1 causing defects in the erythroid and megakaryocyte lineages (Hall et al., 2003; Kallianpur et al., 1994; Shivdasani et al., 1995a). Another example of transcriptional control comes from neutrophil differentiation, which is regulated by the transcription factor Gfi1, with mice that do not have *Gfi1* expression lacking production of neutrophils (Hock et al., 2003; Karsunky et al., 2002). Therefore, investigating how the transcriptional landscape of a given cell changes during haematopoiesis is at the heart of understanding the cell fate decisions in this system.

A major benefit for the haematopoiesis community is that researchers are able to identify cells from different haematopoietic populations based on the levels of surface marker proteins, thereby allowing characterisation of cell types within the HSPC compartment. Surface marker-based strategies have been instrumental in work to establish the structure of the haematopoietic hierarchy, with methods existing for identifying a large number of HSPC populations, including all of the cell types in the haematopoietic tree model displayed in Fig. 1.1.

1.2 Single-cell biology

For maintenance of a healthy tissue, stem and progenitor cells must regulate their functional output at the population level. Yet for multipotent cell types, differentiation is a stochastic process occurring at the level of individual cells (Klein and Simons, 2011; Simons and Clevers, 2011). This realisation, coupled with rapid advances in single-cell profiling technologies, has led to an explosion in single-cell studies across multiple fields, and in particular in haematopoiesis (Hamey et al., 2016).

1.2.1 Evidence of heterogeneity within HSPC populations

Experiments assaying the functional output of HSC populations with seemingly homogeneous surface marker profiles have revealed clear differences in the functional output of these cells upon transplantation, with evidence for HSC subpopulations with a range of lineage biases (Dykstra et al., 2007). Such variation within the classical surface marker-defined HSPC populations underlines the importance of considering the properties of individual cells, rather than relying on measurements that represent population averages (Fig. 1.2). In order to facilitate this, many approaches make use of fluorescence-activated cell sorting (FACS) as a means of isolating cells. By staining cells with combinations of fluorophore-conjugated antibodies that each mark different proteins, single cells can be separated based on their fluorescence, where this is an indicator of surface marker expression levels (Lindström, 2012). This ability to purify specific haematopoietic populations has allowed major advances such as repeated refinement of the HSC compartment to enrich for functional LT-HSCs, which possess the ability to reconstitute the blood system after serial transplantation (Kent et al., 2009; Kiel et al., 2005; Morita et al., 2010). Representing a significant advance in FACS technology, index sorting is a technique that allows measurements for FACS parameters to be recorded as each cell is sorted into a plate before analysis using one of a wide selection of different assays (Osborne, 2011; Schulte et al., 2015). By collecting this information, the results of experimental techniques such as functional and gene expression assays can be related to the FACS profiles of individual cells, which can be used to devise improved purification strategies and link molecular profiles to functional output (Wilson et al., 2015).

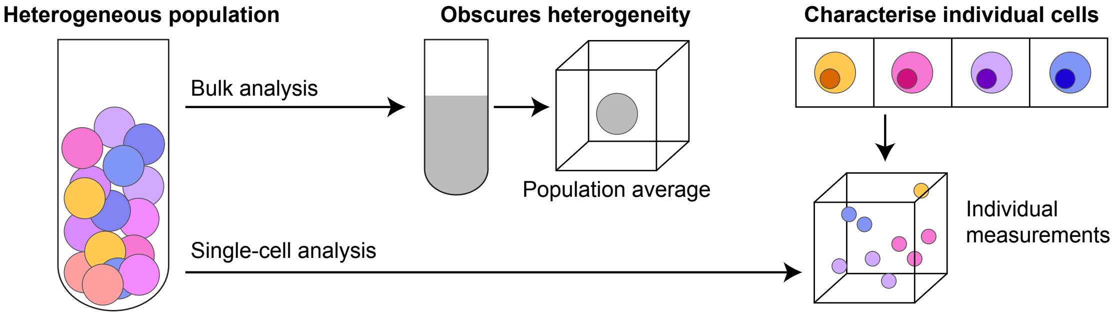


Fig. 1.2. The importance of single-cell analysis. When profiling a population, bulk analysis can only provide information about its average properties. This is particularly a problem for heterogeneous populations, as any variance due to different cell states will be obscured. Single-cell techniques overcome this limitation by allowing individual cells to be characterised.

Single-cell functional studies have also helped to question the classic view of the blood stem and progenitor hierarchy (Laurenti and Göttgens, 2018). Notta et al. (2016) found that whilst there were multipotent cells in human adult haematopoiesis, these were in fact mainly from the stem cell compartment. The majority of downstream progenitors were seemingly unipotent in functional assays, in contrast to the classical idea in the hierarchical model of stepwise loss of potential. Traditional bulk assays could not have revealed this behaviour of individual cells, again emphasising the real need for performing analysis at the single-cell level. Genetic barcoding of cells is another single-cell technique that has been used to reassess the established definitions of haematopoietic progenitor populations (Naik et al., 2014). Here, cells from a population of interest are first isolated, and then each one is labelled with a different genetic barcode. These cells can then be transplanted into lethally irradiated mice. Once the bone marrow has been successfully repopulated, cells from the different blood lineages can be isolated and sequencing performed to reveal the groups of cells with shared barcodes, and by extension a shared origin. This approach allows the lineage output of individual cells to be tested *in vivo*. Perié and colleagues used such a barcoding approach to assess the lineage output of CMPs, a population originally described as producing both erythroid and myeloid cells (Akashi et al., 2000). In this study, barcodes were found to be mostly restricted to either erythroid or myeloid lineages, rather than being seen in a combination of these fates, suggesting that lineage restriction during haematopoiesis largely occurs before the CMP stage (Perié et al., 2015). Again, this type of study emphasises the necessity of performing functional analysis at a single-cell level and highlights some of the limitations that can affect conclusions based on bulk data.

1.2.2 Single-cell expression profiling

After observing large variation in the function of cells within haematopoietic populations, the next question to ask is what lies behind this functional heterogeneity? One way that researchers have attempted to answer this is by profiling gene and protein expression across HSPC populations. Historically, expression profiling was limited to being performed at the population level, due to the need for a sufficient amount of starting material to detect expression of genes or proteins. However, in the past few years technological advances have enabled the levels of multiple genes, or proteins, to be measured simultaneously at the single-cell level. These new technologies have resulted in the generation of increasingly large datasets, and have provided a range of insights into the haematopoietic system.

As discussed in Section 1.2.1, FACS is one way in which researchers can measure the expression of numerous proteins for an individual cells, and has proved a powerful tool in understanding the heterogeneity across single-cell molecular profiles. Whilst FACS can measure up to 30 parameters for each cell (Chattopadhyay and Roederer, 2015), this number is however limited by the availability of dyes with distinct fluorescence and the ability of software to distinguish between the different wavelengths of fluorescence emitted by the dyes. In contrast, mass cytometry (commercially available from Fluidigm as Cytometry by Time of Flight, or CyTOF), labels antibodies targeting proteins of interest with stable isotopes of rare metals, rather than the fluorescent dyes used in FACS. Quantification of these metal-tagged antibodies is then performed using time-of-flight mass spectrometry and allows over 40 parameters to be measured for each cell (Bendall et al., 2012; Spitzer and Nolan, 2016). This technology has been used to measure a wide range of cellular features, such as cell cycle state in human bone marrow progenitors (Behbehani et al., 2012) and to investigate phenomena such as immune signalling in human bone marrow in response to drug inhibition (Bendall et al., 2011).

Studying the expression pattern of high numbers of genes in individual cells is also possible, and has proved a valuable tool for dissecting heterogeneity with haematopoietic populations. Techniques initially developed for generating bulk expression data have been extended to work on single cells. Microarray technology, which can quantify the expression of thousands of messenger RNA (mRNA) transcripts from a sample, consists of a slide printed with a large number of DNA sequences, called probes. Transcripts are converted to complementary DNA (cDNA) that then hybridises to the probes with matching sequences. Material from different samples is labelled with different fluorescent dyes, and the resulting fluorescence intensity allows quantification of gene expression. Development of new cDNA amplification methods allowed generation of sufficient material from individual cells for use with microarray profiling (Kurimoto et al., 2006).

Another method for measuring gene expression is quantitative reverse transcription polymerase chain reaction (qRT-PCR). This technique is targeted to measure expression of selected genes by amplifying specific mRNA transcripts using the polymerase chain reaction (PCR), with the inclusion of a fluorescent reporter that is generated with each amplification cycle. A higher number of starting molecules results in increased concentration of the reporter, enabling quantification of gene expression. Single-cell qRT-PCR allows the expression of multiple genes to be measured at the single-cell level (Sanchez-Freire et al., 2012; Ståhlberg and Bengtsson, 2010). The Fluidigm BioMark™ system provides a commercially available approach for capturing the profiles of up to 96 genes, and it has been shown that this number can be extended using multiplexing approaches (Guo et al., 2013a). As this technology can only detect expression of specific molecules, the genes must be hand-picked by the investigator before carrying out the experiment, which requires knowledge about relevant gene expression in the system. In the context of haematopoiesis, single-cell qRT-PCR has been widely used, including in studies investigating cellular heterogeneity (Guo et al., 2013a; Wilson et al., 2015) and transcriptional regulation (Moignard et al., 2013, 2015; Pina et al., 2012, 2015).

This technique is, however, limited by the requirement to pre-select genes before analysis, which reduces the opportunity to discover novel genes involved in the transcriptional regulation of a system. A transcriptome-wide approach is instead offered by single-cell RNA-sequencing (scRNA-seq), which reverse transcribes the mRNA from a single cell, amplifies the resulting cDNA and then performs library preparation for sequencing (Kolodziejczyk et al., 2015). The first study performing RNA-sequencing (RNA-seq) at the single-cell level profiled only a handful of mouse blastomeres and oocytes, using single-cell sequencing as a means to assess gene expression in the early stages of development where there is very limited material (Tang et al., 2009). After this work, new protocols rapidly followed that vastly improved the cellular throughput and sensitivity of scRNA-seq (Hashimshony et al., 2012; Islam et al., 2011; Ramsköld et al., 2012). Following cell isolation and lysis, the majority of methods use poly(T) priming for capturing and reverse transcribing mRNA, with the exception of DP-seq from Bhargava et al. (2013), which instead uses a library of designed primers that bind the transcripts and initiate reverse transcription. The second strand synthesis step is then performed using either poly(A) tailing, as in the original Tang et al. (2009) protocol (Hashimshony et al., 2012; Sasagawa et al., 2013), or template-switching (Islam et al., 2011; Ramsköld et al., 2012). Different approaches also result in different regions of the transcripts being amplified for sequencing, with many methods opting for either 5' (Fan et al., 2015; Islam et al., 2011) or 3' (Hashimshony et al., 2012; Nakamura et al., 2015; Soumillon et al., 2014) counting, where reverse transcription is restricted to either the 3' or 5' end of the transcript. The SMART-Seq protocol, later appearing in the optimised form of SMART-Seq2, uses an alternative strategy to obtain full length transcripts, which can enable the investigation of features such as alternative splicing (Picelli et al., 2013; Ramsköld et al., 2012). To amplify material before library preparation either PCR (Islam et al., 2011; Nakamura et al., 2015; Ramsköld et al., 2012; Tang et al., 2009), or in vitro transcription (IVT) (Hashimshony et al., 2012; Sasagawa et al., 2013) can be used, with IVT enabling linear amplification, opposed to the exponential amplification of PCR.

Amplification is an essential step in obtaining sufficient material for sequencing, but inevitably introduces biases as not all transcripts are equally amplified. Duplicate reads originating from amplification of the same transcript cannot be distinguished from reads aligning to the same gene but originating from different transcripts. To combat this, several methods now include the addition of barcodes to uniquely mark reads coming from a single mRNA molecule (Fu et al., 2011; Hug and Schuler, 2003). These barcodes, known as unique molecular identifiers (UMIs) are incorporated into the sequencing libraries and allow duplicate reads arising from the amplification process to be identified and corrected for (Kivioja et al., 2012). A number of scRNA-seq protocols now incorporate this feature (Fan et al., 2015; Grün et al., 2014; Hashimshony et al., 2016; Islam et al., 2014; Jaitin et al., 2014; Sasagawa et al., 2018; Soumillon et al., 2014).

Another challenge in performing scRNA-seq comes from the high cost per cell, especially in plate-based methods such as SMART-Seq2, which require the use of large amounts of reagent (Picelli et al., 2013). Several approaches have been developed to address this issue in an attempt to both decrease cost and increase the throughput of scRNA-seq methods. Jaitin et al. (2014) proposed the massively parallel single-cell sequencing framework (MARS-Seq), where cells are index-sorted into 384-well plates and labelled in the first step of reverse transcription with a unique barcode in each well, before being pooled for bulk processing during subsequent steps. This approach reduces the cost of generating sequencing libraries, but results in much shallower sequencing depth per cell. CytoSeq, and more recently Microwell-Seq, use microwell technology to capture cells in a plate with a very high number of wells (Fan et al., 2015; Han et al., 2018). Cells are lysed within wells so that mRNA hybridises to barcoded beads, allowing material to be pooled before the following stages to reduce the amount of reagents needed per cell. Droplet-based microfluidic protocols, which have been used very widely, represent another approach for performing scRNA-seq on high numbers of cells. These work by capturing individual cells in droplets along with barcoded beads. Cells are lysed within droplets so that their transcripts are barcoded before pooling by merging the drops, meaning that tens of thousands of cells can be processed in an experiment. These high-throughput droplet-based methods were first published as Drop-seq (Macosko et al., 2015) and InDrop (Klein et al., 2015), and a commercial platform is also available in the form of the 10x Chromium™ system from 10x Genomics (Zheng et al., 2017). Finally, another high-throughput approach which has been described is SPLiT-seq, which uses multiple rounds of barcoding on pools of cells, where cells are re-pooled between each round of barcoding, so that each cell ends up with a unique barcode without the need for expensive reagents or equipment (Rosenberg et al., 2018). With a range of technologies available, the best choice for a specific project ultimately depends on the research question of interest, with lower-throughput methods such as SMART-Seq2 costing more per cell, but resulting in a more complete picture of a cell's transcriptome than methods that profile tens of thousands of cells.

1.2.3 Technical limitations of single-cell gene expression data

Working with single-cell gene expression data presents a range of technical challenges. The total RNA content of cells can vary widely, prohibiting using the measured gene counts directly when making comparisons between cells. Instead, data must first be normalised to account for the amount of starting material. A common approach in the analysis of single-cell qRT-PCR data is to adjust the expression values in each cell relative to the expression of one or more housekeeping genes, with the assumption that these genes are expressed at a constant level in all cells. For scRNA-seq data, as expression measurements are often very noisy, and as many more genes are measured in comparison to qRT-PCR, normalisation is not carried out based on the values of two or three genes. Instead, a range of normalisation techniques have been suggested, including applying normalisation from methods originally developed for use with bulk expression data, such as the normalisation from the DESeq method (Anders and Huber, 2010). Here, each cell is normalised by a size factor, which is a value that aims to represent the amount of material captured from that cell. DESeq calculates size factors by scaling each gene count by its geometric mean across all cells, and then taking the median of these scaled values in a cell as its size factor. Using the median value attempts to avoid the calculation being influenced by highly expressed differential genes.

However, single-cell data can violate the assumptions necessary for the DESeq size factor normalisation to work, as analysis of heterogeneous populations can mean that a large number of genes are differentially expressed. Additionally, scRNA-seq measurements are also zero-inflated due to so-called dropouts in gene expression (where a gene is in fact expressed at a low level but not detected) and any zero value measured for a gene precludes its scaling by a geometric mean, as this will be zero. Therefore it can be the case that only a handful of genes are actually used in the DESeq size-factor calculation for a single-cell dataset. To combat this, Lun and colleagues suggested a strategy using cell pooling to normalise single-cell data (Lun et al., 2016a). Here, size factors are calculated for pools of cells, and then deconvolution techniques are applied to recover size factors for individual cells. The authors demonstrate that their method improves performance compared to approaches including DESeq, particularly on datasets containing cells with very heterogeneous expression.

A strategy suggested for quantifying the variation in material sequenced for different RNA-seq samples was the use of synthetic spike-ins, such as the set of 92 molecules from the External RNA Controls Consortium (ERCC) (Jiang et al., 2011). These synthetic transcripts are added in the same quantities to each individual sample in an experiment, and so can be used to give an indication in the variance due to technical sources. The expression of spike-in genes is used as an input for the BASiCS method, which models gene counts as Poisson variables and estimates size factors as model parameters based on the expression data (Vallejos et al., 2015). This approach provided accurate quantification on simulated datasets, but due to its complexity the method is limited by the number of cells it can be applied to.

When the transcriptome is measured using scRNA-seq, tens of thousands of genes can be detected across a dataset. However, many of these will only be detected at very low levers, and their variation across the data will mainly be driven by technical, rather than biological, factors. In particular, genes with very low average expression can exhibit the largest variances. To identify genes with variation exceeding a technical level (known as "highly variable genes"), Brennecke and colleagues quantified the relationship between variance and the expression level of technical spike-ins, and used this to identify genes with variation exceeding this threshold (Brennecke et al., 2013). The BASiCS method also allows highly variable genes to be determined, again using spike-in expression (Vallejos et al., 2015). When spike-in genes are not used in the generation of a dataset, highly variable genes can instead be identified as those with extreme variance values in comparison to other genes with similar expression levels (Macosko et al., 2015).

1.3 Resolving heterogeneous populations

Single-cell expression profiling results in complex datasets, measuring the expression of tens to thousands of genes in often hundreds or thousands of cells. Often these cells represent a mixture of heterogeneous populations. This raises computational challenges of how to use the data to understand biological structure within the system.

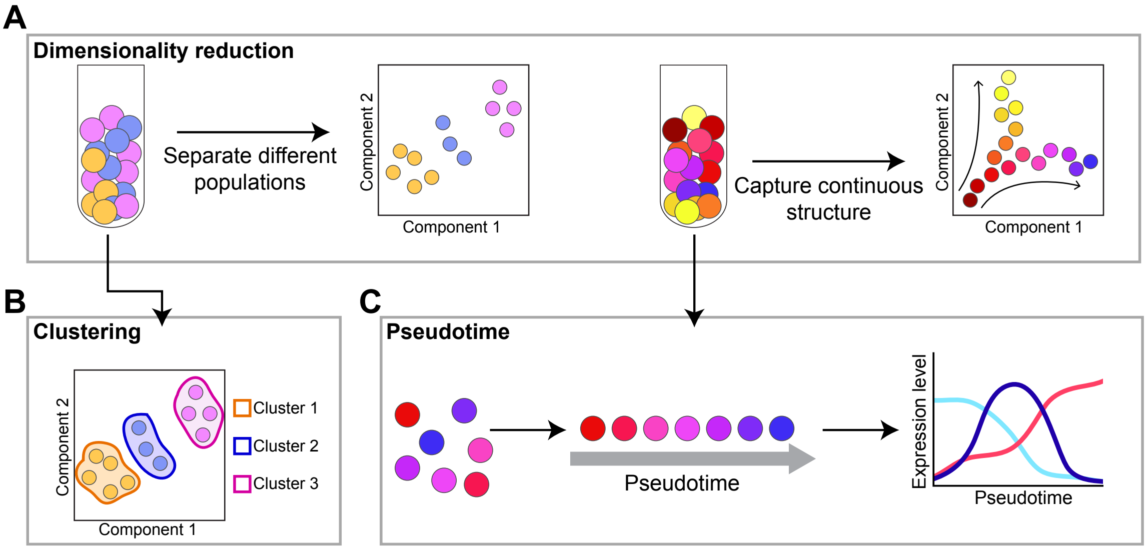


Fig. 1.3. Computational techniques for analysing single-cell expression data. (A) Dimensionality reduction techniques visualise high-dimensional data in a low-dimensional space. This can be useful for separating different groups of cells within the dataset (left panel) or for visualising continuous structure within a dataset related to processes such as differentiation (right panel). (B) Unbiased clustering techniques can be applied to single-cell data to explore similarities between cells and assign cells to different groups. (C) Single-cell expression profiles can be ordered to reconstruct lineage differentiation based on the assumption that the cells closest in the differentiation process will have the most similar molecular profiles. A population containing cells at different stages of maturity can be arranged into this ordering, known as pseudotime. This then allows properties such as changes in the expression of genes or proteins to be investigated along the differentiation trajectory.

1.3.1 Dimensionality reduction

The molecular profiles generated using single-cell technologies such as mass cytometry, qRT-PCR and scRNA-seq form datasets with high numbers of dimensions. Direct interpretation of such data is far from straightforward. Making sense of multidimensional data is not a challenge unique to single-cell biology: a concept from machine-learning and statistics, known as dimensionality reduction, has been widely applied to population expression data to discover similarities and differences between samples from different cell types or conditions. Dimensionality reduction methods enable complex high-dimensional data to be embedded in a low-dimensional space (most frequently two or three dimensions) allowing the differences between groups of cells to be visualised (Fig. 1.3A). Dimensionality reduction methods that have been applied to single-cell data include principal component analysis (PCA) and independent component analysis (ICA) (Pina et al., 2015; Trapnell et al., 2014). PCA applies a linear transformation to the data to calculate positions for each observation in a new coordinate system where the new axes (principal components or PCs) are orthogonal. The PCs are identified so that each component explains the maximum variance within the data. These are ordered so that PC1 has the largest variance, followed by PC2 with the next greatest, meaning that by plotting the data in the first two or three components it is often possible to see separation of different cell states, as this should drive a large part of the variance in the dataset. PCA has been used to help interpret single-cell expression data in numerous studies, for example by revealing differences due to ageing and differentiation in HSC populations profiled using scRNA-seq (Kowalczyk et al., 2015). ICA is another method which applies a linear transformation to the data, but instead of maximising the variance of each component, ICA treats the observed data as a combination of independent signals from several sources, and searches for components that aim to represent these source signals. When ICA is applied to single-cell data, it attempts to find a representation where each component captures a different signal that causes variation within the dataset, which can prove useful in representing structure within the data related to differentiation (Macaulay et al., 2016; Trapnell et al., 2014).

However, these linear techniques can struggle with capturing more complex structures. This means that they do not always provide the most suitable visualisation of expression profiles generated using single-cell technologies, for example data originating from multiple lineages in a differentiating tissue. A widely-used and powerful approach allowing visualisation of highly heterogeneous data is t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton, 2008). t-SNE embeds the data in a low-dimensional space by iteratively searching for a distribution of distances between objects in a new coordinate system that matches the distribution of pairwise distances in the high-dimensional space. This results in the cells with similar expression profiles, and therefore the shortest high-dimensional distances, being positioned close together in the embedding. Many different studies have applied t-SNE to single-cell expression data, with diverse aims ranging from visualising the overlap between different haematopoietic populations (Wilson et al., 2015), to demonstrating how data profiling human haematopoietic cells from different tissues can be integrated (Zheng et al., 2018), to understanding the complex structure present within the developing mouse embryo (Ibarra-Soria et al., 2018; Scialdone et al., 2016). Amir et al. (2013) developed the viSNE algorithm, which is based on t-SNE, specifically for visualising mass cytometry data in order to explore heterogeneities within leukaemic bone marrow. Aided by this visualisation, the authors observed phenotypic differences between wild-type (WT) and cancerous bone marrow samples, and were able to detect the rare phenotype of minimal residual disease, which is linked to relapse in patients who are in remission. Along with those highlighted here, many other studies have demonstrated that t-SNE is a powerful tool for representing data that profile a heterogeneous mixture of cell types. This is also supported by the fact the t-SNE is incorporated into many single-cell analysis pipelines (Butler et al., 2018; Lun et al., 2016b; Wolf et al., 2018).

One of the challenging aspects of studying differentiation in stem cell systems arises from the difficulties in developing strategies for isolating cells at different stages of the differentiation process. Single-cell profiling can be used to capture molecular changes throughout differentiation. To visualise such datasets, it is often useful to apply a dimensionality reduction technique that is suited to representing the continuous nature of differentiation (Fig. 1.3A). Work from Fabian Theis' lab demonstrated that diffusion maps (Coifman et al., 2005) could be adapted for use with single-cell expression data (Haghverdi et al., 2015). This method calculates distances between cells in a low-dimensional space based on the lengths of diffusion-like random walks in the high-dimensional data. A recent study by Moignard et al. (2015) used single-cell qRT-PCR profiling and diffusion maps to visualise the differentiation of cells during early blood development. Their analysis showed that this dimensionality reduction technique could successfully arrange cells in a low-dimensional structure recapitulating the progression from earlier to later time points during embryonic blood development.

Another method which has proved to be very powerful at representing complex data structures is a visualisation technique known as force-directed graphs. To calculate the force-directed graph, cells are first connected based on similarities in their expression profiles, with connections forming the edges in a graph. Edges can be weighted by the strength of the similarity between cells, or be limited to only the strongest connections for each cell. The embedding is then generated by edges in the graph causing an attracting force between cells, which is balanced by a repelling force between the cells to find an arrangement in two dimensions. Due to their scalability to large datasets, this technique has previously been applied to mass cytometry data to visualise connections between different haematopoietic cell types (Levine et al., 2015; Spitzer et al., 2015). Force-directed graphs have also proved adept at representing the branching differentiation structure in scRNA-seq datasets, for example capturing the branching hierarchy of the HSPC compartment (Giladi et al., 2018; Tusi et al., 2018; Zheng et al., 2018). Implementations of force-directed graph algorithms are available in several software programs, including an interactive tool from the Klein lab for applying force-directed graphs to single-cell gene expression data, which allows exploration of gene expression patterns in different regions of the graph (Weinreb et al., 2018a). Another recent algorithm that has been proposed for the visualisation of single-cell expression data is uniform manifold approximation and projection (UMAP) (McInnes and Healy, 2018), which is computationally efficient on large datasets, and has already been applied to provide insightful representations of single-cell data (Wang et al., 2018).

1.3.2 Clustering single-cell profiles

Whilst dimensionality reduction allows cellular heterogeneity to be visualised, it can often also be useful to assign cells to discrete groups, as this allows different cell states present within a sample to be compared. Assigning cells to subpopulations by relying on prior knowledge is not always possible, due to lack of marker genes, or even desirable, as it may miss unexpected or unannotated cell types. Instead, clustering methods can be used to separate cells into groups in an unbiased way, based on information such as their molecular profiles (Fig. 1.3B). The expression of specific genes within each cluster can then be used to identify known cell types or find novel marker genes for the different populations. Well-established clustering methods such as hierarchical clustering have been extensively applied to single-cell data to split samples into groups (Drissen et al., 2016; Grover et al., 2016; Guo et al., 2013a; Moignard et al., 2013; Wilson et al., 2015). These methods work by calculating distance measurements between cells, which represent the similarities in their gene expression profiles. Cells are then assigned to clusters based on their proximities in expression space.

There are also clustering methods that have been developed specifically for partitioning single-cell profiles, such as the approach by Jaitin et al. (2014), who applied a probabilistic mixture model to scRNA-seq data in order to organise cells into groups with distinct gene expression profiles. This approach was applied to scRNA-seq data characterising haematopoietic progenitors from the classical CMP, granuolocyte-macrophage progenitor (GMP) and megakaryocyte-erythroid progenitor (MEP) populations (Paul et al., 2015). By clustering cells based on their transcriptional profiles, the authors were able to assign the majority of cells to either the erythroid-megakaryocyte or the granulocyte-monocyte lineages. Their analysis questioned the existence of common progenitors within the CMP gate, instead arguing that these cells were in fact already committed to one of the two lineages. This study was in agreement with the work measuring the functional output of CMPs using genetic barcoding (Perié et al., 2015).

The outcome of clustering algorithms can be influenced by the gene set on which they are calculated. Many researchers opt for identifying a set of highly variable genes on which to perform this type of downstream analysis, as discussed in Section 1.2.3. Another approach developed for use on scRNA-seq data aimed to find the optimal gene set by using an iterative process to select “guide genes” for the clustering (Olsson et al., 2016). By using the genes with expression patterns corresponding to an initial set of clusters as guides to define a gene set for re-clustering over several iterations, the authors were able to discover clusters corresponding to multiple haematopoietic lineages in expression profiles from mouse bone marrow HSPCs (Olsson et al., 2016). This was used to dig deeper into the specification towards granulocyte and monocyte lineages, and to understand how loss of the genes *Irf8* and *Gfi1* altered this cell fate decision.

Graph-based clustering has proved a powerful tool in identifying different cell states in the large single-cell protein expression datasets generated by mass cytometry, as this approach is easily scaled for use with large numbers of cells and high-dimensional data. Here, a graph is constructed where each cell is a node and edges between nodes represent the similarities in expression profiles of the corresponding cells. PhenoGraph is one such unbiased algorithm that searches for highly connected groups of nodes (cells) within the graph to identify clusters of cell types (Levine et al., 2015). Levine et al. (2015) used this method to investigate intratumour heterogeneity in acute myeloid leukaemia (AML) by obtaining single-cell mass cytometry data to measure protein expression and activation in samples from AML patients and healthy bone marrow donors. Application of PhenoGraph revealed differences in the distribution of cell types in the bone marrow between AML and healthy samples. Graph-based clustering has also been applied to single-cell gene expression data from haematopoiesis to group similar expression profiles from very sparse data where shallow sequencing results in a high dropout rate across cells. Pooling cell profiles using graph-based clustering can reveal changes in gene expression that would be obscured when only considering the sparse data (Giladi et al., 2018).

1.4 Reconstructing differentiation from snapshot data

During haematopoiesis, cells become increasingly specialised as they commit to fates corresponding to the different blood cell types. Isolating and collecting populations at different stages of differentiation, followed by bulk profiling using techniques such as RNA-seq, goes some way to describing changes occurring during cell differentiation, but is limited by the time resolution of the data collected, and must assume that cells are synchronised throughout the differentiation process. As discussed in Section 1.2.1, there is actually a great deal of heterogeneity within the classical HSPC populations, with studies for example identifying evidence of functional biases within the stem cell compartment (Dykstra et al., 2007; Grover et al., 2016) and emphasising the diverse nature of CMPs (Paul et al., 2015; Perié et al., 2015). Single-cell technologies also allow unbiased profiling of systems, such as bone marrow or tumour tissues, which contain cells at multiple stages of differentiation. Researchers realised that by isolating these cells and performing single-cell molecular profiling it was possible to reconstruct the differentiation process *in silico*, thereby providing insights into how features such as gene and protein expression are altered as cells mature (Fig. 1.3C). Early work recognised that dimensionality reduction techniques, such as PCA, could embed cells in a low-dimensional space where coordinates in one component approximately correspond to differentiation, and that these coordinates could be used as a rough measure for the differentiation stage of a cell (Guo et al., 2010). However, dimensionality reduction often does not neatly capture a process such as differentiation in a single component, and in particular will struggle with situations such as branching towards multiple lineages. This led to a range of different techniques for attempting to reconstruct differentiation from single-cell data, which will be discussed in the following text.

1.4.1 Constructing a lineage tree from single-cell data

There are a wide range of approaches available for inferring orderings of single cells that aim to arrange cells into a sequence that represents their progress through a process such as differentiation. Fundamentally, these all work using the same concept: the cells closest together in terms of the biological process will have the most similar expression profiles, and therefore be closest together in the expression space. Even in a system as well-characterised as haematopoiesis, the exact structure of the haematopoietic tree remains under debate (Adolfsson et al., 2005; Guo et al., 2013a; Notta et al., 2016), and several studies have used single-cell expression profiling as a tool to address this question. By measuring their expression state, individual differentiating cells can be clustered into groups and the closest groups connected into a structure representing a lineage hierarchy. The spanning-tree progression analysis of density-normalised events (SPADE) algorithm uses this approach to build lineage hierarchies from flow and mass cytometry data collected from bone marrow cells (Qiu et al., 2011). This method first calculates a density-dependent sample of the data to ensure that rare populations are not obscured. Cells in this sample are then clustered based on their expression profiles, and the most similar clusters are linked into a tree (a graph where the edges form no closed loops) with the aim of representing the lineage hierarchy. SPADE's strengths lie in its inclusion of rare populations in the hierarchy and the fact that is does not require prior information to infer the lineage structure. However, different random density-dependent samples obtained by SPADE lead to different clusters, and can therefore produce alternative tree structures, meaning there is a limitation to the stability of this approach.

The SPADE algorithm was used by Guo et al. (2013a) to construct a lineage tree resembling the haematopoietic differentiation hierarchy, with the aim of investigating the much debated question around the starting point of lineage commitment for HSCs. Recent studies show that, in contrast to the initial beliefs of the field, this likely occurs before the split into CLPs and CMPs (Notta et al., 2016; Paul et al., 2015; Perié et al., 2015). Guo et al. (2013a) also challenged the view that commitment occurs at the CMP stage, collecting more than 1,500 single cells for qRT-PCR profiling and quantifying the expression of 280 genes for commonly-used surface markers across all of these cells. The lineage tree constructed with SPADE showed that CMPs were found in both the megakaryocyte-erythrocyte and the lympho-myeloid lineages. Computational analysis followed by *in vitro* validation identified a new surface marker (CD55) that was able to separate the megakaryocyte-erythrocyte potential from CMP and MPP populations. The authors also observed that the megakaryocyte-erythrocyte cells were closely connected to the LT-HSC branch, and used *in vitro* tracing experiments to show that the megakaryocyte colonies emerged first in HSC cultures, supporting the close connection between these cell types seen in the *in silico* analysis. This study demonstrated how computationally constructing differentiation hierarchies can provide insight into biological systems such as haematopoiesis.

In a more recent study, Spitzer et al. (2015) described a computational method, Scaffold, to arrange immune cells profiled by single-cell mass cytometry into a "reference map" of the murine immune system. This approach involves an initial clustering step followed by the construction of a graph using the clustered cells. Scaffold uses the method of force-directed graphs (discussed in Section 1.3.1) to find a visualisation based on the similarity between cell types. Here, similar clusters are pulled close together in the force-directed graph, whereas dissimilar clusters lack a strong attracting force and consequently lie far apart. The resulting graph links cells in a structure representing the immune system hierarchy. The authors constructed Scaffold maps for cells from several different samples, which enabled comparison of immune system organisation in different tissues, genetic backgrounds and species. Using the Scaffold map revealed that circadian rhythm affected the distribution of immune cells, with some of the immune cell populations fluctuating depending on the time of day.

1.4.2 Ordering cells in pseudotime

An exciting extension of inferring differentiation hierarchies, which link together groups of cells, is to order each individual profile by progress through differentiation. Assuming that gene and protein expression change continuously as cells differentiate, and that a sample contains cells spread at a sufficient density throughout differentiation to cover enough of the process, it was hypothesised that single-cell expression profiles could be arranged in “pseudotime”, where the position of a cell in pseudotime corresponds to its progress through differentiation (Fig. 1.3C). Based on these simple assumptions, several different algorithms have been designed to solve this computational ordering problem and reconstruct differentiation trajectories. Trapnell et al. (2014) designed the Monocle algorithm, which first performs a dimensionality reduction on the data before constructing a graph on cells with edges weighted by the distances in this low-dimensional representation. The minimum spanning tree on the graph (the tree connecting all cells with the lowest total edge weight) is found and cells are ordered in pseudotime based on their position in this tree, allowing changes in gene expression patterns to be investigated. Using this approach, the authors reconstructed the differentiation of human primary myoblasts, and were also able to identify a branching process towards an alternative cell fate present in their data.

Around the same time another algorithm, Wanderlust, was also developed and applied to single-cell mass cytometry data to capture B cell development in human bone marrow (Bendall et al., 2014). Wanderlust constructs a pseudotime ordering by first considering a k-nearest-neighbour graph on the single-cell expression data (each cell is connected to the k most similar other cells, where k is an integer parameter that can be chosen by the user). The ordering of cells is based on the length of paths through this graph originating from a user-defined starting cell, which can be identified either based on experimental metadata or the expression of marker genes. Wanderlust can cope with very large numbers of cells, and uses subsampling methods to obtain a stable ordering, avoiding the possibility of “short circuits” through the data (a route missing out part of the differentiation trajectory). Distances through the graph are calculated based on the distance of a cell from randomly chosen waypoints, another strategy aiming to increase stability, and are calculated in an iterative process. Bendall et al. (2014) used mass cytometry to study 44 parameters across

B-cell lymphopoiesis, collecting cells across B cell development in order to reconstruct a developmental trajectory. Using Wanderlust, the authors confirmed that all of the landmarks of B cell lymphopoiesis were correctly ordered along their inferred trajectory. This ordering revealed that the expression of CD24 and terminal deoxynucleotidyl transferase increased prior to an increase in expression of the canonical B cell markers, and early B cell progenitors were subsequently identified using the expression of these markers. This study also provided additional insights such as observing that changes in signalling occurred in parts of the trajectory corresponding to developmental checkpoints, an observation made possible because the dynamics of different signalling molecules had been reconstructed across differentiation using the concept of pseudotime.

Despite the exciting potential of pseudotime analysis, both the Monocle and Wanderlust algorithms suffer from limitations. The use of the minimum spanning tree in Monocle was unstable and susceptible to short circuits through the data, and Wanderlust could not be used to identify branches where a differentiation trajectory separates towards multiple lineages. Many alternative algorithms for constructing pseudotime orderings have since been suggested. To provide a more robust method that could deal with branching towards different fates, Haghverdi et al. (2016) developed diffusion pseudotime (DPT), which built upon the ideas of diffusion maps to order single-cell profiles. DPT is a distance measure between cells based on the length of random walks through the single-cell expression space, and like diffusion maps is robust when applied to noisy data. Branch points are identified by considering the correlation between orderings from the start and end of the main trajectory. The authors applied DPT to single-cell qRT-PCR of cells collected during mouse developmental haematopoiesis (Moignard et al., 2015) to reconstruct the expression changes of genes throughout this process.

There have also been updates to both the Monocle and Wanderlust algorithms, to overcome some of their earlier limitations. In 2016, the authors of the Wanderlust paper introduced Wishbone, which improves the selection of waypoints compared to the original algorithm to avoid potential problems arising due to the presence of outliers within the data, and also uses these waypoints to identify cells lying a on branch towards an alternative cell fate (Setty et al., 2016). In their study, the authors were able to use Wishbone to correctly identify a branch point in T cell differentiation. The updated version of Monocle, Monocle 2, uses reverse graph embedding (Mao et al., 2017) to learn a low-dimensional representation of the data that reproduces their structure, including any branches (Qiu et al., 2011). The number of branches does not need to be predetermined by the user, which is an advantage of this approach. Cells are then projected onto the low dimensional representation to order them through pseudotime.

1.4.3 Inferring the fate of individual cells

Although assigning cells to discrete branches is in keeping with the classical model of the haematopoietic hierarchy, there have been studies suggesting that haematopoiesis occurs as a continuous, rather than a stepwise, process (Laurenti and Göttgens, 2018). A recent study by Velten and colleagues profiled human bone marrow HSPCs using scRNA-seq (Velten et al., 2017). As there is extensive pre-existing knowledge about the start and end points of individual trajectories from HSCs to mature cell types in this system, the authors chose not to apply one of the existing pseudotime algorithms and instead developed a new approach that did not assume a branching structure. This led to the innovation of the STEMNET algorithm, which uses marker genes that are specifically expressed in the different blood lineages to build a classifier and score the profile of each cell for both its lineage bias and the strength of its commitment. Assessment of these scores across the whole dataset enables cells possessing the potential for multiple fates to be identified, and also the precise combinations of multi- or bipotent fates that exist to be characterised. In human bone marrow, the combined results of *in vitro* culture and xenotransplantation assays along with the STEMNET predictions challenged the classical step-by-step branching mode of haematopoiesis. Strikingly, the authors showed that there was a lack of well-defined hierarchically organised discrete progenitor populations, with the majority of cells in fact appearing to be either multipotent, or committed to just a single fate.

Work from the Grün lab also decided to focus on predicting the fate of individual cells using scRNA-seq data with their algorithm, FateID (Herman et al., 2018). Similar to STEMNET, FateID applies a classification approach to single-cell data by considering the expression states of terminal cells in the trajectories, but FateID instead adopts an iterative process for inferring cell fates by working “backwards” along the trajectories, updating the cell states used to build the classifier at each step. The authors applied their method to scRNA-seq data of mouse HSPCs to investigate lineage priming within the haematopoietic progenitor populations, and could classify cells primed towards several blood cell fates. When comparing with the STEMNET algorithm, they found that FateID could resolve lineage priming in earlier progenitor populations. The authors attributed this to the iterative classification process of their approach, which in contrast to STEMNET does not rely only on the expression of the more mature marker genes.

Another approach has been suggested by Weinreb and colleagues, who developed an algorithm called population balance analysis (PBA), which uses the concept of flux conservation in a homeostatic system to infer cell fates (Weinreb et al., 2018b). By considering biased random walks through the single-cell data, PBA solves equations modelling the differentiation dynamics in order to infer a temporal ordering and assign fate probabilities to each cell. The authors apply this algorithm to InDrop scRNA-seq data of mouse HSPCs to assign cells towards seven different fates (Tusi et al., 2018; Weinreb et al., 2018b), uncovering evidence for a shared basophil (or mast cell) and erythroid progenitor. However, PBA does have limitations due to the prior knowledge that it is necessary to provide as input to the algorithm, with users required to input information about the rates of cell entry to and exit from the system corresponding to the different lineages.

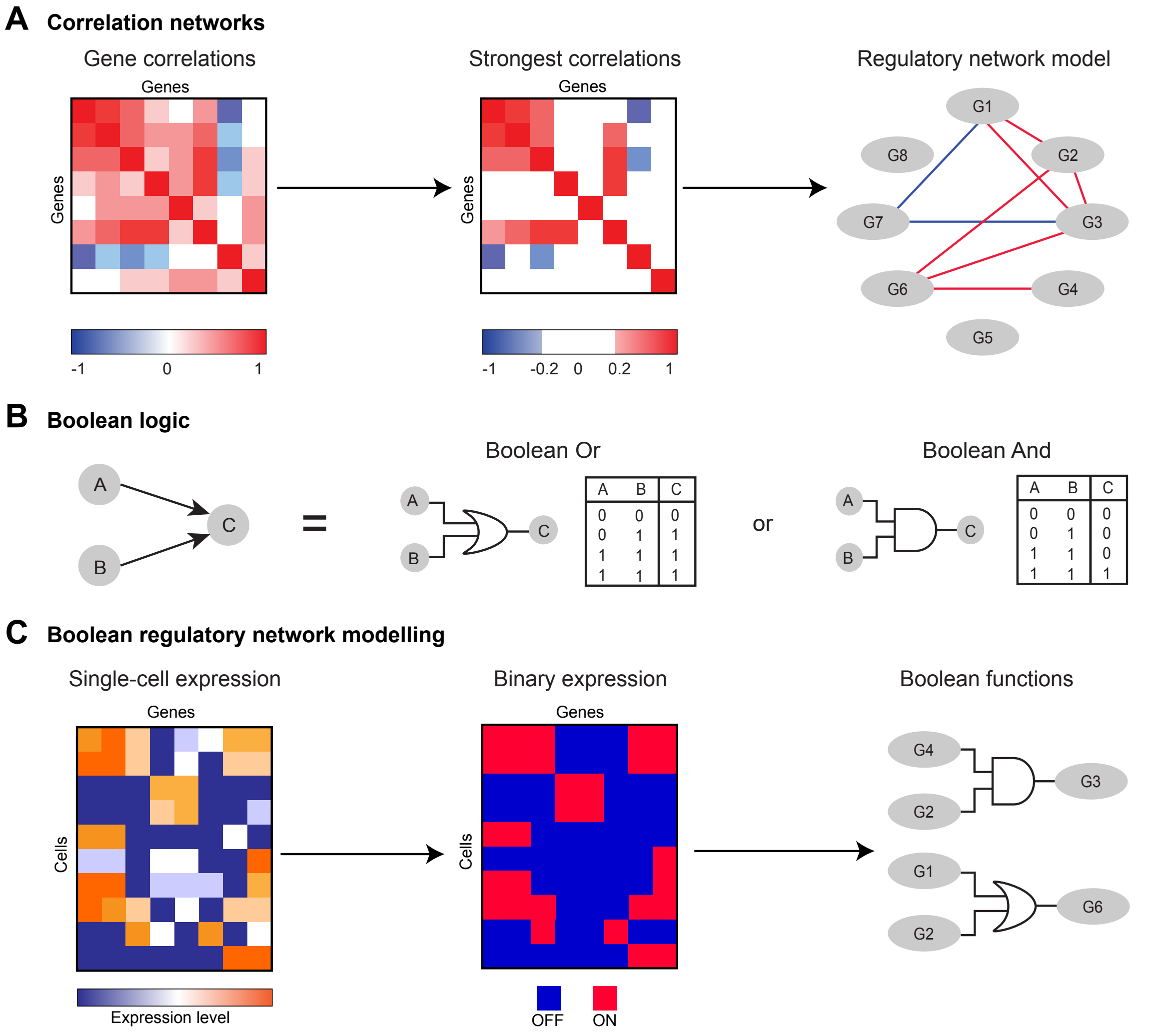
1.5 Identifying and modelling regulatory relationships

To better understand how multipotent cells in a system such as haematopoiesis choose between different fates it is important to define the underlying regulatory programmes governing their cell fate decisions (Göttgens, 2015; Peter and Davidson, 2015). One of the ways in which cell fate decisions are controlled is through transcription factor proteins binding to specific regions of DNA in order to regulate gene expression. Although many transcription factors have been identified as key haematopoietic regulators, they do not act in isolation but, along with the cis-regulatory modules that they bind to, form part of transcriptional regulatory networks. Identifying the interactions in these networks can help to explain how regulatory programmes control differentiation. However, network reconstruction directly from experimental evidence has largely been limited to the simplest forms of life due to the sheer number of possible regulations and the complicated network structures present in complex organisms. Instead, many studies have focused on the more feasible approach of inferring regulatory networks from gene expression data, which requires data to be collected from multiple experimental perturbations or conditions to establish interaction between genes. Network inference from bulk expression data is therefore constrained by both small sample size and also the fact that it masks heterogeneity within cell types. Single-cell data represent a powerful alternative for identifying new regulatory relationships, as each cell provides an observation with its own expression levels, meaning that the number of samples is vastly increased.

1.5.1 Identifying regulatory relationships

Measuring single-cell gene expression provides potentially thousands of observations of the levels of a gene across individual cells. Such large sample sizes can be used to identify potential regulatory relationships by considering the strength of the correlations between genes (Fig. 1.4A). Setting a threshold on correlation can then be used to construct putative networks consisting of connections between the genes with high correlations. A number of studies have taken this approach to calculate correlation between genes using single-cell expression data, and have been able to identify and experimentally validate regulatory relationships between highly correlated genes (Moignard et al., 2013; Pina et al., 2015). These studies take advantage of the statistical power that comes with having numerous observations of gene expression measurements to search for genes with very similar (or in the case of negative regulation very dissimilar) patterns, as such patterns would be expected of regulatory pairs. Moignard et al. (2013) used single-cell qRT-PCR to measure 18 transcription factors that were well-known to be involved in the regulation of haematopoiesis. By profiling several populations from the HSPC compartment, the authors calculated correlations between the transcription factors to reveal two new regulatory links. In the correlation analysis, strong positive correlation was seen between *Gata2* and *Gfi1b*, and *Gata2* and *Gfi1*. Experimental work established that this represented two previously unknown regulatory relationships, with Gata2 activating both Gfi1 and Gfi1b. Another studying also applying correlation techniques to single-cell qRT-PCR data was able to identify *Ddit3* as a key player in the commitment between erythroid and myeloid lineages (Pina et al., 2015).

As well as work considering the correlation of genes in single-cell data, other studies have explored different statistical techniques for inferring regulatory relationships. An investigation into a signalling network in T cells used mass cytometry data to develop an algorithm, DREMI. This calculates the interaction strength between pairs of proteins based on a quantity known as mutual information, which is a measure of the dependence between two variables (Krishnaswamy et al., 2014). Applying DREMI to the mass cytometry measurements was able to identify differences in signalling interactions between different experimental conditions and provided insight into how the signalling network evolves over time after a stimulation.



A	B	C
0	0	0
0	1	1
1	1	1
1	1	1

A	B	C
0	0	0
0	1	0
1	1	0
1	1	1

Fig. 1.4. Inferring regulatory relationships from single-cell expression data. (A) Quantities such as correlation between gene pairs can be calculated using single-cell gene expression measurements. As shown in this gene–gene correlation heatmap, some pairs will exhibit positive correlation and some pairs negative correlation, suggestive of activating or repressing regulatory relationships, respectively. Thresholds can be chosen to select the most strongly correlating gene pairs. (B) If A and B both activate gene C this could correspond to two different scenarios, which are represented here using Boolean logic functions. It could be that either A or B alone will cause activation of C, as shown on the left with the Boolean Or function. Alternatively, it could be that binding of both A and B is required for activation of C, as shown by the And gate and truth table. (C) Regulatory networks can be modelled using Boolean functions. Gene expression measurements for single cells can be converted into binary (ON/OFF) expression by choosing a threshold. Computational methods applied to these binary data allow inference of regulatory relationships, represented here by Boolean And/Or functions.

There are also many methods designed for use on bulk expression data that use different techniques for inferring regulatory relationships, such as GENIE3 (Huynh-Thu et al., 2010), which applies a regression problem to identify the most likely regulators of each gene from the expression data, or relevance networks and the ARACNE algorithm, which use mutual information to identify gene regulatory networks (Butte and Kohane, 2000; Margolin et al., 2006).

1.5.2 Modelling regulatory networks

Identifying the regulatory relationships controlling cell fate decisions can provide important insight into how a system works. But, in the case of a system such as haematopoiesis, one of the reasons we are interested in this regulation is that when it fails it can lead to outcomes such as leukaemia. So in addition to identifying the nature of HSPC regulatory networks, an exciting objective is to construct executable models of the system. These are models where expression states can be simulated so that the outcome of network perturbations can be predicted (Fisher and Henzinger, 2007). Often, literature curation is used as a starting point for constructing these models, where previously established regulatory relationships from the literature are encoded in a framework allowing simulation of cell states. For example, Bonzanni et al. (2013) used literature curation to model a blood stem cell regulatory network. By considering the properties of this network, they found that initially the model did not readily allow differentiation towards the erythroid fate. By searching for additional interactions that would make this differentiation step easier, the researchers identified, and then experimentally validated, repression of Fli1 by Gata1. This demonstrates how executable network modelling can be used identify novel regulation from a starting point of literature curation.

Another approach that can be used to model regulation is Bayesian network modelling, which is a computationally efficient method of network inference that allows perturbations to be simulated. Schütte et al. (2016) applied dynamic Bayesian network modelling to model regulatory relationships between transcription factors, where regulation was identified by transcription factor binding data from HSPCs. This model could be simulated to investigate the effect of knock-down or overexpression of different genes. Differential equation-based modelling can also be used to capture the behaviour of gene regulatory networks, where the interaction between groups of genes is described as a system of differential equations that capture activating or repressing regulation. Such an approach was applied to single-cell data by Ocone et al. (2015), who used the GENIE3 algorithm to identify putative edges in a network model. The edges were then encoded as regulatory relationships using ordinary differential equations, followed by model selection to identify the best fit to a pseudotime ordering constructed from the single-cell data in order to choose the most suitable model. The authors applied this to single-cell qRT-PCR data from Moignard et al. (2013), and were able to reconstruct the regulatory Gata2-Gfi1-Gfi1b triad that had been found and validated by the authors of the original study. However, this type of model is limited to relatively small networks of genes due to the computational requirements of the method.

Of course, regulatory relationships are not always as simple as direct activation from one gene to another. Instead, transcription factors can be involved in combinatorial binding, where the presence of multiple proteins is required to influence the expression of a gene. To capture the logical nature of such relationships, interactions can be abstracted as Boolean functions where expression of a gene is either “on” or “off”, with these functions forming part of a Boolean network (Fig. 1.4B). With this type of abstraction it becomes possible to model and simulate regulatory networks efficiently. Models can be characterised by examining the nature of their attractor states (expression states where gene expression remains stable under the regulatory rules), which can be calculated both under normal and perturbed conditions (Garg et al., 2008). Boolean network modelling has been used to encode literature-curated models in systems ranging from adult haematopoiesis to embryonic development in the sea urchin (Bonzanni et al., 2013; Peter et al., 2012). A Boolean framework has also been used in network inference, such as in the study by Dunn et al. (2014), who used bulk gene expression data to construct a network model of pluripotency in embryonic stem cells. Here, correlations were calculated between genes to describe an ensemble of possible network models, and this was constrained by perturbation data to find the models with the best fit.

Single-cell expression data also offer exciting potential in this area, as gene expression levels can be converted to binary data for each cell, describing a large number of possible Boolean states (Fig. 1.4C). It has been demonstrated that single-cell gene expression data can be used to computationally infer Boolean models in systems including embryonic blood development (Moignard et al., 2015) and embryonic stem cells (Xu et al., 2014). Moignard et al. (2015) used single-cell qRT-PCR data to construct a state transition graph across blood development, where each pair of cells with change in a single gene was linked in the graph. Regulatory networks encoding rules describing the transitions through this graph along the trajectories towards two cell fates were then identified by encoding the problem as Boolean satisfiability problem and searching for solutions (Woodhouse et al., 2015). In their work studying pluripotency in embryonic stem cells, Xu et al. (2014) used single-cell data to add Boolean logic to a regulatory network model constructed from literature curation. Single-cell expression profiles were treated as stable states of the network, and possible regulatory relationships supporting the stable expression of the single-cell profiles were found. Algorithms have also been suggested for refining existing Boolean network models based on the expression space represented by single-cell expression data (Lim et al., 2016). These studies represent a few examples of how Boolean network modelling can be applied to a variety of biological systems, and how it represents a powerful tool for understanding regulatory networks due to the ease with which it allows perturbations to be simulated. However, a drawback of Boolean modelling is the abstraction of gene expression levels to binary on/off states, which discounts any possible influence of quantitative expression differences. To try and address this problem, extensions to the Boolean modelling framework have been suggested that allow more than two discrete expression levels for each gene to be considered (Collombet et al., 2017).

1.6 Understanding shifts in the transcriptional landscape

As well as using single-cell data to reveal heterogeneity and discover regulatory relationships, profiling the expression states of individual cells can reveal how the transcriptional landscape of a system changes in response to a perturbation (Fig. 1.5A). In the context of haematopoiesis this is particularly relevant, as dramatic shifts in cell states can occur in diseases, like in the case of AML where a large expansion of GMP and lymphoid-primed multipotent progenitor (LMPP) cells occurs in many patients (Goardon et al., 2011). To understand how these alterations in cell fate decisions come about it is important to look at changes in cell state across the haematopoietic compartment, particularly as immature cells high up in the haematopoietic hierarchy can have the potential to initiate disease (Bonnet and Dick, 1997). Performing single-cell profiling can reveal changes in the cell type composition of samples from different conditions, be these due to disease, ageing or in response to a drug, and also allow the molecular differences in individual cells to be investigated (Fig. 1.5A).

1.6.1 Comparison with a reference dataset

Whilst single-cell analysis represents an exciting tool for the comparison between data from different conditions, it is often not a straightforward task to make these comparisons. Even

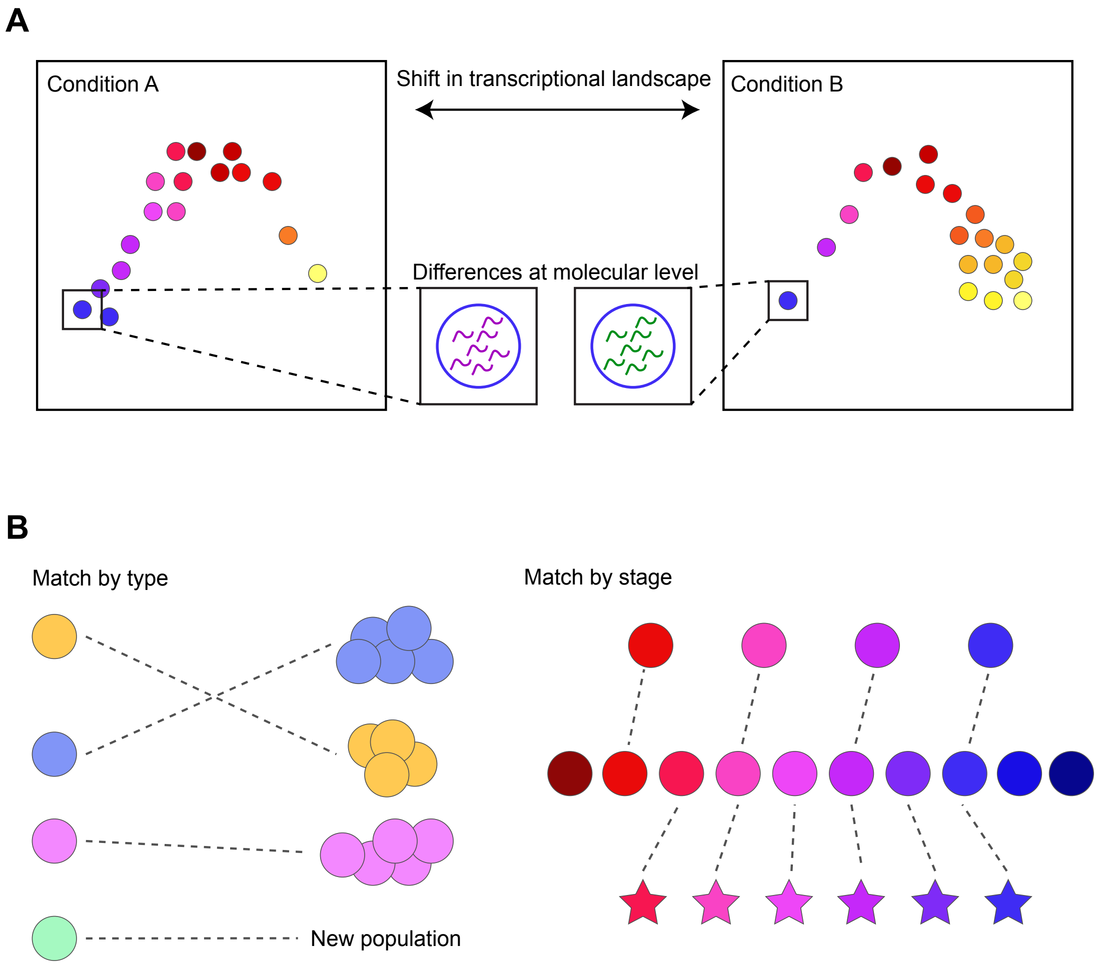


Fig. 1.5. Single-cell expression data can be used to contrast different conditions. (A) Population densities can shift in response to different conditions, such as disease. Such population shifts can be revealed by single-cell analysis, which then also allows comparison of the changes at the molecular level. (B) Computational methods can be used to match up data from multiple experiments or conditions. Left panel, new cells can be matched to groups in an existing reference data set. This can also be used to identify new cell types that do not exist in the reference population. Right panel, cells arranged in ordered sequences can be matched to identify overlapping stages, even when the items originate from different sources, such as different species.

data profiling exactly the same types of cells from the same system can exhibit differences in measured gene expression levels due to technical reasons if these cells are collected on different days, or by different people, or with different equipment. Biological effects such as stress due to disease or perturbation can also cause changes in cell state making the comparison between datasets challenging. Some studies have suggested an approach of comparing new or perturbed data to a reference sample to investigate how cells change across conditions (Fig. 1.5B, left panel). For example, in their analysis of Kit$^+$ mouse bone marrow representing steady-state haematopoiesis, Tusi et al. (2018) generated scRNA-seq data for bone marrow treated with EPO to stimulate erythroid differentiation. Their approach was to “map” the EPO-stimulated bone marrow cells to the reference landscape in order to compare differentially expressed genes along the erythroid differentiation trajectory. First, the EPO-stimulated cells were projected into a reduced dimension PCA space calculated on the untreated bone marrow cells, and then their nearest neighbours in this reference data found. Gene expression was then smoothed across the differentiation landscape so that the expression could be compared between the two conditions in the matched cells.

The concept of creating a “reference” map of a system was also explored in the mouse immune system, with the Scaffold map algorithm discussed in Section 1.3.1 (Spitzer et al., 2015). Here, the use of force-directed graphs allowed intuitive visualisation of a complex system, but also enabled the authors to relate cells from different tissues, mouse strains and even species to the reference cell types found in the mouse bone marrow. As the immune system represents a well-annotated system, the authors first used manual curation to identify cells corresponding to several major populations in the bone marrow. These main groups were used as the backbone for analysis of all of the other samples, with samples first being clustered in an unsupervised fashion into small clusters, and distances between the clusters and reference groups calculated based on similarities in their expression profiles measured on surface proteins and transcription factors using mass cytometry. Positions of the main nodes were calculated using force-directed graphs on the bone marrow data, and these positions were fixed and used as anchors for force-directed graphs calculated in the different datasets. This allowed the authors to visualise how the population distributions changed across the immune system compared to bone marrow in tissues such as the blood, liver and lungs, and also in different mouse strains.

Kiselev et al. (2018) present an approach for mapping cells from a new experiment onto an annotated reference dataset. Their algorithm, scmap-cluster, calculates distances in gene expression space to match cells from a new dataset to their most similar cluster in the reference data. Scmap first identifies a subset of features on which to perform the calculations, as not all of the genes contain relevant information but can instead introduce unwanted noise. Interestingly, the authors find that selecting genes with a higher than expected frequency of zero expression values produces more accurate mappings than selecting highly variable or random genes, an observation that may be useful for other types of scRNA-seq data analysis. Although the algorithm attempts to match cells to a reference set, the cells remain unassigned if they do not show gene expression patterns similar to those in the reference data. This is an essential consideration, as there will be incomplete overlap among the cell types present for many comparisons.

1.6.2 Combining and aligning datasets

Comparing cells to a reference transcriptome is a powerful tool for understanding how the expression landscape of a system changes across different conditions, and will be an important step in utilising the vast amounts of data from initiatives such as the Human Cell Atlas project (Regev et al., 2017; Rozenblatt-Rosen et al., 2017). However, it is sometimes desirable to integrate data from different sources, for example data that profile different mixtures of populations. For this, so-called "batch correction" tools are often necessary to stop technical differences between samples from causing the data from different sources to separate in the analysis. Batch correction is not a phenomenon unique to single-cell data, and methods such as ComBat were originally developed for adjusting the expression values of microarray expression data across experiments (Johnson et al., 2007). ComBat has also been applied to single-cell gene expression, and has been reported as being successful in removing batch effects between data from different days and sources (Büttner et al., 2017).

Methods have now also been developed specifically for use with single-cell data (Butler et al., 2018; Haghverdi et al., 2018). MNN correction, from Haghverdi et al. (2018), explores the concept of using mutual nearest neighbours between two datasets to perform batch correction. These represent pairs of cells across the two datasets that are the amongst the k-nearest neighbours of one another when the nearest neighbours of one dataset are found in the other dataset. If a cell type is present in both of the datasets then these cells should be mutual nearest neighbours of each other. The authors then use the differences in gene expression between mutual nearest neighbour groups to allow batch-corrected expression across the data to be calculated, and the two datasets to be combined. MNN correct was able to combine data from two datasets profiling mouse bone marrow haematopoietic populations, despite the incomplete overlap of cell types and the very different sequencing protocols used for generating these data.

Butler et al. (2018) also describe an algorithm which can be used for combining expression datasets from multiple sources. Here, the authors use a method called canonical correlation analysis to find common structures across the two datasets. This is used to embed each dataset in its own low-dimensional space where the structures match between the datasets.

An approach call dynamic time warping is then used to align the coordinate systems of the two datasets, to correct for "stretching" due to features such as differences in the density of populations. The algorithm can then embed data from different sources in a shared coordinate space, allowing the data to be visualised and clustered together.

Another related approach is relevant for data capturing a continuous differentiation process, to compare between pseudotime orderings on data and ask if, when, and how these processes diverge. Alpert et al. (2018) developed cellAlign, which uses dynamic time warping to align sections of two trajectories with shared expression patterns, thereby enabling the comparison of expression dynamics (Fig. 1.5B, right panel). Excitingly, not only is cellAlign able to compare whole transcriptomes to ask where trajectories diverge, but can also investigate specific genes or gene modules to assess differences between conditions. The authors analysed scRNA-seq data from preimplantation embryos to identify gene modules with different patterns of temporal behaviour across human and mouse development, thereby demonstrating the ability of their algorithm to contrast data from very different sources.

1.7 Aims of this work

To understand the regulation of cell fate decisions in haematopoiesis it is necessary to understand how the heterogeneity in HSPCs is linked to differentiation towards the different blood fates. The idea of this PhD project was to explore the concept of a "transcriptional landscape" of haematopoiesis—that is the gene expression space that cells can occupy during haematopoietic differentiation. This is related to the concept of the epigenetic landscape discussed in 1957 by Conrad Waddington (Waddington, 1957). Waddington described a theoretical potential landscape through which cells rolled from the multipotent hilltops down to the valleys representing more restricted cell fates. Advances in single-cell technologies allowing thousands of cells to be profiled simultaneously provide an opportunity to reconstruct the transcriptional landscape and characterise how cells move through it from HSCs to the different mature blood cell types. The work presented in this thesis uses computational methods to investigate haematopoiesis at the single-cell level with the following aims.

1. To profile the transcriptional landscape of the haematopoietic stem and progenitor cell compartment at a single-cell level

2. Identify regulatory programs controlling differentiation using single-cell expression data
3. To use the transcriptional landscape in order to learn how cells are guided on different routes through differentiation and understand how this control is altered in perturbed haematopoiesis

Chapter 2

Materials and methods

As significant parts of this thesis correspond to published work, sections of this chapter describing data generation and methods have been adapted from the following publications: Dahlin et al. (2018); Hamey et al. (2016); Karamitros et al. (2018); Nestorowa et al. (2016). Where any of the methods in a section were carried out by collaborators this is clearly written at the start of the section, as well as at the opening of the relevant chapter. Work was carried out by F. Hamey unless stated otherwise.

2.1 Cell isolation

Mouse bone marrow progenitors were isolated by Sonia Nestorowa, Nicola Wilson, and Mairi Shepherd. Details can be found at the start of Chapters 3, 4, 6 and 7. Human cord blood progenitors were isolated by Bilyana Stoilova, Dimitris Karamitros and Zahra Aboukhalil.

2.1.1 Isolation of mouse HSPCs for SMART-Seq2 scRNA-seq

Over two consecutive days the bone marrow of 10 female 12-week old C57BL/6 mice was harvested, and from it haematopoietic stem and progenitor cells were collected. On each day, cells from four mice were pooled together, and the cells from the remaining individual mouse were analysed separately. This was performed as a quality control measure to ensure that any populations identified in the analysis were not present in only a single animal. Bone marrow was lineage depleted using the EasySep Mouse Hematopoieitic Progenitor Cell Enrichment Kit (STEMCELL Technologies). Cells were sorted into three different gates: the LSK gate (Lin$^-$ c-Kit$^+$ Sca1$^+$), the Progenitor (Prog) gate (Lin$^-$ c-Kit$^+$ Sca1$^-$), and the LT-HSC gate (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flk2 $^-$ CD34$^-$). Full details of the antibodies used for isolating cells can be found in Nestorowa et al. (2016).

2.1.2 Isolation of mouse HSPCs for single-cell qRT-PCR

Bone marrow cells were isolated from the femurs, tibiae and iliac crest of 8-12 week old C57BL/6 mice and lineage depleted as described above. The gene expression profiles for LT-HSCs, cells from one of the ST-HSC sorting strategies, LMPPs, CMPs, MEPs and GMPs were obtained from the data previously published by Wilson et al. (2015). Additional ST-HSC cells were sorted as Lin$^-$ c-Kit$^+$ Sca1$^+$ IL7Ra$^-$ CD34$^+$ Flt3$^-$, MPPs as Lin$^-$ c-Kit$^+$ Sca1$^+$ IL7Ra$^-$ CD34$^+$ Flt3$^+$, and pre-megakaryocyte-erythroid progenitors (preMegEs) (Pronk et al., 2007) as Lin$^-$ c-Kit$^+$ Sca1$^-$ CD41$^-$ CD150$^+$ FcγR$^{\text{low}}$. Each cell type was sorted onto a separate plate for processing.

2.1.3 Isolation of mouse HSPCs for droplet-based scRNA-seq

Bone marrow cells from six 10 week old C57BL/6 mice were harvested from the crushed femora, tibiae and ilia, and pooled before sorting. Red blood cell lysis was performed, and the sample was lineage depleted using the EasySep Mouse Hematopoietic Progenitor Cell Enrichment Kit. Cells were then sorted into two different gates: the LSK gate (Lin$^-$ c-Kit$^+$ Sca1$^+$) and the LK gate (Lin$^-$ c-Kit$^+$). 3 LK samples and 3 LSK samples were sorted, all from the same pool of cells. In a later experiment, LK samples were similarly sorted from two W^{41}/W^{41} mice and processed separately using the same protocol.

2.1.4 Isolation of human progenitors

Fresh cord blood samples were processed 16-34 hours after collection. First, the mononuclear cells were separated and then CD34$^+$ cells were isolated from these samples using FACS. The sorting strategies for lympho-myeloid progenitor populations were as follows: LMPPs, Lin$^-$ CD34$^+$ CD38$^-$ CD90$^{\text{neg-lo}}$ CD45RA$^+$ CD10$^-$; multi-lymphoid progenitors (MLPs), Lin$^-$ CD34$^+$ CD38$^-$ CD90$^{\text{neg-lo}}$ CD45RA$^+$ CD10$^+$; GMPs, Lin$^-$ CD34$^+$ CD38$^+$ CD45RA$^+$

CD123^{+}. Full details of antibodies for isolating cells can be found in Karamitros et al. (2018).

2.2 Single-cell gene expression profiling

Single-cell gene expression profiling of mouse HSPCs was performed by Sonia Nestorowa and Nicola Wilson, and details can be found at the start of Chapters 3, 4, 6 and 7. Gene expression profiling of human cord blood progenitors was performed by Dimitris Karamitros and Bilyana Stoilova. Sequencing data were aligned and reads counted by Evangelia Diamanti and Rebecca Hannah.

2.2.1 SMART-Seq2 scRNA-seq profiling

scRNA-seq analysis was performed on both mouse bone marrow and human cord blood HSPCs using the SMART-Seq2 protocol as described previously (Picelli et al., 2014; Wilson et al., 2015) with the Illumina Nextera XT DNA preparation kit used for preparation of RNA-seq libraries. Synthetic ERCC spike-ins were added in equal concentrations to each well on the plate to allow quantification of technical variance across the samples (#4456740, Life Technologies). Pooled libraries of mouse HSPCs were initially sequenced using the Illumina HiSeq 2500 system and then resequenced using the Illumina HiSeq 4000 system with single-end 125 base pair reads. Pooled libraries of human progenitors were sequenced using the HiSeq 2000 using 75 base pair paired-end reads. Cells from two cord blood donors were sequenced separately. Reads were aligned using GSNAP (Wu and Nacu, 2010) and assigned to Ensembl genes (release 81) by using HTSeq (Anders et al., 2014). Aligned reads for the resequenced mouse data were combined using the SAM file output from GSNAP before counting using HTSeq. Mouse and human sequencing data were deposited in NCBI's GEO with accession numbers GSE81682 and GSE100618, respectively.

2.2.2 Droplet-based scRNA-seq profiling

Droplet-based scRNA-seq was performed using the 10x Chromium™ system (10x Genomics) (Zheng et al., 2017), with cells sorted and processed according to the manufacturer's protocol. The 6 WT samples were barcoded with a sample barcode and all sequenced across the same 8 lanes of an Illumina HiSeq 4000. W^{41}/W^{41} samples were barcoded and sequenced along with a number of samples for a separate study. Data were deposited in NCBI's GEO, with accession number GSE107727. Sample demultiplexing, barcode processing, and gene counting were all performed using the count command from the *Cell Ranger* v1.3 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome), with experimentally determined expected cell numbers given as input to the count command.

2.2.3 Single-cell qRT-PCR profiling

Gene expression profiling on mouse haematopoietic cells using single-cell qRT-PCR was performed as previously described (Moignard et al., 2013; Wilson et al., 2015). Single-cell qRT-PCR data for 416B cells were obtained from a previously published study (Schütte et al., 2016), and cells from the HoxB8-FL cell line were profiled using the same protocol as for the bone marrow cells. Gene expression profiling on human cord blood progenitors was performed as previously described (Quek et al., 2016) and details of the TaqMan assays can be found in the supplementary material of Karamitros et al. (2018).

2.3 Quality control and normalisation

Quality control and normalisation of mouse single-cell qRT-PCR data was performed by Nicola Wilson and Sonia Nestorowa. All scRNA-seq data and human single-cell qRT-PCR data were processed by F. Hamey.

2.3.1 Processing SMART-Seq2 scRNA-seq data of mouse HSPCs

Quality control measures were applied to the data to filter out low quality profiles, including those that were likely to correspond to empty wells. Cells were removed from further analysis if they had under 200,000 reads mapping to nuclear genes, under 4,000 genes detected (where detection was defined by at least two reads mapping to that gene in a cell), over 10% of mapped reads mapping to mitochondrial genes, or over 50% of mapped reads mapping to the ERCC spike-ins. Cells that passed quality control were then normalised using the *scran* R package with an initial clustering step to group cells with similar expression profiles before normalisation (Lun et al., 2016a). ERCC spike-ins were used to estimate technical variance of the sequencing data following the method of Brennecke et al. (2013). This approach fits a relationships between the variance and mean expression of the spike-ins to represent technical variation, and searches for biological genes with variance exceeding this threshold. These are denoted as the "highly variable genes". Using this method 4,290 genes were identified as having a squared coefficient of variance exceeding the level of technical noise, and were therefore used for downstream analysis such as clustering and dimensionality reduction.

When PCA was performed on the dataset, there were 20 cells that clearly separated from the others in PC3. Genes driving this separation (based on PC3 loadings) were lymphoid genes (genes with the most negative loadings in PC3 were *Iglj3*, *Tnfrsf17*, *Iglc3*, *Iglj1*, *Iglc2*, *Iglc1*, *Slamf7*, *Fcrla* and *Igkj5*). Based on this it was concluded these 20 cells were contaminating mature lymphoid cells and they were removed from the data as a result, to avoid them influencing further analysis.

Alternative normalisation of gene expression counts using the ERCC spike-ins was performed by calculating size factors with the *scran* package *computeSpikeFactors* function. These size factors were then used to normalise each cell by dividing its gene counts by its size factor. Normalised reads were then adjusted to account for batch effect differences in the ERCC concentration across days by using the *ComBat* function (*SVA* R package), with the sorting gate (LSK, Prog or LT-HSC) as an adjustment variable.

2.3.2 Processing SMART-Seq2 scRNA-seq data of human progenitors

Cells were removed from further analysis if they had under 500,000 reads mapping to nuclear genes, more than 20% of mapped reads mapping to mitochondrial genes, more than 20% of mapped reads mapping to ERCC spike-ins, or fewer than 750 high coverage genes (defined as having at least ten counts per cell). 163 out of 166 and 157 out of 249 cells passed quality control from donors 1 and 2, respectively. The method of Brennecke et al. (2013) was used to identify genes exceeding technical variance based on the expression of ERCC spike-ins. In visualisation and clustering analysis, the data showed batch effects between the two donors and so were processed separately. The *Seurat* R package was used to regress out plate effect for each donor, and was applied to set more stringent variable gene thresholds, identifying 1,605 and 1,273 variable genes in donors 1 and 2, respectively. These genes were used for downstream analysis.

2.3.3 Processing droplet-based scRNA-seq data

The default *Cell Ranger* barcode filtering was used to identify 46,447 unique barcodes from combined LSK and LK cells from the WT samples, and 14,675 from the W^{41}/W^{41} samples. This filtering ranks the barcodes by the number of UMI counts associated with each barcode, and calculates a cutoff for the number of UMIs per barcode. Barcodes with more associated UMIs than this cutoff are retained as cells. The cutoff is calculated using the expected number of cells, N, which is given as input to the *Cell Ranger* pipeline. The 99th percentile of the top N barcodes is taken as estimate, m, of the maximum number of UMIs for a cell. The cutoff is then calculated as $m/10$. Plots of UMI counts against ranked barcodes were visually inspected to ensure there was only one distinct mode in the distribution, as recommended in the *Cell Ranger* support documentation. Downstream analysis of data was performed using the python *Scanpy* module (Wolf et al., 2018). Each sample was separately filtered for potential doublets (arising from two cells entering the same droplet during cell capture) by simulating synthetic doublets from pairs of scRNA-seq profiles. Observed profiles were then assigned doublet scores based on a k-nearest neighbour classifier applied to PCA transformed data. This classifier assessed the proximity of the observed profiles to the simulated doublet profiles. Code for performing this simulation and scoring was kindly provided by Samuel Wolock, from Allon Klein's lab and is now available as the *Scrublet* method (Wolock et al., 2018). Based on the distributions of these doublet scores, the 1% and 4.5% of cells with the highest doublet scores from each LSK or LK sample, respectively, were removed before further analysis. 1452 WT and 845 W^{41}/W^{41} cells in total were excluded as potential doublets.

Cells with over 10% of UMI counts mapping to mitochondrial genes, that had fewer than 500 genes detected, or with the total number of UMI counts further than 3 standard deviations from the mean were also excluded. After quality control, 44,802 WT and 13,815 W^{41}/W^{41} cells were retained. Cells were normalised so that the total count for each cell summed to 10,000. All further analysis was performed on these transformed counts. For the WT data, 5,032 variable genes were identified and for the W^{41}/W^{41} data 5,033 variable genes were identified by following the method of Macosko et al. (2015) implemented using the Scanpy function, with minimum cutoffs of expression = 0.001 and dispersion = 0.05. Gene counts were log-transformed using the transformation $x \to \log(x+1)$. For dimensionality reduction and clustering, each gene was scaled so that it was zero-centred.

Initial visualisations of the data using diffusion maps revealed a prominent "gap" in the erythroid branch of the WT data where it looked like cells separated into two routes along the differentiation trajectory. Differentially expressed genes between these cells were found to be significantly enriched in cell cycle related genes. To remove this effect, any of the differentially expressed genes intersecting with a list of cell cycle genes downloaded from Reactome (http:www.reactome.org/), and any other genes that had Pearson correlation coefficient > 0.2 with any gene in this intersection were excluded from visualisation and clustering analysis. 4,664 genes in the WT data and 4,754 genes from the W^{41}/W^{41} were carried forward for this analysis.

2.3.4 Processing single-cell qRT-PCR data

For mouse cells ΔCt (change in cycle-threshold) values were calculated by normalising the expression level for each gene to the mean expression of housekeeping genes *Ubc* and *Polr2a* within the same cell (Guo et al., 2010). Before downstream analysis, all house keeping genes (*Ubc*, *Polr2a*, *Eif2b1* in mouse data) were removed. Additionally, as *Cdkn2a* was not detected in any cells, and as *Egfl7*, *Gfi1* and *Sfpi1* all suffered technical issues, these were also excluded from the analysis in Chapter 4. All quality control and normalisation measures were performed in the R programming language with custom scripts. Single-cell qRT-PCR data from Wilson et al. (2015) and the new ST-HSC, MPP and preMegE populations were normalised together.

For human cells, amplification curves with a Quality Score of < 0.65 as well as any Ct values > 27 were treated as undetected expression. Seven cells lacking the expression of both measured housekeeping genes (*B2M* and *GAPDH*) were removed from further analysis. One additional cell with a large outlying number of detected genes was also removed. Housekeeping gene *ACTB* was measured in the assay but displayed a very different expression profile to *B2M* and *GAPDH*, which both had similar expression patterns, and so was not used for normalisation or further analysis. Normalised ΔCt values were calculated by subtracting the mean of *B2M* and *GAPDH* from the gene expression values in each cell. The housekeeping genes were then excluded from further analysis. Genes detected in fewer than 20 cells, with a variance of less than 1 across all cells, or without expression in any of the MLP, GMP or LMPP ten-cell control samples assayed by qRT-PCR alongside single-cell cells were also removed from further analysis. After quality control, these data measured 74 genes in 919 single cells.

2.4 Index data

Index data for mouse HSPCs were collected by Sonia Nestorowa and Nicola Wilson, and normalised by Blanca Pijuan Sala. Index data for human cells were collected by Dimitris Karamitros, Bilyana Stoilova and Zahra Aboukhalil, and normalised by F. Hamey.

2.4.1 Index data of mouse bone marrow HSPCs

Using index sorting, levels of surface marker proteins (EPCR, CD48, CD150, CD16/32, c-Kit, Sca1 and CD34), DAPI, Lineage markers and forward-scattered light-height (FSC-H) were measured for each cell. The *flowCore* R package was used to extract and compensate the index data, and the *ComBat* function (*SVA* package) used to normalise the data across the two days that cells were sorted on, with the sorting gate (LSK, Prog or LT-HSC) included as an adjustment variable. Scatter plots of normalised data were then used to retrospectively assign cells to populations, with thresholds chosen based on gating strategies in the literature. E-SLAM cells were gated as EPCR$^+$ CD48$^-$ CD150$^+$, rather than EPCR$^+$ CD48$^-$ CD150$^+$ CD45$^+$, as CD45 was not included in the panel of measured surface markers. Two types of retrospective gates were assigned. The first (broad gating) assigned every cell to a population. The second (narrow gating) left gaps between gates to describe more specific populations, and therefore some cells remain unclassified in this gating. Broad gating was performed to ensure each cell was assigned a cell type, and the narrow gating was performed as it provided a more accurate representation of how the gates would be defined for conventional sorting.

2.4.2 Index data of human progenitors

Index sorting was used to record levels of FSC-H, side-scatter (SSC), Hoechst and expression of Lineage markers, CD34, CD38, CD45RA, CD10, CD90 and CD123 for each cell assayed by scRNA-seq, single-cell qRT-PCR, or in single-cell cultures. When cells sorted on different days were analysed together the index data was normalised across batches separately for each cell type (MLP, GMPP, LMPP) using the *ComBat* function from the *SVA* R package. When devising new sorting strategies to enrich for function, thresholds were defined based on the maximum expression of CD10 and CD45RA for LMPPs with myeloid output (LMPP$^{\text{ly}}$ and LMPP$^{\text{mix}}$ cells) and the maximum CD38 expression of GMPs with lymphoid and lympho-myeloid output (CD38$^{\text{hi}}$ GMPs). The percentages of GMPs or LMPPs above or below these maxima were calculated, and the corresponding percentages of cells sorted in the strategies shown in Chapter 5.

2.5 Clustering of single-cell data

2.5.1 Clustering mouse HSPC SMART-Seq2 data

Gene expression profiles were clustered based on the log-transformed normalised expression values ($x \to \log_2(x+1)$) of highly variable genes using the *Scanpy* python module, version 1.0. Profiles were transformed into PCA space, and then the 15 nearest neighbours for each cell were calculated based on cosine distance between cells in this PCA space using the top 20 PCs. The cells were then clustered using louvain clustering, implemented in the *Scanpy louvain* function. An initially high resolution of 2.0 was used with the aim of over-clustering the data, and the resolution iteratively decreased until each pairwise comparison of differential expression between groups (using a t-test) yielded at least 50 differentially expressed genes (Z-score from t-test > 2.576). This resulted in a resolution of 0.9 being used for clustering, assigning each cell to one of seven clusters. To identify highly expressed genes in each cluster, the list of all Ensembl genes was first filtered to those that were annotated as protein coding genes and were expressed in at least 10 cells across the whole dataset. For each cluster, the genes were then filtered to those that were expressed (log-transformed normalised count greater than 4) in at least half of the cells in that cluster, and these were tested for differential expression against cells from all other clusters using a Wilcoxon rank-sum test with Benjamini-Hochberg correction for multiple testing. Genes with false discovery rate < 0.001 were then ranked by fold change, and the top 10 genes with the highest fold change are shown in the heatmap in Fig. 3.2.

2.5.2 Clustering human progenitor single-cell data

To avoid separation due to donor batch effects, the scRNA-seq profiles of human cord blood progenitors were clustered separately for each donor. PCA was performed on the highly variable genes using the *Seurat* package and cells were clustered using the *Seurat FindClusters* function applied to the top 10 PCs. Differential genes between a cluster and the rest of the data were found using the *Seurat FindAllMarkers* function, and the top genes for each cluster are visualised in the heatmaps in Fig. 5.3. Single-cell qRT-PCR profiles were clustered using hierarchical clustering on both genes and cells using the *hclust* R function (*stats* package) with dissimilarity measure 1 - Spearman's correlation and agglomeration method Ward.D2. Cells were assigned to three clusters by applying the *cutree* function (*stats* package) to the results of the hierarchical clustering.

2.5.3 Clustering droplet-based scRNA-seq data

For WT data, clustering of the LSK and LK cells together was performed on diffusion map coordinates calculated using the 4,664 variable genes, by using the *Scanpy louvain* clustering function to cluster cells based on the k-nearest neighbour graph with 15 nearest neighbours. The coarse clustering was performed with resolution = 0.175, resulting in 16 different clusters, and the fine clustering was performed with resolution = 2.0, resulting in 63 different clusters. Fine clusters were assigned a colour for visualisation purposes corresponding to the colour of the coarse cluster to which the majority of their members belonged.

For comparison with the W^{41}/W^{41} data, WT LK cells were clustered using louvain clustering (*Scanpy igraph* method) with 15 nearest neighbours. To assign W^{41}/W^{41} LK cells to clusters, the W^{41}/W^{41} data were projected into the PCA space of the WT data. The nearest WT neighbours of each W^{41}/W^{41} cell were then calculated based on Euclidean distance in the top 50 PCs. W^{41}/W^{41} cells were assigned to the same cluster that the majority of their 15 nearest WT neighbours belonged to. Marker gene expression across these projected clusters was checked to ensure that the clustering assignment made biological sense.

2.6 Dimensionality reduction

2.6.1 Visualisation of mouse HSPC SMART-Seq2 profiles

Mouse bone marrow HSPCs were visualised using the dimensionality reduction method of diffusion maps (Haghverdi et al., 2015). Diffusion map dimensionality reduction was performed using the *DiffusionMap* function from the *destiny* R package (Angerer et al., 2016) with centred cosine distance and $\sigma = 0.16$. The parameter σ controls the probabilities of cell transition in the random walks on the high-dimensional data, with higher σ meaning that cells do not diffuse as far. The input for this function was log-transformed normalised counts of highly variable genes. The first few components of the diffusion map were inspected and the top three components selected for plotting, based on the size of their eigenvalues and the fact that these components captured differentiation towards the three major lineages. 3D plots were generated using the *scatter3D* function from the *plot3D* R package.

2.6.2 Visualisation of mouse HSPC qRT-PCR profiles

The diffusion map dimensionality reduction was calculated on the normalised ΔCt values of the 41 genes that passed quality control using the *DiffusionMap* function from the *destiny* R package, with centred cosine distance and $\sigma = 0.3$.

2.6.3 Visualisation of droplet-based scRNA-seq profiles

PCA was performed on the log-transformed, scaled and normalised variable genes, using the *pca Scanpy* function. For visualisation, a k-nearest neighbour graph with $k = 7$ was constructed, with edge distances calculated from the Euclidean distances between cells in the top 50 principal components. The resulting edge list of the k-nearest neighbour graph was exported into *Gephi* 0.9.1 (https://gephi.org/), where the graph coordinates were calculated using the *ForceAtlas2* layout. When the expression of marker genes was plotted in the landscape, cells with the highest gene expression were plotted on top, as the high number of observations required points to overlap in the plots. With this method it could always be seen which regions of the graph were positive for expression of a given gene.

2.6.4 Visualisation of human progenitor single-cell profiles

Single-cell data were visualised with diffusion maps (Haghverdi et al., 2015) using the *DiffusionMap* function from the *destiny* R package with Euclidean distance.

2.6.5 Projection of qRT-PCR datasets

Single-cell qRT-PCR data for 416B and HoxB8-FL cell lines were projected onto the diffusion map embedding of Section 2.6.2 using the *dm.predict* function from the *destiny* R package.

2.7 Reconstructing differentiation trajectories

2.7.1 Pseudotime ordering of mouse HSPC SMART-Seq2 data

From the diffusion map dimensionality reduction, cells belonging to the stem cell population and to the erythroid, granulocyte-macrophage and lymphoid lineages could be clearly identified. To find cells lying on these three lineage branches, the diffusion map embedding was first used to identify a start cell (within the tip of the E-SLAM population) and end cells for each of the three lineages by choosing cells at the tips of the remaining branches. Cells on three broad trajectories from the stem cells to the three tip cells were then identified following the method of Ocone et al. (2015). For this, a $k = 30$ nearest neighbour graph was constructed using Euclidean distance between cells in the first four diffusion components. The first four components were used based on the magnitude of their corresponding eigenvalues. The shortest path from the start to each end cell was found using Dijkstra's algorithm, and each lineage branch was then formed from the $n = 100$ nearest neighbours of each cell on this backbone path. Cells on a branch were then ordered in pseudotime using custom R code implementing the Wanderlust algorithm with default parameters (Bendall et al., 2014).

To identify up- or downregulated genes along the three trajectories, log-transformed gene expression values were smoothed by calculating the mean expression for a sliding window of 20 cells along the pseudotime ordering. Spearman's correlation between pseudotime values and the smoothed expression was calculated for each gene, and genes with correlation > 0.5 or < -0.5 were identified as up- or downregulated, respectively.

2.7.2 Pseudotime ordering of mouse HSPC qRT-PCR data

Based on the diffusion map coordinates, two lineages branches were identified in the single-cell data linking HSCs to MEPs and HSCs to LMPPs. This was done following the method of Ocone et al. (2015). For this, three cells types, the molecular overlapping population (MolO) HSCs defined by Wilson et al. (2015), MEPs and LMPPs, were highlighted on the diffusion map, and by using this three-dimensional diffusion map visualisation a start and end cell were selected for each trajectory from within the relevant highlighted populations. Branches and trajectories were then constructed as in Section 2.7.1 using the top four diffusion components. The top four components were selected based on the magnitude of their corresponding eigenvalues. The pseudotime ordering was constructed using the expression levels of both transcription factor-encoding and non-transcription factor-encoding genes.

2.7.3 Pseudotime ordering of droplet-based scRNA-seq data

Pseudotime ordering was performed using the *dpt* function from the *Scanpy* python module, with the root cell chosen as the cell with the highest MolO score (see Section 2.11).

2.8 Cell cycle scoring of individual cells

2.8.1 Assigning cell cycle states to SMART-Seq2 data

Cells were assigned to cell cycle categories following the method of Scialdone et al. (2015) implemented in the *cyclone* function (*scran* R package) using the mouse cell cycle markers distributed along with this package. Cells with G1 score > 0.5 were assigned to the "G_0/G_1" phase, cells with G2M score > 0.5 to the "G_2/M " phase and remaining cells to "S" phase. There were no cells with G1 and G2M scores both exceeding 0.5.

2.8.2 Assigning cell cycle states to droplet-based scRNA data

To calculate G_2/M marker gene scores, the set of 200 Hallmark G_2/M checkpoint genes was downloaded from the Molecular Signatures Database (Liberzon et al., 2015). To calculate the score for the set, G, of genes in the droplet-based scRNA-seq landscape, a geometric mean-based score was calculated on the normalised UMI count, x, of each gene in the set. This score was given by $\exp[\sum_{g\in G}\log(x_g+1)/m]$ where $|G|=m$.

2.9 Identifying differentially expressed genes

2.9.1 Finding differential expression in the abstracted graph

The expression of all genes detected in at least two cells was compared between pairs of clusters using the Wilcoxon rank-sum test to calculate a p-value. The fold-change between the average log-transformed expression in each cluster was also calculated. Adjusted p-values were then calculated using the Benjamini-Hochberg correction for multiple testing. Genes were treated as significantly differentially expressed if they had adjusted p-values < 0.05 and $\log_2$ fold-change of > 0.5. For visualisation in the Fig. 6.7C & D heatmaps, the upregulated genes for the first nodes along either the erythroid or megakaryocytic trajectories were filtered to genes unique to each trajectory, and these were ranked by adjusted p-value. The genes with the top 20 most significant adjusted p-values are displayed in the heatmap.

2.9.2 Finding differential expression between WT and W^{41}/W^{41} cells

Differential expression analysis was performed between WT and W^{41}/W^{41} clusters using *edgeR* applied to the UMI counts (Robinson et al., 2010). P-values were corrected for multiple testing (Benjamini-Hochberg correction). Genes were labelled as up- or downregulated based on positive or negative fold-changes and ranked by adjusted p-value. The top 200 up- and downregulated genes between each cluster pair were recorded. Full lists can be found in supplementary table S1 of Dahlin et al. (2018). To calculate the significance of overlap between the differentially expressed genes and annotated gene sets, the top 200 up- or downregulated genes for each cluster were input into the Molecular Signatures Database online tool (http://software.broadinstitute.org/gsea/msigdb/index.jsp) and overlaps between these and the Hallmark gene sets were computed (Liberzon et al., 2015; Subramanian et al., 2005).

2.10 Gene set enrichment analysis

2.10.1 Pseudotime-correlated genes along E, GM and L trajectories

Gene set enrichment analysis was performed using *Enrichr* (Chen et al., 2013). All results with adjusted p-value < 0.05 (Benjamini-Hochberg correction for multiple testing) were treated as significant. The most significant non-overlapping terms are shown in the figures in Chapter 3. When investigating the cell cycle effect, a list of cell cycle genes was downloaded from Reactome (http://www.reactome.org/), and mouse orthologues for each gene found using the Ensembl BioMart online tool. These orthologues were used to filter gene lists before performing enrichment analysis using *Enrichr*.

2.10.2 Analysis on absolute RNA content

The ERCC-normalised counts were summed for each cell to estimate the total RNA content per cell. One-way analysis of variance (ANOVA) tests were applied to calculate the significance of differences in FSC-H and RNA content between cell types. To identify which genes were downregulated in absolute terms along pseudotime, the previously obtained downregulated lists were filtered to remove any genes than had a less than two-fold expression change between the first 10% and final 10% of cells in an ordered pseudotime trajectory. Again, gene set enrichment analysis was performed using *Enrichr*.

2.11 Visualising gene expression signatures

To plot a score for a set of genes, G, in the droplet-based scRNA-seq landscape, a geometric mean-based score was calculated using the normalised UMI count, x, of each gene in the set. This score was given by $\exp[\sum_{g\in G}\log(x_g+1)/m]$ for m genes. To visualise HSC gene expression and calculate the “MolO” score, this score was calculated using the 29 genes previously identified as being enriched in functional HSCs (MolO genes from table S3 from Wilson et al. (2015)).

2.12 Transcriptional regulatory network modelling

2.12.1 Inferring the network

Network inference was performed using the expression of transcription factors measured using single-cell qRT-PCR. The first step in the network inference method is to identify potential regulatory relationships between pairs of transcription factor encoding genes. This was done by calculating pairwise partial correlation coefficients across the whole dataset using the *pcor* function from the *ppcor* R package. Correlation coefficients were filtered to retain only the pairs of genes with significant interaction between them, using a threshold of p-value < 0.01. Gene pairs were then ranked by the magnitude of their correlation coefficients, and the strongest correlations were retained as edges in a gene correlation network. Positive correlation between gene G_1 and gene G_2 was then treated as possible activation of gene G_1 by gene G_2, or of gene G_2 by gene G_1, as the correlation relationship cannot be used to infer the direction of regulation. Negative correlation was treated as potential repression acting in either direction. Self-activation was added to the potential activating or repressing relationships for each gene, as self-activation is a widely-used motif in transcriptional regulation that cannot be revealed by correlation analysis. Combinations of these regulatory relations describe a set of possible Boolean functions governing the expression of each gene, with each rule featuring one or more regulators (Fig. 4.6B).

The next step was to search for the regulatory rules best describing the regulation of each gene. Whilst high correlation between a pair of genes can be an indication of a regulatory relationship, additional information is required to determine if the regulation is direct, establish the direction of regulation between genes, and understand if it is part of a regulatory event that requires the involvement of multiple transcription factors. The pseudotime ordering of cells was used as the basis for a scoring mechanism applied to the set of possible Boolean functions from the correlation network. To reduce the search time of the algorithm, the search was restricted to functions of the form $F = F_1 \wedge \neg F_2$ with each F_i a Boolean function made from AND and OR gates with at most two inputs per gate. F_1 represents the activating part of the function, consisting of at most four activating transcription factors for a gene, and F_2 the repressing part of the function, formed from at most two repressing transcription factors. This restriction approach was used by Moignard et al. (2015) in their Boolean function search.

Gene expression in each cell along the pseudotime trajectory was first discretised into “ON” or “OFF” expression states, by setting any detected values of gene expression to 1, and any undetected values to 0. Each pair of cells positioned k steps apart in the pseudotime ordering was then treated as an input-output pair $P_i = (I_i, O_i)$ for a Boolean function, where $[I_i]_G$ indicates the binary expression of gene G in input cell I_i. Each function F for a gene G was given a score $S(F) = \sum_i s_i(F)$ where

$$s_i = \begin{cases} 1, & \text{if } [F(I_i)]_G = [O_i]_G \\ 0, & \text{otherwise} \end{cases}$$

for the pseudotime pairs $P_i = (I_i, O_i)$. This calculates the number of times the value of the gene G as predicted by F applied to I_i equals the value of G in the corresponding output cell O_i. The step size used for the result in Chapter 4 was $k = 3$ between pseudotime pairs. A random sample of 10 genes was run with different sizes of k to test the sensitivity of the approach to this parameter choice, and this showed that there was good agreement between the pseudotime rule scoring for different k (Fig. A.1). The top-scoring functions for each gene can then be considered the best functions for that gene. To identify the top-scoring functions, the problem was encoded as a Boolean satisfiability problem (de Moura and Bjørner, 2008) using the Python Z3 solver (https://github.com/Z3Prover/z3/). Python code that can be run to identify rules for each gene can be found at https://github.com/fionahamey/Pseudotime-network-inference. The procedure is summarised in Algorithm 2.1.

Algorithm 2.1. Procedure for network inference

```
Set rule agreement threshold $t = T$                         ▷ $T$ is user-defined choice
Construct set of pseudotime input-output pairs $\{P_i\}$
for all genes G in dataset do
    Let $\mathbb{F}_G$ be an empty set of functions describing rules for each gene
    while $\mathbb{F}_G$ is empty do
        Search for existence of function $F$ such that $S(F) > t$ across pairs $\{P_i\}$
        if Such an $F$ exists then
            repeat
                Find $F$ and add to set of functions $\mathbb{F}_G$
                Search for existence of new $F$ such that $S(F) > t$ and $F \notin \mathbb{F}_G$
            until No new $F$ exists
        else
            Let $t = t - \varepsilon$                            ▷ $\varepsilon$ is user-defined choice
        end if
    end while
end for
return $\mathbb{F}_G$ for all genes
```

For many genes the method gave several functions with equally high scores. In this case the results were simplified to the minimum set of simplest functions. For example, if functions “Gata1 $\rightarrow$ Tal1” and “Gata1 $\wedge$ Nfe2 $\rightarrow$ Tal1” had equal scores, then the former would be chosen as it is simpler and contained within the latter. When rules could not be simplified in this way multiple rules were retained. MEP and LMPP network models have been deposited in BioModels and assigned the identifiers MODEL1610060000 and MODEL1610060001, respectively (Chelliah et al., 2015). The full list of rules is available in Appendix A.

2.12.2 Stable state analysis of the network

One way of assessing the behaviour of Boolean network models is by considering their attractor states, or stable states. These represent states that when reached by the network have no further changes in the expression of any genes. For clarification, these will be defined more formally as follows. If the state of a network at time t is given by $\mathbf{X}_t$, where $\mathbf{X}_t = (X_1(t), \dots, X_n(t))$ is a vector of gene expression states $G_1, \dots, G_n$, then consider the transformation $T : \mathbf{X}_t \to \mathbf{X}_{t+1}$, which represents the transition function of the network model. Network transitions can be modelled using either synchronous or asynchronous updates. With synchronous updates $\mathbf{X}_{t+1} = (F_1(X_1(t)), \dots, F_n(X_n(t)))$, so the expression of each gene is updated simultaneously according to a Boolean function F_i. With asynchronous updates, for each transition T a gene G_k is randomly selected so that $\mathbf{X}_{t+1} = (X_1(t), \dots, F_k(X_k(t)), \dots, X_n(t))$, so that the expression of only one gene is altered at a time. This can lead to stochastic behaviour from the network, and is more suited for modelling systems with a range of possible cell fates. A state $\mathbf{X}^*$ is then defined as stable if $\forall t \geq t^*$ where $t^* \in \mathbb{Z}$, $\mathbf{X}_{t+1} = \mathbf{X}_t$ under transformation T.

Stable states of the network were identified using the GenYsis algorithm with asynchronous updates (Garg et al., 2008). All alternative rules for a gene were considered in stable state analysis in the form of OR rules. This ensures that any stable states found are in the intersection of the stable states of all of the possible networks. To identify stable states reachable from MolO expression starting states, the Boolean rules for each network were encoded and simulated with asynchronous updates until the network stabilised and no genes changed in expression. These simulations were carried out in the R programming language with custom scripts. For both MEP and LMPP network models 1000 simulations were run starting from each of the 237 binary expression states corresponding to MolO cells, and the stable states of the simulations were recorded.

To project stable states onto the diffusion map, the stable states were compared to the profiles of bone marrow data converted to binary expression values. The nearest neighbour of each state was identified and highlighted in the diffusion map. If more than one neighbour was the best match for a state then the continuous gene expression levels of these nearest neighbours were averaged and this average expression state was projected onto the diffusion map using the *dm.predict* function.

2.13 Chromatin immunoprecipitation sequencing

Chromatin immunoprecipitation sequencing (ChIP-seq) data were generated and analysed by Nicola Wilson, Sonia Nestorowa and Rebecca Hannah. ChIP-seq assays were performed as previously described (Wilson et al., 2010). Samples were amplified using the Illumina TruSeq ChIP Sample Prep Kit and sequenced using the Illumina HiSeq 2500 System. Sequenced reads were mapped to the mm10 mouse reference genome using *Bowtie2* (Langmead and Salzberg, 2012), converted to a density plot, and displayed as UCSC genome browser custom tracks. Peaks were called using *MACS2* software (Zhang et al., 2008).

2.14 Luciferase assays

Luciferase assays were performed by and analysed by Sonia Nestorowa and Sarah Kinston. Luciferase and LacZ constructs were generated using standard recombinant DNA techniques. The coordinates of chromosomal regions tested were: chr8:122699004-122701098 for the *Cbfa2t3h* promoter, chr8:122699111-122699377 for the *Cbfa2t3h* min promoter, and chr15:103258245-103258850 for the *Nfe2* enhancer. Both WT and GATA2 mutant constructs were generated, where GATA2 binding sites were fully mutated to prevent any binding activity. Luciferase assays were performed as previously described (Bockamp et al., 1995). Luciferase assays were performed in the 416B cell line for stable transfections.

2.15 Single-cell functional assays

Single-cell functional cultures were performed by Dimitris Karamitros, Bilyana Stoilova and Zahra Aboukhalil, who also quantified the functional output of the culture experiments. LMPPs, GMPs and MLPs were cultured in 96 well plates in SGF15/2 culture (containing cytokines SCF, G-CSF, FLT3L, IL-2 and IL-15) or, for one of the LMPP enrichment experiments, SF7b culture (containing SCF, FLT3L and IL-7). Both culture conditions support lymphoid and myeloid output. After 2-2.5 weeks in culture, flow cytometry was performed to assess lineage output, and wells were considered positive if they contained more than 15 CD15$^+$, CD14$^+$, CD56$^+$ or CD19$^+$ cells. Full details can be found in (Karamitros et al., 2018).

2.16 Graph abstraction

The graph abstraction method was developed by Alex Wolf as the result of discussions with F. Hamey. Code for performing the graph abstraction was provided by Alex Wolf as part of the *Scanpy* module. The droplet-based sequenced data were analysed by F. Hamey.

2.16.1 Constructing the abstracted graph

The abstracted graph was calculated on the fine louvain clustering of LSK and LK data together using the *Scanpy paga* function to perform partition-based graph abstraction from *Scanpy* version 1.1 (Wolf et al., 2018, 2017). This works by first connecting cells in a k-nearest neighbour graph, as for performing the louvain clustering. In this case, the number of nearest neighbours used was $k = 15$ and distances between cells were given by Euclidean distance between diffusion map coordinates in the top 20 diffusion components. A confidence c_{ij} is then calculated between clusters i and j to represent the strength of the connection between these two clusters in the abstracted graph. Confidence c_{ij} is given by the ratio of the number of edges between clusters i and j and the geometric mean of the number of edges originating from i and the number of edges originating from j. So $c_{ij} = \frac{1}{2}(e_{ij} + e_{ji})/\sqrt{k^2 n_i n_j}$ where e_{ij} is the number of edges originating from cluster i leading to cluster j, k is the k parameter for the k-nearest neighbour graph, and n_i is the number of nodes in cluster i. This ratio compares how many edges are observed between the two clusters compared to how many would be expected under a random distribution of edges. A confidence threshold of 0.007 was used to filter edges in the *paga* plot function. The abstracted graph node coordinates were positioned using the *Force Atlas 2* algorithm in *Gephi* applied to the filtered, weighted edges.

2.16.2 Plotting average gene expression on the abstracted graph

To plot gene expression (or gene signatures) on the graph nodes, the mean value of the normalised UMI counts for a gene (or the gene signature scores for each cell) was calculated for each cluster. These averages were then visualised on the abstracted graph, with a scale ranging from the minimum to the maximum value for the means of the nodes of the connected component of the graph.

2.16.3 Finding trajectories through the abstracted graph

The node with the highest MolO score (at the top of the abstracted graph) was used as a starting point for the trajectories. End points for erythroid, megakaryocyte, neutrophil, monocyte and lymphoid progenitors were identified based on marker gene expression and the structure of the abstracted graph. The shortest paths through the abstracted graph from the starting node to each lineage end point were found using the *get_shortest_paths* function from the python *igraph* module, with edge weight between nodes i and j given by the number of edges from i to j in the single-cell k-nearest neighbour graph. These shortest paths were taken as the trajectories.

Chapter 3

A single-cell reference map of murine haematopoiesis

Parts of this section have been modified from Nestorowa et al. (2016), on which F. Hamey is joint first author. Experimental work for this project was carried out by Sonia Nestorowa and Nicola Wilson (isolation of primary bone marrow cells, scRNA-seq profiling), and by David Kent and Mairi Shepherd (isolation of primary bone marrow cells). RNA-seq data were aligned by Evangelia Diamanti. Index data normalisation was performed by Blanca Pijuan Sala, who also made the website for interactive plotting of the single-cell data. All computational analysis of aligned scRNA-seq data was carried out by F. Hamey.

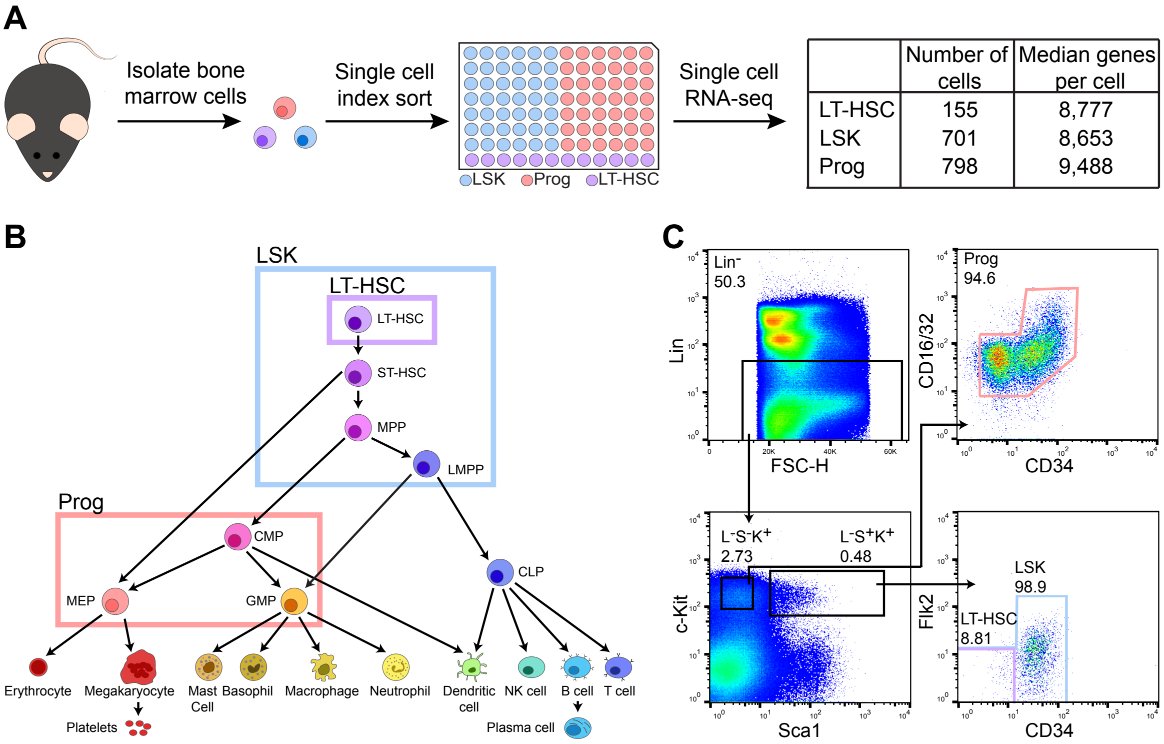
3.1 Background

The haematopoietic system is maintained by differentiation of multipotent cells towards the mature blood cell types. Many changes in gene expression occur during the transition of stem cells and progenitor populations to specialised blood cells, with the dynamics of a number of genes playing important roles in cell fate decision-making during differentiation. To understand how these decisions are regulated, it is necessary to study cells at different stages of maturation, and so over the past three decades researchers have worked to purify populations of haematopoietic cells characterised by their differentiation potential (Beerman et al., 2010; Challen et al., 2010; Kent et al., 2009; Kiel et al., 2005; Morita et al., 2010). Strategies for isolating these cells use the expression of cell surface marker proteins to separate populations using FACS, and have allowed gene expression datasets to be generated across haematopoietic populations.

Although these isolation strategies have led to many significant advances, they remain limited in their purity, as changes in cell state that alter the transcriptome of a cell are not always detectable in the relatively small number of surface markers measured. To combat this, a recent innovation has been to combine measurement of cell surface markers with single-cell gene expression profiling, thereby providing insight into gene expression heterogeneity within the classically defined haematopoietic populations. For example, Paul et al. (2015) used scRNA-seq to profile over 2,700 haematopoietic progenitor cells, including CMPs, GMPs, and MEPs, to investigate heterogeneity within the CMP population by simultaneously capturing cell surface marker levels and gene expression.

However, although studies evaluating the gene expression of several different haematopoietic cell types at the single-cell level existed prior to this work, published data were either limited in the range of cell types they covered, for example by not profiling HSCs alongside more mature progenitors, or were restricted to measuring fewer genes such as by using single-cell qRT-PCR (Grover et al., 2016; Kowalczyk et al., 2015; Moignard et al., 2013; Paul et al., 2015; Wilson et al., 2015). As many HSPC populations exhibit a large degree of heterogeneity at the single-cell level, it was reasoned that a high resolution single-cell dataset allowing comparison of all cell types from early haematopoiesis would provide a valuable tool for the haematopoietic research community, and would enable investigation of gene expression changes occurring during differentiation towards alternative blood lineages.

The focus of this work was to generate and analyse a scRNA-seq dataset with the aim of providing a reference transcriptional landscape of haematopoiesis. This chapter describes the processing of over 1,600 scRNA-seq profiles of HSPCs from mouse bone marrow to visualise diversification of cells from stem cells to three different mature lineages and shows how changes in properties such as gene expression, RNA content, and cell cycle can be investigated using transcriptomic data.



	Number of cells	Median genes per cell
LT-HSC	155	8,777
LSK	701	8,653
Prog	798	9,488

Fig. 3.1. Bone marrow HSPCs can be transcriptionally profiled at the single-cell level. (A) Experimental overview of the study. Index-sorting was used to isolate haematopoietic cells from mouse bone marrow for transcriptional profiling using scRNA-seq. Table summarises the number of cells and number of genes detected per cell after quality control filtering. Cells were sorted in three different gates based on surface marker expression. LSK, Lin$^-$ c-Kit$^+$ Sca1$^+$; Prog, progenitor; LT-HSC, long-term haematopoietic stem cell. (B) Visualisation of overlap between the sorting gates and populations in the classic haematopoietic hierarchy. Boxes indicate which cell types are covered by the sorting gates. (C) Flow cytometry diagrams displaying the gates used to isolate cells. Numbers indicate the percentage of cells within a gate for each plot. L$^-$S$^-$K$^+$, Lin$^-$ Sca1$^-$ c-Kit$^+$; L$^-$S$^+$K$^+$, Lin$^-$ Sca1$^+$ c-Kit$^+$.

3.2 Capturing single-cell transcriptional profiles across haematopoiesis

To obtain a comprehensive map of murine haematopoiesis covering stem and early progenitor populations, cells were isolated from primary mouse bone marrow and index-sorted in three different gates (Fig. 3.1A). As the aim of this work was to characterise gene expression changes occurring during differentiation, cells were sampled from two main broad gates: the LSK gate (Lin$^-$ c-Kit$^+$ Sca1$^+$) and the Progenitor (Prog) gate (Lin$^-$ c-Kit$^+$ Sca1$^-$) (Fig. 3.1B, C). In the bone marrow, cells from the LSK gate are found at a much lower frequency than those in the Prog gate (Fig. 3.1C), therefore cells from these two populations were sorted separately to ensure sufficient coverage of the upper tiers of the haematopoietic hierarchy. Following the same reasoning, additional LT-HSCs (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flk2$^-$ CD34$^-$) were sorted as these represent a rare population within the LSK gate (Fig. 3.1C).

In total, 216 LT-HSC, 852 LSK and 852 Prog index-sorted cells were then profiled using scRNA-seq to measure gene expression, as previously described (Picelli et al., 2014). After quality control filtering to remove empty and low quality profiles, 155 LT-HSC, 701 LSK and 798 Prog profiles were carried forward for further analysis. A high sequencing depth per cell resulted in a median of over 8,600 genes detected for each cell type (Fig. 3.1A). In scRNA-seq experiments it is known that a high number of genes either show low variation across cells, or have very low average expression and therefore their variation can be greatly influenced by technical noise. To identify biologically highly variable genes, synthetic ERCC spike-in controls were used to provide an estimate of technical variance (Brennecke et al., 2013), finding 4,290 genes with variance exceeding the estimated technical threshold. This set of highly variable genes was used for downstream analysis.

3.3 The transcriptional profiles of haematopoietic progenitor cells show priming towards different blood lineages

To investigate heterogeneity within these scRNA-seq data, unbiased clustering of the molecular profiles was performed using the expression of the 4,290 highly variable genes. This grouped cells into seven clusters (Fig. 3.2). Cluster 1 cells showed high expression of genes with previously established expression in HSCs, such as *Procr*, *Trim47*, and *F11r* (Sugano et al., 2008; Wilson et al., 2015). These cells originated from the LT-HSC and LSK gates, indicating that this cluster contained the most immature cells. The majority of cluster 2 cells were found in the LSK gate with a very small number of cells from the LT-HSC gate, suggesting that this cluster contained early, possibly uncommitted, progenitors. Cluster 3 was formed from mostly Prog cells and showed high expression of erythroid genes such as *Pklr* and *Ank1* (Arinobu et al., 2007; Rank et al., 2009). Cluster 5 was formed from almost entirely Prog cells and was revealed to have megakaryocyte progenitor identity based on the expression of genes such as *Pf4*, *Itga2b*, and *F2r* (Rowley et al., 2011). Of the remaining clusters, cluster 6 was also formed from a majority of Prog cells and expressed myeloid marker genes including *Mcpt8*, *Fcgr3* and *Ms4a3*, suggesting specification towards granulocyte/monocyte lineages in these cells (Dwyer et al., 2016; Hulett et al., 2001; Nimmerjahn and Ravetch, 2006). The final cluster, cluster 7, was mostly formed from LSK cells, and expression of *Flt3* and *Dntt* indicated these cells had a lymphoid identity (Rothenberg, 2014), consistent with the observation that the classical LMPP population falls within the LSK sorting gate. Together, this clustering separated cells into groups corresponding to different haematopoietic populations, ranging from stem cells at the top of the haematopoietic hierarchy, down to progenitor cells expressing markers of several different lineages.

3.4 Transcriptional profiling reveals a continuum of differentiation towards three cell types

Whilst clustering is a useful tool for identifying and characterising the gene expression of subpopulations within a dataset, it has the effect of forcing cells into discrete groups, which may not be the most suitable approach for representing cells undergoing a continuous process. Dimensionality reduction techniques can help to visualise the underlying structure within gene expression data and, in particular, the method of diffusion maps has been particularly successful in capturing the branching structure present within single-cell data from differentiating cells (Coifman et al., 2005; Haghverdi et al., 2015; Moignard et al., 2015). The diffusion map calculated on the highly variable genes showed good agreement with the unbiased clustering (Fig. 3.3A) and the cell type identity previously assigned to each group was supported by checking the expression of marker genes across the clusters (Fig. 3.3B). Visualising the expression of these marker genes in the diffusion map landscape demonstrated that the first three diffusion components captured a structure representing haematopoietic differentiation towards three main lineages (Fig. 3.4). Expression of *Procr* and *Hoxb5* highlighted the blood stem cell region at the top of the diffusion map, corresponding with cluster 1 cells (Balazs et al., 2006; Chen et al., 2016). Cells differentiating towards the erythroid lineage were marked by expression of *Klf1* and *Gypa* and could be found in cluster 3 (Dzierzak and Philipsen, 2013). Cluster 6 cells coincided with an area showing high expression of myeloid marker genes *Mpo* and *Ctsg*, therefore representing the granulocyte-macrophage lineage (Olsson et al., 2016). Finally, cluster 7 cells occupied a region of the diffusion map with high expression of lymphoid genes *Ighv1-81* and *Dntt* (Rothenberg, 2014). The scRNA-seq profiles therefore capture differentiation towards the erythroid, granulocyte-macrophage and lymphoid lineages, with this differentiation recapitulated by the coordinates of cells in the first three diffusion components.

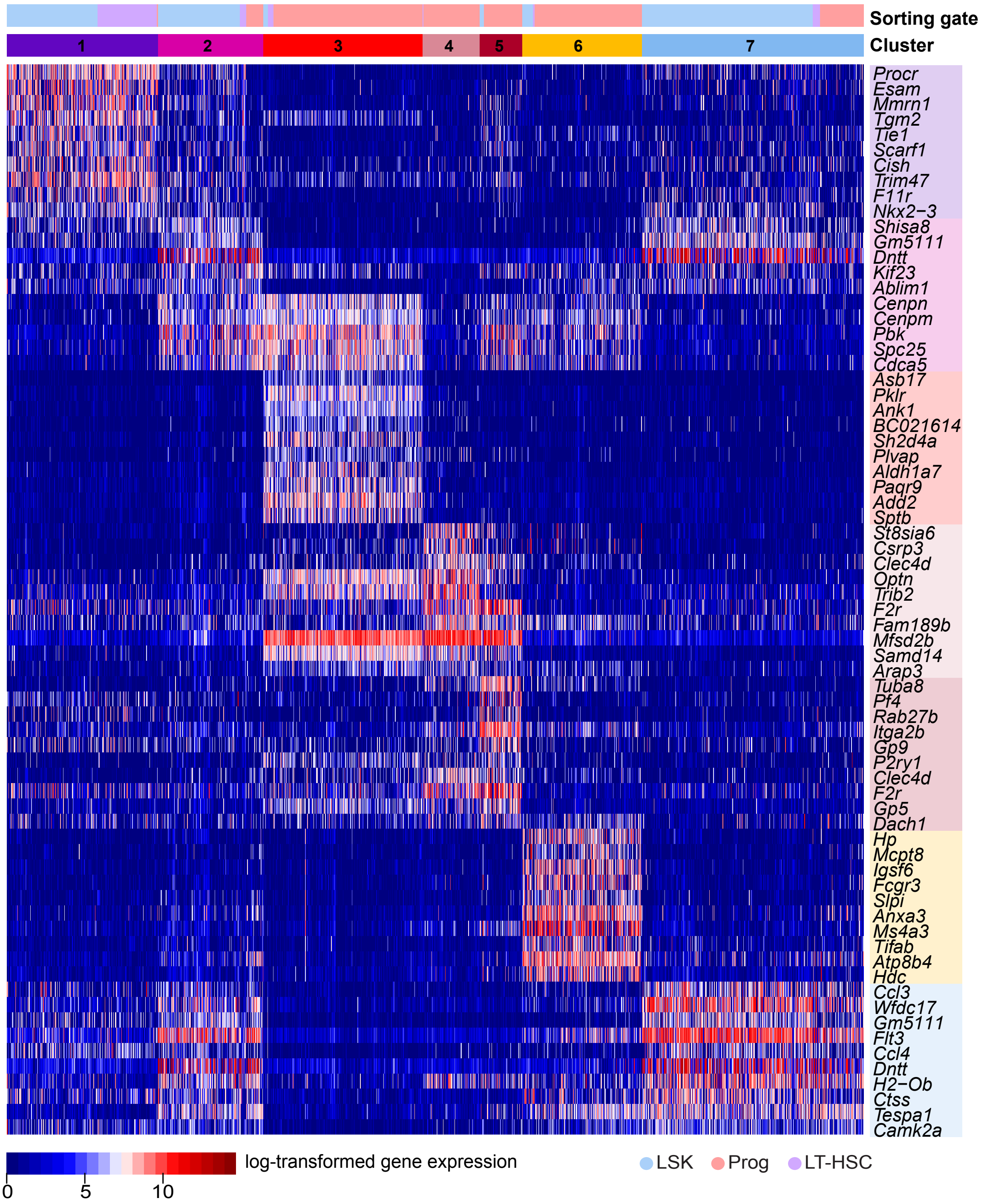


Fig. 3.2. Unbiased clustering reveals transcriptional heterogeneity across haematopoietic stem and progenitor cell populations. Heatmap displaying gene expression across clustered scRNA-seq data. Differential expression analysis was used to rank genes by fold-change between cells within a cluster versus the remaining cells. The top 10 genes most specific for each cluster are shown in the heatmap. Colour bars along the top indicate the cluster identity and the sorting gate for each cell.

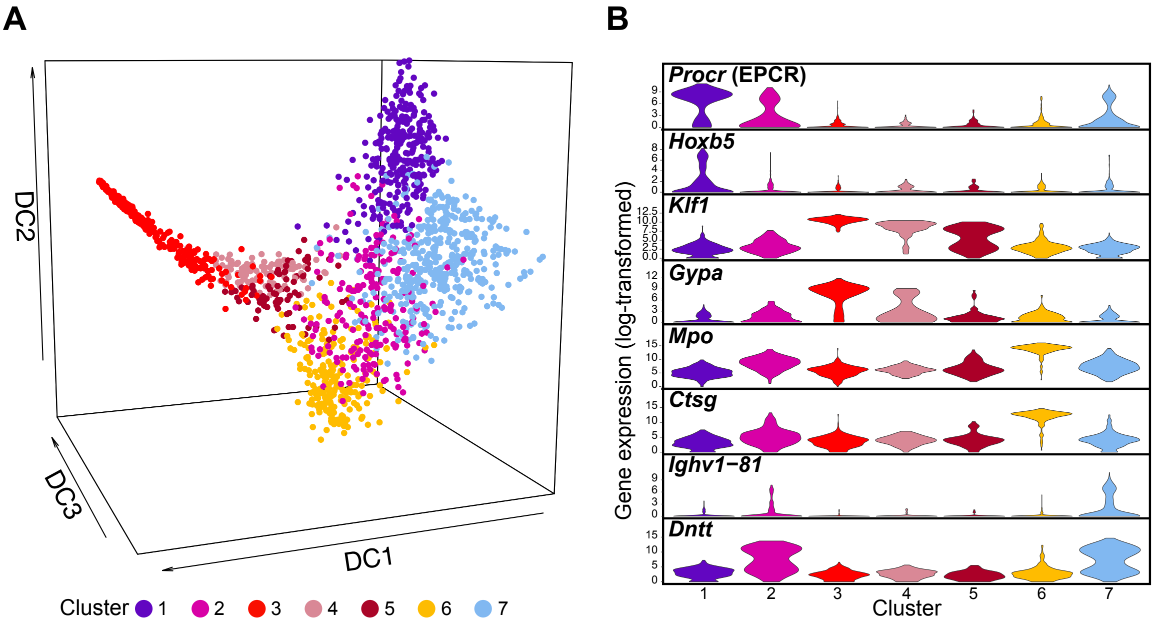


Fig. 3.3. Dimensionality reduction captures continuous specification towards different blood lineages. (A) Diffusion map of the 1,654 scRNA-seq profiles. The embedding was calculated on highly variable genes. Cells are coloured by cluster, with colours corresponding to those in Fig. 3.2. DC, diffusion component. (B) Violin plots of marker gene expression across clusters. Colours match those in panel A.

3.5 Retrospective gating links the transcriptome with cell surface phenotype

When isolated for scRNA-seq profiling, cells were index-sorted, meaning that both transcriptional profiles and surface marker expression were available for each cell. Surface markers Sca1, Flk2, CD34, CD48, CD16/32, CD150 and EPCR all showed expression patterns restricted to specific regions of the diffusion map, consistent with the gene expression-based annotation of the transcriptional landscape. Notably, EPCR was highest in the cells identified as HSCs at the top of the diffusion map, matching the raised expression levels of *Procr* (the gene encoding EPCR) in these cells. Interestingly, the forward-scatter (FSC-H) showed a gradient across the data, with erythroid and myeloid progenitors having higher FSC-H than the HSCs, consistent with the larger size of these more mature cells. So that others could freely view gene and surface marker expression patterns within these data,

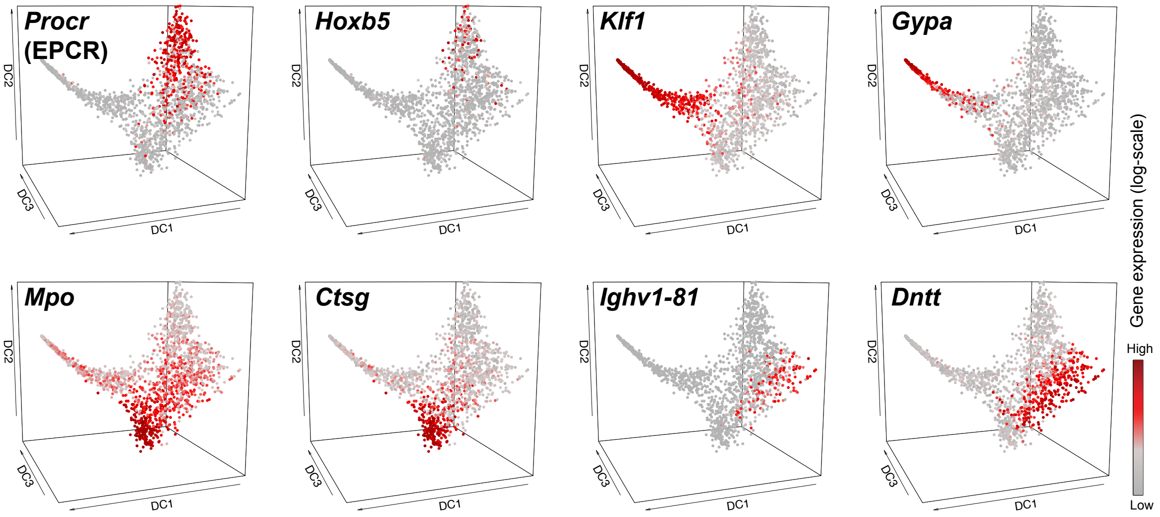


Fig. 3.4. Single-cell transcriptional profiles capture continuous differentiation towards the three main blood lineages. Diffusion map of cells coloured according to the expression of the lineage marker genes shown in Fig. 3.3B. The colour corresponds to a $\log_2$ scale of expression ranging between 0 and the maximum value for each gene. DC, diffusion component.

an interactive website was created by B. Pijuan Sala, based on the analysis described here (http://blood.stemcells.cam.ac.uk/single_cell_atlas.html).

Although cells were sorted in only three broad gates, the additional surface marker information from index-sorting allowed cells to be "retrospectively gated" into conventional haematopoietic populations (Fig. 3.6A). Highlighting the cells falling within different retrospective gates on the diffusion map revealed cell type distributions consistent with the expression patterns of marker genes for different lineages (Fig. 3.6B). Three different LT-HSC sorting strategies were considered in decreasing order of stringency: E-SLAM (CD48$^-$ CD150$^+$ CD45$^+$ EPCR$^+$), L$^-$ S$^+$ K$^+$ CD34$^-$ Flk2$^-$ CD48$^-$ CD150$^+$ and the LT-HSC sorting strategy defined in Fig. 3.1. The varying purity of these sorting strategies (assessed by the enrichment for cells capable of long-term repopulation upon transplantation) is reflected in the heterogeneity of the cell type positioning in the diffusion map, with the most specific strategy (E-SLAM) occupying the least heterogeneous region in the diffusion map. This retrospective gating approach allows the transcriptome to be related to cellular phenotype, and as the cell types in Fig. 3.6 all represent well-established haematopoietic gating strategies this provides a useful resource for comparison with other datasets.

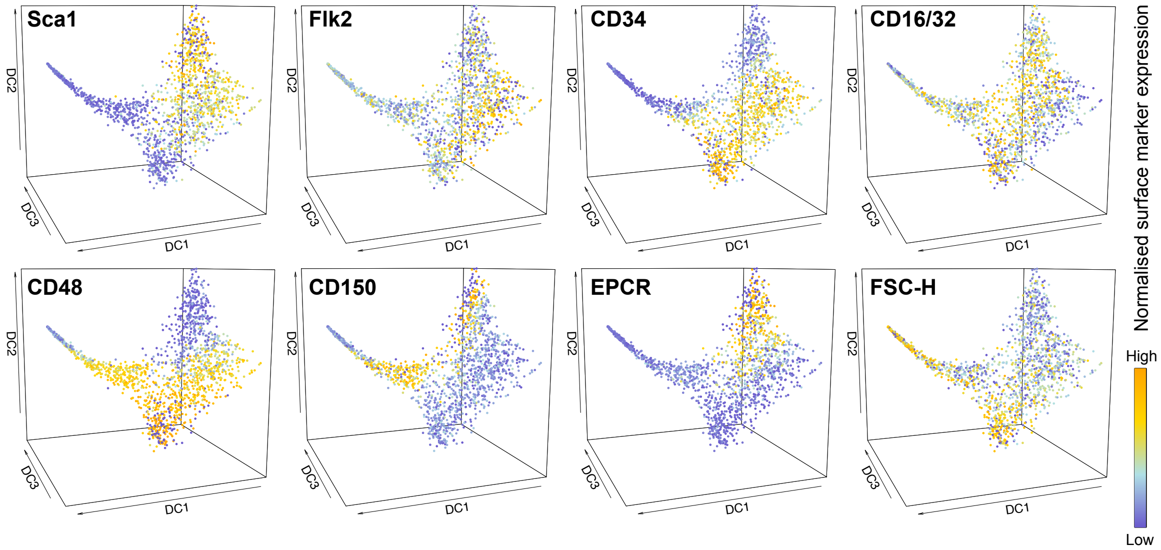


Fig. 3.5. Cell surface markers show expression patterns across the transcriptional landscape. Diffusion map coloured according to index-sorting data for each cell. CD48, CD150 and EPCR were not used for sorting cells. Colour corresponds to normalised expression values ranging between the minimum and maximum value for each marker. Flow cytometry data were normalised across two different sort days. FSC-H, forward-scattered light-height; DC, diffusion component.

3.6 Single-cell profiles can be ordered along differentiation trajectories

Both clustering and dimensionality reduction of the scRNA-seq profiles revealed differentiation of cells towards mature blood lineages. The next step was to investigate whether the recently described concept of pseudotime ordering could be used to capture gene and protein changes during differentiation (Bendall et al., 2014; Trapnell et al., 2014). As the diffusion map embedding captured three lineage branches, differentiation trajectories were identified from HSCs towards erythroid (E), granulocyte-macrophage (GM) and lymphoid (L) lineages (Fig. 3.7A). By ordering cells along these three trajectories, it was possible to visualise how the index-sorting parameters changed along pseudotime (Fig. 3.7B), with several of the surface proteins displaying clear dynamics throughout differentiation. In particular, EPCR, which was not used for isolation of cells, showed a decrease along all three trajectories.

Using the pseudotime orderings, sets of genes showing either up- or downregulation during differentiation were identified for each trajectory (Fig. 3.8A). Gene set enrichment analysis

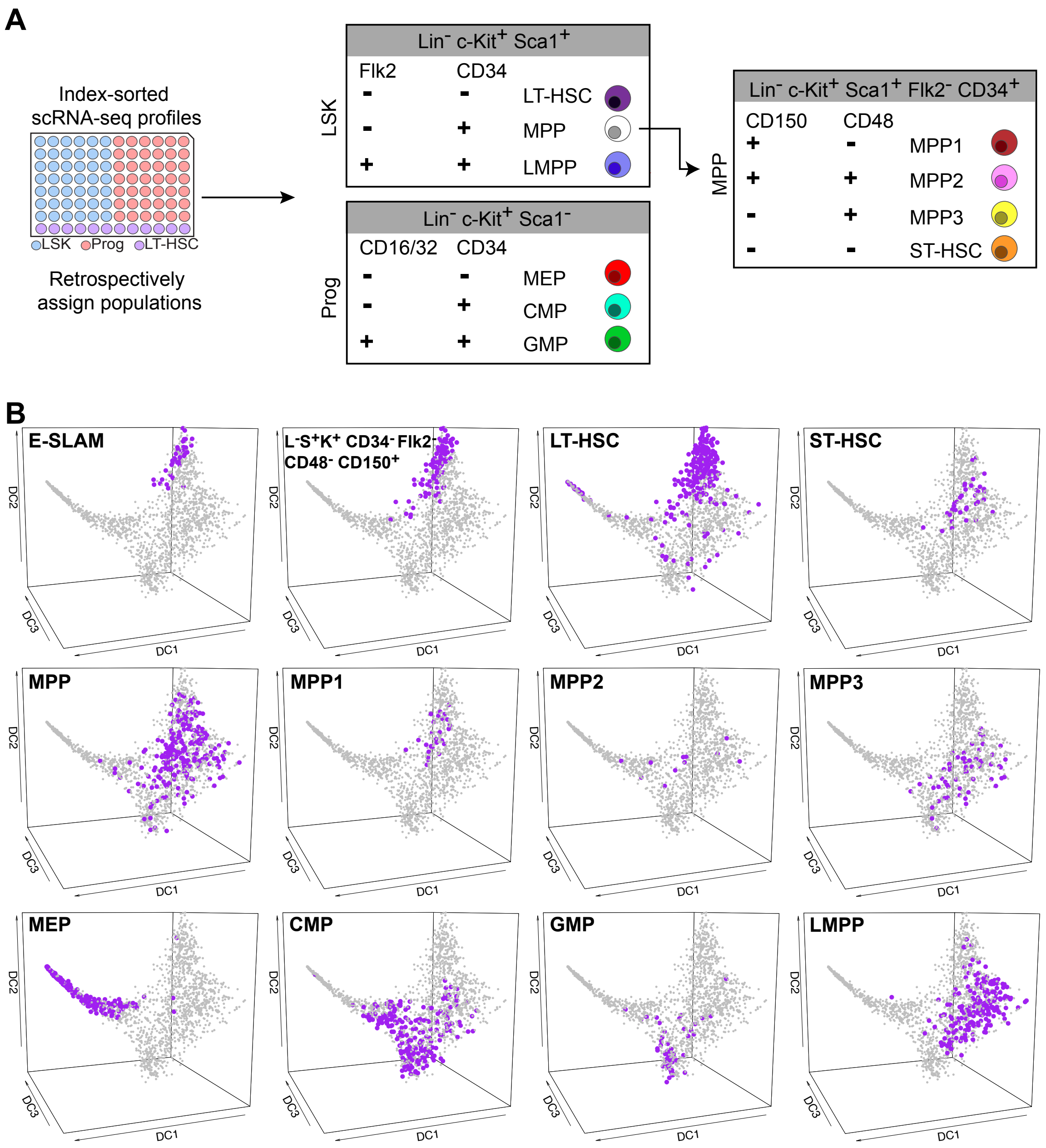


Fig. 3.6. Index-sorting allows scRNA-seq profiles to be retrospectively assigned to haematopoietic progenitor populations. (A) Strategy used for retrospective gating of cells. (B) Diffusion map plots coloured by cell type assigned by the retrospective gating. Each population of interest is highlighted in purple in an individual panel, with all other cells in grey. DC, diffusion component.

was performed to investigate processes related to these genes. Fig. 3.8B displays significant terms related to each set of genes. Many enrichment terms were consistent with their

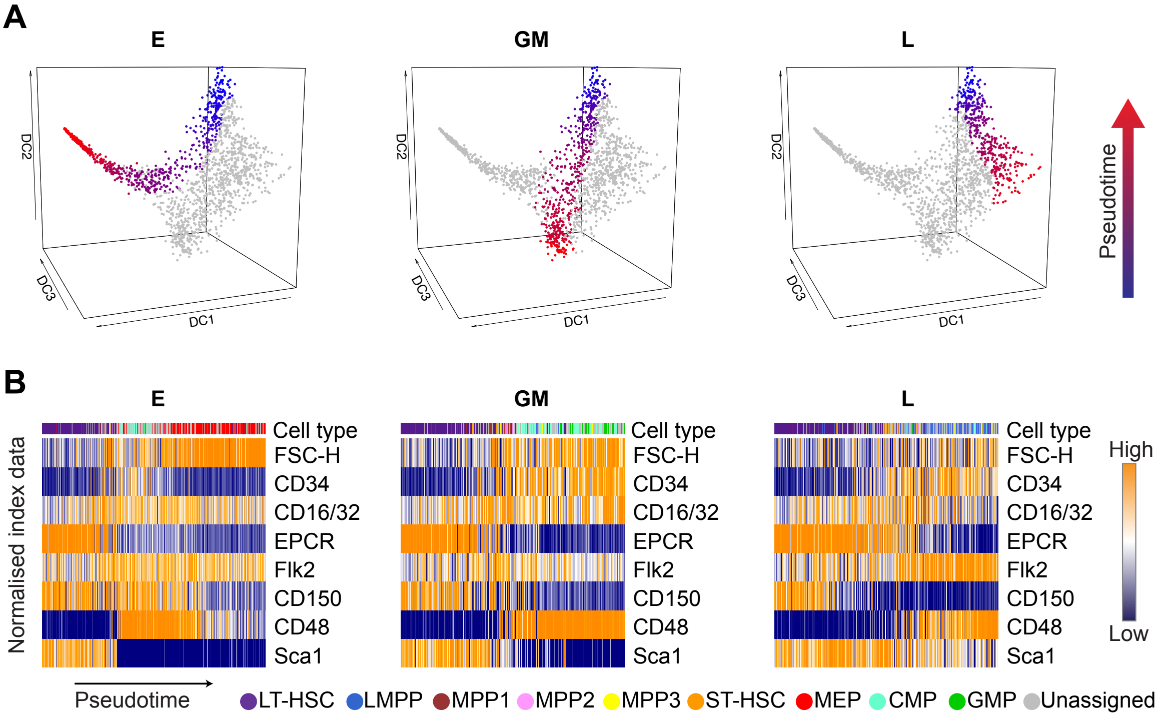
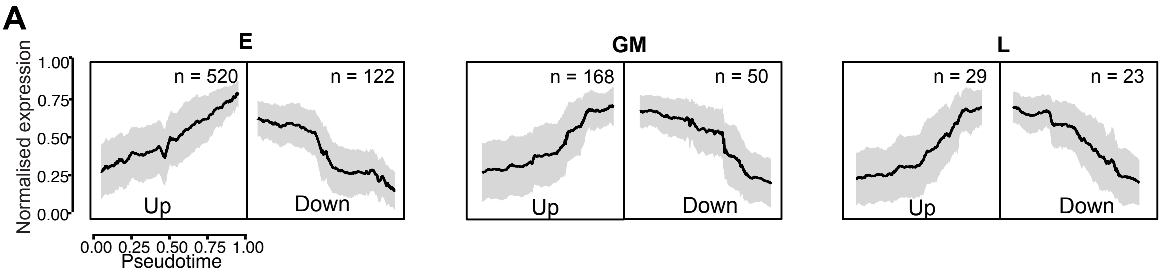


Fig. 3.7. Single-cell molecular profiles can be computationally ordered towards three main blood lineages. (A) Diffusion map highlighting cells in three differentiation trajectories. Cells were ordered from HSCs along erythroid (E), granulocyte-macrophage (GM) and lymphoid (L) trajectories. The pseudotime value for each cell indicates its position in the differentiation trajectory, with blue early in pseudotime and red late in pseudotime. DC, diffusion component. (B) Dynamics of surface marker expression and FSC-H throughout pseudotime. Index data is scaled so that each variable ranges from 0 (low) to 1 (high) in each trajectory. The cell type bar along the top of each heatmap indicates the retrospective gating for each cell. Unassigned cells (in grey) were cells that fell in between the narrow gates drawn for retrospective gating (see Chapter 2 for details).

corresponding trajectory, such as *Neutrophil degranulation* for the genes upregulated along the GM trajectory, and *Megakaryocyte erythrocyte progenitor* for genes upregulated along the E trajectory. Interestingly, both the E and GM upregulated genes had highly significant cell cycle-related terms, which was not the case for genes increasing in expression along the L trajectory.



B

Gene set enrichment analysis

Category	E up	E down	GM up	GM down	L up	L down
Biological processes	Mitotic cell cycle 1.3×10^{-16} Tetrapyrrole biosynthetic process 2.6×10^{-4}	TGFβ receptor signaling pathway 4.2×10^{-5}	DNA replication 2.3×10^{-15} Neutrophil degranulation 6.7×10^{-8}	No significant terms	Hemopoiesis 0.048 B cell lineage commitment 0.048	Endosome to melanosome transport 0.035
Molecular function	Double-stranded RNA binding 5.3×10^{-10}	No significant terms	mRNA binding 1.1×10^{-3}	No significant terms	No significant terms	No significant terms
MGI mammalian phenotype	Abnormal erythrocyte morphology 1.3×10^{-13}	Decreased CD4+ αβ T cell number 3.2×10^{-6}	Abnormal neutrophil physiology 2.9×10^{-4}	Thrombocytopenia 5.9×10^{-4}	Decreased T cell number 0.014	Thrombocytopenia 1.7×10^{-4}
Reactome (Pathways)	Mitotic cell cycle 6.0×10^{-53}	Hemostasis 2.2×10^{-6}	Cell cycle 3.5×10^{-21}	Hemostasis 0.010	No significant terms	Hemostasis 1.7×10^{-4}
Cell types (Mouse gene atlas)	Megakaryocyte erythrocyte progenitor 4.6×10^{-15}	Mast cells 2.3×10^{-4}	Granulocyte monocyte progenitor 4.9×10^{-7}	Hematopoietic stem cells 4.1×10^{-3}	No significant terms	Mast cells 3.0×10^{-5}

Terms shown along with adjusted p-values (Benjamini-Hochberg method for correction for multiple hypotheses testing)

Fig. 3.8. Computational ordering reveals genes dynamically expressed during differentiation. (A) Normalised expression of genes positively (up) or negatively (down) correlated with the pseudotemporal ordering for E, GM and L trajectories. Mean normalised expression (black line) is plotted with ± standard deviation shown by shaded grey regions. For each plot, *n* indicates the number of genes in up or down groups. Gene expression was smoothed by a sliding window of size 20 along the pseudotime ordering. (B) Results of enrichment analysis on the above gene sets. Significant terms are shown along with their adjusted p-value (Benjamini-Hochberg method for correction for multiple hypothesis testing).

3.7 Single-cell data give insight into cell cycle activation during differentiation

As genes upregulated during E and GM differentiation were seen to be significantly related to the cell cycle, it was next decided to investigate the cell cycle status of the stem and progenitor cells. Single-cell transcriptional profiles were assigned to either G_0/G_1, G_2/M or S phases based on their gene expression, by using the method of Scialdone et al. (2015). Calculated for each of the 1,654 profiles, the assigned cell cycle categories were consistent with the

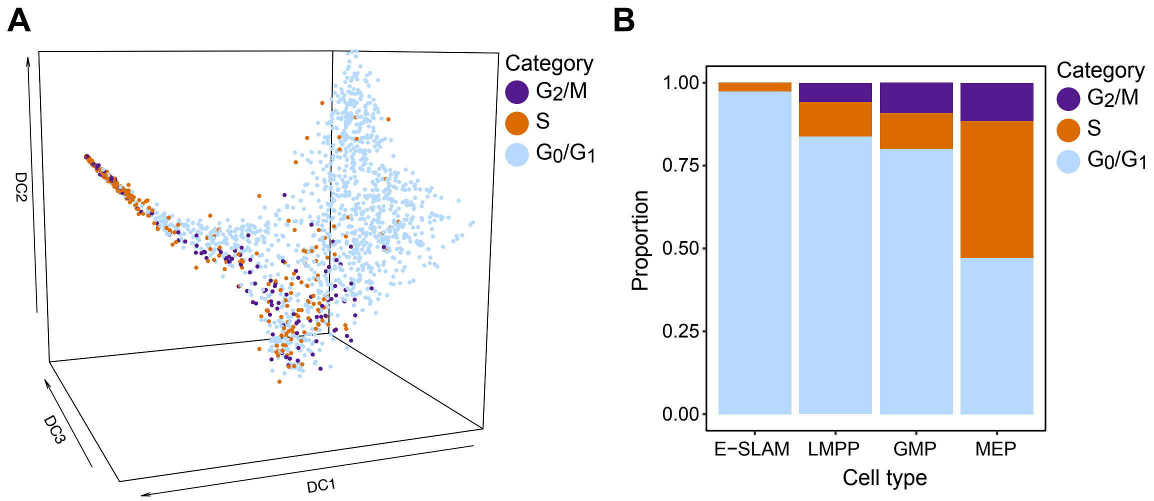
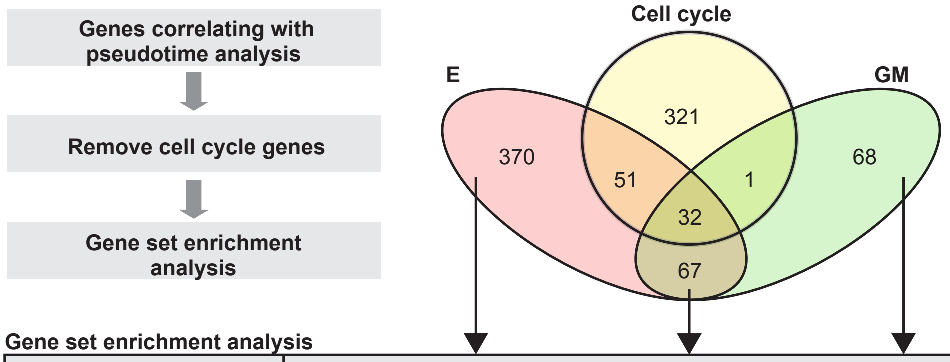


Fig. 3.9. Computational cell cycle assignment highlights changes in cell cycle state along erythroid and granulocyte-macrophage trajectories. (A) Diffusion map with cells coloured by computationally assigned cell cycle category. Cells were assigned to G_0/G_1, S or G_2/M categories based on their transcriptional profiles. DC, diffusion component. (B) Proportion of E-SLAM, LMPP, GMP and MEP cells assigned to each cell cycle category.

observations from the gene set enrichment analysis, with the highest proportion of S and G_2/M category cells in regions corresponding to the E and GM trajectories (Fig. 3.9). Along the L trajectory the vast majority of cells were assigned to the G_0/G_1 category, suggesting that transcriptional diversification and lineage specification occurs prior to widespread cell cycle activation in this lineage.

Further gene set enrichment analysis was then performed to investigate which processes, other than cell cycle, were linked to E and GM differentiation. The upregulated E and GM gene lists were filtered by removing genes overlapping with a curated set of 405 cell cycle genes, and then enrichment analysis was performed. Genes unique to either the E or GM trajectory were annotated with terms related to consistent biological functions such as *Neutrophil degranulation* (GM only). In the Reactome Pathways category, upregulated genes common to both trajectories showed enrichment for terms related to mitochondrial adesnosine triphosphate production, which is compatible with the idea that HSC energy production primarily comes from glycolysis (Simsek et al., 2010; Suda et al., 2011; Takubo et al., 2013), but cells later switch to mitochondrial oxidative phosphorylation as a means of energy production. Mitochondrial oxidative phosphorylation may not be suitable for HSCs as it generates intermediate free oxygen radicals that can damage DNA, potentially


Genes correlating with pseudotime analysis
Remove cell cycle genes
Gene set enrichment analysis
Cell cycle
E
GM
321
370
51
1
68
32
67

Gene set enrichment analysis

Category	E only	E & GM	GM only
Biological processes	**Heme O biosynthetic process** 1.0 x 10^{-6}	**Maintenance of DNA methylation** 1.8 x 10^{-3}	**Neutrophil degranulation** 2.4 x 10^{-12}
Molecular function	**RNA binding** 4.2 x 10^{-7}	**mRNA binding** 1.5 x 10^{-5}	**Protease binding** 0.011
MGI mammalian phenotype	**Reticulocytosis** 4.5 x 10^{-9}	**No significant terms**	**Abnormal neutrophil physiology** 3.6 x 10^{-7}
Reactome (Pathways)	**Folding of actin by CCT/TriC** 3.7 x 10^{-10}	**Metabolism of nucleotides** 6.5 x 10^{-5} **Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins** 0.014	**Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell** 0.032
Cell types (Mouse gene atlas)	**Megakaryocyte erythrocyte progenitor** 3.0 x 10^{-12}	**Bone marrow** 4.5 x 10^{-5}	**Granulocyte monocyte progenitor** 9.0 x 10^{-6}

Terms shown with adjusted p-values (Benjamini-Hochberg method for correction for multiple hypotheses testing)

Fig. 3.10. Excluding cell cycle genes from gene set enrichment analysis gives insight into other processes occurring during differentiation. Cell cycle annotated genes were removed from the gene sets upregulated during E and GM differentiation shown in Fig. 3.8. Gene set enrichment analysis was performed on genes unique to the E trajectory, unique to the GM trajectory, and those shared between both trajectories after the removal of the cell cycle related genes. Terms with adjusted p-value < 0.05 (Benjamini-Hochberg correction for multiple testing) were considered significant.

causing problems in HSCs as they can reside in the bone marrow for a long time (Yu et al., 2013).

Motivated by these observations, the pseudotime dynamics of hydrogen ion transmembrane transport genes and cell cycle genes was investigated (Fig. 3.11). The expression of these gene sets increased along both GM and E trajectories, but showed no substantial increase along the L trajectory. These genes start increasing in the shared E and GM part of the trajectory, indicating a change in these processes occurs before the decision towards either erythroid or granulocyte-macrophage fate is made.

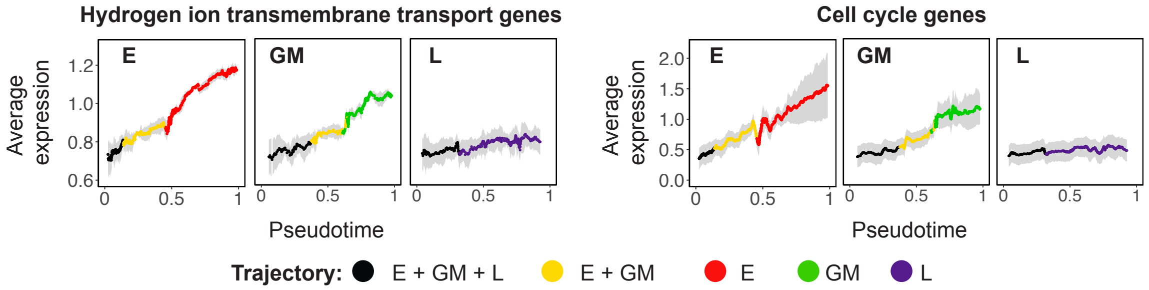


Fig. 3.11. Specific gene modules share upregulation along E and GM trajectories. Average expression of hydrogen ion transmembrane transport genes and cell cycle genes along pseudotime trajectories. Each gene was normalised by the median of all 3 trajectories for plotting. The average expression is coloured by the stage of the trajectory, and means are shown with ± standard deviation in grey.

3.8 Single-cell data relate RNA content to differentiation

An essential step in processing gene expression data is applying a normalisation method to account for differences in sequencing depth between samples of cells. Several approaches for normalising raw sequencing counts exist, and a number of these have been applied to single-cell data. For the analysis in Fig. 3.2-3.11 the commonly used approach of size factor normalisation was applied to the data. This aims to adjust for the amount of starting material that was sequenced in each cell. However, HSCs are known to be smaller and generally more quiescent than progenitor cells such as MEPs, and so are expected to contain less mRNA. The previous analysis identified sets of genes downregulated during differentiation (Fig. 3.8), but differences in mRNA content between HSCs and more mature cells raised the possibility that a decrease in the size factor-normalised expression values for a gene along pseudotime might not correspond to a decrease in the absolute number of mRNA molecules.

To address this question an alternative normalisation method was devised. All cells had been processed for sequencing with synthetic transcripts (ERCC spike-ins) added to each well at equal concentrations. Ideally, spike-in levels in each sample should exactly correspond to the sequencing depth for that cell. However, it was clear that the spike-in concentrations in this experiment were not constant across all lanes (Fig. 3.12A), but instead exhibited a batch effect linked to the day that plates were prepared for sequencing. Therefore, cells were

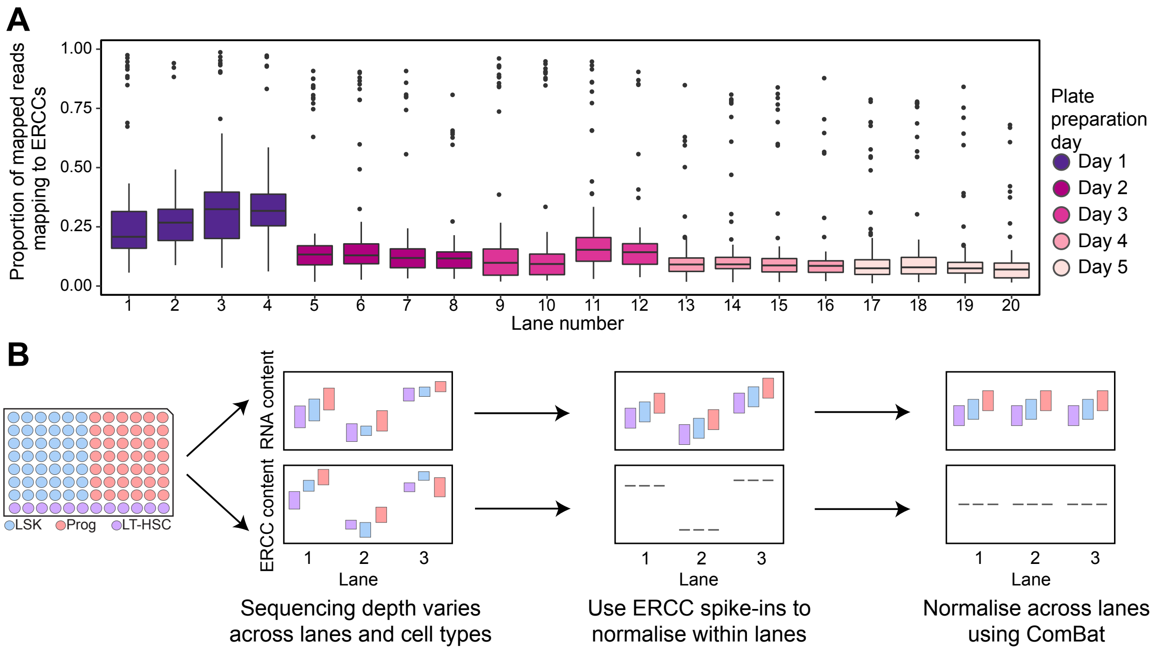


Fig. 3.12. Synthetic spike-ins can be used to estimate RNA content in cells. (A) Proportion of mapped reads mapping to ERCC spike-ins ordered by sequencing lane, with one 96 well plate of cells per lane. Colours indicate the plate preparation day. Boxes summarise the upper and lower quartiles and median. Whiskers extend to maximum/minimum values or upper/lower quartiles ± 1.5 × interquartile range. Points outside this range are shown as outliers. (B) Schematic showing overview of alternative normalisation method. A mixture of cells from different sorting gates were present in each lane in the plate layout shown in the left of the panel. ERCC content per cell can be used to normalise expression values within lanes, and then batch correction adjusts the normalisation across lanes.

first normalised by ERCC spike-in content, and then these normalised values were adjusted for batch effect across the different sequencing lanes (Fig. 3.12B). Each plate contained a mixture of LT-HSC, LSK and Prog cells, and this plate design was vital in making this normalisation possible. Estimates of mRNA content based on this normalisation suggested that cells on the E trajectory had the highest mRNA levels, followed by those in the GM trajectory (Fig. 3.13A). This showed a similar trend to the distribution of FSC-H values, which are an indicator of cell size (Fig. 3.13B-C).

These ERCC-normalised values were then used to investigate whether the absolute expression of genes downregulated in pseudotime actually decreased during E and GM differentiation. Gene lists from Fig. 3.10 were filtered by fold-change in expression difference between the start and end of the pseudotime trajectory to identify those genes showing strong absolute downregulation. The majority of genes were confirmed to have reduced absolute expression,

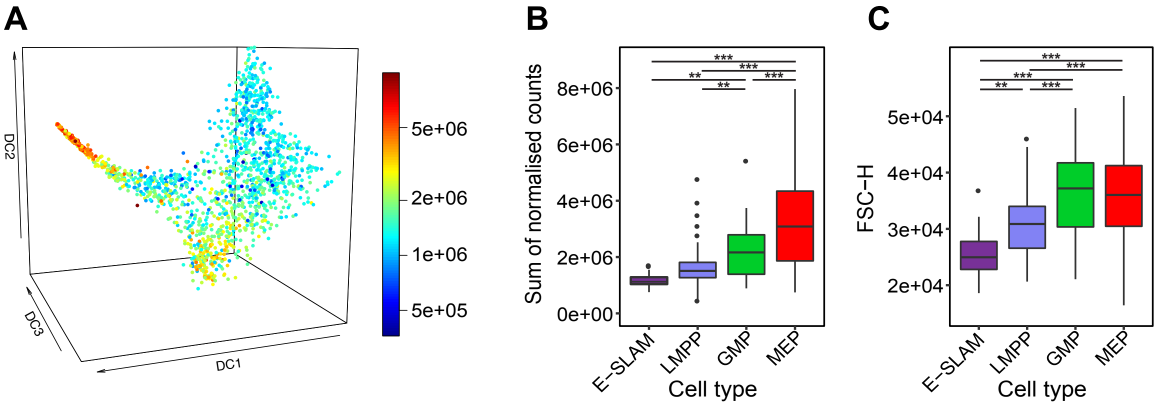
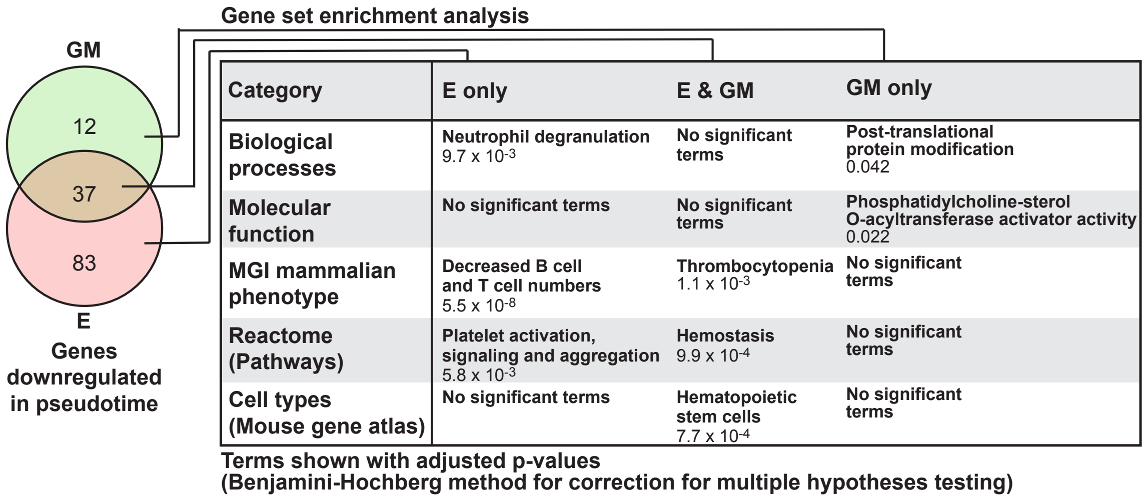


Fig. 3.13. Cells increase in absolute RNA content during differentiation. (A) Diffusion map coloured by estimated RNA content for each cell. RNA content was estimated by summing the ERCC-normalised counts per cell. (B) Sum of normalised counts across different cell types. Significances between cell types were calculated using a one-way analysis of variance test (**, $p < 0.001$; ***, $p < 0.0001$). (C) FSC-H values across different cell types from index sorting data. Significance was calculated using a one-way analysis of variance test (**, $p < 0.001$; ***, $p < 0.0001$). Boxes summarise the upper and lower quartiles and median. Whiskers extend to maximum/minimum values or upper/lower quartiles ± 1.5 × interquartile range. Points outside this range are shown as outliers.

with 120/122 genes for the E trajectory and 49/50 genes for the GM trajectory showing clear downregulation in absolute terms along these trajectories (Fig. 3.14). Genes downregulated along both E and GM trajectories showed enrichment for terms associated with megakaryopoiesis, due to downregulation of genes such as *Procr* and *Mpl*, which are known to have high expression in HSCs as well as megakaryocytes. Terms associated with downregulation along the E trajectory were related to the immune response. The downregulated GM genes did not show highly significant terms as there were only 12 genes in the set. These data demonstrate that single-cell analysis allows identification of genes that are more highly expressed in real terms in individual HSCs than in their downstream progenitors.

3.9 Conclusions

This chapter describes the analysis of scRNA-seq data for 1,654 individual haematopoietic stem and progenitor cells from mouse bone marrow, providing a comprehensive view of the transcriptional landscape of the upper tiers of haematopoiesis. Prior to this work, previous studies had not achieved coverage of such a range of haematopoietic cell types at such high resolution, in terms of both cell and gene numbers. One of the biggest limitations on the


Gene set enrichment analysis
GM
12
37
83
E
Genes downregulated in pseudotime

Category	E only	E & GM	GM only
Biological processes	Neutrophil degranulation 9.7 x 10^{-3}	No significant terms	Post-translational protein modification 0.042
Molecular function	No significant terms	No significant terms	Phosphatidylcholine-sterol O-acyltransferase activator activity 0.022
MGI mammalian phenotype	Decreased B cell and T cell numbers 5.5 x 10^{-8}	Thrombocytopenia 1.1 x 10^{-3}	No significant terms
Reactome (Pathways)	Platelet activation, signaling and aggregation 5.8 x 10^{-3}	Hemostasis 9.9 x 10^{-4}	No significant terms
Cell types (Mouse gene atlas)	No significant terms	Hematopoietic stem cells 7.7 x 10^{-4}	No significant terms

Terms shown with adjusted p-values
(Benjamini-Hochberg method for correction for multiple hypotheses testing)

Fig. 3.14. The majority of genes downregulated in pseudotime show downregulation in absolute terms. Table displaying the most relevant significant terms from gene enrichment expression analysis on genes downregulated in absolute terms in E-only, GM-only, and shared between E and GM trajectories. The numbers of genes showing downregulation in absolute terms are displayed in the Venn diagram. Terms with an adjusted p-value < 0.05 (using Benjamini-Hochberg correction for multiple testing) were considered significant.

number of cells profiled by scRNA-seq in an experiment comes from the substantial cost of single-cell sequencing. Compared to plate-based scRNA-seq, such as the SMART-Seq2 technology used in this work, droplet-based sequencing approaches (Klein et al., 2015; Macosko et al., 2015) or MARS-Seq (Jaitin et al., 2014; Paul et al., 2015) vastly increased the number of cells that could be profiled for a single experiment. These techniques were not readily available at the time the data presented in the chapter were generated, and the increase in cell number comes at the cost of much lower sequencing depth for each cell. It was therefore decided that using SMART-Seq2 would be more useful in generating a reference dataset. The sequencing data in this study detected on average over 8,000 genes per cell, achieving substantial depth. In addition, emerging droplet-based methods did not enable surface marker levels to be simultaneously captured along with transcriptional profiles, meaning that retrospective gating as described here would not have been possible. This would have made the dataset less useful as a resource to the haematopoietic community.

3.9.1 Heterogeneity in haematopoietic populations

Although cells were captured in three sorting gates, the unbiased clustering partitioned cells into seven groups, with cells from each gate spread across the clusters. Progenitor cells mostly separated into erythroid, megakaryocyte, and granulocyte-monocyte clusters. It was also interesting to see that clustering was able to identify a cluster comprised mainly of LT-HSCs, either from the LT-HSC gate or from the LSK gate. A number of genes were identified as differentially expressed between these clusters, helping to explain some of the heterogeneity in the haematopoietic compartment.

However, although clustering revealed transcriptionally distinct groups of cells within the bone marrow, the application of diffusion maps emphasised that the data could also be represented as part of a continuous differentiation landscape. With "snapshot" data sampling differentiating cells this type of continuous representation may be more appropriate, as cells are not isolated in discrete groups and instead represent different stages of a continuous process. Indeed some studies have presented evidence for a continuous structure of haematopoietic differentiation (Velten et al., 2017). The structure realised by the diffusion map was supported by the fact that previously defined haematopoietic populations were restricted to clearly visible regions in the low-dimensional space. The exception to this was the CMP population, which has been suggested to be mainly formed from erythroid or myeloid committed cells (Paul et al., 2015; Perié et al., 2015). The diffusion map supports the idea of CMPs as a highly heterogeneous population, and in particular the overlap of this population with the GMP and MEP regions would agree with a lack of true erythroid-granulocyte-monocyte progenitors in this sorting gate. Capturing the underlying transcriptional structure of haematopoiesis allowed the inference of pseudotime differentiation trajectories through the data from stem cells to erythroid, granulocyte-monocyte, and lymphoid lineages. This type of landscape representation is also useful in representing a reference landscape with which to compare independent datasets.

3.9.2 Gene expression changes throughout differentiation

As well as allowing cells to be grouped and ordered based on their gene expression, transcriptional profiling also enables the investigation of exactly which genes change between these groups or along these orderings. Here, this gave insight into differences in cell cycle behaviour during erythroid, myeloid and lymphoid differentiation. In particular, differentiation towards the lymphoid lineage was observed before an increase in cell cycle activity, suggesting that here cell fate specification occurs independently from cell cycling. By profiling cells at single-cell resolution we can also obtain insight into properties that would be obscured by population level analysis. For example, the alternative ERCC-based normalisation allowed absolute RNA levels to be estimated in each cell. This analysis indicated that E and GM cells contained more mRNA, in agreement with known low *Myc* expression in HSCs (Guo et al., 2009; Laurenti et al., 2008; Wilson et al., 2004), which could be linked with low levels of transcription in keeping with the quiesence (Wilson et al., 2008) and low metabolic activity (Suda et al., 2011; Yu et al., 2013) of these cells. Yet even with this absolute normalisation, it was still possible to find genes with higher absolute expression in HSCs than in more transcriptionally active cells, suggesting that some of these genes may be integral in maintaining haematopoietic stem cell function.

3.9.3 Future directions

One limitation of this work is that the diffusion map analysis focused only on differentiation towards the three major blood lineages (E, GM and L), yet assigning cells to only three trajectories does not fully represent the range of cell types present in the blood. For example, the unbiased clustering identified a group of megakaryocyte progenitors but the pseudotime ordering did not construct a separate trajectory towards these cells. Granulocyte-macrophage progenitors are also treated as one group of cells, although specification towards these two lineages can been seen in the bone marrow at the transcriptional level (Olsson et al., 2016). Therefore, whilst this work represents a useful first step in the analysis of these data, further exploration of pseudotemporal ordering towards more specific blood lineages will be necessary, and forms the starting point for the work discussed in Chapter 6.

Another interesting question is whether single-cell data can be used to go further than identifying dynamic genes along differentiation trajectories, and allow us to investigate the regulatory relationships controlling fate decisions. These ideas motivate the work presented in Chapter 4.

3.9.4 Summary

In summary, the work in this chapter describes the analysis of a scRNA-seq dataset to construct a single-cell reference map of the transcriptional landscape of haematopoiesis.

These data demonstrate how single-cell profiling can give insight into how properties such as gene expression and cell cycle change throughout differentiation, and are freely available as a resource to the community.

Chapter 4

Reconstructing stem cell regulatory network models from single-cell molecular profiles

Parts of this section have been modified from Hamey et al. (2017). Experimental work and some of the subsequent analysis was carried out by Sonia Nestorowa (single-cell qRT-PCR profiling of primary bone marrow cells, quality control and normalisation of single-cell expression data, analysis of ChIP-Seq data, luciferase assays), Nicola Wilson (single-cell qRT-PCR profiling of primary bone marrow and HoxB8-FL cells, generation and analysis of ChIP-Seq data), Sarah Kinston (luciferase assays), Rebecca Hannah (alignment and analysis of ChIP-Seq data) and David Kent (isolation of primary bone marrow cells for qRT-PCR). Computational analysis, including dimensionality reduction, developing the network inference method, and *in silico* modelling of networks was carried out by F. Hamey.

4.1 Background

In blood diseases such as leukaemia, processes regulating the production of haematopoietic cells are often dysregulated, leading to an imbalance in mature cell types. Identifying how blood stem and progenitor cells decide between alternative fates during differentiation is therefore an important part of understanding blood disorders. Transcriptional regulation is one process with a role in determining cell fate decisions and controlling differentiation in the blood (Göttgens, 2015), with transcription factors acting as part of regulatory networks to govern gene expression in cells (Peter and Davidson, 2015). Some early attempts at modelling transcriptional regulation in the blood have used literature-curated regulatory relationships to construct models (Bonzanni et al., 2013; Chickarmane et al., 2009; Krumsiek et al., 2011). One disadvantage of such studies is that they can be limited in their ability to discover new regulatory relationships, as they heavily rely on having good prior knowledge of the system. Other work has used bulk expression data combined with experimental perturbations to identify transcriptional regulation. However, the power of single-cell data in uncovering regulatory relationships is increasingly being recognised. Both Moignard et al. (2013) and Pina et al. (2015) used correlation analysis of single-cell qRT-PCR data to identify novel regulatory relationships between pairs of transcription factors. In systems other than adult haematopoiesis, single-cell expression data have been used to infer Boolean network models in embryonic stem cells (Xu et al., 2014) and embryonic blood development (Moignard et al., 2015).

For reliable network inference based on single-cell data it is important to both profile large numbers of cells and to have confidence in the measured gene expression values. An existing dataset from the Göttgens lab used single-cell qRT-PCR to quantify the expression of 48 genes, including 33 transcription factors, in over 1,600 HSPCs. These included cells from four different HSC sorting strategies, along with cells from ST-HSC and four progenitor populations, therefore capturing a large number of cells in different states from across haematopoietic differentiation.

This chapter describes a transcriptional regulatory network inference method applied to these single-cell haematopoietic data. The method uses correlation between transcription factors combined with pseudotime ordering to infer regulatory relationships, and uncovers differences in regulation between differentiation towards two blood cell types.

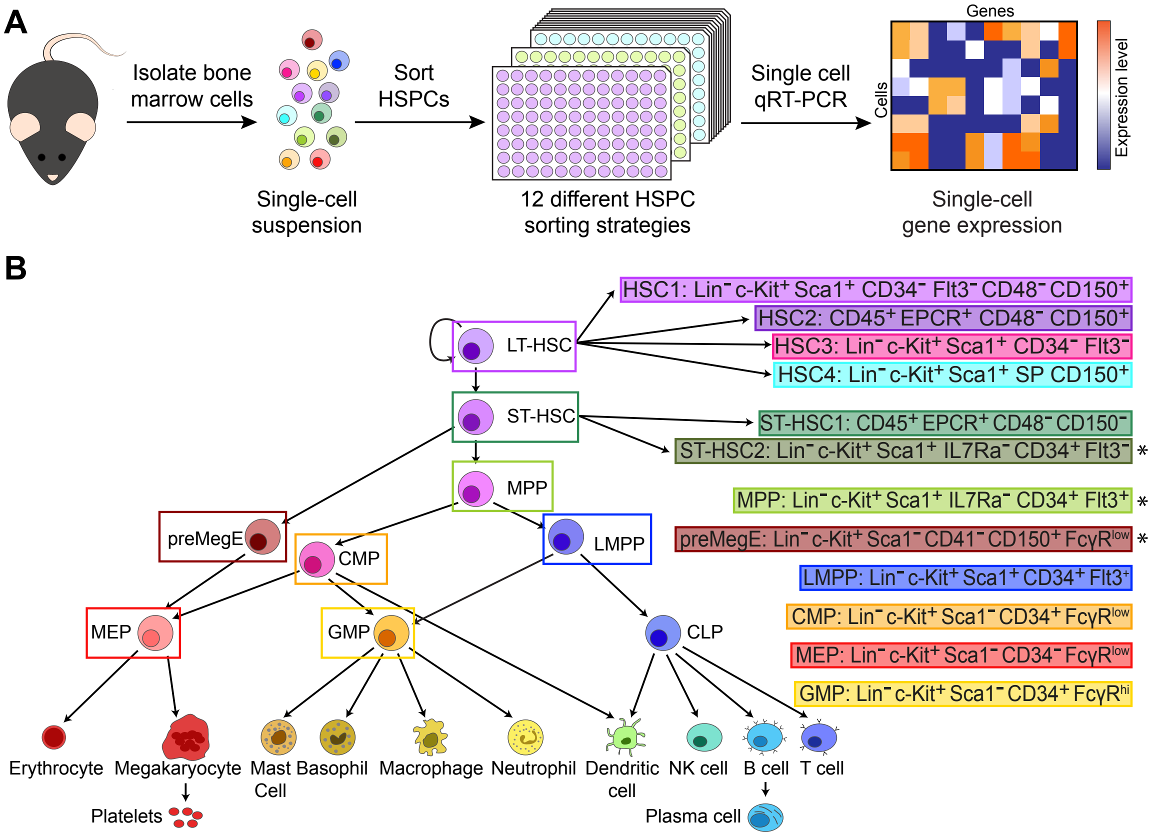

A
Isolate bone marrow cells
Single-cell suspension
Sort HSPCs
12 different HSPC sorting strategies
Single cell qRT-PCR
Genes
Cells
Expression level
Single-cell gene expression
B
LT-HSC
ST-HSC
MPP
preMegE
CMP
LMPP
MEP
GMP
CLP
HSC1: Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^-$ Flt3$^-$ CD48$^-$ CD150$^+$
HSC2: CD45$^+$ EPCR$^+$ CD48$^-$ CD150$^+$
HSC3: Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^-$ Flt3$^-$
HSC4: Lin$^-$ c-Kit$^+$ Sca1$^+$ SP CD150$^+$
ST-HSC1: CD45$^+$ EPCR$^+$ CD48$^-$ CD150$^-$
ST-HSC2: Lin$^-$ c-Kit$^+$ Sca1$^+$ IL7Ra$^-$ CD34$^+$ Flt3$^-$ *
MPP: Lin$^-$ c-Kit$^+$ Sca1$^+$ IL7Ra$^-$ CD34$^+$ Flt3$^+$ *
preMegE: Lin$^-$ c-Kit$^+$ Sca1$^-$ CD41$^-$ CD150$^+$ FcγR^{low} *
LMPP: Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^+$ Flt3$^+$
CMP: Lin$^-$ c-Kit$^+$ Sca1$^-$ CD34$^+$ FcγR^{low}
MEP: Lin$^-$ c-Kit$^+$ Sca1$^-$ CD34$^-$ FcγR^{low}
GMP: Lin$^-$ c-Kit$^+$ Sca1$^-$ CD34$^+$ FcγR^{hi}
Erythrocyte
Megakaryocyte
Platelets
Mast Cell
Basophil
Macrophage
Neutrophil
Dendritic cell
NK cell
B cell
T cell
Plasma cell

Fig. 4.1. Profiling the haematopoietic compartment using single-cell qRT-PCR. (A) Schematic showing the experimental overview. (B) Different sorting strategies used to isolate haematopoietic populations for gene expression profiling. Data from 9 of the sorting strategies were previously published (Wilson et al., 2015), and "*" indicates the three new populations added specifically for the work in this chapter.

4.2 Single-cell qRT-PCR profiling captures the structure of haematopoietic differentiation

A previously published dataset from the Göttgens laboratory used single-cell qRT-PCR to profile gene expression in 1,626 cells from HSC, ST-HSC, GMP, CMP, LMPP, and MEP mouse bone marrow populations (Wilson et al., 2015), with the aim of dissecting heterogeneity in the stem cell compartment. As the selected genes were heavily biased towards transcription factors it was reasoned that these data could be used to learn about haematopoietic regulatory relationships. Because the focus of the original study was to devise a sorting strategy to enrich for functional HSCs, the majority of cells were from the LT-HSC and ST-HSC populations. To ensure that the data captured intermediate populations from differentiation, the first step of this project was to sort and profile additional populations (Fig. 4.1). Cells from the pre-megakaryocyte-erythroid progenitor (preMegE), MPP and an alternative sorting strategy for ST-HSC populations were isolated using FACS and profiled by single-cell qRT-PCR measuring the same panel of 48 genes as in Wilson et al. (2015). Combining all of the data together resulted in 2,167 murine HSPC single-cell profiles, capturing cells at multiple stages of differentiation towards mature blood cell types. The 12 sorting strategies are indicated in Fig. 4.1B. After quality control to remove failed genes, and the exclusion of housekeeping genes, the expression of 41 genes including 31 transcription factor encoding genes was retained.

Dimensionality reduction techniques were applied to establish how well these data could represent the transcriptional landscape of haematopoiesis, and ensure that the new data integrated with the previously published dataset. Diffusion maps, a non-linear dimensionality reduction method recently adapted for use with single-cell data (Coifman et al., 2005; Haghverdi et al., 2015), revealed a low-dimensional landscape consistent with the classic haematopoietic hierarchy (Fig. 4.2A). This structure was supported by t-SNE and PCA embeddings, and all three visualisations demonstrated good integration between the new and old datasets (Fig. 4.2).

4.3 Differentiation trajectories towards two blood lineages can be constructed from single-cell expression profiles

Wilson et al. (2015) identified a subset of HSCs with increased probability of long-term multilineage reconstitution upon single-cell transplantation, which they named the molecular overlapping population (MolO) cells. As well as providing a visualisation of the data, the diffusion map analysis demonstrated that the HSCs separated from progenitor populations (Fig. 4.3A), with the MolO cells occupying a distinct region of the diffusion map. Additionally, MEPs, which generate both megakaryocyte and erythroid cells, and LMPPs, producing both lymphoid and myeloid cells, also occupied separate regions of the diffusion map, with intermediate cell populations present between the HSCs and more mature progenitors. This suggested that the diffusion map representation could help to capture cells on differentiation trajectories from stem cells towards different mature cell types (Fig. 4.3B).

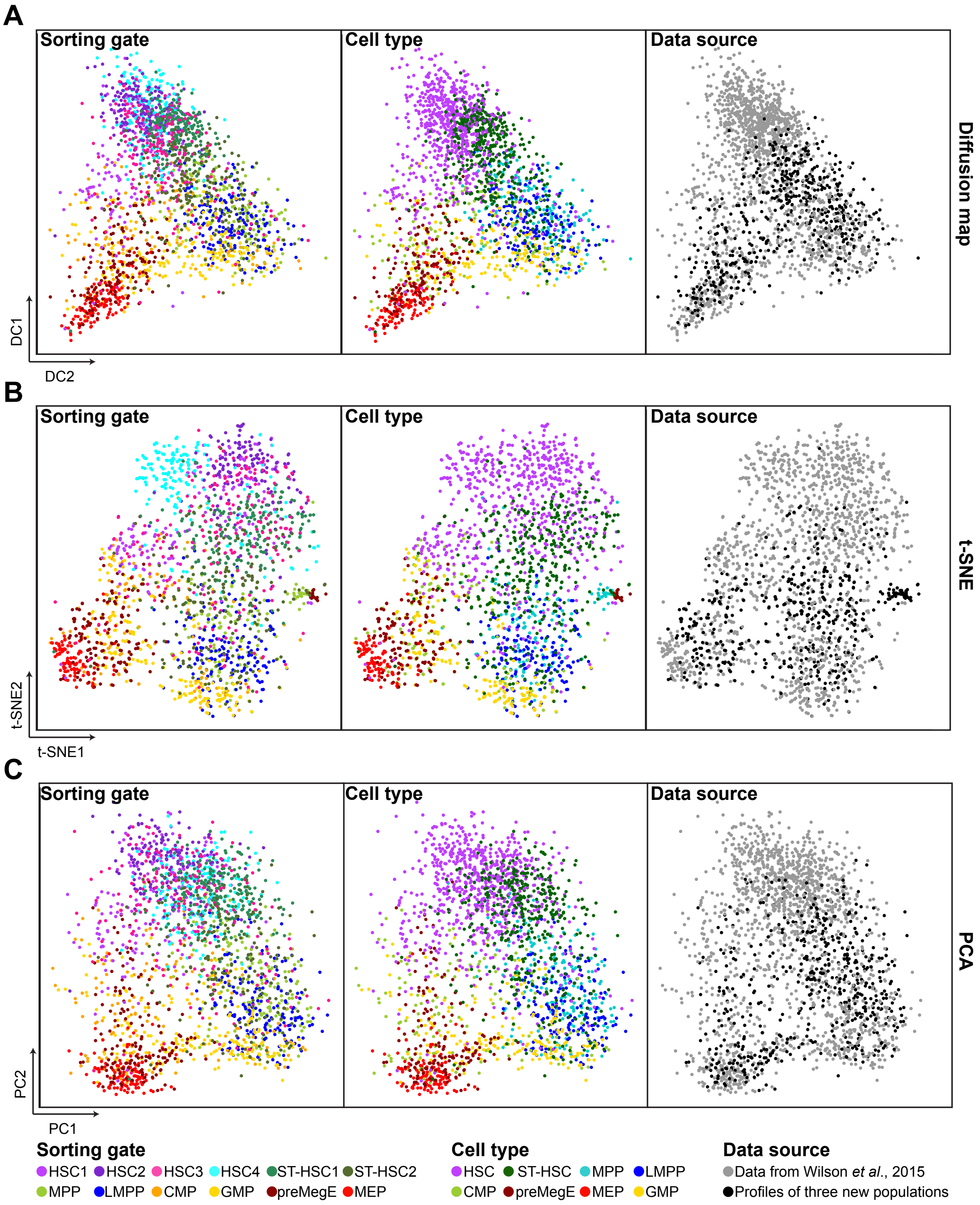


Fig. 4.2. Single-cell expression profiling of haematopoietic genes captures differentiation structure. (A) Diffusion map dimensionality reduction calculated on expression of 41 genes measured by qRT-PCR. Panels from left to right are coloured by sorting gate, phenotypic cell type, and the integration of new data into the data from Wilson et al. (2015). DC, diffusion component. (B) t-distributed stochastic neighbour embedding (t-SNE) showing the same cells, with colours as above. (C) Principal component analysis (PCA) showing the same cells, with colours as above. PC, principal component.

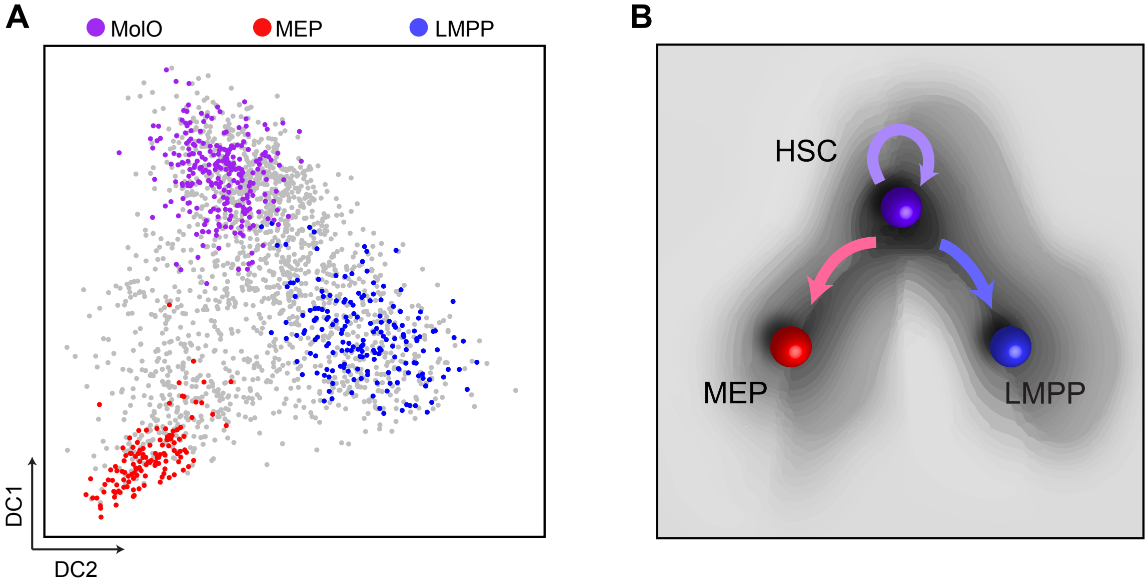


Fig. 4.3. Single-cell gene expression data capture differentiation decision between MEPs and LMPPs. (A) Diffusion map from Fig. 4.2A highlighting cells from three specific populations. Cells outside these populations are coloured grey. MolO, molecular overlapping population from Wilson et al. (2015); MEP, megakaryocyte-erythroid progenitor; LMPP, lymphoid-primed multipotent progenitor; DC, diffusion component. (B) Schematic highlighting the fate decision from stem cells to MEP or LMPP progenitor populations.

Motivated in part by the diffusion map structure, the next aim was to use the single-cell data to understand how transcription factor expression varies throughout differentiation. Coordinates in the diffusion map space were used to identify cells on differentiation branches from HSCs to MEPs and from HSCs to LMPPs, following the method of Ocone et al. (2015). As shown in Fig. 4.4A, this assigned cells to two broad differentiation branches, which were then ordered by progress through differentiation using the Wanderlust algorithm (Bendall et al., 2014). The algorithm assigns each cell a pseudotime value, which can then be used to investigate how properties of the cells change throughout differentiation. Several transcription factors displayed strong expression dynamics, often exhibiting different behaviour between the two trajectories (Fig. 4.4B). For example, *Notch* expression increased through the trajectory to LMPPs but was almost entirely unexpressed during differentiation towards MEPs. In contrast, expression of *Gata1* was only activated during differentiation towards MEPs. Other genes, such as *Myb*, demonstrated increased expression along both trajectories.

Calculating pairwise correlation between transcription factor expression in the two trajectories revealed groups of genes with similar expression patterns during differentiation (Fig. 4.5). From this analysis it was apparent that genes with strong correlation in one trajectory could

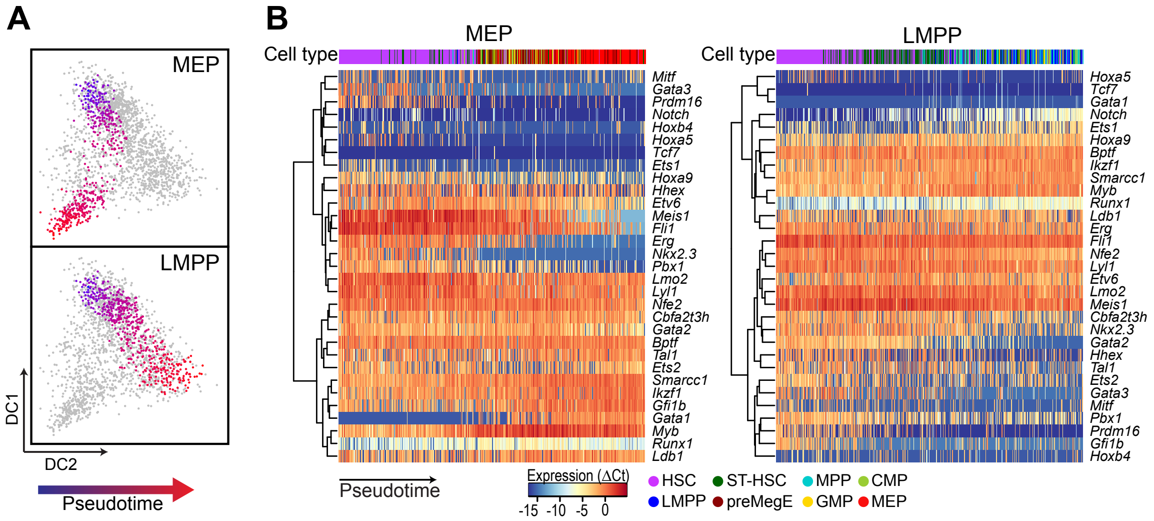


Fig. 4.4. Cells can be computationally ordered along differentiation trajectories towards two progenitor cell types. (A) Differentiation trajectories from stem cells to MEPs (top panel) and to LMPPs (bottom panel) highlighted in the diffusion map. Cells in the trajectory are coloured by pseudotime value ranging from blue (early in pseudotime) to red (late in pseudotime). Cells not included in the trajectory are shown in grey. (B) Heatmaps showing expression of transcription factor encoding genes measured by qRT-PCR. Cells are ordered by pseudotime along the two trajectories, and the dendrogram for each heatmap indicates the results of hierarchical clustering on the genes. Colourbars at the top of each heatmap indicate the cell types along each trajectory, matching the colour key displayed below the heatmaps.

show a very different relationship in the other. For example, in the MEP trajectory *Erg* and *Myb* were highly negatively correlated, but were positively correlated in the LMPP trajectory. This raised the possibility that these data could be used to uncover differences in regulatory relationships between the two trajectories.

4.4 Gene regulatory network models can be inferred from the pseudotime dynamics of single-cell data

Differences in the behaviour of transcription factors between the two trajectories suggested that these data could be used to help understand transcriptional regulation linked to haematopoietic differentiation. One of the challenges in inferring transcriptional regulation from gene expression data is determining the direction of regulation. Correlation based approaches, such as used by Moignard et al. (2013) and Pina et al. (2015), require additional

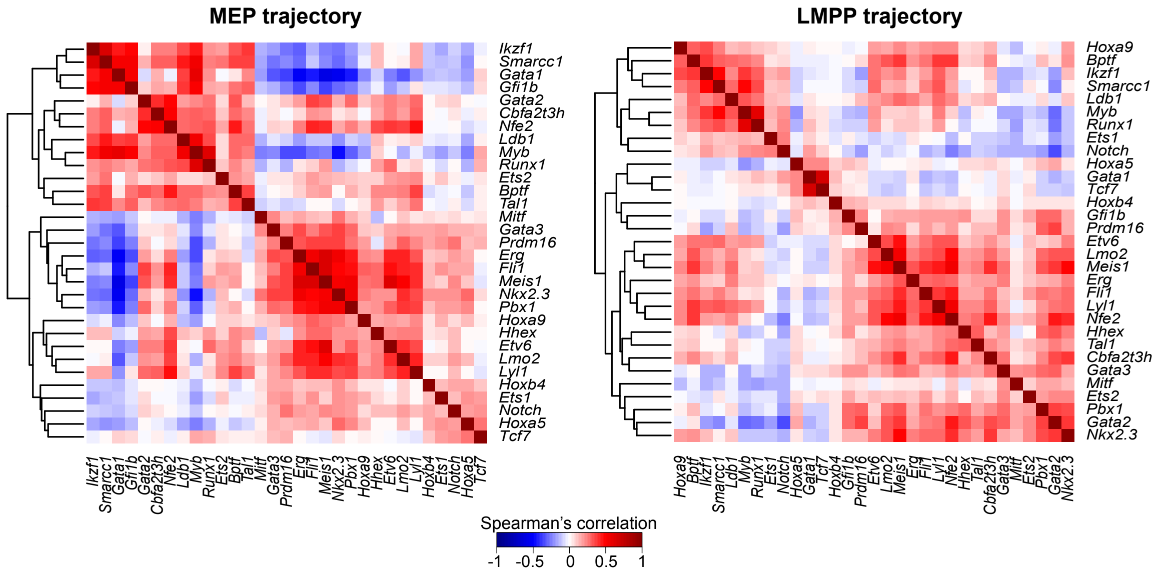


Fig. 4.5. Groups of transcription factors with similar expression patterns change between MEP and LMPP trajectories. Heatmaps showing Spearman's correlation of pairs of genes along the two pseudotime trajectories. Dendrograms indicate results of hierarchical clustering on genes.

information to discover which are the source and target genes in an activating pair. This type of information can be obtained from sources such as transcription factor binding assays, perturbation information, or time course data. Here, it was hypothesised that the dynamics from the pseudotime ordering could be used to identify regulatory relationships between transcription factors.

Studies in HSCs (Bonzanni et al., 2013), embryonic stem cells (Dunn et al., 2014; Xu et al., 2014) and embryonic blood development (Moignard et al., 2015) have all successfully used Boolean abstraction to capture the behaviour of their respective systems by modelling transcriptional regulatory networks. However, there are limitations with Boolean modelling, most notably that gene expression can only take binary values in these models. In an attempt to address this limitation, the decision was made to develop a hybrid network inference method based on information about continuous gene expression levels (Fig. 4.6A). Firstly, potential regulation between genes was identified by taking the gene pairs with the highest pairwise partial correlations across the data. For each gene, this potential activation or repression was then abstracted to Boolean functions, giving an ensemble of possible Boolean functions for each gene (Fig. 4.6B).

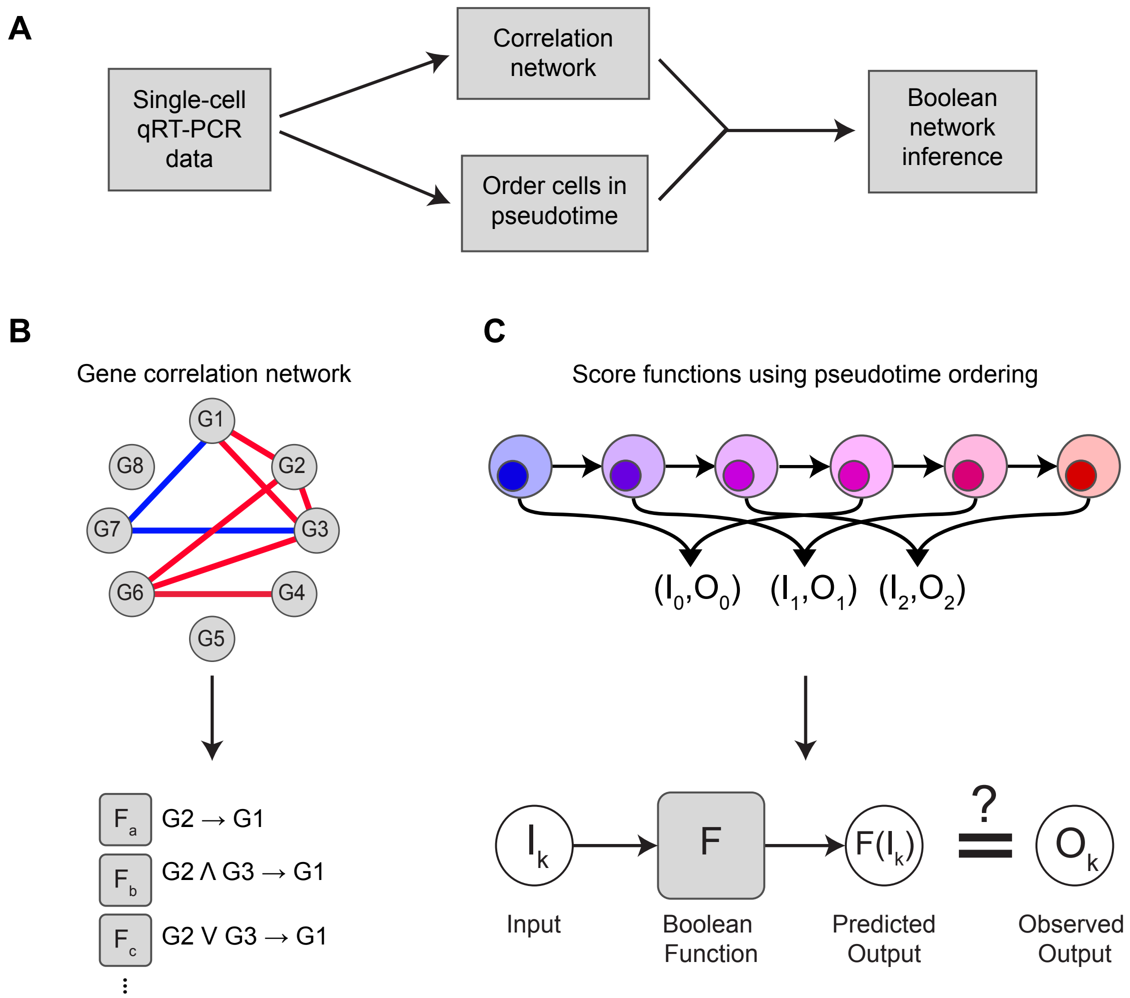


Fig. 4.6. Gene regulatory network models can be inferred from single-cell gene expression profiles. (A) Overview of steps in the network inference method. (B) Possible regulatory rules for each gene are identified using a gene-gene correlation network as input. Gene pairs showing the strongest positive or negative correlations are linked together in the network, with positive correlation shown in red and negative correlation in blue. These are then treated as activating or repressing relationships, respectively. The regulators of each gene define a set of potential Boolean functions governing the expression of that gene. Three of the possible functions for G1 are shown here. $\wedge$, AND; $\vee$, OR. (C) The pseudotime trajectory is used to identify the most suitable Boolean functions. Cells are ordered in pseudotime (based on continuous expression data) and converted to binary expression profiles. Pairs of cells a fixed distance apart then represent input–output pairs to the Boolean function. These pairs are used to score a Boolean function F by comparing $F(I_k)$ to O_k for a pair (I_k, O_k). The highest scoring function is the one where these values agree for the greatest number of pairs.

In previous studies where single-cell data have been used to infer Boolean regulation each single-cell profile was considered to be an allowed state of the Boolean network (Moignard et al., 2015; Xu et al., 2014). Here, the aim was to use the inferred pseudotime dynamics to describe permitted transitions between states, and use this to score the ensemble of functions. Pairs of cells throughout pseudotime were treated as an input to and output from a Boolean function (Fig. 4.6C), and this was used to investigate which of the functions from the correlation analysis best fitted the data. A detail description of the method can be found in Chapter 2. This method was applied to the single-cell qRT-PCR data to identify two sets of functions: one to construct a network for the HSC to MEP trajectory and another for the HSC to LMPP trajectory. Fig. 4.7 displays simplified regulatory relationships showing activation and repression between pairs of genes. The full list of regulatory relationships can be found in Appendix A. Code for performing the network inference is freely available from GitHub (https://github.com/fionahamey/Pseudotime-network-inference).

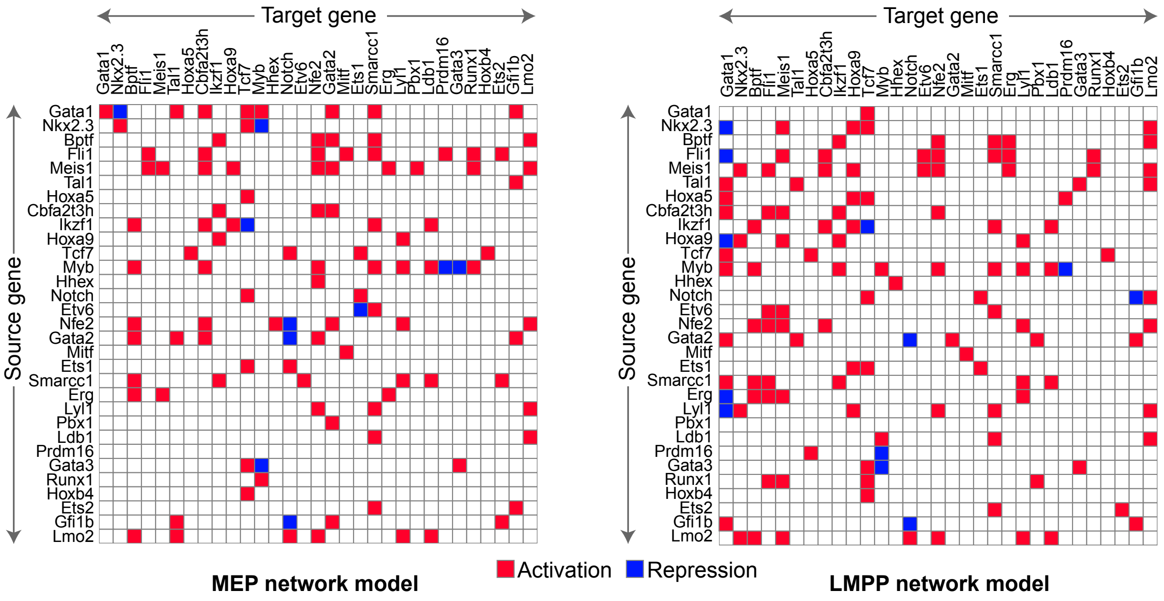


Fig. 4.7. Activating and repressing relationships between transcription factors identified by the network inference. Activation (red) and repression (blue) between pairs of transcription factors in the MEP and LMPP network models. The source and target gene for a relationship (e.g. source → target) are represented by the rows and columns, respectively.

4.5 Network analysis identifies biologically meaningful stable states

The Boolean network models inferred from the single-cell data showed complex structures (Fig. 4.7), with an average of four regulating transcription factors per gene, often as part of composite Boolean functions (Table A.1 in Appendix A). To assess whether these networks recapitulated HSPC biology, analysis was performed to identify stable states of both networks (Fig. 4.8A). Importantly, when the binary stable states of the models were compared to the qRT-PCR data (converted to binary expression values), it could be seen that the stable states of the MEP network model matched only with binary MEP expression profiles and not those of LMPPs, whereas for the LMPP network model the stable states matches were found within the LMPP and not MEP cells (Fig. 4.8B). To visualise where the stable states were positioned in the transcriptional landscape, cells with the observed qRT-PCR profiles most closely matching the stable states were highlighted in the diffusion map (Fig. 4.9). For the stable states corresponding to an observed binary expression state in the primary bone marrow data the matching cell was simply highlighted in the diffusion map. When a stable state was not present in the binary qRT-PCR data instead the closest matching binary expression profile (differing in the expression of the smallest number of genes) was identified and highlighted in the diffusion map. If multiple binary expression states were equally good matches then their expression in the continuous high-dimensional space was averaged and projected onto the diffusion map using the *destiny dm.predict* function. This process revealed that the stable states of the network models corresponded most closely to expression profiles found along the relevant trajectories for each network.

A limitation of this analysis is that *all* stable states of the network are identified, irrespective of whether they can be reached from a biologically meaningful starting point. To see whether the states matching with MEP and LMPP binary expression profiles could be reached from a starting point corresponding to stem cell gene expression, trajectories were simulated starting from MolO expression profiles. This confirmed that simulations originating from the MolO states could stabilise on either MEP or LMPP expression states, depending on the network model used. Together, these results demonstrate that the network models recapitulate the behaviour of HSCs differentiating towards two different haematopoietic progenitor populations.

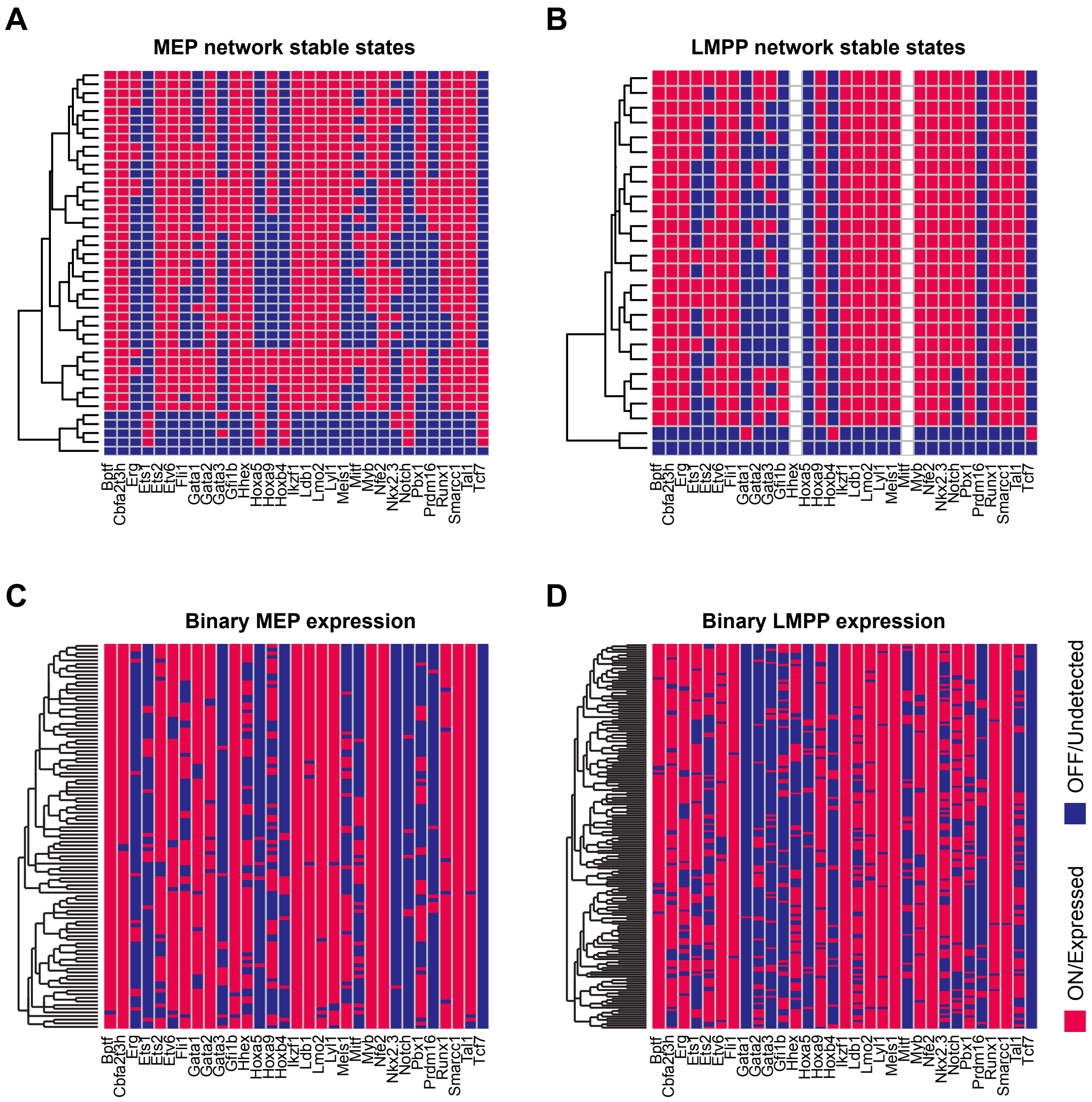


Fig. 4.8. Stable states of network models can be identified and compared to binary expression profiles from MEP and LMPP cells. (A) Stable states of MEP network model showing "ON/OFF" states of each gene. (B) Stable states of LMPP network model. Hhex and Mitf did not feature in the inferred LMPP network model, other than with a self-activation link, and so were excluded from the stable state analysis. (C) Gene expression profiles of MEPs measured by qRT-PCR and converted to binary expression based on whether a gene was detected in a cell. (D) Binary gene expression profiles of LMPP cells. Dendrograms indicate hierarchical clustering of stable states or cellular expression profiles.

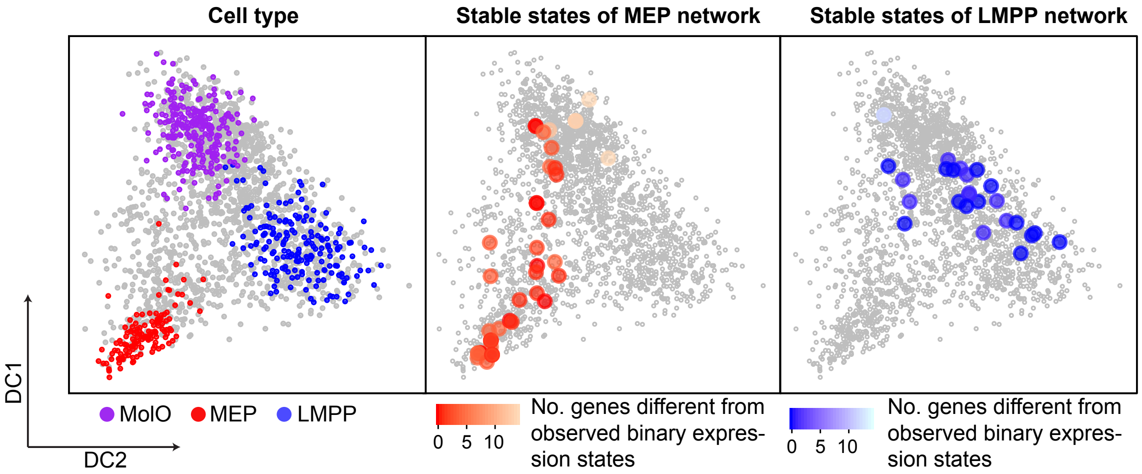


Fig. 4.9. Network models exhibit biologically meaningful stable states. Diffusion map of qRT-PCR profiles indicating three specific cell types (left panel), and with projected stable states of MEP (centre panel) and LMPP network models (right panel). Colour of the projected states indicates how closely the stable state matches with a binary expression profile observed in the single-cell qRT-PCR data, with the darkest colours corresponding to exact matches and paler colours representing disagreement in one or more genes. DC, diffusion component; MolO, molecular overlapping population of stem cells; MEP, megakaryocyte-erythroid progenitor; LMPP, lymphoid-primed multipotent progenitor.

4.6 Differences in network connectivity are supported by transcription factor binding

When comparing the Boolean models it was noted that two regulatory relationships involving Gata2 were different between the two networks. Positive regulation of genes *Nfe2* and *Cbfa2t3h* by Gata2 was present in the MEP network model, but was absent in the LMPP network model (Fig. 4.7, 4.10A). As validation of these regulatory differences would be challenging with primary cells, two haematopoietic model cell lines were used instead to investigate the relationships. These, the 416B cell line (Dexter et al., 1979) and the HoxB8-FL cell line (Redecke et al., 2013) have been described as having megakaryocyte and lympho-myeloid potential, respectively. Single-cell qRT-PCR data from these cell lines measuring the same genes as in the primary HSPCs were projected onto the diffusion map, and this confirmed that 416B cells resembled cells on the MEP trajectory, and the HoxB8-FL cells those on the LMPP trajectory (Fig. 4.10B).

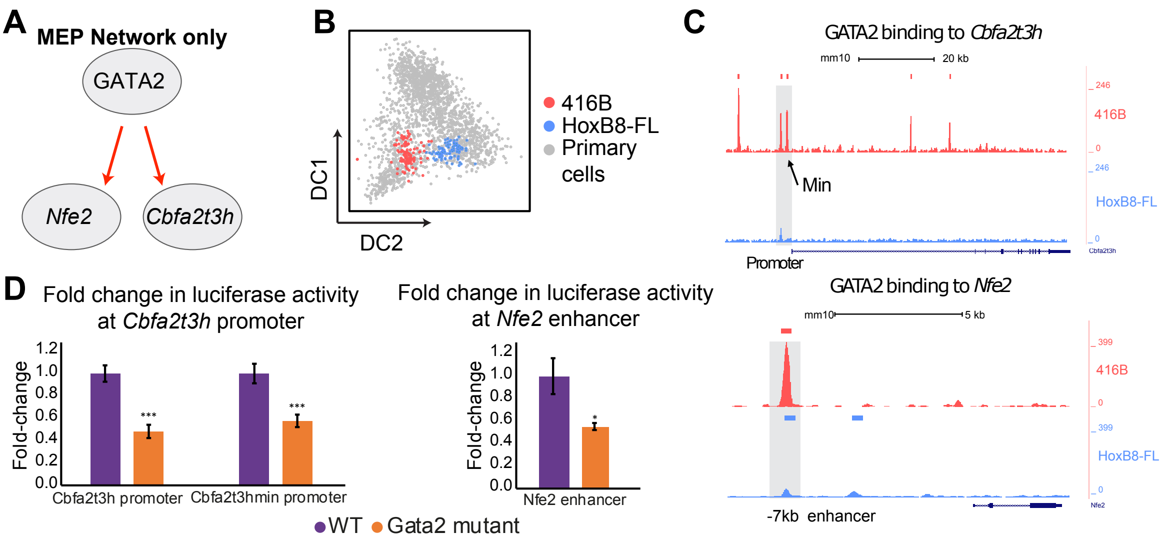


Fig. 4.10. Gata2 regulation unique to the MEP network model is supported by transcription factor binding. (A) Activating relationships found in the MEP but not LMPP network models. (B) Diffusion map projection of single-cell qRT-PCR data from haematopoietic cell lines. Primary bone marrow HSPCs are shown in grey. (C) ChIP-seq analysis of GATA2 in 416B and HoxB8-FL cell lines. (D) Luciferase assays showing activity at the *Cbfa2t3h* promoter and *Nfe2* enhancer in wild-type and GATA2 mutant regulatory regions. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; two-tailed unpaired t-test, $n = 3$. Error bars show ± standard deviation. WT, wild-type.

Chromatin immunoprecipitation sequencing (ChIP-seq) data from the two cell lines revealed strong binding of GATA2 at the *Cbfa2t3h* promoter and the *Nfe2* enhancer in the 416B cell line, with only very limited binding in the HoxB8-FL cells (Fig. 4.10C). Luciferase assays were performed in the 416B cell line to confirm that this binding corresponded to positive regulation. When the GATA2 binding sites in the *Cbfa2t3h* promoter or *Nfe2* enhancer regions were mutated, a significant reduction in luciferase activity was observed compared to the activity in non-mutated cells (Fig. 4.10D), supporting the regulation predicted by the network modelling.

4.7 Conclusions

The work in this chapter describes a transcriptional regulatory network inference method applied to single-cell qRT-PCR data from murine haematopoietic populations to infer regulatory networks for two alternative differentiation processes. Comparison of the networks revealed differential gene regulation between the two differentiation trajectories.

4.7.1 Network inference and modelling method

The concept of inferring gene regulatory network models from expression data is well-established, with many examples from a wide range of systems (Peter and Davidson, 2015). A variety of approaches have been used to infer and encode these networks, including the Boolean modelling used for the work described in this chapter. Boolean models have been successfully used in studies in the blood (Bonzanni et al., 2013; Moignard et al., 2015) and other biological systems (Dunn et al., 2014; Peter et al., 2012; Xu et al., 2014). Boolean models remain a popular technique due to the interpretability of encoding regulatory functions as logical relationships, and their power in allowing simulation of possible network states (Fisher and Piterman, 2010). Indeed, several studies in the stem cell field have used Boolean modelling techniques effectively to infer networks from gene expression data. Similar to Dunn et al. (2014) and Moignard et al. (2015), the work here used observed expression measurements to encode the solution to the network, but with some key differences. Dunn et al. (2014) rely on bulk perturbation data for their model, which is much harder to generate in *in vivo* systems such as haematopoiesis, and also obscures heterogeneity linked to cell fate decisions. Moignard et al. (2015) on the other hand used single-cell data, but only consider binary gene expression, and do not explicitly model a pseudotime ordering, instead constructing state transitions between cells based on a change in one gene. By not considering continuous data they may lose accuracy in modelling how cells move between states along differentiation. Again using single-cell data, Xu et al. (2014) used scRNA-seq data to score possible Boolean rules, similar to the idea of using the pseudotime ordering, but here they treated each cell as a stable state of the network, which is not applicable to a differentiating system such as haematopoiesis.

However, an obvious limitation of Boolean models is their reduction of gene expression to a binary state, where each gene must be considered “ON” or “OFF”. To overcome this problem, whilst still maintaining the benefits of a logical modelling formalism, some studies have chosen to extend Boolean modelling to multilevel modelling, where instead of only two allowed values gene expression can take one of several discrete levels (Collombet et al., 2017). This could be interesting to explore in a network inference approach from single-cell data, but would present a challenge in deciding how to define the thresholds for multiple levels of gene expression. An alternative network representation is offered by differential equation models, as used by the method of Ocone et al. (2015), which used pseudotime ordering of single-data to infer regulatory relationships. The strength of such models lies in considering the actual level of gene expression, which could cause different responses at different thresholds (Wolpert, 1969), but there are computational restrictions on how many genes can be included in the model. Bayesian models are another alternative, as demonstrated by Schütte et al. (2016), but these can only consider networks of a certain topology, excluding cyclical relationships and therefore ruling out motifs such as self-activation. The method described in this chapter, does not have these restrictions, which is the compensation for restricting the model to binary expression states.

4.7.2 Examples of known regulation captured by the models

It was reassuring to see that evidence for several examples of transcriptional regulation inferred by the network models could be found in the literature. One such relationship was the activation of Gfi1b by Gata2, which was present in both networks. This regulation was identified by Moignard et al. (2013) as a gene pair with highly correlating expression, and subsequently validated using ChIP-seq and luciferase assays. A different type of regulatory relationship with support from experimental studies was seen in the form of the self-activation of Fli1 in the MEP network model. Donaldson et al. (2005) experimentally validated a small network of transcription factor interactions, which included a feed-forward activation loop for Fli1. Prior work also showed that Gata2 activates the transcription factor Tal1 as part of initiating the blood program (Göttgens et al., 2002), and activation of Tal1 by Gata2 is seen in both the LMPP and MEP network models.

4.7.3 MEP network specific regulation

Network inference on the single-cell HSPC expression data and subsequent experimental validation led to the identification of Gata2 regulation unique to the MEP network model, with this transcription factor activating *Cbfa2t3h* and *Nfe2*. The primary role of Gata2 is in regulating haematopoietic stem and progenitor cell function, playing a part in the formation of haematopoietic cells and in maintaining stem cells (Lim et al., 2012; Rodrigues et al., 2005; Tsai and Orkin, 1997). Its importance in the haematopoietic system has been confirmed by observations such as severe haematopoietic defects in *Gata2* knock-out mice

(Tsai et al., 1994). *Cbfa2t3h* is a known regulator of differentiation towards the erythroid and megakaryocyte lineages (Fujiwara et al., 2010; Goardon et al., 2006; Hamlett et al., 2008) and encodes the transcription factor ETO2 (Schuh et al., 2005). It has been shown that GATA2 binding to *Cbfa2t3h* activates this gene, which is then followed by ETO2 binding and repressing its own promoter, as part of the initiation of transcriptional programs driven by Gata1 during erythroid differentiation (Fujiwara et al., 2009). *Nfe2* is an upstream regulator of globin genes and is necessary for generation of megakaryocytes (Ney et al., 1993; Shivdasani et al., 1995b). By linking an early stem cell regulator in the form of Gata2 to the erythroid-megakaryocyte related genes *Nfe2* and *Cbfa2t3h* in the MEP network model this provides an insight into how cells can switch from general gene expression programs to activating lineage specific genes.

4.7.4 Limitations from using qRT-PCR data

Although single-cell qRT-PCR data provide a sensitive way of measuring gene expression in individual cells, any network inferred using these data is restricted to the genes measured by the assay. These are hand-picked and therefore limit the potential for discovery of regulatory relationships involving novel genes, inevitably leading to incomplete network models. This may miss key regulatory relationships and also may lead to the misinterpretation of indirect regulation between genes as a direct regulatory relationship, due to intermediate regulators being absent from the assay. In the specific context of this work, the data were previously published by Wilson et al. (2015), with the aim of dissecting heterogeneity within the stem cell compartment. The authors therefore chose genes that were biased towards regulators of stem cells maintenance and, in particular, this choice lacked key regulators of lymphoid development. There were also technical issues with some genes, including *Gfi1*, which had to be excluded from the analysis. Together this means that some important regulators of haematopoietic differentiation were missing from the network models. Related to the issue of hand-selecting genes, there was another limitation from using the qRT-PCR data in terms of incomplete lineage resolution when constructing differentiation trajectories. Whilst a clear separation between LMPP and MEP progenitors could be observed with the measured genes, Fig. 4.2 shows that other progenitor populations were not as well resolved, for example GMP and LMPP populations had substantial overlap in the dimensionality reduction plots. The work from Chapter 3 demonstrates that these different cell types can be separated based on the full transcriptome, suggesting that the qRT-PCR gene panel does not fully capture the transcriptional heterogeneity linked to differentiation that is present within these cells.

However, avoiding the bias due to hand-picking genes for single-cell qRT-PCR would require use of technologies such as scRNA-seq to measure gene expression. This would come with its own challenges for network inference. Inferring a network on hundreds or thousands of genes would have issues for scalability, and exclude some types of network modelling due to the number of genes involved. The interpretability of such models would also be challenging, particularly if simulation of these networks was not possible due to their scale. Another problem, particularly when considering using Boolean network modelling, arises from the issue of so-called "dropouts" in scRNA-seq data — when a gene is present in a cell but not detected in the sequencing data. Dropouts make categorising expression into binary states more challenging. Single-cell qRT-PCR data do not suffer from these limitations to the same extent, and therefore represent a useful starting point for network inference as demonstrated in this chapter.

4.7.5 Future directions

Future work would aim to investigate ways in which regulatory relationships between transcription factors could be identified in a less biased way, based on scRNA-seq data. Whilst a full network inference method may not be feasible for the reasons discussed above, this work could be important in finding previously unknown genes involved in haematopoietic fate decisions. This idea provided the motivation behind the work discussed in Chapter 6. Work on extending network inference methods to scRNA-seq data would also explore the potential of using computational methods to impute missing values in these datasets to address the issue of dropouts, as this could allow techniques such as Boolean modelling to be applied to these data (Eraslan et al., 2018; van Dijk et al., 2018).

4.7.6 Summary

In summary, this chapter describes applying a computational network inference method to single-cell gene expression data to compare differences in transcriptional regulation between differentiation towards two haematopoietic lineages.

Chapter 5

Dissecting heterogeneity within human lympho-myeloid progenitors

Parts of this section have been modified from Karamitros et al. (2018), on which F. Hamey carried out the bioinformatics analysis of single-cell data. Single-cell gene expression analysis experiments were performed by Dimitris Karamitros and Bilyana Stoilova and scRNA-sequencing data were aligned by Evangelia Diamanti. Single-cell functional assays were performed by Dimitris Karamitros, Bilyana Stoilova and Zahra Aboukhalil, who also analysed and summarised the functional output of these experiments. F. Hamey performed analysis of single-cell gene expression data (except for the alignment of sequencing data) and analysis of surface marker expression linked to the functional assays.

5.1 Background

Despite study over many years, the human blood progenitor compartment still remains incompletely understood, with debate surrounding the exact hierarchy leading to the production of mature blood cells (Laurenti and Göttgens, 2018). In particular, in relation to the lymphoid and myeloid lineages, several progenitor populations with a mixture of lymphoid and myeloid potential have been described in the human haematopoietic system. LMPPs have been defined as Lin$^-$ CD34$^+$ CD38$^-$ CD90$^{\text{neg-lo}}$ CD45RA$^+$ CD10$^-$ cells, and it has been shown these can give rise to granulocytes, monocytes, B cells and T cells, but do not produce erythroid or megakaryocytic output (Goardon et al., 2011). Differing in their expression of

CD10, multi-lymphoid progenitors (MLPs) are another progenitor population with lympho-myeloid output. These Lin$^-$ CD34$^+$ CD38$^-$ CD90$^{\text{neg-lo}}$ CD45RA$^+$ CD10$^+$ cells are capable of producing lymphoid cells (B cells, T cells and NK cells) as well as monocytes and dendritic cells. However, MLPs do not have the potential to produce granulocytes (Doulatov et al., 2010). Progenitor populations with stronger bias towards the myeloid lineages also exist, including GMPs (defined as Lin$^-$ CD34$^+$ CD38$^+$ CD45RA$^+$ CD123$^+$), which have residual lymphoid potential but produce mainly cells belonging to the myeloid lineages (Goardon et al., 2011; Lee et al., 2015).

Work in the past decade has provided support for a hierarchical model of human haematopoiesis with LMPPs upstream of both MLP and GMP progenitors (Görgens et al., 2013). However, recent studies have questioned the idea of this hierarchical organisation, instead arguing for a model where HSCs differentiate into unilineage blood progenitors without passing through populations with stepwise loss of lineage potential (Notta et al., 2016; Velten et al., 2017). Such debate highlights the need for single-cell analysis of blood progenitors. The work in this chapter focuses on single-cell gene expression profiling of GMPs, LMPPs and MLPs to examine the transcriptional programs present within these human lympho-myeloid progenitors, and also considers the results of single-cell functional assays to examine their functional heterogeneity (Fig. 5.1). The single-cell functional assays identified bipotent lympho-myeloid progenitors and, by linking the surface marker expression of cells to their functional output in single-cell assays, new sorting strategies were defined to enrich for function within the conventional GMP and LMPP sorting gates. Both transcriptional and functional analyses support the idea of a continuum between commitment to lymphoid and myeloid cell fates, rather than a discrete hierarchy.

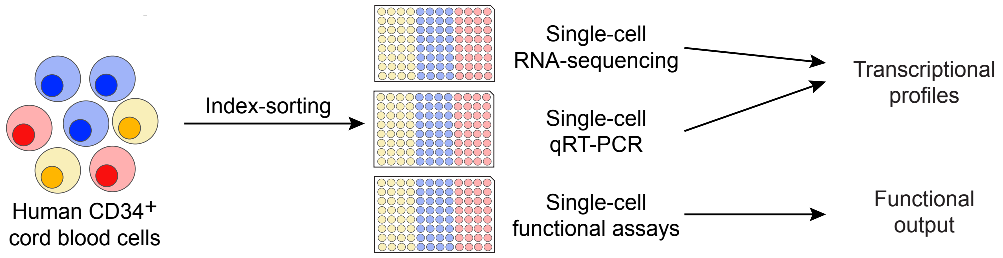
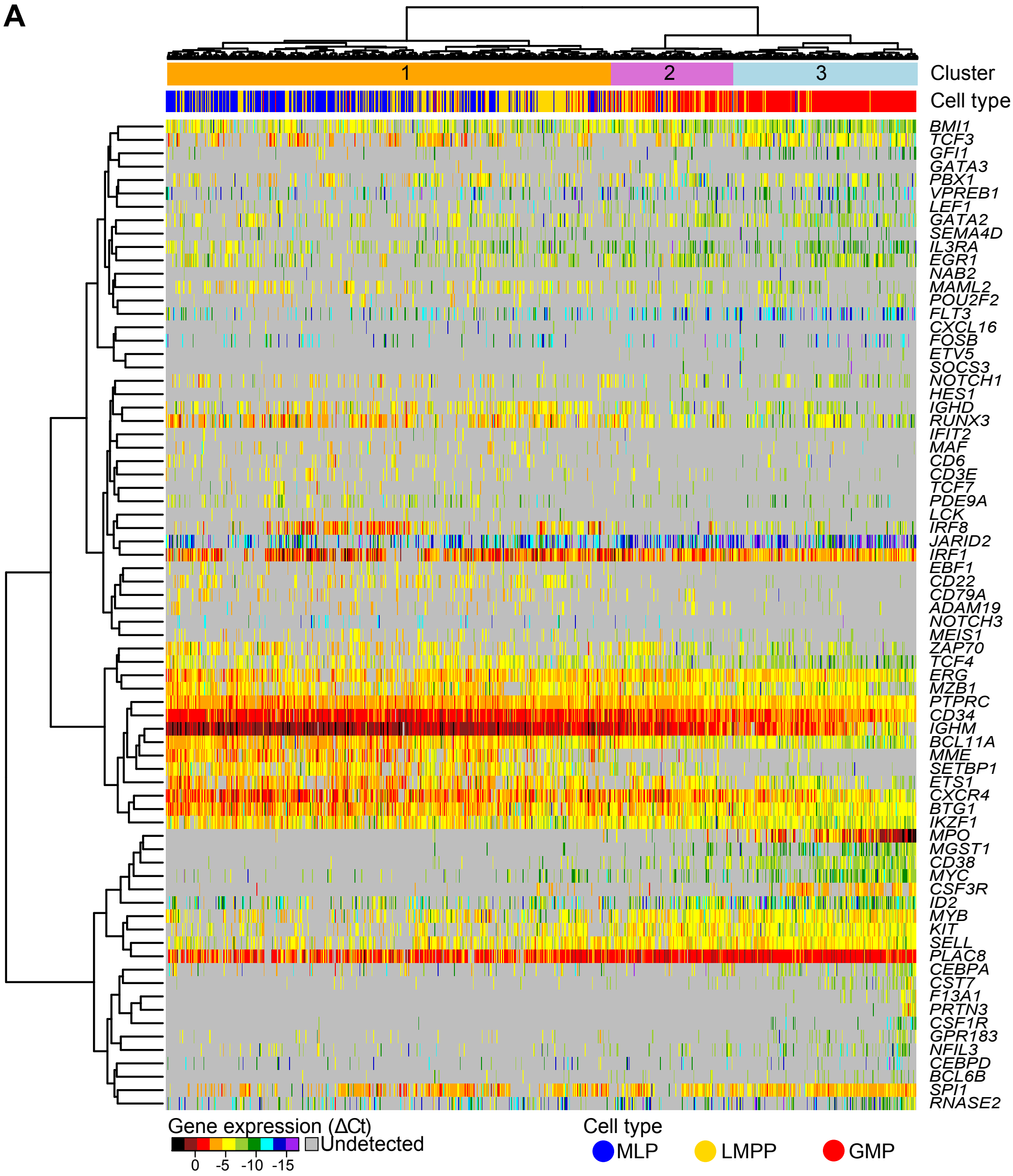


Fig. 5.1. Strategy for assaying heterogeneity within human cord blood progenitors. Schematic shows how human cord blood progenitors were isolated and assayed at the single-cell level to measure the gene expression and functional output from individual cells.

5.2 Single-cell profiling reveals a continuum of gene expression in human lympho-myeloid progenitors

To investigate transcriptional heterogeneity within lympho-myeloid progenitor populations, the expression of genes linked to myeloid and lymphoid cell fates was measured in GMPs, LMPPs and MLPs using single-cell qRT-PCR. After quality control, this resulted in 919 transcriptional profiles of cells from four cord blood donors. Whilst many genes in this set were only detected at very low levels, hierarchical clustering was able to separate cells into three clusters (Fig. 5.2A). Cluster 1 was formed from a mixture of mostly LMPP and MLP cells (Fig. 5.2B), and had higher expression of genes such as *MME* (CD10) and *IGHM* (immunoglobulin heavy constant mu). Cluster 2 had a different cell type composition, containing mostly GMP and LMPP cells along with a small number of MLPs. Cluster 3 displayed the highest expression of myeloid genes such as *MPO* and *CSF3R* and also contained a small group of cells with expression of more mature myeloid markers *PRTN3* and *CSF1R*. Almost 90% of cells in this cluster were GMPs. Together, this analysis was able to separate cells into a cluster with higher expression of lymphoid genes, a cluster with expression of myeloid genes, and a cluster without strong expression of either set of lineage markers. It was also interesting to see that cells did not simply separate based on FACS phenotype, instead showing considerable overlap between these sorted populations.

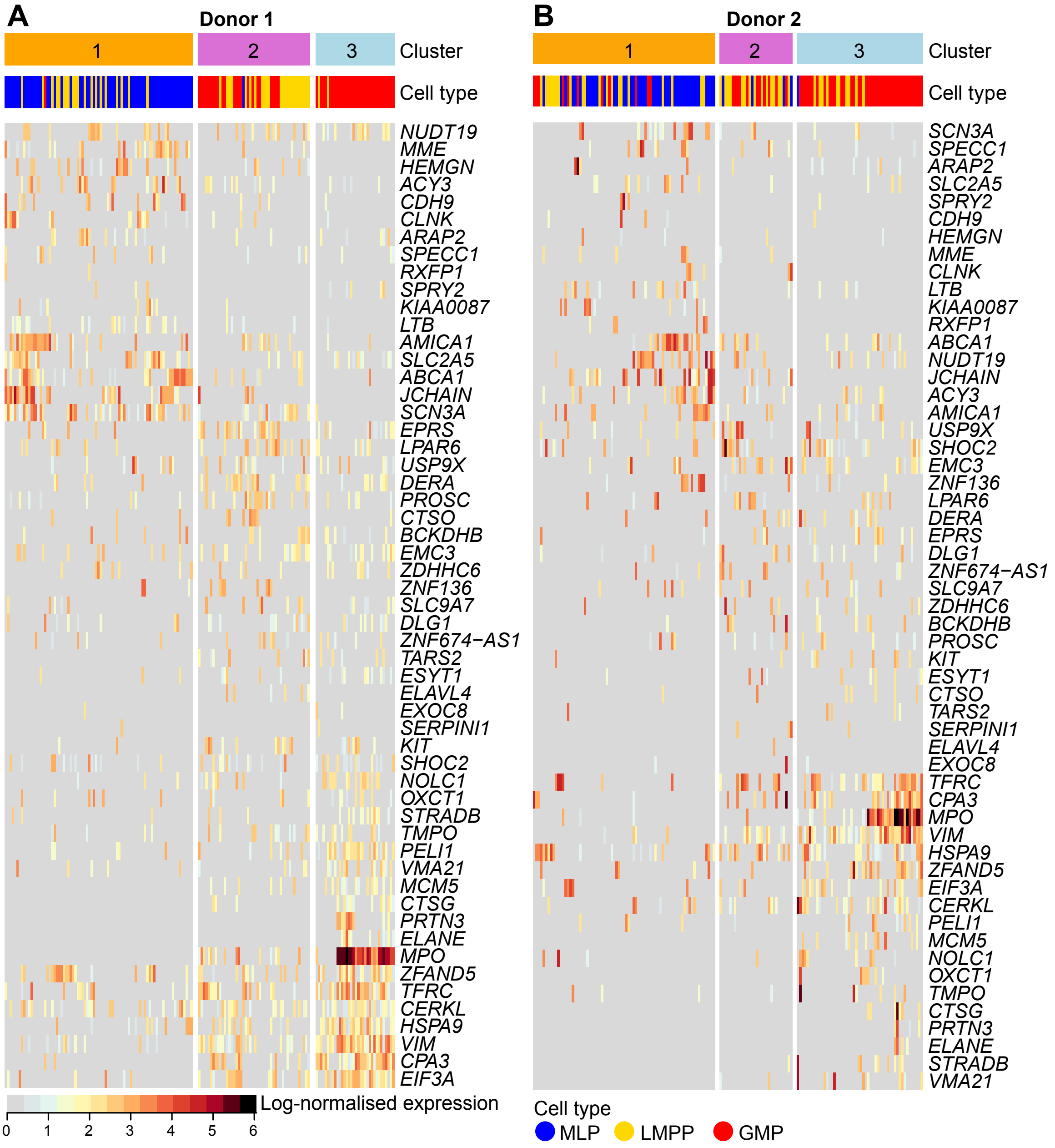
As many of the genes selected for profiling by single-cell qRT-PCR were detected in very few cells, it was decided to perform less biased analysis using scRNA-seq on the same three progenitor populations. Cord blood cells were isolated from two donors and processed separately (Fig. 5.3A, B). Clustering of the single-cell profiles using graph-based clustering identified three groups with similar cell type composition to the qRT-PCR clusters: cluster 1 composed of mostly LMPP and MLP cells, cluster 2 with more balanced composition between GMP and LMPP cells, and cluster 3 formed from mostly GMPs (Fig. 5.3C, D). Differential expression between cells in a cluster and the remaining cells was performed using the Wilcoxon rank sum test, and the top 10 most significant genes for each cluster displayed in the heatmaps in Fig. 5.3A, B. This was able to identify lymphoid genes such as *MME*, *JCHAIN* and *LTB* with highest expression in cluster 1, and myeloid genes including *ELANE*, *MPO*, and *PRTN3* with highest expression in cluster 3 cells.



B

Cell type	Cluster 1	Cluster 2	Cluster 3
GMP	25	80	203
LMPP	230	57	19
MLP	288	13	4

Fig. 5.2. Single-cell gene expression profiles capture the transcriptional heterogeneity of human lympho-myeloid cord blood progenitors. (A) Heatmap of gene expression in human cord blood progenitors measured by single-cell qRT-PCR. Dendrograms indicate results of hierarchical clustering on genes (rows) and cells (columns). Clustering on cells separated cells into three clusters. (B) Cell type compositions of clusters shown in heatmap.



C

Donor 1

Cell type	Cluster 1	Cluster 2	Cluster 3
GMP	1	23	32
LMPP	20	24	2
MLP	60	1	0

D

Donor 2

Cell type	Cluster 1	Cluster 2	Cluster 3
GMP	11	11	41
LMPP	20	15	10
MLP	44	4	1

Fig. 5.3. The clustering of qRT-PCR profiles is supported by scRNA-seq data. (A, B) Heatmaps of gene expression in human cord blood progenitors measured by scRNA-seq. Cells are from one donor in each panel. (C, D) Cell type composition of the clusters shown in heatmaps.

Next, dimensionality reduction was performed to visualise the structure present within the gene expression datasets. PCA on qRT-PCR and RNA-seq data captured a continuum from MLPs through LMPPs to GMPs along PC1 (Fig. 5.4A). Visualising the clusters from Fig. 5.2 and 5.3 on the PCA showed that cluster 2 was positioned between clusters 1 and 3, supporting the idea of cluster 2 having an intermediate gene expression pattern compared to the lymphoid or myeloid expression of clusters 1 and 3.

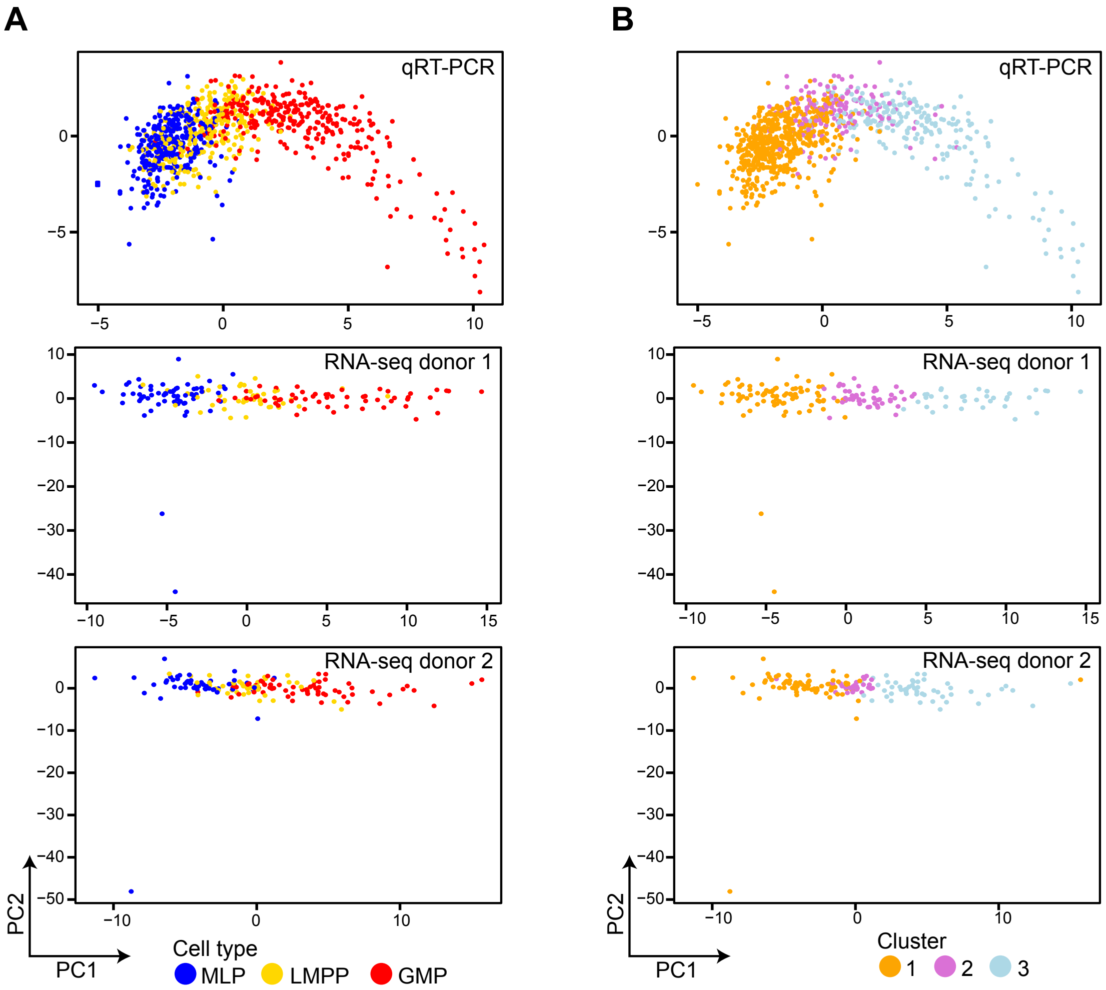


Fig. 5.4. Single-cell data reveal a continuum of lympho-myeloid gene expression. (A) Principal component analysis dimensionality reductions on human cord blood progenitors coloured by cell type. Top, middle and bottom panels show qRT-PCR data, RNA-seq data from one donor, and RNA-seq data from a second donor, respectively. PC, principal component. (B) Principal component plots coloured by clusterings shown in Fig. 5.2 and 5.3.

5.3 Conventional sorting strategies can be refined to enrich for function

As well as assaying transcriptional heterogeneity within lympho-myeloid progenitor populations, GMPs, LMPPs and MLPs were also isolated for single-cell culture assays to assess their functional output (Fig. 5.5A). Individual cells were sorted into culture conditions supporting both lymphoid and myeloid output, and the cultures scored for granulocyte, monocyte, B cell and NK cell output after 14 days (see Chapter 2 for details). Using FACS to investigate functional output allowed individual cells to be categorised as having lymphoid, myeloid, or mixed lympho-myeloid output (Fig. 5.5B). In keeping with the reported lympho-myeloid biases of these cell types, GMPs produced predominately myeloid cells, and LMPPs and MLPs had mostly lymphoid output, with MLPs almost entirely giving rise to lymphoid colonies. Interestingly, a number of cells were found to produce both lymphoid and myeloid output, showing that these progenitors were not restricted to only one of the myeloid or lymphoid lineages.

As cells were index-sorted for the functional experiments this meant that cell surface marker levels were recorded for each individual cell. Dimensionality reduction was performed on the surface markers measured by index-sorting to see whether cells separated based on their functional output (Fig. 5.5C-E). Whilst considerable overlap between the functionally different cells was seen, some regions of the PCA were occupied by predominantly myeloid or lymphoid cells in the GMP and LMPP populations, respectively (Fig. 5.5C, D). To determine whether these surface marker levels could be used to separate function within the conventional sorting gates, pairwise comparisons between the level of each surface marker were made between different functional outputs (Table 5.1). In particular, this highlighted significantly different levels of CD38 between all three functional categories amongst the GMP cells. Closer inspection revealed that GMPs with purely myeloid output had the highest levels of CD38 (Fig. 5.6A, B). Within the LMPP population both CD10 and CD45RA levels were significantly higher in lymphoid output cells compared to both lympho-myeloid and pure myeloid producing cells (Table 5.1, Fig. 5.6C-E). Scatter plots of these surface markers suggested that GMP and LMPP gating strategies could be further refined to enrich for function based on these surface marker levels (Fig. 5.6B, E).

Table 5.1. Linking differences in surface marker levels to functional output. Significance in differences between surface marker levels across different functional groups within a cell type. Values in table represent p-values from Wilcoxon rank-sum test. Ly, lymphoid output; Ly-My, mixed lymphoid and myeloid output; My, myeloid output.

Marker	Ly vs My		Ly vs Ly-My		My vs Ly-My	
	GMP	LMPP	GMP	LMPP	GMP	LMPP
CD38	1.6×10^{-8}	0.041	3.6×10^{-3}	0.89	1.6×10^{-11}	0.082
CD34	1.2×10^{-3}	0.72	0.46	0.38	0.10	0.64
CD123	0.98	0.20	0.87	0.023	0.75	0.42
CD10	2.4×10^{-8}	5.1×10^{-4}	0.091	2.1×10^{-3}	0.19	0.70
CD90	0.84	0.77	0.20	0.019	0.063	0.29
CD45RA	2.0×10^{-3}	0.023	5.3×10^{-4}	1.8×10^{-6}	0.026	0.16

Motivated by the significant differences in CD38, CD10 and CD45RA between GMPs and LMPPs with different functional output, it was decided to investigate whether subsets of the GMP and LMPP cell types could be enriched for function using new sorting strategies based on these markers (Fig. 5.7A, B). CD10^{hi} CD45RA^{hi} LMPPs and CD10^{lo} CD45RA^{lo} LMPPs, denoted as LMPP^{ly} and LMPP^{mix}, respectively, were isolated from cord blood (Fig. 5.7A). Based on the previous functional analysis it was hypothesised that the LMPP^{ly}, but not the LMPP^{mix}, population would enrich for lymphoid output, compared to conventionally sorted LMPPs. This hypothesis was supported by the functional output of single-cell cultures of these progenitors in two different culture conditions (Fig. 5.7Ci, ii). In these assays LMPP^{ly} cells had almost no myeloid output and indeed significantly lower myeloid output than LMPP and LMPP^{mix} cells ($p = 0.0496$ and $p = 0.0280$, respectively, Fisher's exact test). Similarly, in an attempt to enrich for myeloid function within GMPs, cord blood cells were also sorted based on their levels of CD38. CD38^{hi} cells were further sorted based on CD10, CD45RA and CD123 to isolate the GMP CD38^{hi} population (Fig. 5.7B). CD38^{int} CD45RA^{+} CD10^{-} cells, denoted as CD38^{mid}, and CD38^{lo} LMPPs were also sorted. As the CD38^{lo} population represented only 1 in 10^8 mononuclear cells it was too rare to be assessed in single-cell cultures. However, the functional output of CD38^{hi} GMPs and the CD38^{mid} population was assessed at the single-cell level (Fig. 5.7D). The lymphoid and lympho-myeloid output of the CD38^{hi} GMPs was significantly lower than that of conventionally purified GMPs ($p < 0.0001$ and $p = 0.0115$, respectively, Fisher's exact test), showing that CD38^{hi} GMPs were enriched for myeloid output.

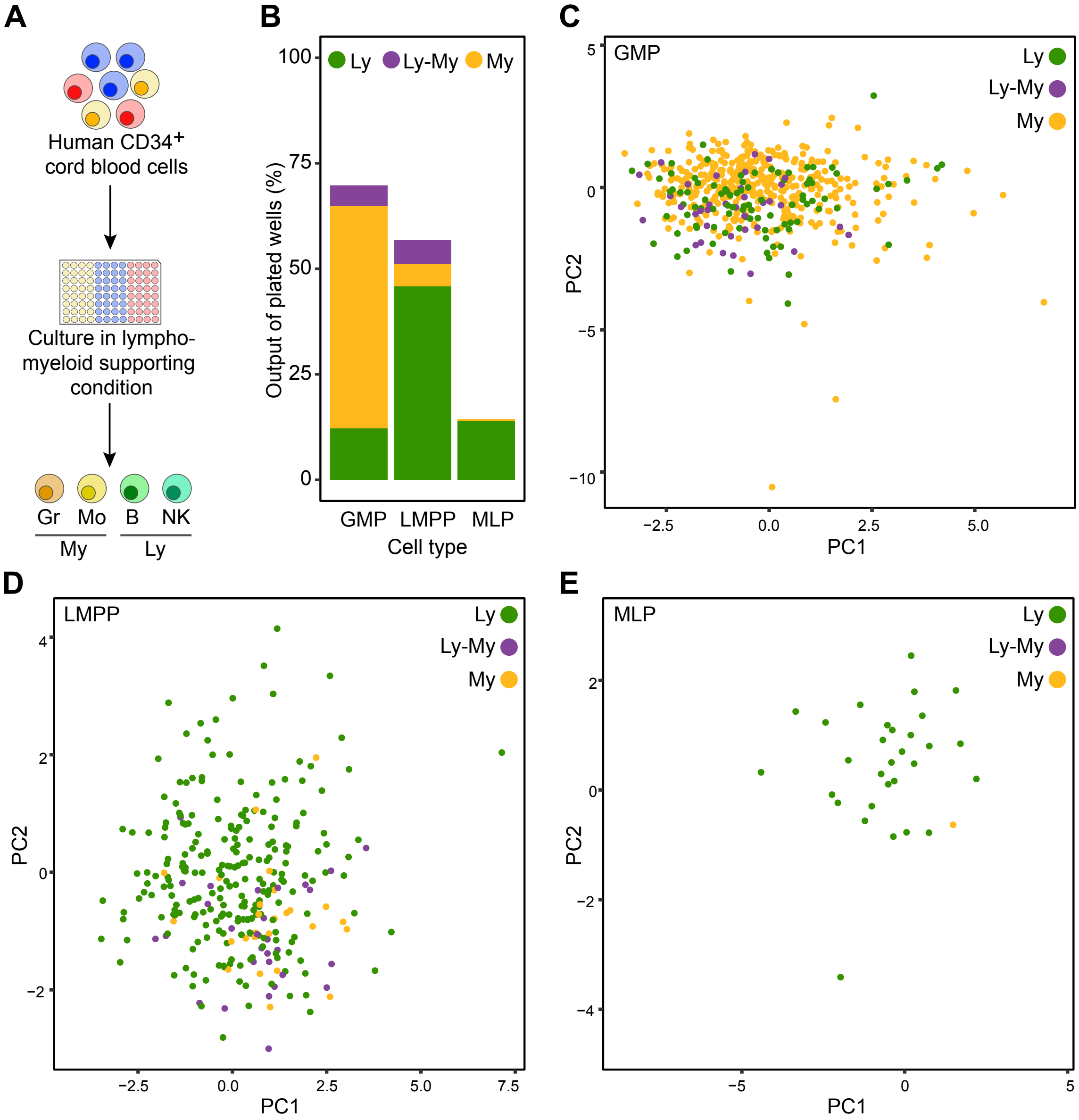


Fig. 5.5. LMPP and GMP cells produce a mixture of lymphoid and myeloid output in single-cell assays (A) Experimental overview for single-cell cultures. Individual cells were isolated and after culturing assessed for lympho-myeloid output. Total number of cells plated were $n = 760$ (GMP), $n = 514$ (LMPP) and $n = 215$ (MLP). Gr, granulocyte; Mo, monocyte; B, B cell; NK, NK cell. (B) Functional output of human cord blood progenitors in SF7b/Dox culture as a percentage of total plated cells. Ly, lymphoid output; Ly-My, mixed lymphoid and myeloid output; My, myeloid output. (C, D, E) Principal component analysis of colony-producing GMPs, LMPPs or MLPs based on the expression of surface markers measured by index-sorting. Cells are coloured based on the cell types present within the colony they produced.

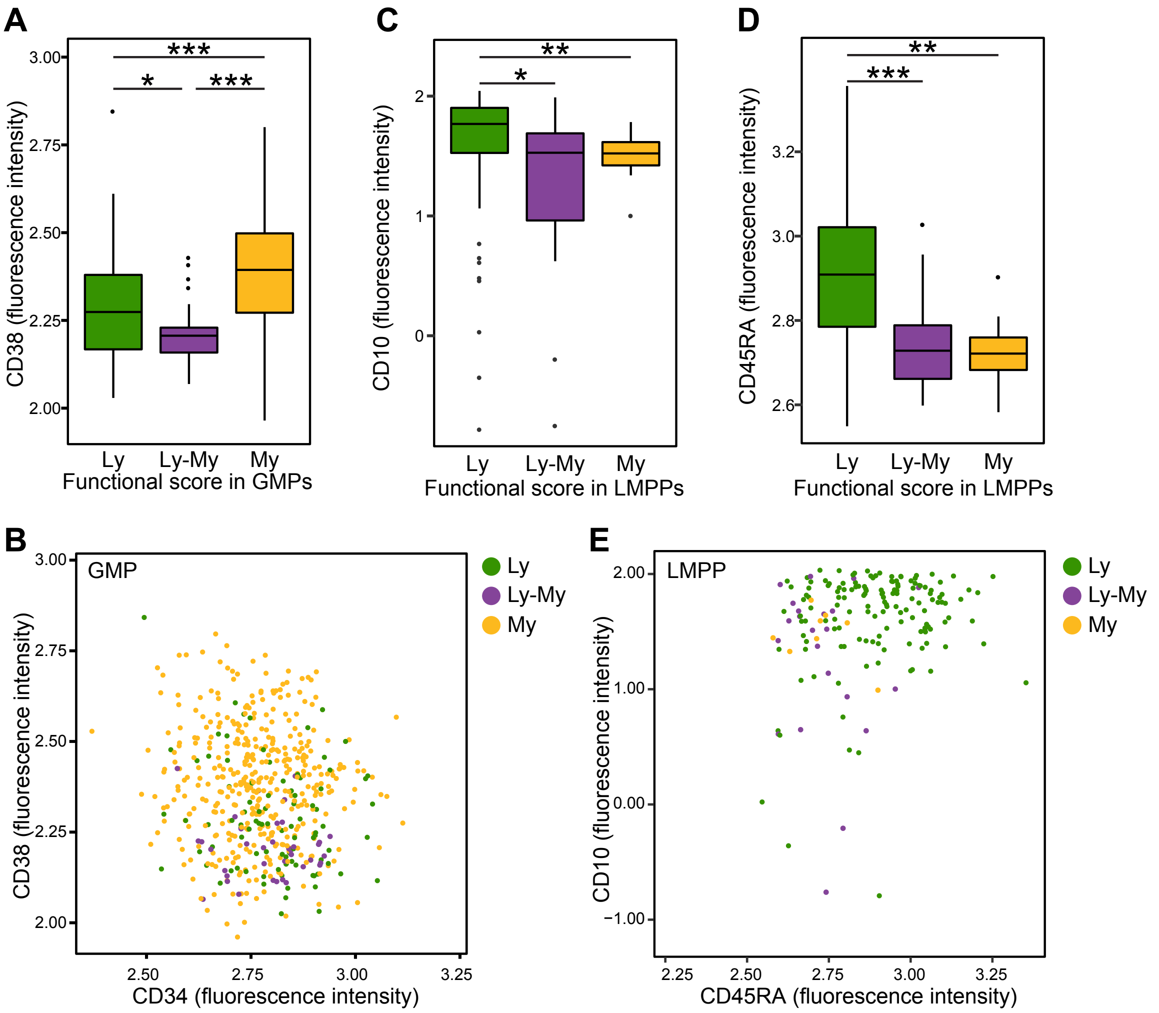


Fig. 5.6. Single-cell assays reveal link between surface markers and function in blood progenitors. (A) CD38 surface marker expression of GMPs grouped by functional output. $p = 1.6 \times 10^{-11}$ (myeloid vs lymphoid), $p = 1.6 \times 10^{-8}$ (myeloid vs lympho-myeloid), and $p = 3.6 \times 10^{-3}$ (lymphoid vs lympho-myeloid), Wilcoxon rank-sum test. (B) CD38 vs CD34 surface expression in GMPs. (C) CD10 surface marker expression of LMPPs grouped by functional output. $p = 2.1 \times 10^{-3}$ (lymphoid vs lympho-myeloid) and $p = 5.1 \times 10^{-4}$ (lymphoid vs myeloid), Wilcoxon rank-sum test. (D) CD45RA surface marker expression of LMPPs grouped by output in single-cell functional assays. $p = 1.8 \times 10^{-6}$ (lymphoid vs lympho-myeloid) and $p = 0.023$ (lymphoid vs myeloid), Wilcoxon rank-sum test. (E) CD10 vs CD45RA surface expression in LMPPs. All fluorescence intensities were transformed using the *logicle* transform function from the *flowCore* package and normalised across sort days within each cell type. Boxes indicate median, upper and lower quartile, and whiskers extend to either maximum/minimum values, or upper/lower quartiles ± interquartile range. Points outside this range are shown as outliers.

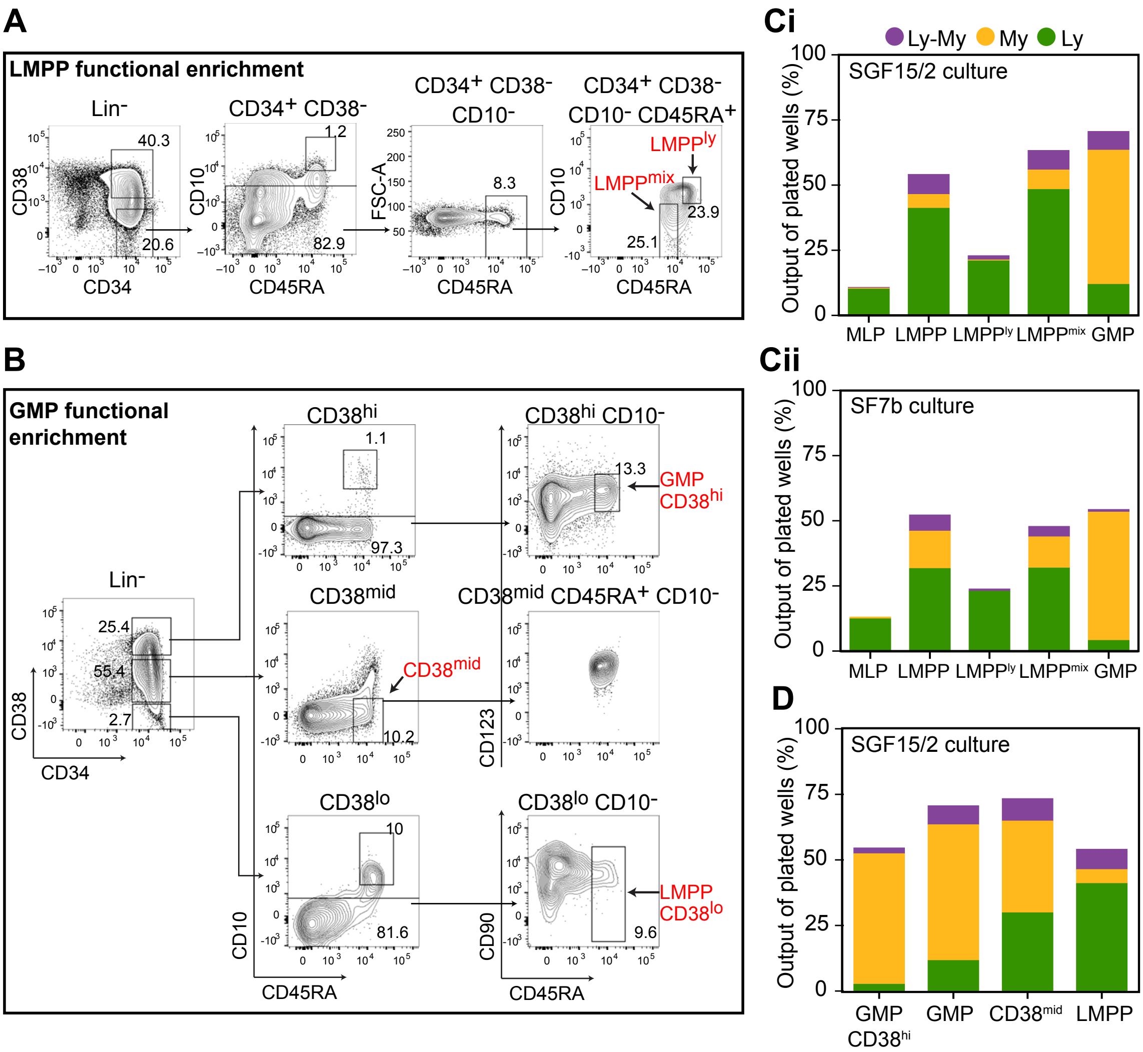


Fig. 5.7. New flow-cytometry sorting strategies can purify for function in LMPP and GMP compartments. (A) Revised sorting strategy of cord blood LMPP cells based on CD45RA and CD10 expression. Numbers indicate percentages of cells in the parental gate, which has the sorting strategy written above each panel. (B) Revised sorting strategy of cord blood GMPs based on levels of CD38 expression. (Ci, ii) Functional output of single cells from original and revised sorting strategies in two different culture conditions. Sorting strategies for LMPP^{ly} and LMPP^{mix} cells are indicated in panel A. (D) Functional output of single cells from original and revised sorting strategies based on CD38 expression. GMP CD38^{hi} and CD38^{mid} strategies are indicated in panel B.

To relate the functional output of lympho-myeloid progenitors to their transcriptional profiles, the levels of CD10 and CD38 were investigated in the cells profiled using scRNA-seq (Fig. 5.8A). CD10 levels showed significant differences between the cluster 1 and cluster 3 LMPPs, with cluster 3 LMPPs having significantly lower CD10 expression. As cluster 3 cells exhibited expression of mainly myeloid genes this was consistent with the functional enrichment seen in the revised LMPP sorting strategies. Additionally, GMPs in cluster 3 displayed significantly higher CD38 levels than both cluster 2 and cluster 1 GMPs (Fig. 5.8B), again consistent with the gene expression programs of these cells.

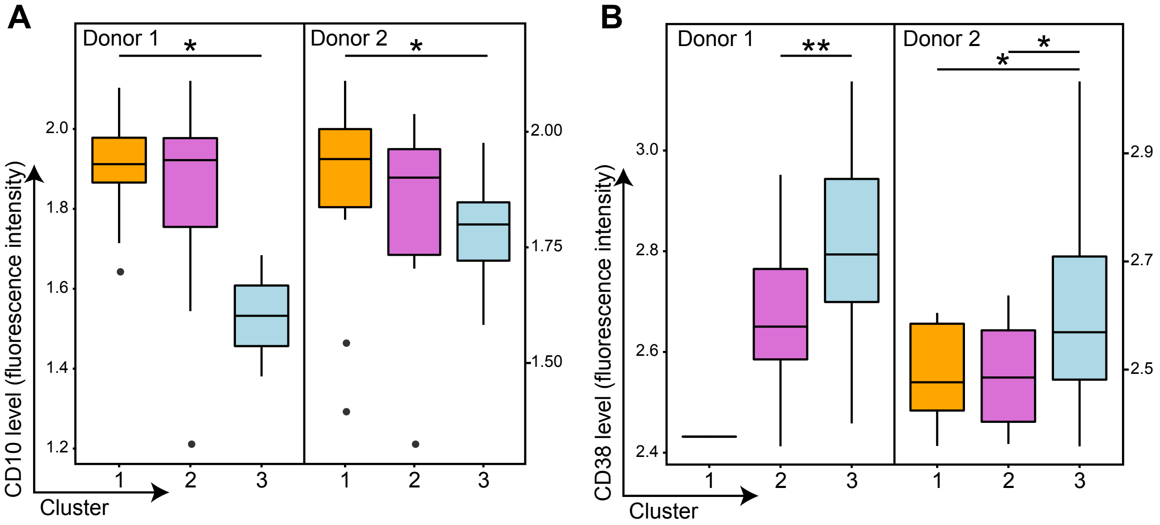


Fig. 5.8. RNA-seq clusters display differences in CD10 and CD38 surface marker expression. (A) CD10 surface marker expression in LMPP scRNA-seq profiles grouped by the clustering in Fig. 5.3. Donor 1, $p = 0.017$; Donor 2, $p = 0.029$, Wilcoxon rank-sum test. (B) CD38 surface marker expression of GMPs profiled by scRNA-seq and grouped by the clustering in Fig. 5.3. Donor 1, $p = 0.0044$; Donor 2, $p = 0.046$ (1 vs 3) and $p = 0.046$ (2 vs 3). Boxes indicate median, upper and lower quartile, and whiskers extend to either maximum/minimum values, or upper/lower quartiles ± interquartile range. Points outside this range are shown as outliers.

5.4 Conclusions

The work in this chapter presents an analysis of data describing heterogeneity in human cord blood lympho-myeloid progenitors at the single-cell level. GMP, LMPP and MLP populations were profiled using single-cell qRT-PCR, scRNA-seq and single-cell cultures to assess functional output.

5.4.1 Single-cell data support a continuum of lympho-myeloid progenitors

The analysis of single-cell gene expression of GMP, LMPP and MLP progenitors revealed a continuum of expression from lymphoid to myeloid transcriptional programmes. Profiles with expression of lymphoid genes largely belonged to MLP and LMPP sorting gates, and those with a myeloid signature mostly originated from the GMP gate, with a small contribution from LMPP cells. These observations are in agreement with reported lymphoid/myeloid biases of these progenitors populations (Doulatov et al., 2010; Goardon et al., 2011; Lee et al., 2015). However, it was also notable that the three populations showed considerable overlap in their transcriptional states.

New sorting strategies were devised to enrich for lymphoid and myeloid function within conventional sorting gates based on the surface marker levels of cells with different functional output in single-cell assays. CD38^{hi} GMPs were significantly enriched for myeloid output, and CD10^{hi} CD45RA^{hi} LMPPs were found to give a significant increase in lymphoid output compared to conventional LMPPs. Together with the gene expression analysis these findings could suggest a continuum of lympho-myeloid bias in the human progenitor populations, rather than a strict hierarchical structure (Laurenti and Göttgens, 2018).

5.4.2 Functional assays identify single cell with both lymphoid and myeloid potential

Within both LMPP and GMP compartments there were a number of cells (10% and 7% of positive wells, respectively) giving rise to mixed lymphoid and myeloid colonies. This is particularly interesting in light of recent work that has supported the idea of human progenitors being mostly unilineage (Notta et al., 2016; Velten et al., 2017). Whilst the frequency of bipotent progenitors was fairly low, there are limitations to functional culture assays as a unilineage output from a cell does not necessarily indicate it was restricted to only one lineage: if specification occurred before cell division in the culture then only one output would be seen.

The question of whether it was possible to enrich for bipotent lympho-myeloid cells within the conventional gates was considered, as if these bipotent cells could be isolated then the transcriptional signatures specific to these cells could be examined. However, these cells did not clearly separate from unipotent cells based on their surface marker profiles, perhaps due to the difference between true potential and functional output, or alternatively because the limited surface markers measured were not sufficient to distinguish between these functionally distinct cells.

5.4.3 Choice of cord blood to study progenitor behaviour

The experiments here measuring single-cell function and gene expression were performed on human progenitors isolated from cord blood. However, differences in the behaviour of progenitors from cord blood and adult bone marrow have previously been observed (Notta et al., 2016). Cord blood rather than bone marrow progenitors were used to assay the function of the lympho-myeloid human progenitors for this work as the LMPP and MLP cells were too rare in bone marrow (Karamitros et al., 2018). Further experiments would be needed to investigate whether similar functional output and gene expression patterns could be seen in bone marrow progenitors.

5.4.4 Future directions

Further investigation of differential expression between the RNA-seq clusters could be explored to see whether genes encoding surface marker proteins could be used to separate these more immature cells from more mature populations. This would allow better investigation of the gene expression programs activated during lymphoid and myeloid differentiation. In particular, the maturation of LMPP cells to more mature progenitors would be an interesting avenue to explore, as in acute myeloid leukaemia there is a substantial increase in the number of LMPP-like cells (Goardon et al., 2011), indicating dysregulation of the control of progenitor numbers in this disease state.

5.4.5 Summary

In summary, this chapter describes the analysis of functional and transcriptional data from human cord blood progenitors supporting the existence of a continuum in the differentiation of cells towards either myeloid or lymphoid lineages.

Chapter 6

Characterising transcriptional changes in the haematopoietic landscape

Parts of this section have been modified from Dahlin et al. (2018), on which F. Hamey is joint first author. Experimental work for this project was carried out by Nicola Wilson (isolation of primary bone marrow cells and scRNA-seq profiling) and Mairi Shepherd (isolation of primary bone marrow cells). Joakim Dahlin and Nicola Wilson provided assistance with finding marker genes to identify the different haematopoietic lineages. The abstracted graph algorithm was developed by Alex Wolf. Computational analysis of the data was carried out by F. Hamey.

6.1 Background

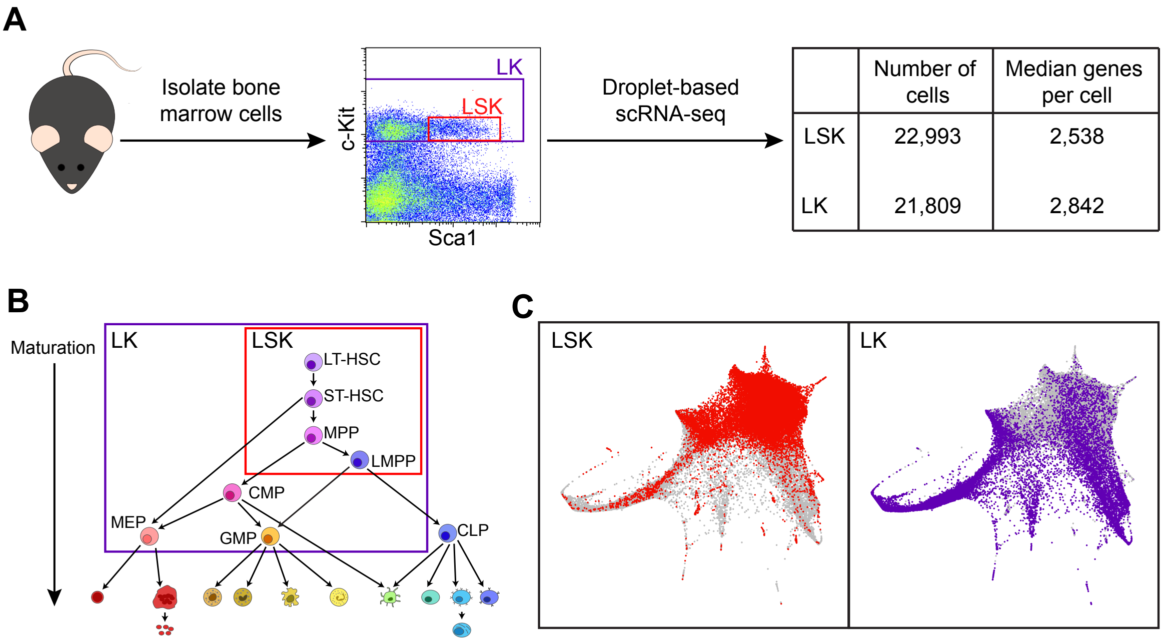
Chapters 3-5 presented analysis demonstrating that single-cell gene expression data can be used to reconstruct haematopoietic differentiation. Depending on the number of cells, genes and populations profiled, each dataset provided insights focusing on different aspects of haematopoiesis. In this previous work, murine bone marrow HSPCs were profiled with both scRNA-seq and qRT-PCR, yet due to constraints on the number of cells or genes profiled these data remain limited in their ability to resolve rare progenitor populations. In addition, both datasets were generated by isolating specific haematopoietic populations, and therefore were unable to capture representative proportions of the different cells types within the bone marrow HSPC compartment.

Recent technological advances have now made it possible to perform single-cell gene expression assays on increased numbers of cells at reduced cost (Zheng et al., 2017), enabling transcriptional landscapes to be generated based on much larger sample sizes. It was reasoned that this ability to profile so many cells could be used to capture a more representative snapshot of haematopoietic populations at their true densities within the HSPC compartment, which would allow insights into how the transcriptional landscape is altered in perturbed states.

The aim of this work was to generate a high quality unbiased single-cell landscape with data generated using a droplet-based scRNA-seq method. By capturing high numbers of cells these data would be able to cover the HSPC populations without the gap between Lin$^-$ c-Kit$^+$ Sca1$^-$ and Lin$^-$ c-Kit$^+$ Sca1$^+$ gates seen in Chapter 3. It was hoped that from these data it would be possible to construct differentiation trajectories towards a higher number of lineages than before, including rare cell types, so that the gene expression changes around branch points towards different cell fates could be investigated.

6.2 Entry points to eight haematopoietic lineages are resolved by droplet-based scRNA-seq profiling

A commercial droplet-based scRNA-seq platform (10x Chromium™ from 10x Genomics) was used to generate single-cell gene expression profiles of HSPCs from mouse bone marrow (Zheng et al., 2017). To capture progenitor cells at representative densities in the landscape, bone marrow samples were lineage depleted and sorted based on c-Kit marker expression in the Lin$^-$ c-Kit$^+$ (LK) gate (Fig. 6.1A). After quality control filtering to remove low quality profiles and potential doublets (see Chapter 2 for details) this dataset detected an average of around 2,800 genes in 21,809 cells. The LK sorting gate contains both stem cells and more specialised progenitor populations (Fig. 6.1B), but the most mature progenitors are found at much higher frequency in the bone marrow (Fig. 6.1A). As stem and early progenitors are so rare, Lin$^-$ c-Kit$^+$ Sca1$^+$ (LSK) cells, a subset of the LK fraction, were also isolated from the same pooled bone marrow samples and profiled alongside the LK cells (Fig. 6.1A, B). Together, 44,802 single-cell gene expression profiles were captured with an average of over 2,500 genes detected per cell. Interestingly, fewer genes were detected on average in the LSK cells than in LK cells, in keeping with the more quiescent nature of the immature cells at the top of the hierarchy. This observation is also in agreement with the quantification of the SMART-Seq2 scRNA-seq data seen in Chapter 3, which showed that the stem and early progenitors had lower RNA content than the more mature progenitors.



	Number of cells	Median genes per cell
LSK	22,993	2,538
LK	21,809	2,842

Fig. 6.1. Droplet-based scRNA-seq can be used to capture over 40,000 molecular profiles from the HSPC compartment. (A) Schematic of how cells were isolated for scRNA-seq profiling based on c-Kit and Sca1 expression. Table shows post quality-control summary of the data. LK, Lin$^-$ c-Kit$^+$; LSK, Lin$^-$ c-Kit$^+$ Sca1$^+$. (B) Diagram depicting which conventional haematopoietic populations are within the sorting gates. (C) Force-directed graph visualisations calculated on the k-nearest neighbour graph of LSK and LK profiles processed together. Left-hand panel highlights the LSK cells in red, right-hand panel highlights the LK cells in purple.

The gene expression data were then visualised using a force-directed layout calculated on the k-nearest neighbour graph of single-cell profiles, where edges represent the connections between the most similar cells (Weinreb et al., 2018a). As cells from a single lineage are closely connected they are positioned closely in the two-dimensional embedding, hence this method has previously been successfully applied to datasets to represent complex branching differentiation structures. Highlighting LSK and LK cells separately showed overlap between the two populations, with LSK cells more tightly positioned at the top of the graph and more LK cells lying on the branches visible in the two-dimensional structure (Fig. 6.1C). Plotting gene expression for markers of several haematopoietic lineages on the force-directed graph demonstrated that cells from different progenitor populations occupied distinct regions of the landscape (Fig. 6.2). HSCs were marked by expression of *Procr* and *Fgd5*, and resided at the top of the structure (Balazs et al., 2006; Gazit et al., 2014). Cells from the lymphoid lineage were highlighted by genes such as *Dntt* and *Flt3* (Rothenberg, 2014), erythroid progenitors by *Klf1* and *Epor* (Dzierzak and Philipsen, 2013), and megakaryocytic progenitor cells by *Pf4* and *Itga2b* (Rowley et al., 2011). Within the myeloid lineage, *Elane* and *Cebpe* expression identified neutrophil progenitors and *Irf8* and *Ly68* monocyte progenitors (Olsson et al., 2016).

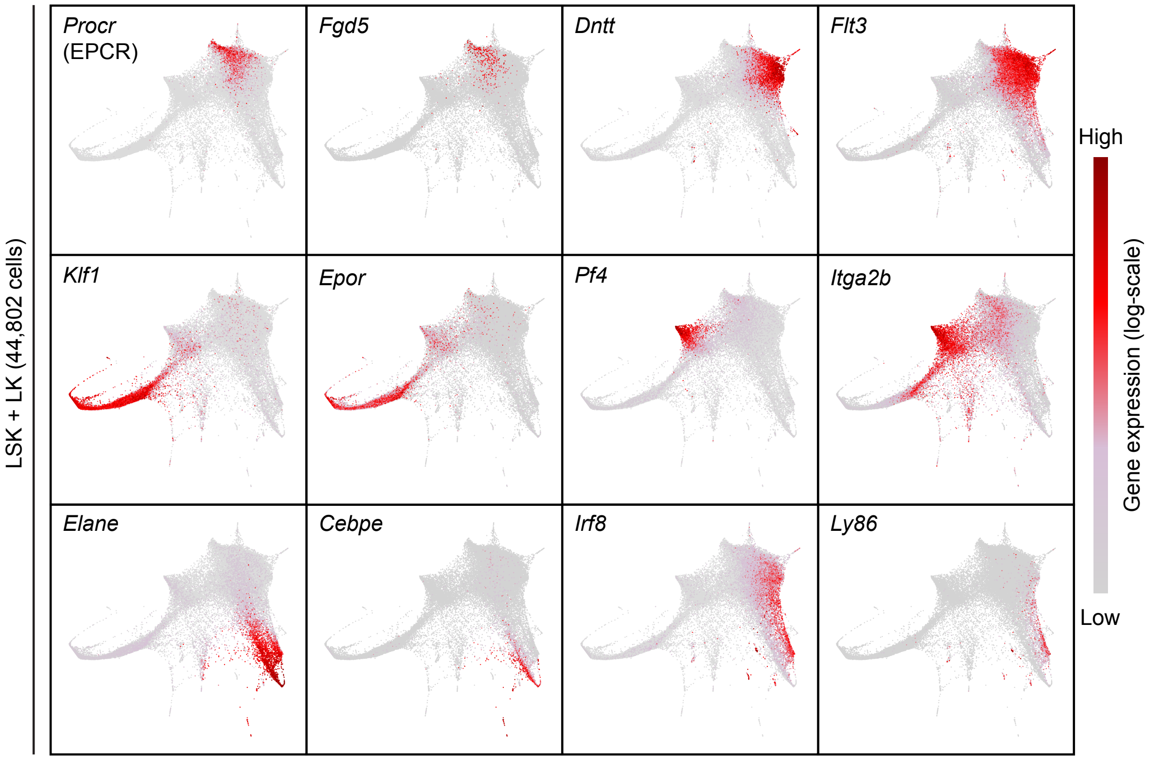


Fig. 6.2. Dimensionality reduction of single-cell profiles visualises branches belonging to different blood lineages. Force-directed graph embedding of LSK and LK cells coloured by log-transformed expression of marker genes highlighting the different blood lineages.

As such a high number of cells had been profiled it was also possible to identify rare populations, including entry points to mast cell, basophil and eosinophil lineages (Fig. 6.3). Genes *Ms4a2* and *Cpa3* were expressed in both basophil and mast cell progenitors (Dwyer et al., 2016), *Gzmb* and *Cma1* marked the mast cell progenitor population (Dwyer et al., 2016; Pardo et al., 2007) and *Prss34* and *Mcpt8* highlighted cells in the basophil lineage (Dwyer et al., 2016; Ugajin et al., 2009). A small group of eosinophil progenitors was also identified by the expression of genes such as *Prg2* and *Prg3* (de Graaf et al., 2016; Olsson et al., 2016).

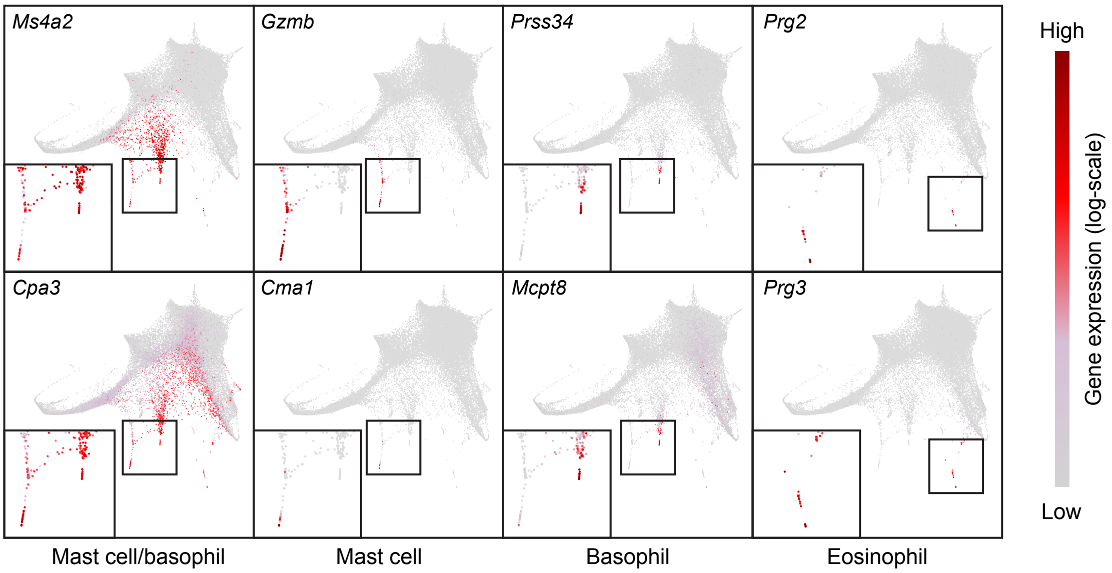


Fig. 6.3. Generating a high number of single-cell profiles allows rare progenitor populations to be identified. Force-directed graph embedding of LSK and LK cells coloured by log-transformed expression of marker genes for mast cell, basophil and eosinophil lineages. Regions of interest are enlarged for each plot.

Analysis of the LSK and LK cells together revealed considerable overlap between the two populations (Fig. 6.1C). To see whether the overall structure of the data was similar in the LSK or LK populations on their own these two sets of cells were processed separately and visualised using force-directed graphs (Fig. 6.4A, B). Cells from the LSK gate were more tightly connected and, whilst regions expressing different marker genes were visible including populations expressing *Klf1* and *Elane*, many of these were formed from relatively few cells compared to the corresponding regions in the combined graph (Fig. 6.4A). The graph calculated on only the LK cells had a more similar shape to the combined graph, but was less dense in the central region (Fig. 6.4B). Cells expressing HSC gene *Procr* and lymphoid gene *Dntt* were mainly seen in the LSK graph. LSK cells were closer in high dimensional space to their LSK neighbours than LK cells were to other LK cells, in agreement with the differences in heterogeneity suggested by the visualisations (Fig. 6.4C).

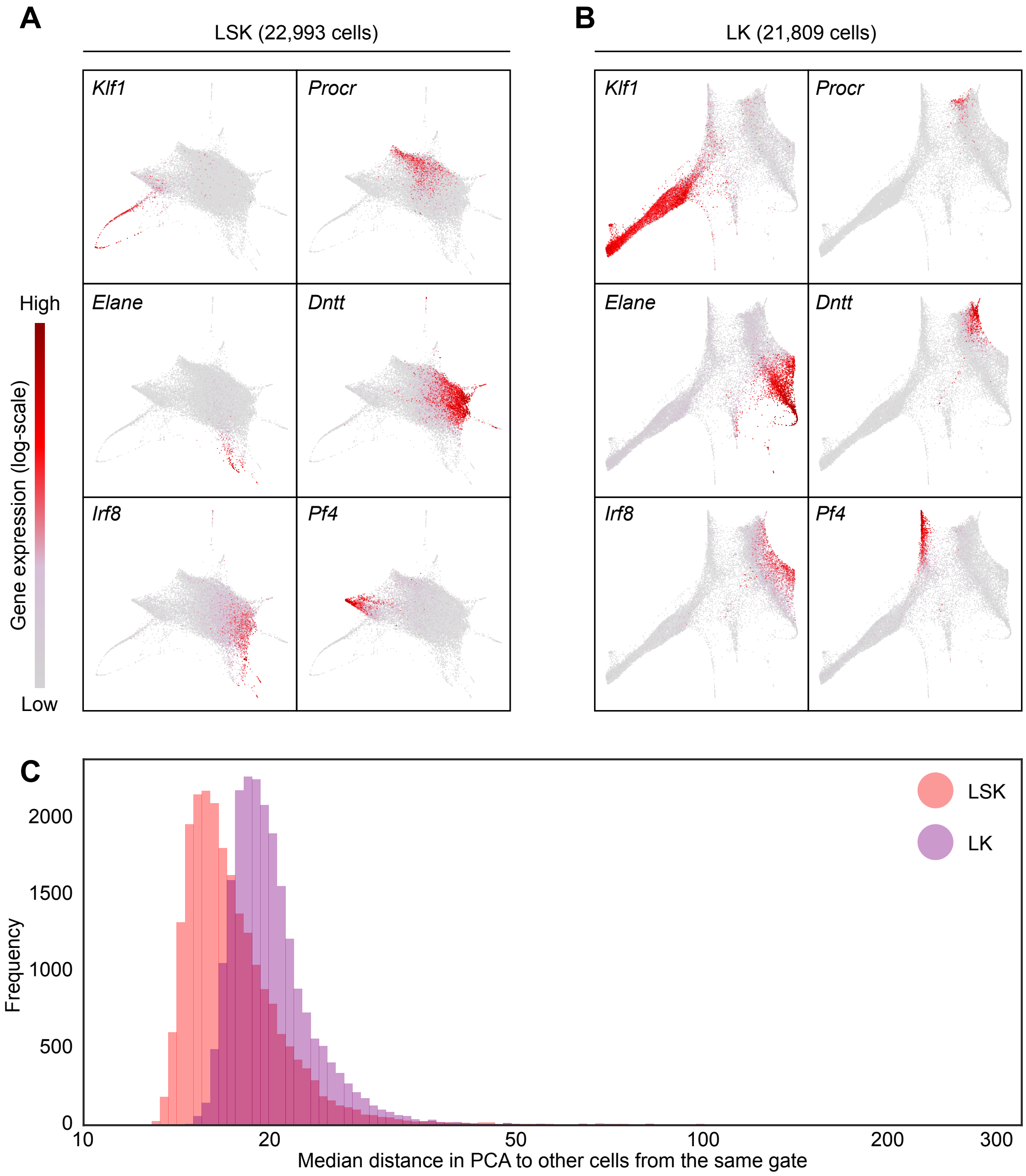


Fig. 6.4. LSK cells represent a more homogeneous population than LK cells. (A) Force-directed graph calculated on only LSK cells coloured by marker gene expression with log scale. (B) Force-directed graph calculated on only LK cells coloured by marker gene expression with log scale. (C) Histograms showing distributions of the median distance to other cells within the same gate for either LSK or LK populations. Distances were calculated between the PCA coordinates of cells, with PCA performed on LSK and LK cells together.

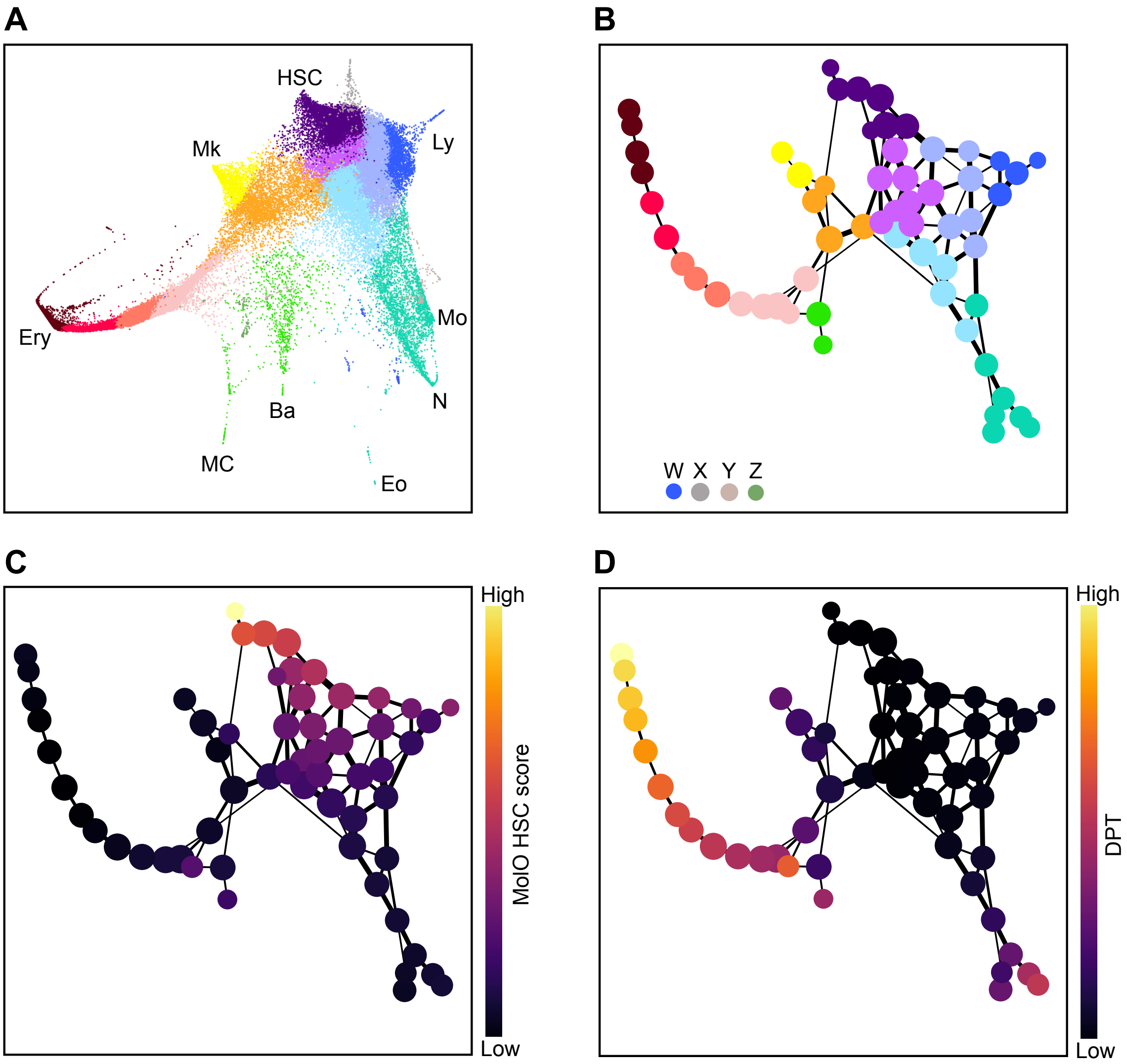


Fig. 6.5. Graph abstraction groups cells along paths through differentiation. (A) Force-directed graph embedding coloured by coarse louvain clustering. This was calculated using a k-nearest neighbour graph constructed based on the diffusion map coordinates. Ery, erythroid; Mk, megakaryocyte; MC, mast cell; Ba, basophil; HSC, stem cell; Eo, eosinophil; N, neutrophil; Mo, monocyte; Ly, lymphoid. (B) Results of graph abstraction calculated on high resolution louvain clustering of LSK and LK cells. These fine clusters are coloured by the coarse cluster that the majority of their cells belong to. Edge weights are proportional to the confidence of a connection between clusters. Only edges with confidence above a certain threshold are shown. Four clusters W, X, Y & Z had no connections with confidence above this threshold and are therefore disconnected from the main graph. These clusters were excluded from further analysis. (C) Abstracted graph with each node coloured by its mean MolO HSC score. The MolO HSC score for each cell represents the mean of MolO HSC genes from Wilson et al. (2015). (D) Abstracted graph coloured by the average pseudotime value for the cells in each node, with pseudotime ordering calculated from the cell with the highest MolO score. DPT, diffusion pseudotime.

6.3 The haematopoietic tree can be reconstructed from single-cell data using graph abstraction

As the location of both stem cells and the entry points to multiple lineages could be detected within the scRNA-seq landscape, the next question was to ask how these populations were connected, with the aim of reconstructing differentiation trajectories. Due to the high number of lineage entry points it was not a simple task to assign cells to differentiation pathways, as many existing algorithms were only able to cope with simple branching leading to at most two fates (Haghverdi et al., 2016; Setty et al., 2016). Clustering the data using louvain clustering based on the k-nearest neighbour graph between cells showed good agreement with the force-directed graph visualisation (Fig. 6.5A), and was able to assign many of the cells belonging to the main lineages to separate clusters, including megakaryocyte and lymphoid progenitors. The clustering also split the trajectory towards erythroid differentiation into several clusters. However, at this resolution, some of the clusters contained cells that were known to belong to more than one lineage. One such example was neutrophil and monocyte progenitor cells (Fig. 6.5A). To separate these progenitors, clustering at a higher resolution was performed that assigned cells into 63 different clusters.

The next aim was then to reconcile the discrete clustering approach with the concept of pseudotime analyses, which identify continuous differentiation trajectories through the data. For this, an algorithm called graph abstraction was applied (Wolf et al., 2017). This tests connectivity between clusters, based on the number of edges lying between cells from different cluster pairs in the k-nearest neighbour graph of cells. Connections with high confidence compared to a random distribution of edges are retained, and those with low confidence are removed, resulting in a graph where the nodes correspond to clusters and an edge suggests a strong similarity between the cells in the linked cluster pair (Fig. 6.5B). As some clusters only have low confidence connections to the rest of the graph, these nodes are not connected to the main graph structure (Fig. 6.5B). These disconnected clusters could be annotated by considering marker gene expression. Cluster W appeared to contain common dendritic cell precursors, as this group of cells highly expressed genes including *Ctss*, *Fcer1g*, and *Pld4*. Interferon response genes such as *Iigp1* and *Ifi1*, rather than genes corresponding to a specific haematopoietic lineage, were high in cluster X. Cluster Y had elevated expression of *Tcf7*, *Tox*, *Zbtb16*, *Gata3* and *Il7r*, which are genes that have been described as being expressed in innate lymphoid progenitors (Yu et al., 2016). Finally, cluster Z cells exhibitied high expression of B cell marker genes including *Cd79a*, *Ebf1* and *Vpreb3*. These four clusters may represent contaminating populations or progenitors with rare intermediate steps that are not captured in the scRNA-seq data, and were removed from the subsequent analysis as they are not closely related to any other clusters in the abstracted graph. Constructing the abstracted graph means that the nodes, which represent groups of cells with similar transcriptional profiles, are linked based on their proximity in the differentiation landscape. To understand how this clustering partitioned the gene expression space, a score was calculated for each cell, representing the average expression of the MolO HSC genes from Wilson et al. (2015) (Fig. 6.5C). MolO genes are a set of genes highly expressed in the HSC population, and so a high MolO score is expected in the part of the graph corresponding to the stem cells. The node at the top of the abstracted graph had the highest average MolO score, with a decreasing score for nodes further into the landscape. To obtain a measure of the distance from HSCs, pseudotime ordering from the stem cells was calculated by using DPT (Haghverdi et al., 2016) and visualised along the branches of the abstracted graph (Fig. 6.5D). Higher pseudotime values could be seen towards the tips of the branches, in support of graph abstraction capturing the structure of haematopoietic differentiation.

Average expression of different lineage marker genes was calculated for each node in the abstracted graph, demonstrating that the clusters were arranged on branches towards the erythroid, megakaryocyte, neutrophil, monocyte and lymphoid lineages (Fig. 6.6A). As a demonstration of how the abstracted graph structure can be used as a starting point to computationally reconstruct differentiation trajectories, shortest paths through the weighted graph from the HSC node (with the highest MolO score) towards the five different lineages were found (Fig. 6.6B). With this approach, each trajectory was formed from between 11 and 22 clusters, splitting cells into groups at different stages of differentiation.

6.4 Identification of genes with lineage-specific dynamics around differentiation branch points

As a proof of concept for an application of constructing trajectories using the abstracted graph, it was decided to focus on the branching point between erythroid (Ery) and megakaryocyte (Mk) differentiation. To reveal dynamic genes at this point in the landscape, the first nodes unique to either the Ery or Mk trajectories were compared to the preceding node shared between the two trajectories (Fig. 6.7A). Differential expression between these

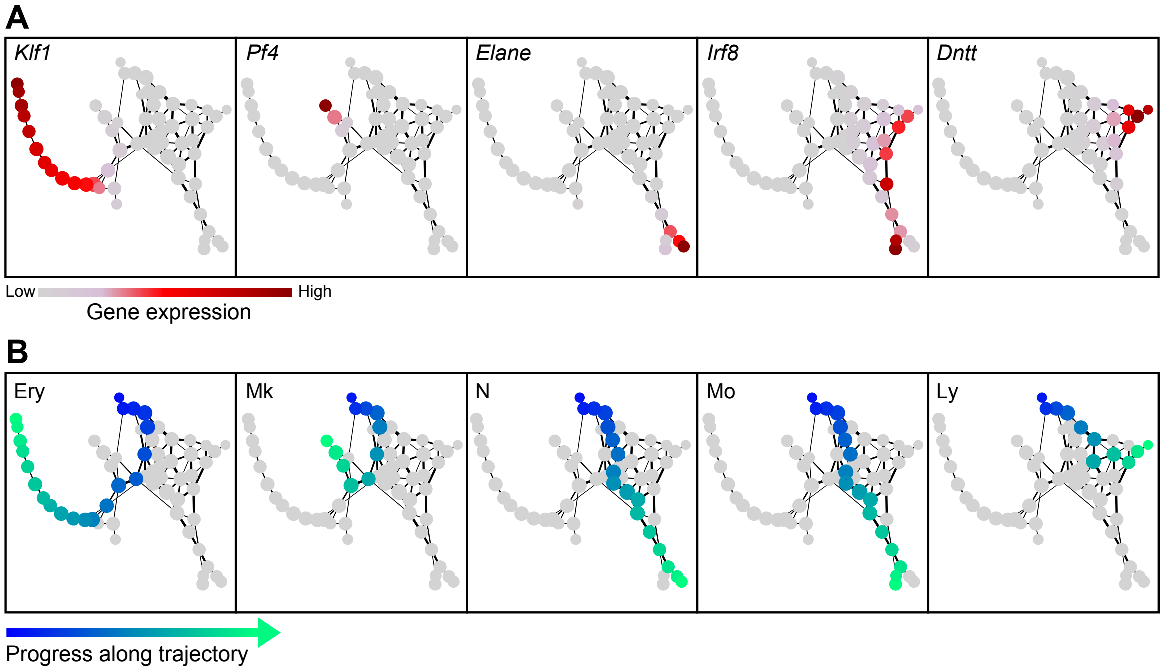


Fig. 6.6. Connections between abstracted graph nodes can be used to build differentiation trajectories. (A) Abstracted graph with nodes coloured by the mean expression of lineage-specific marker genes. (B) Trajectories from haematopoietic stem cells to five different lineages in the abstracted graph. Nodes lying along a trajectory are coloured from blue to green in order of differentiation. Ery, erythroid; Mk, megakaryocyte; N, neutrophil; Mo, monocyte; Ly, lymphoid.

nodes identified 86 genes upregulated for this step of Ery differentiation, and 26 for Mk differentiation (Fig. 6.7B). The dynamics for the most significant of these genes were then visualised around the branch point (Fig. 6.7C, D). Several of the genes were known regulators of differentiation towards erythrocytes or megakaryocytes. For example, *Klf1* was amongst the most significant genes upregulated during Ery but not Mk differentiation. Towards Mk differentiation *Pf4*, *Gp5* and *Cd9*, *F2r* and *F2rl2* were all significantly upregulated and are known to be expressed during megakaryopoiesis. Amongst both gene sets were genes with clear upregulation in only one trajectory, such as *Lpin2* in the Ery trajectory and *Rap1b* in the Mk trajectory. Therefore, this approach both identifies known regulators and provides new candidate genes potentially involved in the fate decision towards these two lineages.

To delve deeper into the gene dynamics around the Ery-Mk branch point, a search was carried out to identify genes showing expression patterns consistent with lineage priming (Fig. 6.8A). These were defined as genes displaying intermediate expression in the shared node, along with uprgulation towards one lineage *and* downregulation towards the other. 12 genes

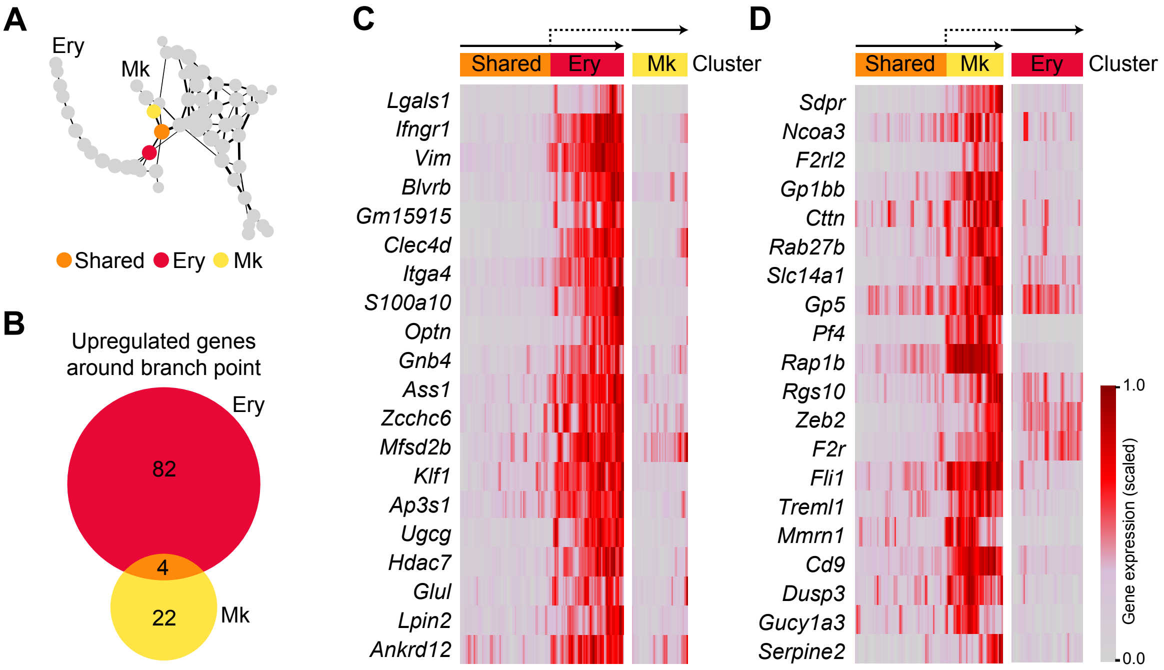


Fig. 6.7. Graph abstraction can be used to discover lineage-specific gene sets around erythroid-megakaryocyte differentiation branching point. (A) Abstracted graph highlighting the final shared node lying on both erythroid and megakaryocyte trajectories (orange), and the succeeding nodes on only the erythroid (red) or megakaryocyte (yellow) trajectories. (B) Venn diagram showing numbers and overlap of genes upregulated between the shared and succeeding nodes around the branch point. (C) Heatmap showing expression of the top 20 most significant genes upregulated between shared node and erythroid branch point. Cells are ordered by pseudotime within each cluster and gene expression is smoothed by a running average of 50 cells. Expression is scaled between 0 and 1 across the union of the Shared, Ery and Mk nodes. (D) Heatmap showing the expression of top 20 most significant genes upregulated between the shared and succeeding nodes around the megakaryocyte branch point. Ery, erythroid; Mk, megakaryocyte.

displayed expression consistent with Ery priming, and 24 genes had dynamics in keeping with Mk priming (Fig. 6.8B, C). Interestingly, several transcription factor encoding genes were amongst those with "primed" expression patterns. For example *Bcl11a* and *Myb* were both upregulated during Ery differentiation and downregulated during Mk differentiation, and *Lmo2*, *Pbx1*, *Fli1* and *Cited2* were amongst the Mk priming genes. Searching for genes with these patterns aimed to identify genes involved in regulating the Ery-Mk fate decision, and such a list therefore represents candidates for future experimental validation.

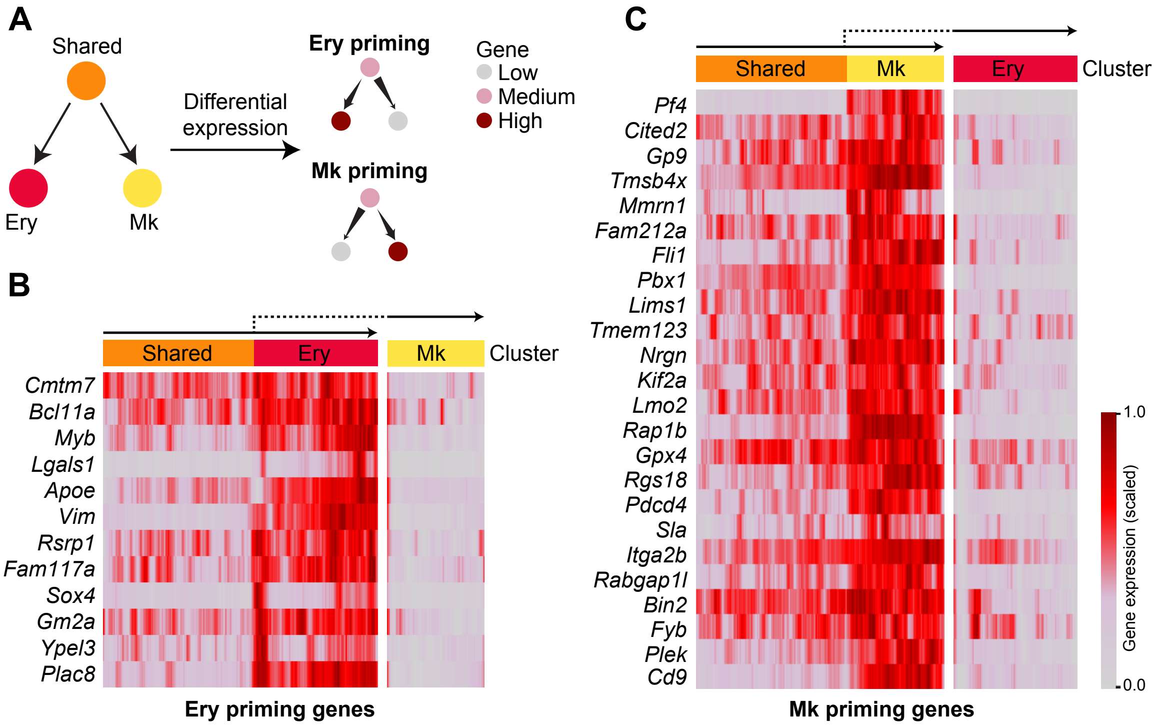


Fig. 6.8. Genes exhibit expression dynamics consistent with lineage priming. (A) Diagram explaining how differential expression analysis between nodes was used to search for genes with lineage-priming patterns. (B) Heatmap showing expression of genes with upregulation along the Ery branch and downregulation along the Mk branch, consistent with Ery priming. Cells are ordered by pseudotime within each cluster and gene expression is smoothed by a running average of 50 cells. Expression is scaled between 0 and 1 across the union of the Shared, Ery and Mk nodes. (C) Heatmap showing expression of genes with upregulation along the Mk branch and downregulation along the Ery branch, consistent with Mk priming. Ery, erythroid; Mk, megakaryocytic.

6.5 Conclusions

The work in this chapter presents analysis of a reference transcriptomic landscape of the HSPC compartment formed from over 40,000 scRNA-seq profiles. Visualising the expression of marker genes on force-directed graphs of the data identified a complex branching structure with entry points to eight haematopoietic lineages. An approach combining the concepts of discrete clustering with continuous pseudotime ordering was then used to assign cells to several differentiation trajectories starting from stem cells, and to characterise gene expression changes around a branch point between two haematopoietic lineages.

6.5.1 The challenges of sampling large cell numbers

Mouse bone marrow is formed from a complex mixture of cells, with frequencies ranging from very abundant, for example erythroid progenitors, to cells found in much smaller numbers, such as the very rare LT-HSCs, which represent less than 1 in 1,000 of the LK bone marrow cells and less than 1 in 20,000 of all nucleated bone marrow cells. Due to the rarity of such populations, it is vital to profile large enough samples to capture these cells in a dataset. As HSCs occur in much lower frequencies than many more differentiated progenitors it was decided to supplement the LK cells with 23,000 cells captured from the LSK gate. LT-HSCs represent around 1% of LSK cells, so this sample provides greater coverage of the upper parts of the haematopoietic hierarchy. However, although these additional data were incorporated into the analysis in this chapter, it is important to note that it is still possible to perform analysis on just the LK data, which are sampled at representative density across the transcriptional landscape. This will provide an important reference for comparisons with perturbation models where haematopoietic populations have shifted, as will be discussed in Chapter 7.

In switching to droplet-based sequencing to obtain higher cell numbers at a reasonable cost there is a compromise in the sequencing depth achieved per cell. For example, for the data in this chapter an average of 2,500 genes were detected per LSK cell, whereas 8,600 genes per cell were detected for the same population in Chapter 3 when LSK cells were profiled using SMART-Seq 2. Yet, despite this reduced sequencing depth, it was still possible to separate cells belonging to different lineages using dimensionality reduction and clustering. Also, compared to other published datasets profiling similar populations with scRNA-seq methods allowing such high throughput, the sequencing depth here was comparatively high (Giladi et al., 2018; Paul et al., 2015; Tusi et al., 2018). This is important if the aim is to use the data not just to understand the structure of haematopoietic differentiation but also to identify novel regulators of cell fate decisions.

Another limitation of using droplet-based rather than plate-based sequencing was that retaining surface marker information for the individual cells was not possible. This makes relating the cells to the conventional haematopoietic populations as in Section 3.5 more challenging, so could make it more difficult to isolate potentially interesting subpopulations from the data for further analysis. New techniques are now emerging to allow simultaneous measurement of the transcriptome and a selection of surface markers for each cell using barcoded antibodies (Stoeckius et al., 2017), which could be used in future work to address this limitation of droplet-based scRNA-seq data.

6.5.2 Viewing the data at different resolutions

An interesting observation from this Chapter is how the ability to view data at a range of resolutions can prove useful in understanding different features of the data. Using the single-cell profiles allows the whole landscape to be visualised and annotated using marker gene expression. Alternatively, the abstracted graph represents the data at a different resolution, which provides a method for identifying strongly connected groups of cells and enables differential expression to be calculated in specific regions of the graph by grouping similar cells together. Despite these different resolutions, it is clear that both the single-cell data and abstracted graphs capture similar structure within the data, as can be seen in Fig. 6.5.

6.5.3 Using graph abstraction to infer the differentiation structure

Graph abstraction is applied to the clustered single-cell data and does not require any supervision to decide which groups of cells are connected. This is an advantage compared to some other methods for finding complex differentiation trajectories, such as the PBA algorithm presented by (Weinreb et al., 2018b), which requires knowledge of the proportion of different lineages produced by stem cells as an input parameter. When used with the high resolution clustering, graph abstraction was able to identify branching towards the majority of haematopoietic lineages known to be present in the data, although it did fail to separate the very rare eosinophil progenitor population, as this represents a very small number of cells that did not form their own cluster. A challenge in analysing this type of data covering a mixture of cell types is that the highly variable genes will be dominated by those genes varying across the most abundant cell types, and therefore these genes will probably prove more successful in subdividing these populations. For example, the erythroid progenitor trajectory contains a very high number of cells and is divided into the highest number of different clusters. Another interesting observation from the graph abstraction is that some small populations of cells were not part of the connected graph structure, as they had very low confidence connections to other groups of cells. In some cases these could correspond to contaminating populations of cells, or perhaps their preceding progenitor populations are very rare so do not represent a high enough proportion of a cluster to form any connections.

6.5.4 Characterisation of gene expression around a branch point

Several of the genes identified as differentially expressed between differentiation towards the erythroid or megakaryocyte lineages are known to be expressed in one of the two progenitor populations, demonstrating that this approach can find established lineage-specific genes. For the remaining genes, further investigation will be required to establish whether they play a role in the cell fate decision making between the two lineages, and therefore this gene set should be treated as a list of candidates for possible validation. Searching Ery or Mk priming patterns revealed a number of genes, including several transcription factors, that had interesting expression dynamics around the branch point between these lineages. These genes will need to be investigated for their role in the decision making process by perturbing progenitors upstream to this branching and assessing any changes in the lineage output of these cells. Overall, applying graph abstraction to these data demonstrates that this approach can be useful in understanding gene expression changes at points in the transcriptional landscape that were previously inaccessible using the conventional strategies for isolating haematopoietic progenitors. Additionally, combining the differential expression analysis with pseudotime ordering within clusters could provide a framework for establishing the order of activation of different genes during the differentiation process.

6.5.5 Future work

Future work will first focus on working to identify potential targets for experimental validation of the erythroid-megakaryocyte branch point-related genes, based on investigation into the known roles of these genes and how specifically they are expressed in cells around the branch point. Characterisation of the genes involved in different lineage fate decisions and at different stages in individual differentiation trajectories could also be carried out to further understand which genes are dynamic at specific regions in the landscape. It would also be interesting to see whether the branch points themselves could be validated by looking for surface marker combinations expressed by these populations to isolate these cells by FACS for experimental assays. Additionally, as well as using these data to characterise normal haematopoiesis, they also have exciting potential for use as a reference dataset for comparison with perturbed haematopoiesis. This concept forms the basis for the work presented in Chapter 7.

6.5.6 Summary

In summary, the work in this chapter has used single-cell gene expression data to identify entry points to eight different blood lineages, and has shown how graph abstraction can be used to construct differentiation trajectories within this complex branching dataset. This allowed characterisation of gene expression changes at specific points in the transcriptional landscape.

Chapter 7

How does a *Kit* mutation alter the haematopoietic landscape?

Parts of this chapter have been modified from Dahlin et al. (2018), on which F. Hamey is joint first author. Experimental work for this project was carried by Nicola Wilson (isolation of primary bone marrow cells and scRNA-seq profiling), Joakim Dahlin (isolation of primary bone marrow cells and FACS analysis), and Mairi Shepherd (isolation of primary bone marrow cells). Computational analysis was carried out by F. Hamey.

7.1 Background

As discussed in the previous chapter, characterising the haematopoietic landscape using droplet-based scRNA-seq was motivated by the desire to generate a reference with which to compare haematopoiesis in perturbed states. To investigate how the landscape can be altered by a genetic perturbation, it was decided to consider a mouse model with a mutation in the *Kit* gene. This gene is widely expressed across the haematopoietic compartment and is involved in the maintenance of HSCs. The mouse model chosen was the W^{41}/W^{41} mouse, which has a V831M mutation in the *Kit* gene, leading to impaired c-Kit kinase activity (Nocka et al., 1990). These mice have mostly normal haematopoiesis, but suffer from mild anaemia and a mast cell deficiency (Geissler and Russell, 1983a,b). The aim of this chapter was to analyse scRNA-seq data from the HSPC compartment of W^{41}/W^{41} mice to study how the single-cell landscape changes in this defective signalling environment.

This work discusses scRNA-seq data profiling of over 13,000 LK cells from the bone marrow of W^{41}/W^{41} mice, and compares this with the reference transcriptional landscape of Chapter 6. Computational analysis of the transcriptional profiles revealed differences in cell type composition across the samples, as well as providing insights into changes occurring at the molecular level.

7.2 W^{41}/W^{41} *Kit* mutant mice lack a distinct mast cell trajectory in the single-cell landscape

To see how the haematopoietic progenitor compartment was altered in the presence of reduced c-Kit signalling, bone marrow HSPCs were isolated from two W^{41}/W^{41} mice and processed for droplet-based scRNA-seq, following the same protocol as in Chapter 6. As the aim was to compare with the existing WT landscape from this earlier chapter, only LK cells were isolated from W^{41}/W^{41} bone marrow for comparison with the 21,809 WT LK cells. Restricting the analysis to only LK cells ensured that the sorting gates sampled the same populations at representative frequencies (Fig. 7.1A). Cells from W^{41}/W^{41} bone marrow were processed using the same quality control criteria as the WT cells, leading to a dataset consisting of 13,815 transcriptional profiles. To visually inspect differences between the perturbed and reference datasets, dimensionality reduction in the form of a force-directed layout was calculated on the W^{41}/W^{41} cells (Fig. 7.1B). Even though this embedding was generated independently from the one on the WT cells, the two structures still showed very similar shapes. Visualisation of marker gene expression on both graphs confirmed the similarity between their structures. However, one noticeable exception was the lack of a distinct mast cell branch in the W^{41}/W^{41} mice, along with reduced expression of mast cell marker genes such as *Cma1* in cells from these animals (Fig. 7.1B, lower panels). These observations are in agreement with the severe mast cell deficiency seen in the W^{41}/W^{41} mice (Ingram et al., 2000). Basophil progenitors, marked by *Prss34*, were still clearly present in the W^{41}/W^{41} bone marrow.

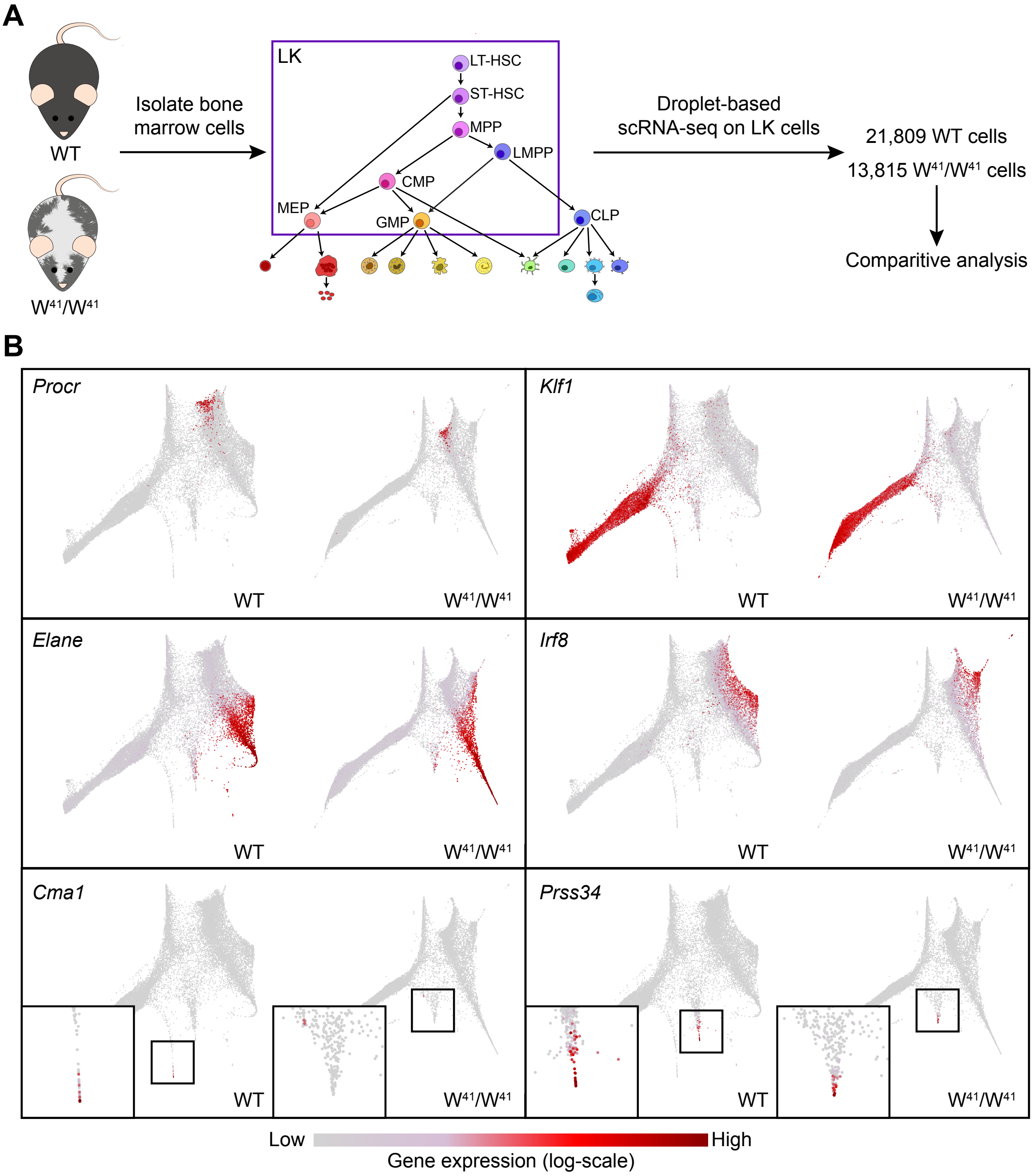


Fig. 7.1. Droplet-based scRNA-seq can be used to compare the transcriptional landscape of a *Kit* mutant mouse to a WT reference dataset. (A) Schematic showing an overview of the experiment. WT LK cells were from the dataset discussed in Chapter 6. (B) Force-directed layouts on the WT and W^{41}/W^{41} LK cells. Layouts were calculated independently based on the highly variable genes for each dataset. Gene expression was normalised separately for each dataset, log-transformed, and then visualised for each pair of plots using the same colour scale. Insets in lower panels show magnified regions of the plots.

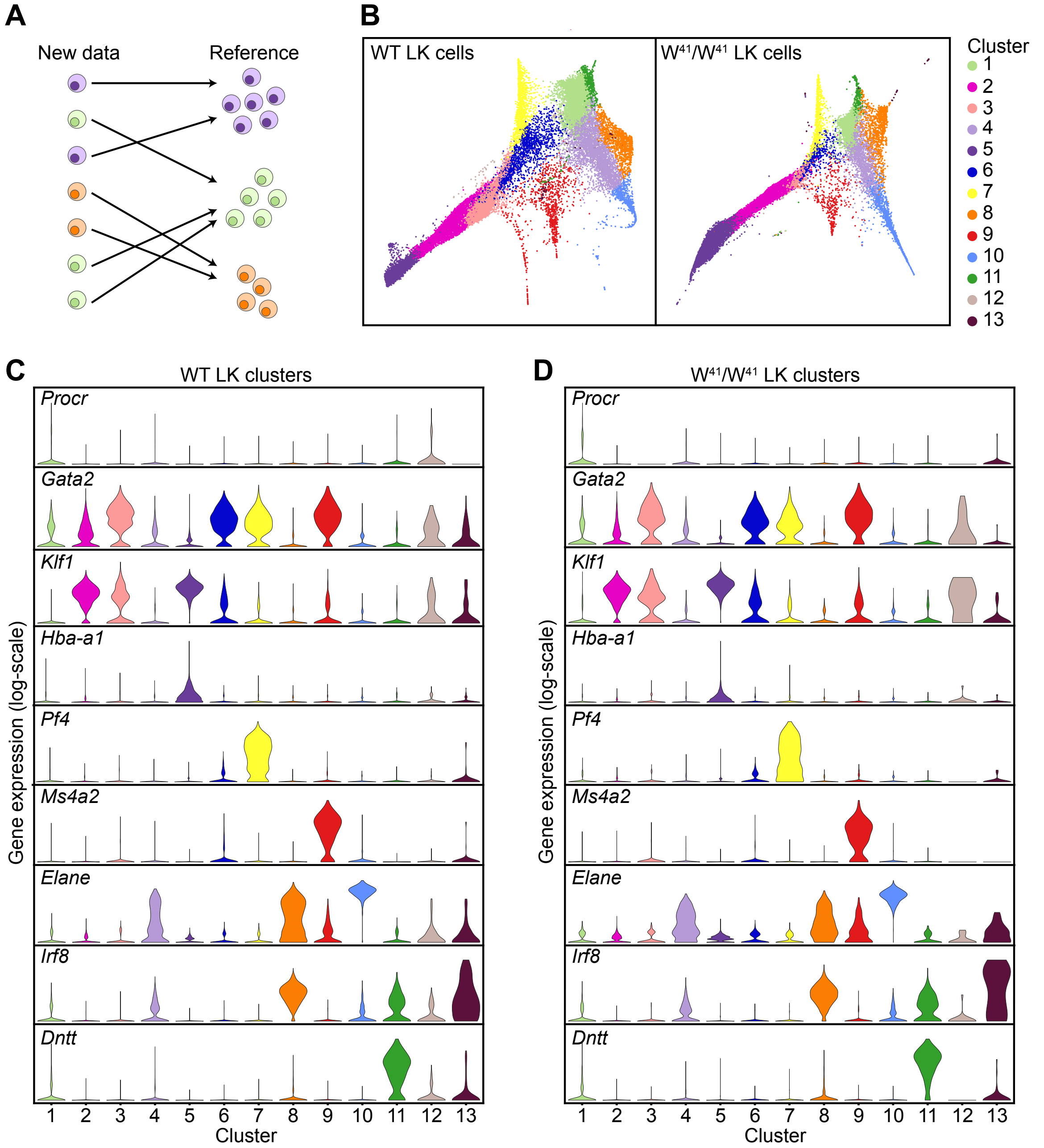


Fig. 7.2. *Kit* mutant cells can be assigned to clusters in the WT reference landscape. (A) Schematic illustrating mapping W^{41}/W^{41} LK data to WT clusters. (B) Force-directed layouts of WT and W^{41}/W^{41} cells coloured by cluster assignment. Left panel shows the result of louvain clustering on the WT data. Right panel shows the results of mapping W^{41}/W^{41} cells to WT clusters. (C, D) Violin plots of marker gene expression in the WT and W^{41}/W^{41} clusters. Genes were normalised separately in each dataset and are shown with a log-transformed scale.

7.3 Transcriptional landscape of W^{41}/W^{41} mice displays a shift towards more mature progenitor populations

Whilst dimensionality reduction was able to reveal qualitative changes in the W^{41}/W^{41} landscape, it could not be used to quantify any changes in the proportion of different HSPCs. scRNA-seq is a valuable tool for performing this quantification as it is not necessary to pre-define haematopoietic populations for isolation using surface-marker levels. In order to make this comparison it was first necessary to relate W^{41}/W^{41} cells to their WT counterparts, to assign cells to a set of progenitor groups common across the two datasets (Fig. 7.2A). Processing the data together showed a considerable batch effect, with large separation between the genotypes in dimensionality reductions. Such batch effects are common in single-cell data from different conditions, with several suggestions for alternative ways of combining data from different sources (Hamey and Göttgens, 2018; Wen and Tang, 2018). Here, the decision was made to initially cluster WT cells into groups, to provide a characterisation of the transcriptional states in normal haematopoiesis (Fig. 7.2A, B). Then W^{41}/W^{41} cells were mapped to this partitioning by assigning each W^{41}/W^{41} cell to its most similar WT cluster, based on the cluster identity of its closest WT neighbours (Fig. 7.2B). It was reassuring to see that W^{41}/W^{41} cells were assigned to all 13 clusters of the WT data (Fig. 7.2B). Plotting the expression of haematopoietic marker genes allowed cell type identities to be assigned to the clusters (Fig. 7.2C). For example, cluster 1 contained the HSCs, as indicated by high expression of *Procr*, whereas erythroid progenitor clusters 2, 3 and 5 were highlighted by *Klf1* expression. Distributions of marker gene expression were similar between the WT and W^{41}/W^{41} clusters, supporting the cluster assignment of the W^{41}/W^{41} cells (Fig. 7.2D). Clusters 12 and 13, the two smallest clusters, displayed less consistent expression patterns but also contained very few cells (96 and 91 cells in the WT data, respectively), and so more variable distributions were expected for these populations.

Next, to quantify changes in the composition of the HSPC compartment, the fold-change between the proportion of cells in each cluster for the WT and the W^{41}/W^{41} datasets was calculated (Fig. 7.3A). As cells from two different W^{41}/W^{41} mice were sequenced, this provided two repeats of how the cell type proportions were altered in the *Kit* mutant model in comparison with the WT reference. This analysis revealed some dramatic changes in the perturbed environment. Cluster 3 showed the biggest change in the proportion of cells, with around a 4-fold decrease in the percentage of cluster 3 cells in the W^{41}/W^{41} data. Inspection of marker gene expression indicated that this cluster corresponded to early erythroid progenitors

(Fig. 7.2C, D). Further along the erythroid trajectory, increases were seen in the proportion of cells in both clusters 2 and 5, with the more mature cluster 5 erythroid progenitors showing an over two-fold increase in W^{41}/W^{41} mice. Other populations exhibiting a relative expansion in the W^{41}/W^{41} mice were clusters 8 and 11, which corresponded to monocyte and neutrophil progenitors, respectively. The proportion of neutrophil progenitors roughly doubled and the monocyte progenitors displayed a more modest increase in the mutant bone marrow. Populations that were under-represented in the mutant landscape included cluster 1, which contained the stem and early progenitors, and cluster 9, which contained the mast cell and basophil progenitors. Together, these data suggested an overall reduction in the proportion of LK populations at the top tiers of the haematopoietic hierarchy, and highlighted a shift along the erythroid trajectory towards more mature erythroid progenitors.

A recently-described gating strategy was then applied (Tusi et al., 2018) to use FACS analysis to investigate the changes seen in the transcriptomic data (Fig. 7.3B). In the study describing this strategy, the authors defined five populations, denoted P1-P5, to isolate different haematopoietic cell types within the c-Kit^{+} compartment. Populations P5 $\rightarrow$ P2 $\rightarrow$ P1 represent a progression along the erythroid trajectory, with P5 containing the multipotent progenitor population. P3 cells were described as basophil progenitors, and P4 cells as megakaryocyte progenitors. For each of these populations its proportion as a percentage of the LK cells was calculated in both WT and W^{41}/W^{41} mice (Fig. 7.3C). These results validated the shift towards more mature cells along the erythroid trajectory, and also confirmed reductions in megakaryocyte and basophil progenitor populations similar to the gene expression-based analysis.

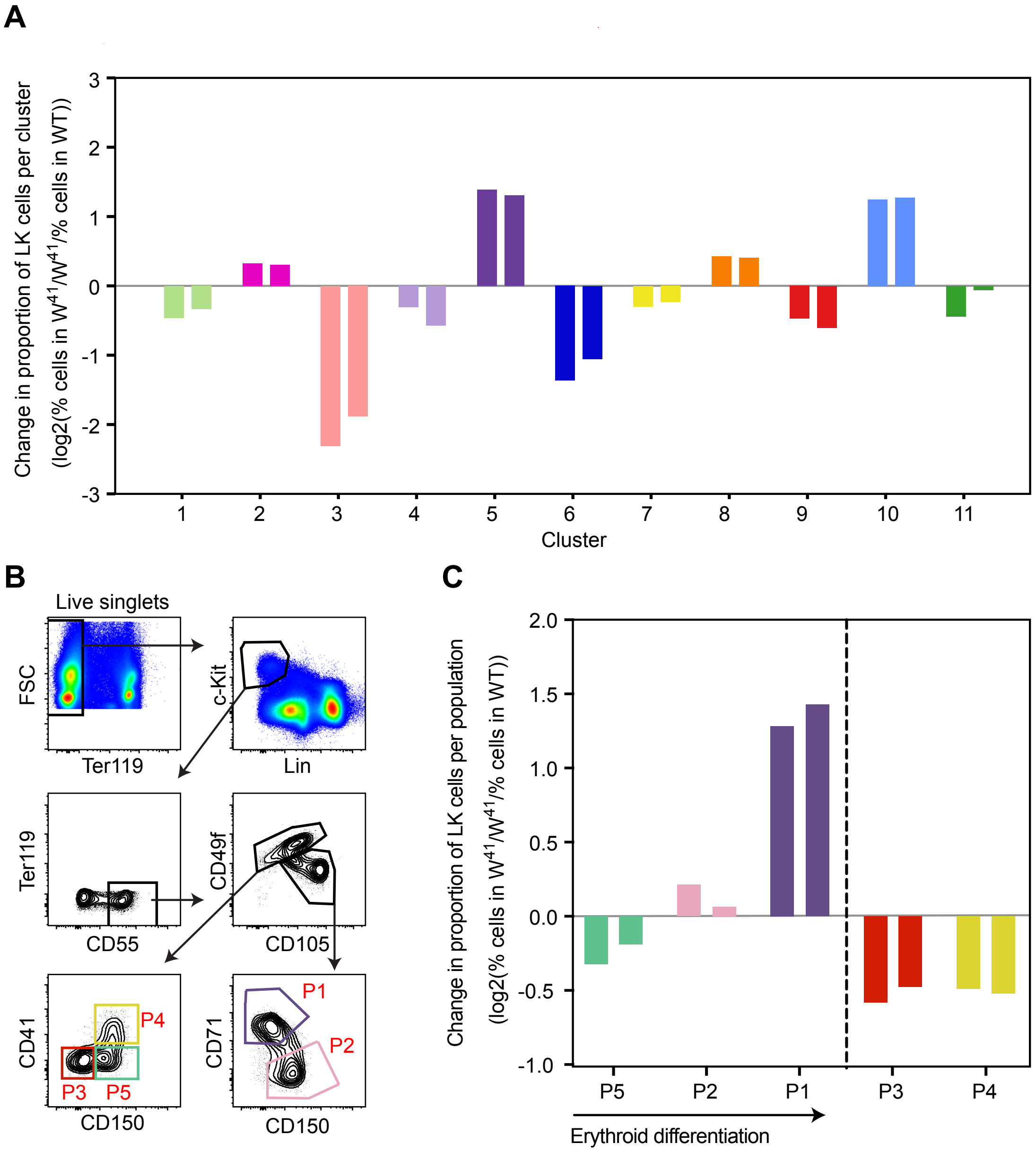


Fig. 7.3. Reproducible changes in the proportion of different progenitor types can be seen between the W^{41}/W^{41} and WT landscapes. (A) Log_2 fold-change of the percentage of cells in a cluster in the W^{41}/W^{41} data divided by the percentage of cells in the corresponding cluster in the WT data. Paired bars indicate fold-changes for samples from 2 separate mice. Fold changes are only displayed for WT clusters with at least 100 cells. (B) Gating strategy for the isolation of populations P1-P5 from bone marrow defined in Tusi et al. (2018). (C) Log_2 fold-change of the proportion of a cell population in the W^{41}/W^{41} mice compared with WT mice as a percentage of the LK population (as measured by FACS using the gating in panel B). Paired bars indicate fold-changes for samples from two separate W^{41}/W^{41} mice. The mean of samples from two separate WT mice was used for the comparison.

7.4 Identifying molecular effects of impared c-Kit signalling

Single-cell analysis of the W^{41}/W^{41} bone marrow composition highlighted changes at the tissue level of the haematopoietic system. However, as well as being used to assign cells to groups, single-cell transcriptomics data can simultaneously be used to measure the expression of individual genes. A benefit of performing gene expression analysis in single cells is that cells can be grouped based on the similarity of their expression profiles before differential expression is performed between the corresponding WT and W^{41}/W^{41} clusters. This approach therefore provides the opportunity to search for genes with altered behaviour in specific regions of the landscape. Differential expression analysis between pairs of clusters identified *Myc* as one of the most significantly downregulated genes across all clusters, consistent with reduced c-Kit signalling in the W^{41}/W^{41} animals (Fig. 7.4A, Table 7.1). Computing the overlap between genes upregulated in W^{41}/W^{41} clusters and annotated genes sets from the Molecular Signatures Database Hallmark Gene Set Collection (Liberzon et al., 2015) identified a significant overlap with unfolded protein response genes, which suggests an induction of a stress response programme in the *Kit* mutant cells (HALLMARK_UNFOLDED_PROTEIN_RESPONSE; FDR = 5.74×10^{-5}). Individual stress response genes included *Atf4*, *Psat1*, *Mthfd2* and *Imp3* (Fig. 7.4B).

Table 7.1. Significance of differentially expressed genes in Fig. 7.4. Adjusted p-values (Benjamini-Hochberg correction) are shown for the pairwise comparisons between cells in WT and W^{41}/W^{41} clusters. Differential expression was performed using the *EdgeR* package.

	Adjusted p-value for differential expression of gene across WT and W^{41}/W^{41} cluster								
Cluster	*Myc*	*Atf4*	*Psat1*	*Mthfd2*	*Imp3*	*Casp3*	*Casp6*	*Pdcd4*	*Bid*
1	2.7×10^{-149}	4.8×10^{-30}	2.5×10^{-03}	9.4×10^{-04}	3.4×10^{-92}	1.7×10^{-21}	0.31	3.1×10^{-58}	3.1×10^{-03}
2	0.0	2.4×10^{-288}	1.7×10^{-46}	1.1×10^{-24}	0.00	0.00	1.0	9.7×10^{-300}	3.1×10^{-06}
3	1.4×10^{-57}	1.4×10^{-04}	0.81	0.96	8.0×10^{-27}	1.2×10^{-24}	1.0	6.8×10^{-35}	0.044
4	3.5×10^{-156}	0.0	2.0×10^{-05}	0.39	1.8×10^{-114}	5.0×10^{-14}	1.0	2.5×10^{-109}	0.033
5	0.0	5.9×10^{-10}	2.6×10^{-118}	6.7×10^{-80}	0.0	0.0	1.1×10^{-15}	0.0	2.7×10^{-27}
6	7.7×10^{-30}	3.0×10^{-36}	1.0	0.97	5.1×10^{-28}	1.9×10^{-12}	1.0	2.8×10^{-35}	0.59
7	8.5×10^{-60}	3.3×10^{-31}	7.8×10^{-04}	7.6×10^{-03}	4.8×10^{-62}	9.0×10^{-10}	1.0	4.0×10^{-98}	7.7×10^{-04}
8	4.1×10^{-85}	1.1×10^{-04}	0.033	1.0	2.9×10^{-70}	2.0×10^{-04}	0.025	1.6×10^{-67}	0.031
9	1.5×10^{-17}	4.2×10^{-22}	0.20	0.81	9.8×10^{-23}	7.7×10^{-09}	1.0	1.5×10^{-25}	0.36
10	4.4×10^{-38}	2.5×10^{-04}	2.4×10^{-06}	1.0	5.1×10^{-60}	1.2×10^{-05}	0.27	8.6×10^{-80}	1.0
11	8.5×10^{-22}	2.5×10^{-04}	0.031	1.00	2.6×10^{-14}	0.042	1.0	1.62×10^{-03}	0.25

As it was the erythroid clusters that displayed the strongest changes in relative numbers, focus was then shifted to search for differentially expressed genes between these clusters. The distribution of cytokine receptors such as *Epor* was unchanged between WT and W^{41}/W^{41} clusters (Fig. 7.4C), suggesting that lineage-specific receptors can compensate for reduced c-Kit signalling in these lineage-restricted cells to maintain the relatively normal haematopoiesis seen in these mice (Wu et al., 1997, 1995). Genes downregulated in the later erythroid clusters (2 and 5) in the W^{41}/W^{41} mice possessed significant overlap with apoptosis related genes (HALLMARK_APOPTOSIS; FDR = 1.19×10^{-2}). In particular, *Casp3* was amongst these genes and specifically downregulated in the later erythroid clusters (Fig. 7.4D), along with other examples of apoptosis-related genes including *Casp6*, *Pdcd4* and *Bid*.

In an attempt to look further into the mechanisms of how cluster proportions were altered in the W^{41}/W^{41} bone marrow, the scRNA-seq data were used to investigate cell cycle activity across the single-cell landscape. If differences in cell cycle state were seen between the WT and W^{41}/W^{41} populations this could provide an explanation of how HSPC numbers changed between the two genotypes. Potential changes in cell cycle state were examined by considering the expression of G_2/M marker genes in individual cells. Scoring for the combined expression of these genes and visualising in the force-directed layout demonstrated a variation in cell cycle states between different regions of the single-cell landscape (Fig. 7.5A). In particular, the highest cell cycle activity (as indicated by higher G_2/M gene expression) was seen in the erythroid lineage; in contrast the stem and early progenitor region appeared more quiescent. When the distribution of G_2/M scores was visualised for each cluster it was clear that both the WT and W^{41}/W^{41} clusters displayed very similar patterns, rather than showing differences between the genotypes (Fig. 7.5B). From this analysis it initially seemed that it was not differences in cell cycle regulation causing an increase in the relative number of late erythroid progenitors. However, on closer inspection comparing the G_2/M scores across WT clusters indicated that cluster 3, containing early erythroid progenitors, had much lower G_2/M gene activity than the later erythroid clusters, 2 and 5. To confirm that the erythroid clustering was indeed due to erythroid maturation, and not driven by cell cycle effects, the expression of erythroid marker genes was checked across the erythroid clusters (Fig. 7.5C), confirming a difference in maturity between clusters 3 and 2. Cluster 3 was also the cluster with the most dramatic change in its proportion, with a very large decrease in this population seen in the W^{41}/W^{41} bone marrow (Fig. 7.3, 7.6A). In contrast, the later erythroid clusters 2 and 5, which had higher G_2/M gene activity, were both seen to increase in proportion in the *Kit* mutant mice. Taken together, these data suggest that the W^{41}/W^{41} LK fraction has a reduction in the proportion of more quiescent erythroid progenitors, which could indicate a compensatory mechanism by which the proportion of later erythroid progenitors is increased (Fig. 7.6A).

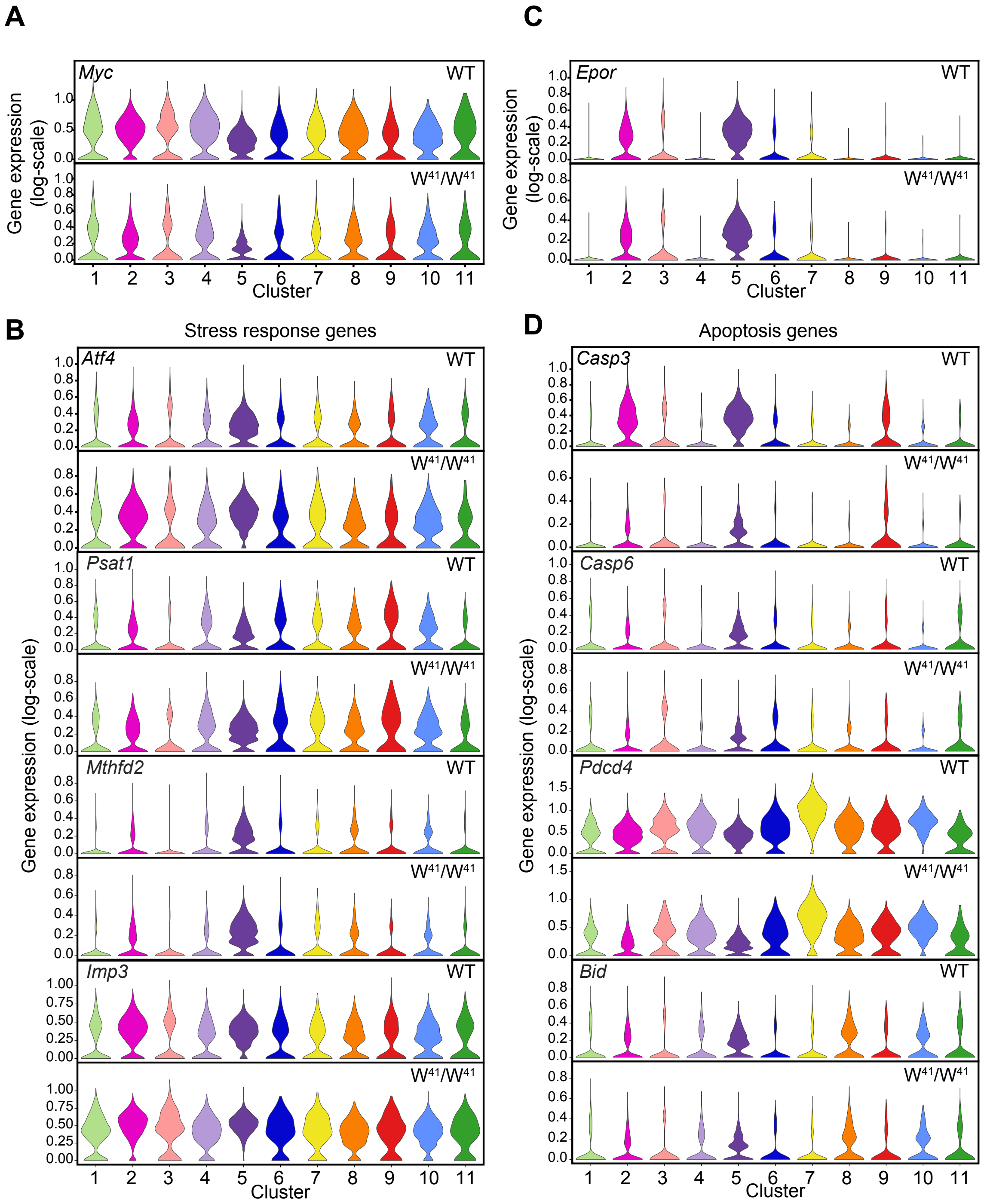


Fig. 7.4. Local and global differences in signalling programs of c-Kit mutant mice are revealed by differential expression of scRNA-seq data. (A–D) Violin plots showing the distribution of selected genes in WT and corresponding W^{41}/W^{41} clusters, as measured by scRNA-seq. Distributions are shown for clusters containing at least 100 WT cells. Data from WT and W^{41}/W^{41} mice were normalised independently.

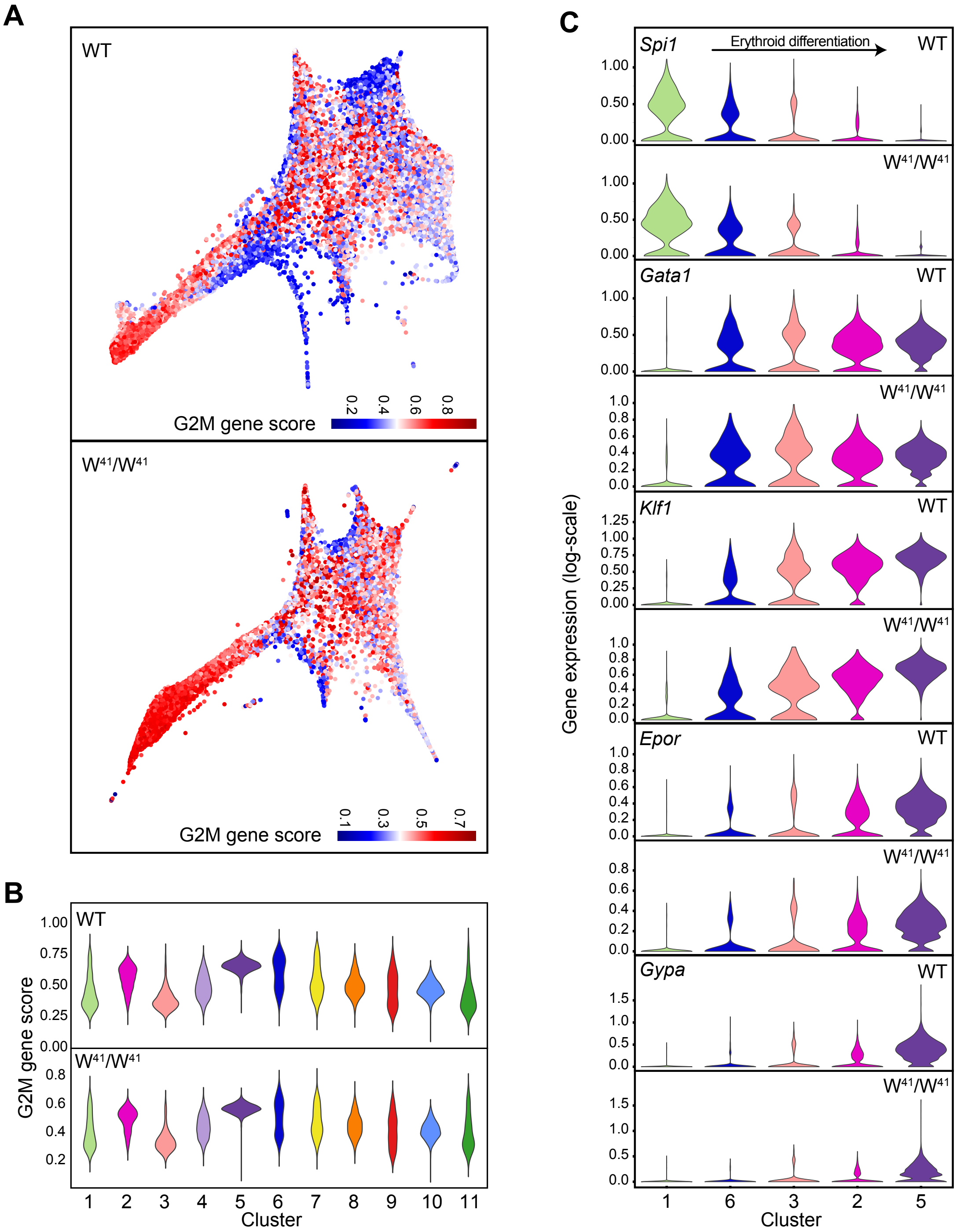


Fig. 7.5. Computational cell cycle scoring reveals uneven distribution of G_2/M gene expression along the erythroid trajectory in *Kit* mutant mice. (A) Expression of G_2/M marker genes in the WT and W^{41}/W^{41} droplet-based scRNA-seq data of LK cells. The colour of cells indicates the combined expression level of these genes, with blue being the lowest value and red the highest value. (B) Violin plots of G2M score from panel A in WT and W^{41}/W^{41} clusters. (C) Violin plots of erythroid genes marking different stages of differentiation in WT and W^{41}/W^{41} clusters 1, 6, 3, 2 and 5.

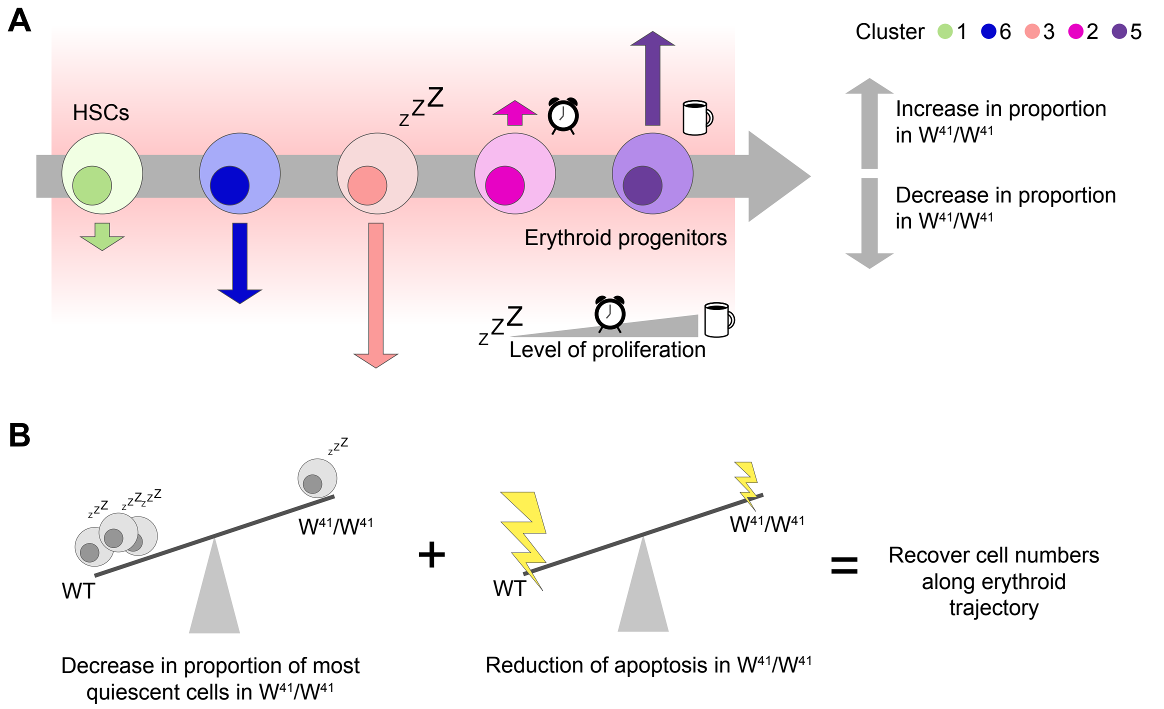

A
Cluster ●1 ●6 ●3 ●2 ●5
HSCs
zzZ
Increase in proportion in W^{41}/W^{41}
Decrease in proportion in W^{41}/W^{41}
Erythroid progenitors
zzZ
Level of proliferation
B
zzZ
zzZzZ
zzZ
W^{41}/W^{41}
WT
+
W^{41}/W^{41}
WT
=
Recover cell numbers along erythroid trajectory
Decrease in proportion of most quiescent cells in W^{41}/W^{41}
Reduction of apoptosis in W^{41}/W^{41}

Fig. 7.6. Overview of the potential compensatory mechanisms along the erythroid trajectory. (A) Summary of the change between the proportion of different progenitor populations in WT and W^{41}/W^{41} bone marrw, along with different levels of cell cycle activity of these populations. Populations are ordered along the erythroid trajectory. Arrow length corresponds to the magnitude of fold-changes displayed in Fig. 7.3. Symbols represent the relative cell cycle activity of the later erythroid progenitor populations indicated by the G_2/M scores in Fig. 7.5. (B) Suggested model for how W^{41}/W^{41} mice attempt to recover cell numbers during erythropoiesis, despite defects at the top of the haematopoietic hierarchy due to disrupted c-Kit signalling.

7.5 Conclusions

To demonstrate how single-cell gene expression analysis can be used as a tool to gain insights into perturbed haematopoiesis, this chapter compared scRNA-seq profiles of over 13,000 HSPCs from a *Kit* mutant mouse model to the dataset representing normal haematopoiesis discussed in Chapter 6. By matching mutant cells with their closest WT counterparts, this analysis revealed changes in the proportions of HSPC populations, and identified potential compensatory processes based on differential gene expression.

7.5.1 Single-cell transcriptomics as a tool for unbiased comparisons

An important outcome of this work is the demonstration of how single-cell transcriptomics can be used to better understand perturbed biological systems. The data-driven approach used here to characterise bone marrow composition is a powerful technique. In particular, it represents an advance over traditional methods that rely on flow cytometry as these can only be applied to pre-defined populations. Here, analysis is not limited to conventional haematopoietic populations, and the whole HSPC compartment can be measured at once. These data also provide a simultaneous measurement of both changes in cell type composition and changes at the molecular level. This type of approach will be applicable in many different systems, including in helping to understand global and molecular alterations in disease states.

7.5.2 Overcoming batch effects between data from different sources

One challenge in this work was deciding on the best way to integrate the datasets from different experiments in order to allow comparison between the WT and W^{41}/W^{41} cells. Approaches have been suggested for both performing “batch correction” on data from different sources (Butler et al., 2018; Haghverdi et al., 2018) and for assigning new cells to an existing annotation in a reference dataset (Kiselev et al., 2018). Here, it was decided to assign the W^{41}/W^{41} cells to clusters in the WT landscape, to reveal how populations changed in comparison to normal haematopoiesis. However, it could be interesting to try different approaches, as these could potentially reveal new insights, such as the existence of novel populations in the *Kit* mutant model. Another difficulty that arose during this work surrounded the batch effects due to differences in gene detection between the two datasets. In particular, a higher number of genes were detected for the W^{41}/W^{41} samples compared to the WT samples, even though WT cells were sequenced with higher depth. This led to challenges in normalisation of the data, and made it more complex to interpret the results of differential expression analysis performed with standard tests.

7.5.3 Altered bone marrow composition of the W^{41}/W^{41} mice

The approach of mapping W^{41}/W^{41} cells to WT clusters revealed changes in the cell type composition of the LK bone marrow fraction, which was subsequently validated by FACS analysis. A reduction in the percentage of more immature progenitor cells is consistent with the idea that c-Kit signalling plays a more important role in the upper tiers of the haematopoietic hierarchy, whereas during differentiation lineage-specific cytokines reduce the impact of a deficiency in c-Kit signalling (Wu et al., 1997, 1995). Additionally, the shift towards a greater proportion of more mature progenitors could be seen as a system-wide response to adapt to the defects in HSCs and immature progenitors caused by reduced c-Kit signalling in order to try and recover the numbers of progenitor cells.

It is important to note that a change in progenitor populations as a proportion of the LK compartment does not necessarily correspond to a change in the actual numbers of these cells. Unfortunately, due to differences in factors such as lineage depletion between the experiments, the data collected for this study did not measure the absolute numbers of c-Kit$^+$ cells. However, it has been reported that W^{41}/W^{41} mice have a decrease in the number of LK cells, so even populations exhibiting a *proportional* increase could still exist in smaller numbers in the W^{41}/W^{41} bone marrow. In particular, this explains how a shift towards more mature erythroid progenitors could still be consistent with the phenotype of the W^{41}/W^{41} mice, as they exhibit mild macrocytic anaemia (Geissler and Russell, 1983a,b). Also, as the changes measured are proportional it should be noted that some populations will have differences in proportion due to the effects of changes in the size of other populations.

7.5.4 Potential compensatory mechanisms identified in response to the *Kit* mutation

Differential gene expression analysis was applied to the single-cell data to examine changes at the molecular level in the *Kit* mutant cells. The system-wide changes in transcription factor *Atf4* suggest activation of an integrated stress response in the W^{41}/W^{41} cells (B'chir et al., 2013; Pakos-Zebrucka et al., 2016), and could be a possible coping mechanism of how the system responds to c-Kit signalling defects. Specifically in the erythroid trajectory, a reduction of apoptosis-related genes including caspase-3 was seen. During normal erythropoiesis there are high levels of apoptosis, and so a reduction of *Casp3* could be consistent with the perturbed cells promoting a reduction of apoptosis in order to overcome defects in earlier progenitors at the top of the haematopoietic hierarchy and minimise the effect on the erythroid system (Fig. 7.6B). The single-cell analysis also revealed that an erythroid progenitor population with low G_2/M cell-cycle gene activity was depleted in the W^{41}/W^{41} mice. This raises the possibility of the W^{41}/W^{41} cells being "pushed" out of this less proliferative population in favour of the more rapidly cycling progenitors in order to produce more erythroid cells and “catch up” with normal erythropoiesis (Fig. 7.6B).

7.5.5 Future work

One aspect of future work will be to see if the graph abstraction method from Chapter 6 can be used to give insights into how gene expression differences between different stages of differentiation are altered in a perturbed state. In particular, focusing the analysis around specific regions of the landscape where population frequencies shift could help to reveal whether changes in the activity of particular genes or signalling pathways is linked to these changes. Additionally, it will be exciting to extend this approach to different mouse models, particularly to mouse models with a relevance to disease.

7.5.6 Summary

In summary, this chapter discusses an unbiased investigation of tissue- and molecular-level changes in a *Kit* mutant mouse model using scRNA-seq data, to understand how the transcriptional landscape changes in a perturbed signalling environment.

Chapter 8

Discussion

Detailed discussions of the results from this thesis are included at the ends of Chapters 3-7. This chapter will focus on a more general discussion of the work, describe the challenges facing single-cell biology and discuss future directions for the field.

8.1 Single-cell characterisation of the haematopoietic transcriptional landscape

8.1.1 Single-cell landscapes recapitulate haematopoietic differentiation

Three different single-cell profiling techniques were used in this work to measure gene expression across the haematopoietic stem and progenitor compartment: single-cell qRT-PCR, plate-based scRNA-seq and droplet-based scRNA-seq. Analysis using dimensionality reduction techniques and pseudotime methods demonstrated that all of these datasets could be used to generate representations of single-cell transcriptional landscapes that recapitulated haematopoietic differentiation, but at a selection of different resolutions. For example, the arrangement of cell types within dimensionality reductions, identified based on either surface-marker profiles or marker gene expression, showed good agreement with current models of the haematopoietic hierarchy. Gene expression dynamics along the pseudotime orderings constructed using both qRT-PCR and scRNA-seq measurements were consistent with known biology about lineage-specific transcription factor behaviour. Together, this work highlights how single-cell gene expression profiling can provide insight into how the transcriptome of a cell changes during differentiation, and how this can be used to characterise properties such as cell cycle state.

8.1.2 Discovering transcriptional regulation

As well as providing a description of how gene expression changes during differentiation, this work also investigated approaches for identifying how haematopoietic fate decisions are transcriptionally regulated. The analysis in Chapter 4 focused on constructing a transcriptional regulatory network model for haematopoietic differentiation based on single-cell expression data. This was made possible due to the high number of transcription factor genes measured in the qRT-PCR dataset. However, although this method was able to identify differences in regulatory relationships between differentiation towards two cells fates, it nevertheless remains limited by the choice of genes. Profiling using scRNA-seq avoids this issue, but is not suitable for use with the Boolean network inference method due to the high dropout rate in these data. Searching for transcriptional regulation of cell fate decisions using the scRNA-seq data has more scope for discovering novel regulation, and so a different approach was taken to identify dynamic genes around a region of the landscape where the trajectories branched towards multiple lineages. The genes found by this analysis will be investigated and validated in future work.

8.1.3 Insights from examining heterogeneity within conventional haematopoietic populations

An obvious strength of single-cell profiling, and a driving force behind its popularity, is that it can be used to characterise the heterogeneity within a population. The work in Chapter 5 focused on variation within human lympho-myeloid progenitor populations to devise a new sorting strategy to enrich for functional output. This was done by examining differences in surface marker levels between cells giving rise to different mature cell types in single-cell cultures. The ability of single-cell analysis to characterise heterogeneous populations was also exploited in Chapter 7, in order to understand how the transcriptional landscape changes in response to disrupted Kit signalling. This strategy will be relevant for studying changes in HSPC populations for many different scenarios, for example in the context of infection and in genetic perturbations related to disease.

8.1.4 Choosing an appropriate gene expression assay

Increases in both the number of genes and the number of cells measured in a dataset resulted in improved resolution in terms of distinguishing cells belonging to different lineages, but at the expense of accuracy in gene expression measurements. In Chapter 6, droplet-based scRNA-seq data were used to identify entry points to eight different blood lineages, where these entry points could be used to reconstruct differentiation trajectories through the data. This offered much better discrimination between cell fates than the single-cell qRT-PCR data in Chapter 4, where considerable overlap was seen between progenitor populations such as GMPs and LMPPs, which do separate based on their transcriptional profiles in the scRNA-seq data. However, measurements of gene expression from the qRT-PCR data are more accurate and suffer less from the issue of dropouts, particularly in comparison to the much sparser droplet-based scRNA-seq data. These observations highlight the importance of choosing the most suitable assay for an experiment. For interrogating the regulatory relationships between a specific set of genes qRT-PCR offers clear advantages over RNA-seq data. In contrast, for defining trajectories through the data and identifying novel regulators a more suitable choice is scRNA-seq. For scRNA-seq there is also the decision about whether to profile cells using a higher throughput technology, such as the droplet-based sequencing of Chapters 6 and 7, or to obtain more in depth measurements of lower numbers of cells with a method such as SMART-Seq2. The generation of a reference transcriptome in Chapter 3 was carried out using the more expensive (on a per cell basis) SMART-Seq2 technology, as the aim was to make a high-quality dataset that could be related to index-sorting profiles in order to allow conventional haematopoietic populations to be identified within the dataset. However, the high-throughput droplet-based data were more suitable for understanding how the transcriptional landscape changed in response to a perturbation, as this approach could capture many cells in an unbiased manner to reveal how populations were altered in the perturbed bone marrow.

8.2 Challenges in the analysis of single-cell expression data

8.2.1 Sparsity and noise in scRNA-seq data

Although scRNA-seq provides a transcriptome-wide view into gene expression, it is important to recall that only a sample of the mRNA molecules present within a cell are captured for sequencing. Generally, due to the prohibitive cost, samples are not sequenced to saturation for scRNA-seq. In high-throughput methods, such as the droplet-based sequencing from Chapters 6 and 7, the sampling depth is normally even shallower as profiling more cells in a sample means fewer sequencing reads are generated for each cell. As discussed in Section 1.2.3, this results in a high frequency of dropouts, with zero-inflation particularly a problem for lowly-expressed genes. Noise in gene expression measurements limits the potential of using scRNA-seq gene expression values as input for methods such as transcriptional regulatory network inference as these rely on accurate reporting of the levels of each gene. These issues have led to the recent development of imputation methods, where algorithms are used to infer "missing" values in gene expression data. A range of approaches have been suggested for this task, including using neural networks (Eraslan et al., 2018), diffusion-based smoothing (van Dijk et al., 2018), and regression techniques (Li and Li, 2018). These methods use information from cells across the dataset to try and recover gene expression values representing transcriptional profiles that are unaffected by dropout. This may become an essential step of scRNA-seq data analysis in the future. However, one major challenge with such approaches is knowing how to avoid overfitting, where the gene expression values are smoothed too far, obscuring biological heterogeneity within the sample.

8.2.2 Limitations in trajectory inference

Over the past four years a wide range of algorithms have been suggested for inferring differentiation trajectories based on single-cell expression profiles, with several of these reviewed in Section 1.4. Pseudotime inference techniques provide a powerful approach for examining gene expression changes in previously inaccessible systems, particularly for *in vivo* cells such as the bone marrow cells used to construct haematopoietic differentiation trajectories for this work. However, trajectory inference remains limited by factors such as parameter choices (Weinreb et al., 2018b). It is also important to remember that these methods only allow the inference of *pseudo*time, not real time. Densities of cells in gene expression space, and consequently in pseudotime trajectories, do not necessarily correspond to closeness of these cells in the time it actually takes them to differentiate. This is important to consider when examining the dynamics of transcriptional changes during differentiation. Some studies currently investigate "real time" dynamics in a low-throughput manner using *in vitro* imaging combined with fluorescent reporters (Etzrodt and Schroeder, 2017), and it will be interesting to see how this can be related to pseudotime approaches in the future. Flux through the haematopoietic compartments has also been estimated using cell labelling and mathematical modelling techniques (Busch et al., 2015; Sawai et al., 2016), and this could be useful information to parametrise pseudotime analysis with real time.

8.2.3 Matching transcriptional state with function

Profiling the transcriptional state of a cell using techniques such as qRT-PCR and RNA-seq necessitates its destruction. It is therefore not possible to obtain measurements of both the transcriptome and future functional output of the same cell with these methods. Several studies have taken advantage of index-sorting to link molecular profiles of cells to function (Tusi et al., 2018; Velten et al., 2017; Wilson et al., 2015). Here, index-sorting is applied to both the cells selected for expression profiling and those used in functional assays, so that those with overlapping surface marker levels can be related. Chapter 5 investigated this approach for human cord blood progenitors, and identified a relationship between the levels of several surface markers (CD38, CD10 and CD45RA) and the functional output in single-cell cultures. However, whilst these surface markers could enrich for functional output, they could not be used to completely purify for function. This could be a limitation of the surface markers measured, and it is possible that a different set of markers could be used to accurately classify the different cell types. This highlights the difficultly of investigating the function of populations defined based on their transcriptome.

8.2.4 Coping with high cell numbers

Another challenge facing the field is ensuring that computational techniques can be applied to the increasingly large datasets that are being generated. This is particularly pertinent in the light of initiatives such as the Human Cell Atlas (Regev et al., 2017), which are helping to drive the generation of vast amounts of data. Several methods that were developed when the majority of datasets were at most a few hundred cells have high computational complexity, and so cannot be applied to thousands of cells. With the existence of datasets containing upwards of one million cells (10x Genomics, 2017), scalability is an essential consideration when designing new algorithms.

8.3 Future directions for single-cell biology

8.3.1 Measuring multiple properties of the same cell

As discussed in Section 8.2.3, the expression of cell surface marker proteins is an important way of being able to interrogate the functional properties of transcriptionally distinct cells that are identified using gene expression analyses. For plate-based methods such as SMART-Seq2 (Picelli et al., 2013) and MARS-Seq (Jaitin et al., 2014), it is possible to perform index-sorting to retain information about the surface markers measured by FACS for each cell. However, with droplet-based scRNA-seq protocols this is not possible. Instead, a range of alternative approaches have been suggested for obtaining simultaneous measurements of the transcriptome along with expression of a number of proteins of interest for individual cells. CITE-seq (Stoeckius et al., 2017) and REAP-seq (Peterson et al., 2017) are methods that use DNA-barcoded antibodies in conjunction with droplet-based scRNA-seq to measure both gene and selected protein expression in a sequencing approach by adding barcodes to antibodies that get transcribed. As strategies using cell-surface markers are extensively used for cell type identification across the haematopoietic system these should prove to be powerful techniques.

This thesis has focused on characterising transcriptional heterogeneity in HSPC populations to learn more about haematopoietic differentiation. However, transcriptional regulation is only one aspect influencing cell fate decisions. Techniques measuring a number of cellular properties at the single-cell level are now available, for example genomic DNA from individual cells can be profiled to assess variation in DNA copy number and the mutations present within a population (Gawad et al., 2016). Characterising the epigenome of a cell is also possible, with methods existing for interrogating DNA methylation (Guo et al., 2013b; Smallwood et al., 2014), histone modifications (Rotem et al., 2015), chromatin accessibility (Buenrostro et al., 2015; Cusanovich et al., 2015) and chromatin arrangement (Nagano et al., 2013) at the single-cell level. Buenrostro et al. (2018) used a single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) to profile chromatin accessibility in human HSPCs. These data were able to recapitulate the single-cell haematopoietic landscape by measuring regions of open chromatin. Assessing epigenetic changes across the haematopoietic landscape will help to reveal how cell fate decisions are controlled, as the arrangement and accessibility of chromatin is a mechanism used for regulating gene expression. In this study, the authors also performed single-cell transcriptional profiling on the same populations, and used computational techniques to try and match up the scATAC-seq and scRNA-seq profiles to understand how the transcriptome and epigenome were related across the haematopoietic landscape.

Excitingly, there are now also numerous protocols for performing single-cell multiomics techniques, where different properties are measured within the same cell (Chappell et al., 2018; Macaulay et al., 2017). Genetic mutations occurring in stem and progenitor cells are a driving factor in many blood disorders including myeloproliferative neoplasms and AML, and so new techniques allowing simultaneous genome and transcriptome profiling could prove useful for studying haematopoiesis in the context of disease (Dey et al., 2015; Han et al., 2014; Macaulay et al., 2015). Several methods have also been described that enable the measurement of the methylation and gene expression states of the same cell (Angermueller et al., 2016; Cheow et al., 2016; Hou et al., 2016; Hu et al., 2016), or of both methylation and gene expression along with chromatin accessibility (Clark et al., 2018). Having access to both epigenetic and transcriptional states of a cell could help understanding of how the transcriptional changes in cell state are regulated during differentiation. In particular, epigenetics may provide information about the potential of a cell (what it "can do") whereas transcriptomics provides a window into its current state (what the cell "is doing"). Due to limitations in the amount of starting material available from individual cells, and the challenges associated with separating cellular components for performing these parallel analyses, a number of these methods have restricted sensitivity, or in practice prove very costly for profiling large numbers of cells. It is likely that it the next few years innovations in technology will help to address these issues, and that single-cell multiomics datasets will become more widespread. As a result, it will be important to focus on developing new computational approaches for integrating and analysing these data.

8.3.2 Combining perturbations with single-cell gene expression measurements

Single-cell technologies also have powerful potential in understanding how perturbations can affect the transcriptional landscape. The role of a particular gene in a system can be tested by knocking it out to observe its effect. Traditionally this would have been a laborious and time-consuming process. With the advent of CRISPR (clustered regularly interspaced short palindromic repeats)-associated nuclease Cas9 (Cong et al., 2013; Qi et al., 2013) it became possible to perform genome-wide knockout screens (Shalem et al., 2014; Wang et al., 2014), where individual cells from a population are randomly perturbed by guide RNAs from a pooled library that targets thousands of genes. After application of the CRISPR library, cells can be cultured for a period of time. When genes essential for survival have been knocked out in some of the cells, these cells will die. Sequencing the surviving cells identifies the nature of their perturbation to reveal the essential genes. Yet, initially, this approach did not provide a read-out of how the molecular state of a cell was altered after a genetic perturbation. This led to the development of several methods combining CRISPR perturbation with high-throughput scRNA-seq. Both the Perturb-Seq (Adamson et al., 2016; Dixit et al., 2016) and the CRISP-seq (Jaitin et al., 2016) methods use a library of barcoded guide RNAs targeting different genes, where the barcodes for the guide RNAs are transcribed and read during sequencing in order to identify which guide RNA was present in a cell. In an alternative method, CROP-seq, the authors altered a CRISPR-screening construct so that the guide RNA itself is transcribed during single-cell transcriptional profiling using Drop-seq (Datlinger et al., 2017). This removed the need for creating a barcoded guide RNA library as the guide is read directly. These approaches should be able to provide exciting insights into how the gene expression states of cells across many systems, including haematopoiesis, are altered in response to loss of different genes. However, there are still challenges to overcome in the analysis of data generated from these combined CRISPR screening and transcriptional profiling methods. In particular, the presence of a guide in a cell does not always mean that the gene has been successfully targeted. Reliably determining whether this has happened will require computational or experimental innovation.

8.3.3 Revealing the past and the future of individual cells

Profiling a sample using single-cell omics technologies, such as scRNA-seq, aims to describe the state of each cell at the moment it was captured. Time-course experiments or computational pseudotime approaches attempt to add dynamics to these type of data, but cannot actually identify the past or future states of an individual cell. Recently there has been promising development of new methods offering glimpses into these states, which is particularly exciting considering the debate surrounding the structure of the haematopoietic hierarchy.

In a new approach for investigating the organisation of haematopoietic differentiation, Rodriguez-Fraticelli et al. (2018) used transposon tagging to label cells in steady state haematopoiesis. Cells in adult mice were labelled during a doxycycline pulse, and after a number of weeks different blood lineage and progenitor populations were isolated. Sequencing was then performed to reveal which barcodes were present across the different populations, and in what combinations of cell types they appeared. The authors saw that very few barcodes within the megakaryocyte populations were shared across other cell fates, in contrast to the presence of the MEP population within the classical haematopoietic hierarchy.

The concept of genetically labelling single cells to reconstruct lineage hierarchies has also been applied in zebrafish, but here has excitingly been combined with single-cell transcriptomic approaches. LINNAEUS (Spanjaard et al., 2018), ScarTrace (Alemany et al., 2018) and scGESTALT (Raj et al., 2018) all take advantage of the CRISPR-Cas9 system. It was realised that this system could be used for lineage tracing as Cas9 can be used to generate random small insertions and deletions in a target gene, known as scars. These scarred regions can then be sequenced along with the transcriptomes from single cells, allowing both unbiased cell type identification (based on the transcriptome) and lineage reconstruction (based on the combinations of scars observed across cells). These methods were used to scar early embryos in order to investigate lineage separation during development, reveal the clonal structure of organs in the adult fish and to study zebrafish brain development. Combining lineage tracing with transcriptional profiling has huge potential for understanding the differentiation structure of a tissue, and has very recently been extended to mammalian organisms in a study investigating lineage hierarchies in the gastrulating mouse embryo (Chan et al., 2018). Nevertheless, considerable challenges arise from the fact that the scars in these methods are not reliably detected in all cells, and so improvements in the computational methods for inferring lineage relationships will be needed.

In contrast to lineage tracing techniques which seek to reveal the history of a cell, very different work has demonstrated the potential of using single-cell transcriptomics for predicting a cell's future. This approach, called RNA velocity, considers the ratio of spliced to unspliced mRNAs within a cell, reasoning that this ratio changes as cells undergo differentiation at different speeds (La Manno et al., 2018). The authors applied this algorithm to identify differentiation trajectories in the developing mouse hippocampus by calculating current and future states for cells in this system. In work using single-cell transcriptomics to reconstruct the lineage hierarchy of a whole organism, the flatworm planaria, RNA velocity estimates showed good agreement with lineage relationships inferred on gene expression counts, and helped to add direction to connections seen between cell states in this hierarchy (Plass et al., 2018).

8.4 Concluding remarks

Over the past few years there has been an explosion in single-cell data generation and analysis techniques. Studies using single-cell approaches have offered unprecedented insight into processes such as differentiation and as such have been widely applied across the stem cell field, amongst many other areas of biology. Work focusing on haematopoiesis has ranged from modelling transcriptional regulation to probing the structure of the haematopoietic differentiation hierarchy, with single-cell biology instrumental in revealing a huge amount about this complex system.

Investigating how the gene expression state of a cell changes during haematopoiesis can advance our understanding of how this process is regulated. This PhD project therefore focused on characterising the transcriptional landscape within the haematopoietic stem and progenitor cell compartment using single-cell gene expression profiling. Datasets generated as part of this work have been made publicly available, representing a valuable resource for the haematopoietic community. Analysis of these data allowed computational reconstruction of differentiation towards several blood lineages, revealing changes in gene expression and in processes such as cell cycle along these differentiation trajectories. Inferred transcription factor dynamics were used to discover differences in regulatory relationships between cells differentiating towards two alternative blood fates. Heterogeneity in the blood progenitor compartment was also studied using single-cell molecular profiling. Data from human blood progenitors revealed a continuum of gene expression states in agreement with the functional heterogeneity observed within these populations. In the mouse, cellular heterogeneity in the bone marrow HSPC compartment was considered when investigating how its composition is altered in response to a genetic perturbation. Together, all of this work highlights the importance of performing analysis at the single-cell level. Future steps will focus on validating the potential regulators of cell fate decisions identified using the scRNA-seq landscape, and then expanding this approach to consider the mechanisms behind how the landscape is altered in perturbed states. The techniques discussed in this thesis should be widely applicable to investigation of other haematopoietic perturbations, which will be particularly interesting in the context of disease.

References

10x Genomics (2017). Transcriptional profiling of 1.3 million brain cells with the chromium single cell 3' solution. https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons.

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., and Weissman, J. S. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C. T., Bryder, D., Yang, L., Borge, O.-J., Thoren, L. A. M., 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(6774):193–197.

Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J., and van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108–112.

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

Amir, E.-a. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*, 31(6):545–552.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106.

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Angerer, P., Haghverdi, L., Büttner, M., Theis, F. J., Marr, C., and Buettner, F. (2016). *destiny*: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*, 13(3):229–232.

Arinobu, Y., Mizuno, S.-i., Chong, Y., Shigematsu, H., Iino, T., Iwasaki, H., Graf, T., Mayfield, R., Chan, S., Kastner, P., and Akashi, K. (2007). Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell*, 1(4):416–427.

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.

B'chir, W., Maurin, A.-C., Carraro, V., Averous, J., Jousse, C., Muranishi, Y., Parry, L., Stepien, G., Fafournoux, P., and Bruhat, A. (2013). The eIF2α/ATF4 pathway is essential for stress-induced autophagy gene expression. *Nucleic Acids Res*, 41(16):7683–7699.

Beerman, I., Bhattacharya, D., Zandi, S., Sigvardsson, M., Weissman, I. L., Bryder, D., and Rossi, D. J. (2010). Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A*, 107(12):5465–5470.

Behbehani, G. K., Bendall, S. C., Clutter, M. R., Fantl, W. J., and Nolan, G. P. (2012). Single-cell mass cytometry adapted to measurements of the cell cycle. *Cytometry*, 81A(7):552–566.

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.

Bendall, S. C., Nolan, G. P., Roederer, M., and Chattopadhyay, P. K. (2012). A deep profiler's guide to cytometry. *Trends Immunol*, 33(7):323–332.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., and Nolan, G. P. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696.

Bhargava, V., Ko, P., Willems, E., Mercola, M., and Subramaniam, S. (2013). Quantitative transcriptomics using designed primer-based amplification. *Sci Rep*, 3:1740.

Bockamp, E., McLaughlin, F., Murrell, A., Göttgens, B., Robb, L., Begley, C., and Green, A. (1995). Lineage-restricted regulation of the murine SCL/TAL-1 promoter. *Blood*, 86(4):1502–1514.

Bonnet, D. and Dick, J. E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med*, 3(7):730–737.

Bonzanni, N., Garg, A., Feenstra, A., Schütte, J., Kinston, S., Miranda-Saavedra, D., Heringa, J., Xenarios, I., and Göttgens, B. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics*, 29(13):i80–i88.

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. *Nat Methods*, 10(11):1093–1095.

Bryder, D., Rossi, D. J., and Weissman, I. L. (2006). Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *Am J Pathol*, 169(2):338–346.

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.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

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. *Nat Biotechnol*, 36(5):411–420.

Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 5:418–429.

Büttner, M., Miao, Z., Wolf, A., Teichmann, S. A., and Theis, F. J. (2017). Assessment of batch-correction methods for scRNA-seq data with a new test metric. *bioRxiv*, doi: 10.1101/200345.

Challen, G. A., Boles, N. C., Chambers, S. M., and Goodell, M. A. (2010). Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-β1. *Cell stem cell*, 6(3):265–278.

Chan, M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Meissner, A., and Weissman, J. S. (2018). Molecular recording of mammalian embryogenesis. *bioRxiv*, doi: 10.1101/384925.

Chappell, L., Russell, A. J. C., and Voet, T. (2018). Single-cell (multi)omics technologies. *Annu Rev Genomics Hum Genet*, 19(1), doi: 10.1146/annurev-genom-091416-035324.

Chattopadhyay, P. K. and Roederer, M. (2015). A mine is a terrible thing to waste: high content, single cell technologies for comprehensive immune analysis. *AM J Transplant*, 15(5):1155–1161.

Chelliah, V., Juty, N., Ajmera, I., Ali, R., Dumousseau, M., Glont, M., Hucka, M., Jalowicki, G., Keating, S., Knight-Schrijver, V., Lloret-Villas, A., Natarajan, K. N., Pettit, J.-B., Rodriguez, N., Schubert, M., Wimalaratne, S. M., Zhao, Y., Hermjakob, H., Le Novère, N., and Laibe, C. (2015). Biomodels: ten-year anniversary. *Nucleic Acids Res*, 43(D1):D542–D548.

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. *BMC bioinformatics*, 14(1):128.

Chen, J. Y., Miyanishi, M., Wang, S. K., Yamazaki, S., Sinha, R., Kao, K. S., Seita, J., Sahoo, D., Nakauchi, H., and Weissman, I. L. (2016). Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature*, 530(7589):223–227.

Cheow, L. F., Courtois, E. T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R. Z., Tan, D. S. W., Robson, P., Loh, Y.-H., Quake, S. R., and Burkholder, W. F. (2016). Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat Methods*, 13(10):833–836.

Chickarmane, V., Enver, T., and Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol*, 5(1):e1000268.

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, 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. *Nat Commun*, 9(1):781.

Coifman, R. R., Lafon, S., Lee, a. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci U S A*, 102(21):7426–7431.

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., and Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc Natl Acad Sci U S A*, 114(23):5792–5799.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914.

Dahlin, J. S., Hamey, F. K., Pijuan-Sala, B., Shepherd, M., Lau, W. W. Y., 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. *Nat Methods*, 14(3):297–301.

de Graaf, C. A., Choi, J., Baldwin, T. M., Bolden, J. E., Fairfax, K. A., Robinson, A. J., Biben, C., Morgan, C., Ramsay, K., Ng, A. P., Kauppi, M., Kruse, E. A., Sargeant, T. J., Seidenman, N., D'Amico, A., D'Ombrain, M. C., Lucas, E. C., Koernig, S., Baz Morelli, A., Wilson, M. J., Dower, S. K., Williams, B., Heazlewood, S. Y., Hu, Y., Nilsson, S. K., Wu, L., Smyth, G. K., Alexander, W. S., and Hilton, D. J. (2016). Haemopedia: an expression atlas of murine hematopoietic cells. *Stem Cell Reports*, 7(3):571–582.

de Moura, L. and Bjørner, N. (2008). Z3: An Efficient SMT Solver. In Ramakrishnan, C. R. and Rehof, J., editors, *Tools and algorithms for the construction and analysis of systems*, Lecture notes in computer science, pages 337–340. Springer, Berlin, Heidelberg.

Dexter, T. M., Allen, T. D., Scott, D., and Teich, N. M. (1979). Isolation and characterisation of a bipotential haematopoietic cell line. *Nature*, 277(5696):471–474.

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol*, 33(3):285–289.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., 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 screens. *Cell*, 167(7):1853–1866.

Donaldson, I. J., Chapman, M., Kinston, S., Landry, J. R., Knezevic, K., Piltz, S., Buckley, N., Green, A. R., and Göttgens, B. (2005). Genome-wide identification of *cis*-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet*, 14(5):595–601.

Doulatov, S., Notta, F., Eppert, K., Nguyen, L. T., Ohashi, P. S., and Dick, J. E. (2010). Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat Immunol*, 11(7):585–593.

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. *Nat Immunol*, 17(6):666–676.

Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S., and Smith, A. G. (2014). Defining an essential transcription factor program for naïve pluripotency. *Science*, 344(6188):1156–1160.

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. *Nat Immunol*, 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 Harb Perspect Med*, 3(4):a011601.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2018). Single cell RNA-seq denoising using a deep count autoencoder. *bioRxiv*, doi: 10.1101/300681.

Etzrodt, M. and Schroeder, T. (2017). Illuminating stem cell transcription factor dynamics: long-term single-cell imaging of fluorescent protein fusions. *Curr Opin Cell Biol*, 49:77–83.

Fan, H. C., Fu, G. K., and Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222):1258367.

Fisher, J. and Henzinger, T. A. (2007). Executable cell biology. *Nat Biotechnol*, 25(11):1239–1249.

Fisher, J. and Piterman, N. (2010). The executable pathway to biological networks. *Brief Funct Genomics*, 9(1):79–92.

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. *Proc Natl Acad Sci U S A*, 108(22):9026–9031.

Fujiwara, T., Lee, H.-Y., Sanalkumar, R., and Bresnick, E. H. (2010). Building multifunctionality into a complex containing master regulators of hematopoiesis. *Proc Natl Acad Sci U S A*, 107(47):20429–20434.

Fujiwara, T., O'Geen, H., Keles, S., Blahnik, K., Linnemann, A. K., Kang, Y.-A., Choi, K., Farnham, P. J., and Bresnick, E. H. (2009). Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell*, 36(4):667–681.

Garg, A., Cara, A. D., Xenarios, I., Mendoza, L., and Micheli, G. D. (2008). Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics*, 24(17):1917–1925.

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nat Rev Genet*, 17(3):175–188.

Gazit, R., Mandal, P. K., Ebina, W., Ben-Zvi, A., Nombela-Arrieta, C., Silberstein, L. E., and Rossi, D. J. (2014). Fgd5 identifies hematopoietic stem cells in the murine bone marrow. *J Exp Med*, 211(7):1315–1331.

Geissler, E. N. and Russell, E. S. (1983a). Analysis of the hematopoietic effects of new dominant spotting (W) mutations of the mouse. I. Influence upon hematopoietic stem cells. *Exp Hematol*, 11(6):452–460.

Geissler, E. N. and Russell, E. S. (1983b). Analysis of the hematopoietic effects of new dominant spotting (W) mutations of the mouse. II. Effects on mast cell development. *Exp Hematol*, 11(6):461–466.

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. *Nat Cell Biol*, 20(7):836–846.

Goardon, N., Lambert, J. A., Rodriguez, P., Nissaire, P., Herblot, S., Thibault, P., Dumenil, D., Strouboulis, J., Romeo, P.-H., and Hoang, T. (2006). ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J*, 25(2):357–366.

Goardon, N., Marchi, E., Atzberger, A., Quek, L., Schuh, A., Soneji, S., Woll, P., Mead, A., Alford, K. A., Rout, R., Chaudhury, S., Gilkes, A., Knapper, S., Beldjord, K., Begum, S., Rose, S., Geddes, N., Griffiths, M., Standen, G., Sternberg, A., Cavenagh, J., Hunter, H., Bowen, D., Killick, S., Robinson, L., Price, A., Macintyre, E., Virgo, P., Burnett, A., Craddock, C., Enver, T., Jacobsen, S. E. W., Porcher, C., and Vyas, P. (2011). Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell*, 19(1):138–152.

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 Rep*, 3(5):1539–1552.

Göttgens, B. (2015). Regulatory network control of blood stem cells. *Blood*, 125(17):2614–2620.

Göttgens, B., Nastos, A., Kinston, S., Piltz, S., Delabesse, E. C. M., Stanley, M., Sanchez, M.-J., Ciau-Uitz, A., Patient, R., and Green, A. R. (2002). Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *EMBO J*, 21(12):3039–3050.

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. *Nat Commun*, 7:11075.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat Methods*, 11(6):637–640.

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. *Dev 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. (2013a). Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell*, 13(4):492–505.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013b). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res*, 23(12):2126–2135.

Guo, Y., Niu, C., Breslin, P., Tang, M., Zhang, S., Wei, W., Kini, A. R., Paner, G. P., Alkan, S., Morris, S. W., Diaz, M., Stiff, P. J., and Zhang, J. (2009). c-Myc-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. *Blood*, 114(10):2097–2106.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). 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. *Nat 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. *Nat Biotechnol*, 36(5):421–427.

Hall, M. A., Curtis, D. J., Metcalf, D., Elefanty, A. G., Sourris, K., Robb, L., Göthert, J. R., Jane, S. M., and Begley, C. G. (2003). The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12. *Proc Natl Acad Sci U S A*, 100(3):992–997.

Hamey, F. K. and Göttgens, B. (2017). Demystifying blood stem cell fates. *Nat Cell Biol*, 19(4):261–263.

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

Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K., and Göttgens, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc Natl Acad Sci U S A*, 114(23):5822–5829.

Hamey, F. K., Nestorowa, S., Wilson, N. K., and Göttgens, B. (2016). Advancing haematopoietic stem and progenitor cell biology through single-cell profiling. *FEBS Lett*, 590(22):4052–4067.

Hamlett, I., Draper, J., Strouboulis, J., Iborra, F., Porcher, C., and Vyas, P. (2008). Characterization of megakaryocyte GATA1-interacting proteins: the corepressor ETO2 and GATA1 interact to regulate terminal megakaryocyte maturation. *Blood*, 112(7):2738–2749.

Han, L., Zi, X., Garmire, L. X., Wu, Y., Weissman, S. M., Pan, X., and Fan, R. (2014). Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. *Sci Rep*, 4:6485.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G.-C., Chen, M., and Guo, G. (2018). Mapping the mouse cell atlas by Microwell-seq. *Cell*, 172(5):1091–1107.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K. J., Rozenblatt-Rosen, O., Dor, Y., Regev, A., and Yanai, I. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol*, 17(1):77.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*, 2(3):666–673.

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

Hock, H., Hamblen, M. J., Rooke, H. M., Traver, D., Bronson, R. T., Cameron, S., and Orkin, S. H. (2003). Intrinsic requirement for zinc finger transcription factor Gfi-1 in neutrophil differentiation. *Immunity*, 18(1):109–120.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., and Peng, J. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res*, 26(3):304–319.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol*, 17(1):88.

Hug, H. and Schuler, R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J Theor Biol*, 221(4):615–624.

Hulett, M. D., Pagler, E., and Hornby, J. R. (2001). Cloning and characterization of a mouse homologue of the human haematopoietic cell-specific four-transmembrane gene HTm4. *Immunol Cell Biol*, 79(4):345–349.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9):e12776.

Ibarra-Soria, X., Jawaid, W., Pijuan-Sala, B., Ladopoulos, V., Scialdone, A., Jörg, D. J., Tyser, R. C. V., 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. *Nat Cell Biol*, 20(2):127–134.

Ingram, 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. *J Exp Med*, 191(1):181–188.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*, 21(7):1160–1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*, 11(1):163–166.

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(6172):776–779.

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 167(7):1883–1896.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Res*, 21(9):1543–1551.

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

Kallianpur, A., Jordan, J., and Brandt, S. (1994). The SCL/TAL-1 gene is expressed in progenitors of both the hematopoietic and vascular systems during embryogenesis. *Blood*, 83(5):1200–1208.

Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J., Usukhbayar, B., Calvo, J., Taylor, S., Goardon, N., Six, E., Pflumio, F., Porcher, C., Majeti, R., Göttgens, B., and Vyas, P. (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat Immunol*, 19(1):85–97.

Karsunky, H., Zeng, H., Schmidt, T., Zevnik, B., Kluge, R., Schmid, K. W., Dührsen, U., and Möröy, T. (2002). Inflammatory reactions and severe neutropenia in mice lacking the transcriptional repressor Gfi1. *Nat Genet*, 30(3):295–300.

Kent, D. G., Copley, M. R., Benz, C., Wöhrer, S., Dykstra, B. J., Ma, E., Cheyne, J., Zhao, Y., Bowie, M. B., Zhao, Y., Gasparetto, M., Delaney, A., Smith, C., Marra, M., and Eaves, C. J. (2009). Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. *Blood*, 113(25):6342–6350.

Kiel, M. J., Yilmaz, O. H., Iwashita, T., Yilmaz, O. H., Terhorst, C., and Morrison, S. J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell*, 121(7):1109–1121.

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods*, 15(5):359–362.

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. *Nat Methods*, 9(1):72–74.

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.

Klein, A. M. and Simons, B. D. (2011). Universal patterns of stem cell fate in cycling adult tissues. *Development*, 138(15):3103–3111.

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. *Mol 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(5):661–672.

Kowalczyk, M. S., Tirosh, I., Heckl, D., Nageswara Rao, T., Dixit, A., Haas, B. J., Schneider, R., 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 Res*, 25(12):1860–1872.

Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C., Litvin, O., Stone, E., Pe'er, D., and Nolan, G. P. (2014). Conditional density-based analysis of T cell signaling in single-cell data. *Science*, 346(6213):1250689.

Krumsiek, J., Marr, C., Schroeder, T., and Theis, F. J. (2011). Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS One*, 6(8):e22649.

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 Res*, 34(5):e42.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, doi: 10.1038/s41586-018-0414-6.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359.

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

Laurenti, E., Varnum-Finney, B., Wilson, A., Ferrero, I., Blanco-Bose, W. E., Ehninger, A., Knoepfler, P. S., Cheng, P.-F., MacDonald, H. R., Eisenman, R. N., Bernstein, I. D., and Trumpp, A. (2008). Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. *Cell Stem Cell*, 3(6):611–624.

Lee, J., Breton, G., Oliveira, T. Y. K., Zhou, Y. J., Aljoufi, A., Puhr, S., Cameron, M. J., Sékaly, R.-P., Nussenzweig, M. C., and Liu, K. (2015). Restricted dendritic cell and monocyte progenitors in human cord blood and bone marrow. *J Exp Med*, 212(3):385–399.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, 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.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun*, 9(1):997.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6):417–425.

Lim, C. Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., and Göttgens, B. (2016). BTR: training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics*, 17(1):355.

Lim, K.-C., Hosoya, T., Brandt, W., Ku, C.-J., Hosoya-Ohmura, S., Camper, S. A., Yamamoto, M., and Engel, J. D. (2012). Conditional Gata2 inactivation results in HSC loss and lymphatic mispatterning. *J Clin Invest*, 122(10):3705–3717.

Lindström, S. (2012). *Flow cytometry and microscopy as means of studying single cells: a short introductional overview.*, volume 853, pages 13–15. Humana Press.

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

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*, 5, doi: 10.12688/f1000research.9501.2.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods*, 12(6):519–522.

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends Genet*, 33(2):155–168.

Macaulay, I. C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S. A., and Cvejic, A. (2016). Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep*, 14(4):966–977.

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.

Mao, Q., Wang, L., Tsang, I. W., and Sun, Y. (2017). Principal graph and structure learning based on reversed graph embedding. *IEEE Trans Pattern Anal Mach Intell*, 39(11):2227–2241.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Supp 1):S7.

McInnes, L. and Healy, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, arXiv:1802.03426.

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., Jacobsen, S. E., de Bruijn, M. F., and Göttgens, B. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol*, 15(4):363–372.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S.-I., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J., and Göttgens, B. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol*, 33(3):269–276.

Morita, Y., Ema, H., and Nakauchi, H. (2010). Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J Exp Med*, 207(6):1173–1182.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.

Naik, S. H., Schumacher, T. N., and Perié, L. (2014). Cellular barcoding: a technical appraisal. *Exp Hematol*, 42(8):598–608.

Nakamura, T., Yabuta, Y., Okamoto, I., Aramaki, S., Yokobayashi, S., Kurimoto, K., Sekiguchi, K., Nakagawa, M., Yamamoto, T., and Saitou, M. (2015). SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Res*, 43(9):e60.

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.

Ney, P. A., Andrews, N. C., Jane, S. M., Safer, B., Purucker, M. E., Weremowicz, S., Morton, C. C., Goff, S. C., Orkin, S. H., and Nienhuis, A. W. (1993). Purification of the human NF-E2 complex: cDNA cloning of the hematopoietic cell-specific subunit and evidence for an associated partner. *Mol Cell Biol*, 13(9):5604–5612.

Nimmerjahn, F. and Ravetch, J. V. (2006). Fcγ receptors: old friends and new family members. *Immunity*, 24(1):19–28.

Nocka, K., Tan, J. C., Chiu, E., Chu, T. Y., Ray, P., Traktman, P., and Besmer, 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. *EMBO J*, 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):aab2116.

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.

Ogawa, M. (1993). Differentiation and proliferation of hematopoietic stem cells. *Blood*, 81(11):2844–2853.

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.

Osborne, G. W. (2011). *Recent advances in cytometry, part A*, volume 102 of *Methods in cell biology*, chapter 21 - Recent advances in flow cytometric cell sorting, pages 533–556. Academic Press.

Pakos-Zebrucka, K., Koryga, I., Mnich, K., Ljujic, M., Samali, A., and Gorman, A. M. (2016). The integrated stress response. *EMBO Rep*, 17(10):1374–1395.

Pardo, J., Wallich, R., Ebnet, K., Iden, S., Zentgraf, H., Martin, P., Ekiciler, A., Prins, A., Müllbacher, A., Huber, M., and Simon, M. M. (2007). Granzyme B is expressed in mouse mast cells *in vivo* and *in vitro* and causes delayed cell death independent of perforin. *Cell Death Differ*, 14(10):1768–1779.

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.

Perié, L., Duffy, K. R., Kok, L., de Boer, R. J., and Schumacher, T. N. (2015). The branching point in erythro-myeloid differentiation. *Cell*, 163(7):1655–1662.

Peter, I. and Davidson, E. H. (2015). *Genomic control process: development and evolution*. Academic Press, 2nd edition.

Peter, I. S., Faure, E., and Davidson, E. H. (2012). Predictive computation of genomic logic processing functions in embryonic development. *Proc Natl Acad Sci U S A*, 109(41):16434–16442.

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S., and Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol*, 35(10):936–939.

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*, 10(11):1096–1098.

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. *Nat Protoc*, 9(1):171–181.

Pina, C., Fugazza, C., Tipping, A., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. *Nat Cell Biol*, 14(3):287–294.

Pina, C., Teles, J., Fugazza, C., May, G., Wang, D., Guo, Y., Soneji, S., Brown, J., Edén, P., Ohlsson, M., Peterson, C., and Enver, T. (2015). Single-cell network analysis identifies DDIT3 as a nodal lineage regulator in hematopoiesis. *Cell Rep*, 11(10):1503–1510.

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaaq1723.

Pronk, C. J. H., Rossi, D. J., Månsson, R., Attema, J. L., Norddahl, G. L., Chan, C. K. F., Sigvardsson, M., Weissman, I. L., and Bryder, D. (2007). Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell*, 1(4):428–442.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., and Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183.

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*, 29(10):886–891.

Quek, L., Otto, G. W., Garnett, C., Lhermitte, L., Karamitros, D., Stoilova, B., Lau, I.-J., Doondeea, J., Usukhbayar, B., Kennedy, A., Metzner, M., Goardon, N., Ivey, A., Allen, C., Gale, R., Davies, B., Sternberg, A., Killick, S., Hunter, H., Cahalin, P., Price, A., Carr, A., Griffiths, M., Virgo, P., Mackinnon, S., Grimwade, D., Freeman, S., Russell, N., Craddock, C., Mead, A., Peniket, A., Porcher, C., and Vyas, P. (2016). Genetically distinct leukemic stem cells in human CD34$^-$ acute myeloid leukemia are arrested at a hemopoietic precursor-like stage. *J Exp Med*, 213(8):1513–1535.

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

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. *Nat Biotechnol*, 30(8):777–782.

Rank, G., Sutton, R., Marshall, V., Lundie, R. J., Caddy, J., Romeo, T., Fernandez, K., McCormack, M. P., Cooke, B. M., Foote, S. J., Crabb, B. S., Curtis, D. J., Hilton, D. J., Kile, B. T., and Jane, S. M. (2009). Novel roles for erythroid Ankyrin-1 revealed through an ENU-induced null mouse mutant. *Blood*, 113(14):3352–3362.

Redecke, V., Wu, R., Zhou, J., Finkelstein, D., Chaturvedi, V., High, A. A., and Häcker, H. (2013). Hematopoietic progenitor cell lines with myeloid and lymphoid potential. *Nat Methods*, 10(8):795–803.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., and Yosef, N. (2017). The human cell atlas. *Elife*, 6:e27041.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rodrigues, N. P., Janzen, V., Forkert, R., Dombkowski, D. M., Boyd, A. S., Orkin, S. H., Enver, T., Vyas, P., and Scadden, D. T. (2005). Haploinsufficiency of GATA-2 perturbs adult hematopoietic stem-cell homeostasis. *Blood*, 106(2):477–484.

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.

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*, 360(6385):176–182.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol*, 33(11):1165–1172.

Rothenberg, E. V. (2014). Transcriptional control of early T and B cell developmental choices. *Annu Rev Immunol*, 32(1):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):e101–e111.

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., and Teichmann, S. A. (2017). The Human Cell Atlas: from vision to reality. *Nature*, 550(7677):451–453.

Sanchez-Freire, V., Ebert, A. D., Kalisky, T., Quake, S. R., and Wu, J. C. (2012). Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat Protoc*, 7(5):829–838.

Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., Kurisaki, A., and Nikaido, I. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol*, 19(1):29.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., and Ueda, H. R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol*, 14(4):R31.

Sawai, C. M., Babovic, S., Upadhaya, S., Knapp, D. J. H. F., Lavin, Y., Lau, C. M., Goloborodko, A., Feng, J., Fujisaki, J., Ding, L., Mirny, L. A., 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.

Schuh, A. H., Tipping, A. J., Clark, A. J., Hamlett, I., Guyot, B., Iborra, F. J., Rodriguez, P., Strouboulis, J., Enver, T., Vyas, P., and Porcher, C. (2005). ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Mol Cell Biol*, 25(23):10235–10250.

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

Schütte, J., Wang, H., Antoniou, S., Jarratt, A., Wilson, N. K., Riepsaame, J., Calero-Nieto, F. J., Moignard, V., Basilico, S., Kinston, S. J., Hannah, R. L., Chan, M. C., Nürnberg, S. T., Ouwehand, W. H., Bonzanni, N., de Bruijn, M. F. T. R., and Göttgens, B. (2016). An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *Elife*, 5:e11469.

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.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol*, 34(6):637–645.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87.

Shivdasani, R. A., Mayer, E. L., and Orkin, S. H. (1995a). Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*, 373(6513):432–434.

Shivdasani, R. A. and Orkin, S. H. (1996). The transcriptional control of hematopoiesis. *Blood*, 87(10):4025–4039.

Shivdasani, R. A., Rosenblatt, M. F., Zucker-Franklin, D., Jackson, C. W., Hunt, P., Saris, C. J., and Orkin, S. H. (1995b). Transcription factor NF-E2 is required for platelet formation independent of the actions of thrombopoietin/MGDF in megakaryocyte development. *Cell*, 81(5):695–704.

Simons, B. D. and Clevers, H. (2011). Strategies for homeostatic stem cell self-renewal in adult tissues. *Cell*, 145(6):851–862.

Simsek, T., Kocabas, F., Zheng, J., Deberardinis, R. J., Mahmoud, A. I., Olson, E. N., Schneider, J. W., Zhang, C. C., and Sadek, H. A. (2010). The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. *Cell Stem Cell*, 7(3):380–390.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods*, 11(8):817–820.

Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., and Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, doi: 10.1101/003236.

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. *Nat Biotechnol*, 36(5):469–473.

Spitzer, M. H., Gherardini, P. F., Fragiadakis, G. K., Bhattacharya, N., Yuan, R. T., Hotson, a. N., Finck, R., Carmi, Y., Zunder, E. R., Fantl, W. J., Bendall, S. C., Engleman, E. G., and Nolan, G. P. (2015). An interactive reference framework for modeling a dynamic immune system. *Science*, 349(6244):1259425.

Spitzer, M. H. and Nolan, G. P. (2016). Mass cytometry: single cells, many features. *Cell*, 165(4):780–791.

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

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. *Nat Methods*, 14(9):865–868.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.

Suda, T., Takubo, K., and Semenza, G. L. (2011). Metabolic regulation of hematopoietic stem cells in the hypoxic niche. *Cell Stem Cell*, 9(4):298–310.

Sugano, Y., Takeuchi, M., Hirata, A., Matsushita, H., Kitamura, T., Tanaka, M., and Miyajima, A. (2008). Junctional adhesion molecule-A, JAM-A, is a novel cell-surface marker for long-term repopulating hematopoietic stem cells. *Blood*, 111(3):1167–1172.

Takubo, K., Nagamatsu, G., Kobayashi, C. I., Nakamura-Ishizu, A., Kobayashi, H., Ikeda, E., Goda, N., Rahimi, Y., Johnson, R. S., Soga, T., Hirao, A., Suematsu, M., and Suda, T. (2013). Regulation of glycolysis by Pdk functions as a metabolic checkpoint for cell cycle quiescence in hematopoietic stem cells. *Cell Stem Cell*, 12(1):49–61.

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. *Nat Methods*, 6(5):377–382.

Tenen, D. G. (2003). Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer*, 3(2):89–101.

Till, J. E. and McCulloch, E. A. (1961). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res*, 14(2):213–222.

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. *Nat Biotechnol*, 32(4):381–386.

Tsai, F. Y., Keller, G., Kuo, F. C., Weiss, M., Chen, J., Rosenblatt, M., Alt, F. W., and Orkin, S. H. (1994). An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature*, 371(6494):221–226.

Tsai, F. Y. and Orkin, S. H. (1997). Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood*, 89(10):3636–3643.

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., Kawano, Y., Minegishi, Y., Eishi, Y., Yokozeki, H., and Karasuyama, H. (2009). Basophils preferentially express mouse Mast Cell Protease 11 among the mast cell tryptase family in contrast to mast cells. *J Leukoc Biol*, 86(6):1417–1425.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6):e1004333.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *J Mach Learn Res*, 9:2579–2605.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.

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. G., and Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol*, 19(4):271–281.

Waddington, C. (1957). *The strategy of the genes*. Allen and Unwin.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, 343(6166):80–84.

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*, 361(6400):eaat5691.

Weinreb, C., Wolock, S., and Klein, A. M. (2018a). SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):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. *Proc Natl Acad Sci U S A*, 115(10):E2467–E2476.

Wen, L. and Tang, F. (2018). Boosting the power of single-cell analysis. *Nat Biotechnol*, 36(5):408–409.

Wilson, A., Laurenti, E., Oser, G., van der Wath, R. C., Blanco-Bose, W., Jaworski, M., Offner, S., Dunant, C. F., Eshkind, L., Bockamp, E., Lió, P., Macdonald, H. R., and Trumpp, A. (2008). Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell*, 135(6):1118–1129.

Wilson, A., Murphy, M. J., Oskarsson, T., Kaloulis, K., Bettess, M. D., Oser, G. M., Pasche, A.-C., Knabenhans, C., Macdonald, H. R., and Trumpp, A. (2004). c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev*, 18(22):2747–2763.

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P. M., Kinston, S., Ouwehand, W. H., Dzierzak, E., Pimanda, J. E., de Bruijn, M. F., and Göttgens, B. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, 7(4):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.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*, 19(1):15.

Wolf, F. A., Hamey, F., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2017). Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv*, doi: 10.1101/208819.

Wolock, S. L., Lopez, R., and Klein, A. M. (2018). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *bioRxiv*, doi: 10.1101/357368.

Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. *J Theor Biol*, 25(1):1–47.

Woodhouse, S., Piterman, N., Köksal, A. S., and Fisher, J. (2015). Synthesizing executable gene regulatory networks from single-cell gene expression data. In *Computer Aided Verficiation*. Springer.

Wu, H., Klingmüller, U., Acurio, A., Hsiao, J. G., and Lodish, H. F. (1997). Functional interaction of erythropoietin and stem cell factor receptors is essential for erythroid colony formation. *Proc Natl Acad Sci U S A*, 94(5):1806–1810.

Wu, H., Klingmüller, U., Besmer, P., and Lodish, H. F. (1995). Interaction of the erythropoietin and stem-cell-factor receptors. *Nature*, 377(6546):242–246.

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.

Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I. R., and Ma'ayan, A. (2014). Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput Biol*, 10(8):e1003777.

Yu, W.-M., Liu, X., Shen, J., Jovanovic, O., Pohl, E. E., Gerson, S. L., Finkel, T., Broxmeyer, H. E., and Qu, C.-K. (2013). Metabolic regulation by the mitochondrial phosphatase PTPMT1 is required for hematopoietic stem cell differentiation. *Cell Stem Cell*, 12(1):62–74.

Yu, Y., Tsang, J. C. H., Wang, C., Clare, S., Wang, J., Chen, X., Brandt, C., Kane, L., Campos, L. S., Lu, L., Belz, G. T., McKenzie, A. N. J., Teichmann, S. A., Dougan, G., and Liu, P. (2016). Single-cell RNA-seq identifies a PD-1^{hi} ILC progenitor and defines its development pathway. *Nature*, 539(7627):102–106.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137.

Zheng, G. X. Y., 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. *Nat Commun*, 8:14049.

Zheng, S., Papalexi, E., Butler, A., Stephenson, W., and Satija, R. (2018). Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol Syst Biol*, 14(3):e8041.

Appendix A

Transcriptional regulatory network model rules from Chapter 4

Table A.1. Boolean rules governing the expression of each gene in both the MEP and LMPP networks. Network rule columns describe the highest scoring rules for each gene. Agreement columns show agreement level of the rule with pseudotime input-output pairs. For example, an agreement of 0.98 means that these Boolean rules agreed with 98% of the pseudotime pairs. ∧, AND; ∨, OR; ¬, NOT.

Gene	MEP network rules	LMPP network rules	MEP agreement	LMPP agreement
Bptf	Myb ∨ Gata2 ∨ Ikzf1 ∨ Lmo2	Ikzf1	0.98	0.97
	Erg ∨ Ikzf1	Nfe2		
	Smarcc1	Lmo2		
	Nfe2	Erg ∨ Smarcc1		
		Myb ∨ Smarcc1		
Cbfa2t3h	Nfe2	Fli1	0.95	0.82
	Myb ∨ Meis1 ∨ Ikzf1	Nfe2		
	Gata2 ∨ Fli1 ∨ Gata1 ∨ Meis1	Meis1		
	Ikzf1 ∨ Fli1 ∨ Gata2 ∨ Myb	Ikzf1		
Erg	Erg ∧ Meis1	Bptf	0.72	0.88
		Meis1		

		Fli1		
Ets1	(Notch ∧ Tcf7) ∧ ¬(Etv6)	Ets1 ∧ Notch	0.77	0.57
Ets2	Smarcc1 ∨ Gfi1b ∨ Fli1	Ets2	0.78	0.58
Etv6	Smarcc1	Fli1 Meis1	0.89	0.90
Fli1	Fli1 ∨ Meis1	Meis1 Nfe2 Runx1 ∨ Erg ∨ Cbfa2t3h Smarcc1 ∨ Etv6 ∨ Cbfa2t3h	0.91	0.99
Gata1	Gata1	Smarcc1 ∧ ¬Fli1 Tcf ∧ ¬Fli1 Tcf7 ∧ ¬Erg Gfi1b ∧ ¬Fli1 Gfi1b ∧ ¬Lyl1 Tcf7 ∧ ¬Nkx2.3 Tcf7 ∧ ¬Lyl1 Tcf7 ∧ ¬Hoxa9 Gata2 ∧ ¬Fli1 Myb ∧ ¬Fli1 Tal1 ∧ ¬Fli1 Hoxa5 ∧ ¬Fli1 Cbfa2t3h ∧ ¬Fli1 Gata2 ∧ ¬Lyl1 Gfi1b ∧ Tcf7 Hoxa5 ∧ ¬(Hoxa9 ∨ Nkx2.3) Gfi1b ∧ ¬(Hoxa9 ∨ Nkx2.3) Hoxa5 ∧ ¬(Erg ∨ Nkx2.3) Myb ∧ ¬(Erg ∨ Hoxa9) Smarcc1 ∧ ¬(Hoxa9 ∨ Lyl1) Hoxa5 ∧ ¬(Erg ∨ Hoxa9)	0.88	0.98

		Cbfa2t3h ∧ ¬(Erg ∨ Hoxa9) Tal1 ∧ ¬(Erg ∨ Hoxa9) Hoxa5 ∧ Myb ∧ ¬Hoxa9 Gata2 ∧ ¬(Erg Or Hoxa9) Smarcc1 ∧ ¬(Erg ∨ Lyl1) Hoxa5 ∧ ¬(Lyl1 ∨ Nkx2.3) Myb ∧ ¬(Lyl1 ∨ Nkx2.3) Smarcc1 ∧ ¬(Lyl1 ∨ Nkx2.3) Hoxa5 ∧ ¬(Erg ∨ Lyl1) Myb ∧ ¬(Erg ∨ Lyl1) Tal1 ∧ ¬(Lyl1 ∨ Nkx2.3) Cbfa2t3h ∧ ¬(Lyl1 ∨ Nkx2.3) Cbfa2t3h ∧ ¬(Erg ∨ Lyl1) Hoxa5 ∧ ¬(Hoxa9 ∨ Lyl1) Smarcc1 ∧ ¬(Erg ∨ Hoxa9) Gfi1b ∧ Hoxa5 ∧ ¬Erg Tal1 ∧ ¬(Erg ∨ Lyl1) Cbfa2t3h ∧ ¬(Hoxa9 ∨ Lyl1) Tal1 ∧ ¬(Hoxa9 ∨ Lyl1)		
Gata2	Nfe2 Bptf Gfi1b ∨ Meis1 Gata1 ∨ Meis1 Cbfa2t3h ∨ Meis1 Cbfa2t3h ∨ Pbx1	Gata2	0.97	0.79
Gata3	Gata3 ∧ ¬Myb	Tal1 ∧ Gata3	0.72	0.65
Gfi1b	Gata1 ∨ Ets2 ∨ Gata2 ∨ Tal1	Gata2 ∧ Gfi1b ∧ ¬Notch	0.86	0.74
Hhex	Nfe2	Hhex	0.74	0.56

Hoxa5	Tcf7	Prdm16 ∧ Tcf7	0.87	0.85
Hoxa9	Ikzf1 ∧ Meis1	Meis1 Ikzf1 Nkx2.3 ∨ Ets1 ∨ Lyl1 ∨ Hoxa5	0.65	0.87
Hoxb4	Tcf7	Tcf7	0.84	0.80
Ikzf1	Smarcc1 Bptf Cbfa2t3h ∨ Hoxa9	Bptf Hoxa9 ∨ Smarcc1 ∨ Myb ∨ Cbfa2t3h	0.93	0.96
Ldb1	Smarcc1 Myb ∨ Lmo2 ∨ Ikzf1	Ikzf1 Lmo2 Myb ∨ Smarcc1	0.87	0.81
Lmo2	Bptf Nfe2 Lyl1 Ldb1 ∨ Meis1	Bptf Meis1 Nfe2 Lyl1 ∨ Notch Ldb1 ∨ Lyl1 Lyl1 ∨ Tal1 Lyl1 ∨ Nkx2.3	0.96	0.98
Lyl1	Smarcc1 Nfe2 Myb ∨ Hoxa9 ∨ Lmo2	Nfe2 Lmo2 Erg ∨ Smarcc1 ∨ Hoxa9 Myb ∨ Smarcc1 Erg ∨ Lmo2	0.95	0.98
Meis1	Meis1 ∨ Erg	Lmo2 Nfe2 Fli1 Nkx2.3 ∨ Runx1 Cbfa2t3h ∨ Etv6 Erg ∨ Hoxa9 Cbfa2t3h ∨ Erg ∨ Runx1	0.88	0.99
Mitf	Fli1 ∧ Mitf	Mitf	0.67	0.66

Myb	(Gata1 ∨ Runx1) ∧ ¬(Gata3 ∧ Nkx2.3)	Myb ∨ Ldb1 ∧ ¬(Gata3 ∧ Prdm16)	0.83	0.69
Nfe2	Bptf Lyl1 Fli1 ∨ Myb ∨ Lmo2 ∨ Meis1 Cbfa2t3h ∨ Lmo2 ∨ Gata2 Gata2 ∨ Myb Cbfa2t3h ∨ Fli1 Hhex ∨ Lmo2 Cbfa2t3h ∨ Meis1	Bptf Fli1 Lmo2 Meis1 Cbfa2t3h ∨ Lyl1 ∨ Myb ∨ Fli1	0.99	0.99
Nkx2.3	Nkx2.3 ∧ ¬Gata1	Lmo2 Meis1 Hoxa9 ∨ Lyl1	0.85	0.77
Notch	Tcf7 Ets1 ∧ ¬Gata2 Ets1 ∧ ¬Nfe2 Lmo2 ∧ ¬Nfe2 Lmo2 ∧ ¬(Gata2 ∨ Gfi1b)	Lmo2 ∧ ¬(Gata2 ∧ Gfi1b)	0.84	0.66
Pbx1	Meis1	Gata2 ∨ Runx1	0.75	0.71
Prdm16	Fli1 ∧ ¬Myb	Hoxa5 ∧ ¬Myb	0.75	0.71
Runx1	Fli1 ∨ Myb ∨ Meis1	Fli1 Meis1	0.90	0.88
Smarcc1	Bptf Lyl1 Etv6 ∨ Ldb1 ∨ Ikzf1 ∨ Fli1 Fli1 ∨ Myb ∨ Ikzf1 ∨ Ets2 Ets2 ∨ Etv6 Fli1 ∨ Gata1	Ikzf1 Fli1 Bptf Lyl1 ∨ Etv6 ∨ Ldb1 Lyl1 ∨ Myb ∨ Ets2	0.99	0.98
Tal1	Lmo2 ∨ Gata1 ∨ Gfi1b ∨ Gata2	Gata2 ∨ Tal1	0.85	0.64
Tcf7	Notch ∧ ¬Ikzf1	Hoxb4 ∧ ¬Ikzf1	0.97	0.97

	Hoxa5 ∧ ¬Ikzf1	Hoxa5 ∧ ¬Ikzf1		
	Nkx2.3 ∧ ¬Ikzf1	Gata1 ∧ ¬Ikzf1		
	Gata3 ∧ ¬Ikzf1	Notch ∧ ¬Ikzf1		
	Hoxb4 ∧ ¬Ikzf1	Nkx2.3 ∧ ¬Ikzf1		
	Ets1 ∧ ¬Ikzf1	Ets1 ∧ ¬Ikzf1		
	Gata1 ∧ ¬Ikzf1	Runx1 ∧ ¬Ikzf1		
	Gata1 ∧ Hoxa5	Gata3 ∧ ¬Ikzf1		

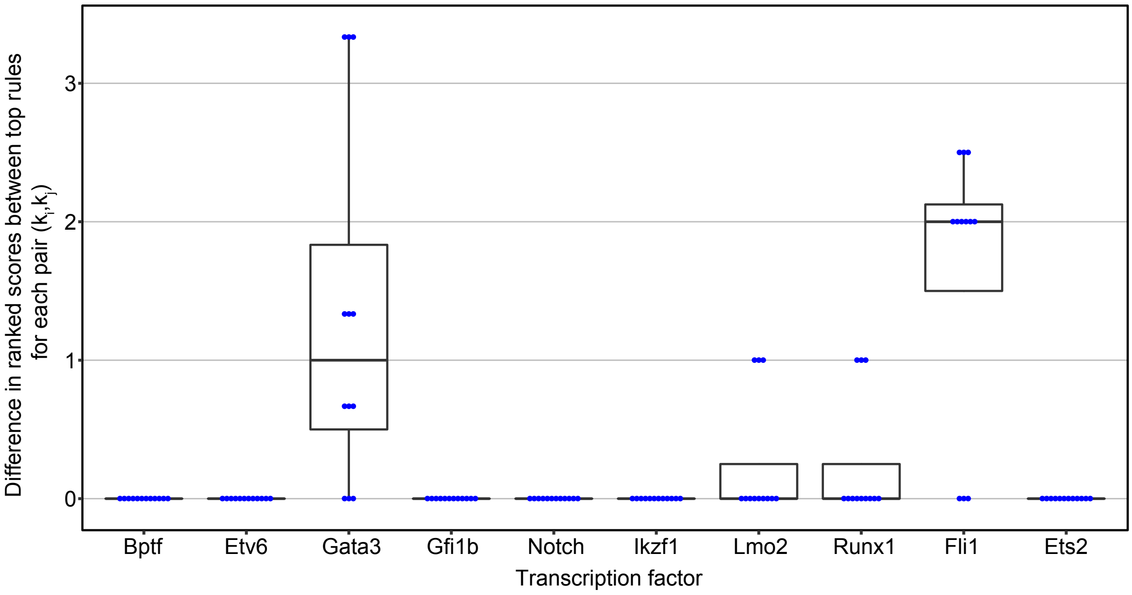


Fig. A.1. Influence of choice of parameter k on rule scoring. Network inference was run on 10 randomly selected genes for the MEP model inference using $k \in \{3,5,7,9\}$. The rules for each k were ranked by their pseudotime-agreement score, so the top-scoring rule(s) would get rank 0, the next best scoring rule(s) rank 1, and so on. The top-scoring rules for each k_i were then compared to the results of rule scoring with each other $k_j \neq k_i$. For each top-scoring k_i rule, its rank was found for each of the other k_j, and this rank averaged across all of the top-scoring k_i rules for each k_i vs k_j comparison. These 12 averages (corresponding to the 12 (k_i, k_j) pairs) are shown for each gene. An average of 0 means that the top-scoring rules for the k_i were also in the top-scoring rules for k_j. The majority of comparisons had this level of agreement. When the top-scoring rules did not match, this analysis shows that they were still amongst the high-ranking rules for the other k_i.