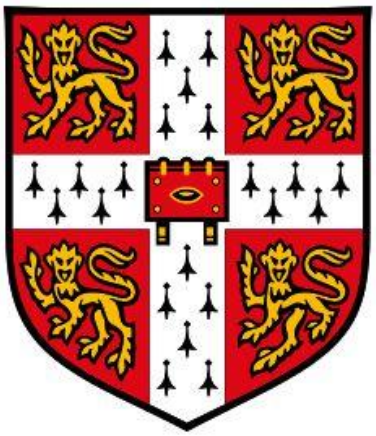
Mapping the transcriptional landscape of haematopoietic stem and progenitor cells



Sonia Shaw

(née Nestorowa)

Department of Haematology

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Pembroke College

January 2019

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. Specific details of work done in collaboration are given at the start of relevant chapters. The contents of this dissertation are 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 50,345 words and therefore does not exceed the limit of 60,000 words for such a dissertation.

Sonia Shaw

January 2019

Mapping the transcriptional landscape of haematopoietic stem and progenitor cells

Sonia Shaw

Maintenance of the blood system requires balanced cell-fate decisions of haematopoietic stem and progenitor cells (HSPCs). Individual haematopoietic stem cells (HSCs) decide between self-renewal and differentiation and can generate all mature cell types. Cell-fate decisions are made at the single-cell level and are governed by regulatory networks. Dysregulation in this balanced process could lead to serious blood disorders such as leukaemia; therefore, it is important to understand how individual cells make these cell-fate decisions.

To investigate HSPC populations, 1,654 cells were profiled by single-cell RNA-sequencing. Index sorting made it possible to sort HSPCs using broad sorting gates and retrospectively assign them to common HSPC populations, retaining all information about specific functionally pure populations while also capturing any intermediate cells normally excluded by conventional gating. Reconstruction of differentiation trajectories revealed dynamic expression changes associated with early lineage differentiation from HSCs. This transcriptional atlas of HSPC differentiation was further used to identify candidate genes for a CRISPR screen investigating genes implicated in HSC biology. These candidate gene perturbations were interrogated for changes in the expression of the HSC marker EPCR, as well as changes in apoptosis and lineage output.

Transcription factors play a key role in regulating cell-fate decisions and operate within organized regulatory programs. To study relationships between transcription factors in HSPC populations, qRT-PCR was used to profile the expression of 41 genes, including 31 transcription factors, in HSPCs at the single-cell level. This approach confirmed known aspects of haematopoiesis and made deeper investigation of HSPC heterogeneity possible. Regulatory networks were reconstructed using Boolean network inference models and recapitulated differentiation of HSCs towards megakaryocyte–erythrocyte progenitors and lymphoid-primed multipotent progenitors. By comparing these two models, a rule specific to the megakaryocyte-erythrocyte progenitor network was identified, in which GATA2 positively regulated *Nfe2* and *Cbfa2t3h*. This was subsequently validated using transcription factor binding profiles and *in vitro* luciferase assays using a model cell line.

Overall, the work presented in this thesis confirmed known aspects of HSPC biology using single-cell gene expression analysis and demonstrated how *in silico* approaches can be used to guide *in vitro* and *in vivo* investigations. In addition, the single-cell RNA-sequencing data was developed into an intuitive web interface that can be used to visualise the gene expression for any gene of choice at single-cell resolution across the HSPC atlas, providing a powerful resource for the haematopoietic community.

Acknowledgements

First and foremost, I would like to thank my supervisor, Berthold Göttgens, for his mentorship and support over the last four years, as well as Nicola Wilson, who has been a great source of support, both academically and otherwise. I am very grateful to Fiona Hamey for our collaborative work, her patience with my endless questions, and for being a joy to know and work with. I would also like to thank Iwo Kucinski and Joakim Dahlin, for always being available to advise me, support me, and for generally making the lab brighter. I am greatly indebted to everyone in the Göttgens lab for always being willing to help and for being great people to work with throughout my time in the lab.

I would also like to thank David Rubinsztein and Scottie Robinson for accepting me onto the CIMR PhD programme, as well as the Medical Research Council for funding my PhD. A huge thank you also goes to Martin Dawes, who always had the answers to all my questions, and has always gone above and beyond to help. Thank you to the Flow Cytometry Team at the CIMR for being a huge support during my PhD and making me look forward to otherwise very long sorting days!

I would like to thank my parents, who have been an endless source of love and fierce support and have been invaluable in the toughest times. My wonderful dogs, Trixie and Huxley, are two of the greatest treasures in my life and have provided endless distraction when most needed, as well as unwavering love and emotional support in the form of licks and big, breath-taking cuddles (literally).

Finally, I would like to thank my husband Ben Shaw, without whom this PhD and the last four years may have looked very different. He has supported me in every way imaginable and I will be eternally grateful to him for making me a better, stronger, happier person, and for helping me achieve all my goals and more.

Table of Contents

Declaration....i

Abstract.... ii

Acknowledgements v

Table of Contents vii

List of Figures....x

List of Tables.... xiii

Abbreviations xiv

Papers arising from this PhD....xvi

Chapter 1: Introduction....1

1.1. The adult haematopoietic system....1

1.2. Cell fate decision making 5

1.3. Dysregulation of transcriptional control in haematopoiesis 8

1.4. Mammalian genome editing 9

1.5. Single cell biology 11

1.6. Aims....25

Chapter 2: Materials & Methods 26

2.1. Cell culture 26

2.2. Cell Biology....27

2.3. Molecular Biology....39

2.4. HSC retroviral transduction analysis: genotyping....45

2.5. Single Cell Gene Expression Analysis 48

2.6. Computational Analysis 53

Chapter 3: A single-cell atlas of adult murine haematopoiesis 57

3.1. Background....57

3.2. Isolation of haematopoietic stem and progenitor cells for single-cell analysis....59

3.3. Single-cell gene expression analysis reveals distinct HSPC clusters 60
3.4. Transcriptional profiling reveals a differentiation continuum 63
3.5. Visualising gene expression on the HSPC differentiation continuum 64
3.6. Linking cell phenotypes with the transcriptome 67
3.7. Capturing changes in gene and protein expression using pseudotime ordering 71
3.8. Ordering cells along differentiation trajectories using STREAM 73
3.9. Visualising HSPCs and capturing rare populations using SPRING 78
3.10. Conclusions 82
Chapter 4: CRISPR screening of HSC-enriched genes 86
4.1. Background 86
4.2. Study design to investigate genes implicated in HSC biology 88
4.3. Identifying candidate genes important to HSC characteristics 89
4.4. CRISPR screen candidate genes 95
4.5. Isolation and analysis of E-SLAM HSCs using flow cytometry 96
4.6. Effect of candidate gene perturbations on EPCR expression in E-SLAM cells 99
4.7. Candidate gene perturbation does not influence apoptosis in E-SLAM cells 102
4.8. Changes in lineage output after candidate gene perturbation in E-SLAM cells 105
4.9. Genotyping confirms CRISPR gRNAs are correctly targeting the candidate genes 108
4.10. Conclusions 109
Chapter 5: Resolving heterogeneity in HSPC populations 113
5.1. Background 113
5.2. Isolation of haematopoietic stem and progenitor cell populations 114
5.3. Selection of a gene set for HSPC analysis 117
5.4. Processing single cells using the Fluidigm BioMark™ platform 122
5.5. Resolving populations using multidimensionality analysis 123
5.6. Single-cell gene expression analysis reveals cell population clusters 128

5.7. Pairwise correlation analysis reveals putative relationships between genes 129

5.8. Reconstructing differentiation trajectories from single-cell gene expression profiles 133

5.9. Conclusions 136

Chapter 6: Validating regulatory networks models 140

6.1. Background 140

6.2. Boolean network modelling reveals regulatory relationships within MEP and LMPP differentiation networks 143

6.3. Linking ChIP-seq data and regulatory rules to identify relationships to validate *in vitro* 146

6.4. *In vitro* validation supports differences in network model connectivity 152

6.5. Conclusions 153

Chapter 7: Discussion 156

7.1. Thesis Overview 156

7.2. Future directions for single-cell biology 163

7.3. Concluding remarks 171

References 172

List of Figures

Figure 1.1. The haematopoietic differentiation hierarchy. 2

Figure 1.2. Single-cell analysis. 13

Figure 1.3. Single-cell profiling enables the exploration of cell population heterogeneities. 18

Figure 1.4. Reconstructing lineage differentiation using single-cell expression profiles. 21

Figure 1.5. Inferring regulatory relationships from single-cell expression data. 23

Figure 3.1. Isolation and profiling of HSPCs at the single cell level. 60

Figure 3.2. Unbiased hierarchical clustering reveals heterogeneity in the gene expression of HSPC clusters. 62

Figure 3.3. Multidimensional analysis shows a differentiation continuum towards different blood lineages. 64

Figure 3.4. Gene expression in the HSPC differentiation atlas. 66

Figure 3.5. Surface marker expression on the HSPC differentiation atlas. 68

Figure 3.6. Visualising HSPC populations on the HSPC differentiation atlas. 70

Figure 3.7. scRNA-seq profiles can be computationally ordered along pseudotime trajectories. 72

Figure 3.8. Pseudotime ordering reveals changes in gene expression during differentiation. 73

Figure 3.9. STREAM analysis reveals information about pseudotime ordering in the HSPC differentiation landscape. 77

Figure 3.10. Visualising HSPC populations using the SPRING interface. 79

Figure 3.11. SPRING analysis captures rare populations in the HSPC differentiation atlas. 81

Figure 4.1. Study design to investigate genes implicated in HSC biology. 89

Figure 4.2. Expression plots of potential candidate genes for the CRISPR study. 91

Figure 4.3. Violin plots for selected candidate genes. 92

Figure 4.4. Violin plots for selected and discarded candidate genes (continued). 93

Figure 4.5. Expression plots of MolO and SuMO genes eliminated as candidates for the CRISPR study. 94

Figure 4.6. Flow cytometry sorting and analysis strategies for E-SLAM cells.....98

Figure 4.7. Changes in the percentage of EPCR+ cells after candidate gene perturbation. 100

Figure 4.8. Changes in median EPCR expression after candidate gene perturbation. 102

Figure 4.9. Candidate gene perturbation does not influence apoptosis in E-SLAM cells.....104

Figure 4.10. Changes in lineage output after candidate gene perturbation. 107

Figure 4.11. Genotyping analysis indicates the CRISPR gRNAs are correctly targeting the candidate genes. 109

Figure 5.1. Isolation of HSPCs for single-cell gene expression analysis. 116

Figure 5.2. Visualisation of single-cell qRT-PCR data using principal component analysis. 125

Figure 5.3. Visualisation of single-cell qRT-PCR data using t-distributed stochastic neighbour embedding. 126

Figure 5.4. Visualisation of single-cell qRT-PCR data using diffusion maps. 127

Figure 5.5. Unsupervised clustering of HSPC populations reveals clusters in single-cell gene expression data. 129

Figure 5.6. Pairwise correlations reveal regulatory relationships between transcription factors.131

Figure 5.7. Pairwise correlations for each HSPC population. 133

Figure 5.8. Pseudotime ordering reveals two differentiation trajectories in the single-cell HSPC data. 135

Figure 6.1. Single-cell molecular profiles allow inference of regulatory network models. 141

Figure 6.2. Transcriptional regulatory network models for differentiation from HSCs to MEPs or LMPPs. 142

Figure 6.3. Transcriptional profiles of model haematopoietic cell lines occupy territories on MEP and LMPP trajectories on the HSPC qRT-PCR dataset. 146

Figure 6.4. ChIP-seq data reveals transcription factor binding patterns in 416B and HoxB8-FL cell lines.....148

Figure 6.5. Integrating transcription factor binding with network rules to identify regulatory relationships to validate in vitro. 149

Figure 6.6. Regulatory relationships unique to the MEP network model are supported by transcription factor binding. 151

Figure 6.7. In vitro validation of MEP regulatory network rules. 153

List of Tables

Table 2.1. Mouse strain, age, sex, and number used for FACS. 29

Table 2.2. List of antibodies used for FACS and flow cytometry. 31

Table 2.3. Antibody panels used for FACS and flow cytometry. 32

Table 2.4. MolO and SuMO genes. 33

Table 2.5. Sequences of gRNAs used for the HSC CRISPR screen. 36

Table 2.6. Primer sequences used to sequence the pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W vector. 38

Table 2.7. List of plasmids used as backbones for various applications. 45

Table 2.8. Primer overhang sequences 46

Table 2.9. Genotyping primers 46

Table 2.10. Thermocycler conditions for synthesis and specific target amplification of cDNA from single cells. 48

Table 2.11. List of TaqMan® assays used for single cell gene expression analysis. 49

Table 2.12. Thermocycler conditions for reverse transcription (Smart-seq2* protocol). 51

Table 2.13. Thermocycler conditions for preamplification (Smart-seq2* protocol). 51

Table 2.14. Oligo sequences. 51

Table 2.15. Thermocycler conditions for amplification of cDNA libraries. 52

Table 4.1. List of all MolO and SuMO genes identified by Wilson *et al.* (2015). 87

Table 4.2. Properties of candidate genes selected for the CRISPR screen. 95

Table 5.1. Properties of genes selected for single-cell gene expression analysis. 117

Table 5.2. Table indicating number of cells included in this study. 123

Table 6.1. Simplified rules: activating relationships in MEP and LMPP networks. 143

Table 6.2. Simplified rules: repressive relationships in MEP and LMPP networks. 145

Table 7.1. Single-cell methods interrogating the epigenome. 167

Abbreviations

7-AAD	7-amino-actinomycin
ALL	Acute lymphoid leukaemia
AML	Acute myeloid leukaemia
ANOVA	Analysis of variance
BFU-E	Burst forming unit-erythroid
BSA	Bovine serum albumin
cDNA	Complementary DNA
CFU	Colony forming unit
CFU-GEMM	Colony forming unit-granulocyte-erythrocyte-macrophage-megakaryocyte
CFU-GM	Colony forming unit-granulocyte-macrophage
ChIP-seq	Chromatin immunoprecipitation sequencing
CITE-seq	Cellular indexing of transcriptomes and epitopes by sequencing
CLP	Common lymphoid progenitor
CML	Chronic myeloid leukaemia
CMP	Common myeloid progenitor
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
Ct	Cycle-threshold
CyTOF	Cytometry by time of flight
DAPI	4',6-Diamidino-2-Phenylindole
DC	Diffusion component
DMEM	Dulbecco's Modified Eagle Medium
DPT	Diffusion pseudotime
DR-seq	gDNA-mRNA sequencing
E	Erythroid lineage
EPCR	Endothelial protein C receptor
$EPCR^+$ cells	GFP^+ Lin^- $EPCR^+$ $Sca1^+$ cells (Chapter 4)
ERCC	External RNA controls consortium
E-SLAM	$EPCR^+$ HSC population ($CD48^-$ $CD150^+$ $EPCR^+$)
FACS	Fluorescence-activated cell sorting
FBS	Foetal bovine serum
FCS	Foetal calf serum
FSC-H	Forward-scattered light-height
FSR-HSC	Finite self-renewal haematopoeitic stem cell
FMO	Fluorescence Minus One
G&T-seq	Genome and transcriptome sequencing
gDNA	Genomic DNA
GFP	Green fluorescent protein
GM	Granulocyte-macrophage lineage
GMP	Granulocyte-macrophage progenitor
GRE	Gene regulatory element
gRNA	guide RNA
GSEA	Gene set enrichment analysis
HSC	Haematopoietic stem cell
HSPC	Haematopoietic stem and progenitor cell
HSPC gate	Lin^- $Sca1^+$ $c\text{-}Kit^+$ sorting gate
L	Lymphoid lineage
LB	Lysogeny broth
Lin	Lineage
LINNAEUS	Lineage tracing by nuclease-activated editing of ubiquitous sequences

LMPP	Lymphoid multipotent progenitor
LSK	Lin- Sca1+ c-Kit+ phenotype
LT-HSC	Long-term haematopoietic stem cell
MARS-seq	Massively parallel single-cell RNA sequencing
MEP	Megakaryocyte-erythroid progenitor
MLLE	Modified local linear embedding
MolO	Molecularly overlapping HSCs/genes (N. K. Wilson et al. 2015)
MPN	Myeloproliferative neoplasms
MPP	Multipotent progenitor
mRNA	Messenger RNA
P/S	Penicillin-streptomycin
PBS	Phosphate-buffered saline
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PreMegE	Pre-megakaryocyte-erythrocyte progenitor
Prog	Progenitor (Lin^- $Sca1^-$ c-Kit^+) sorting gate
QC	Quality control
qRT-PCR	Quantitative real-time polymerase chain reaction
REAP-seq	RNA expression and protein sequencing
RNA-seq	RNA sequencing
RT	Reverse transcription
scATAC-seq	Single-cell assay for transposase-accessible chromatin
scBS	Single-cell bisulfite sequencing
sc-GEM	Single-cell analysis of genotype, expression and methylation
scGESTALT	Single-cell genome editing of synthetic target arrays for lineage tracing
scMT-seq	Single-cell genome-wide methylome and transcriptome sequencing
scNMT-seq	Single-cell nucleosome, methylation and transcription sequencing
scRNA-seq	Single-cell RNA-sequencing
scRRBS	Single-cell reduced representation bisulfite sequencing
SLAM	Signalling lymphocytic activation molecule ($CD48^-$ $CD150^+$)
SPADE	Spanning-tree progression analysis of density-normalised events
SPLiT-seq	Split pool ligation-based transcriptome sequencing
STREAM	Single-cell trajectories reconstruction, exploration and mapping
SuMO	Surface marker overlap population/genes (N. K. Wilson et al. 2015)
TALEN	Transcription activator-like effector nuclease
TBE	Tris/Borate/EDTA buffer
t-SNE	t-distributed stochastic neighbour embedding
ZFN	Zinc finger nuclease
ΔCt	Change in cycle-threshold

Papers arising from this PhD

All publications were published under Sonia Nestorowa (maiden name)

* denotes equal contribution of both authors

1. Azzarelli, Roberta, Christopher Hurley, Magdalena K. Sznurkowska, Steffen Rulands, Laura Hardwick, Ivonne Gamper, Fahad Ali, Laura McCracken, Christopher Hindley, Fiona McDuff, Sonia Nestorowa, et al. 2017. "**Multi-Site Neurogenin3 Phosphorylation Controls Pancreatic Endocrine Differentiation.**" *Developmental Cell* 41 (3): 274–286.e5.
2. Azzarelli, Roberta, Steffen Rulands, Sonia Nestorowa, John Davies, Sara Campinoti, Sébastien Gillotin, Paola Bonfanti, et al. 2018. "**Neurogenin3 Phosphorylation Controls Reprogramming Efficiency of Pancreatic Ductal Organoids into Endocrine Cells.**" *Scientific Reports* 8 (1): 15374.
3. Dahlin, Joakim S.*, Fiona K. Hamey*, Blanca Pijuan-Sala, Mairi Shepherd, Winnie W. Y. Lau, Sonia Nestorowa, Caleb Weinreb, et al. 2018. "**A Single-Cell Hematopoietic Landscape Resolves 8 Lineage Trajectories and Defects in Kit Mutant Mice.**" *Blood* 131 (21): e1–11.
4. Hamey, Fiona K*, Sonia Nestorowa*, Sarah J Kinston, David G Kent, Nicola K Wilson, and Berthold Göttgens. 2017. "**Reconstructing Blood Stem Cell Regulatory Network Models from Single-Cell Molecular Profiles.**" *Proceedings of the National Academy of Sciences of the United States of America* 114 (23): 5822–29.
5. Hamey, Fiona K*, Sonia Nestorowa*, Nicola K Wilson, and Berthold Gottgens. 2016. "**Advancing Haematopoietic Stem and Progenitor Cell Biology through Single Cell Profiling.**" *FEBS Letters* 590(22): 4052-4067.
6. Nestorowa, Sonia*, Fiona K Hamey*, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. 2016. "**A Single Cell Resolution Map of Mouse Haematopoietic Stem and Progenitor Cell Differentiation.**" *Blood* 128(8): e20-31.
7. Sznurkowska, Magdalena K., Edouard Hannezo, Roberta Azzarelli, Steffen Rulands, Sonia Nestorowa, Christopher J. Hindley, Jennifer Nichols, et al. 2018. "**Defining Lineage Potential and Fate Behavior of Precursors during Pancreas Development.**" *Developmental Cell* 46 (3): 360–375.e5.
8. Wilkinson, Adam C., David J. Ryan, Iwo Kucinski, Wei Wang, Jian Yang, Sonia Nestorowa, Evangelia Diamanti, et al. 2019. "**Expanded Potential Stem Cell Media as a Tool to Study**

Human Developmental Hematopoiesis in Vitro." *Experimental Hematology* https://doi.org/10.1016/j.exphem.2019.07.003.

Chapter 1: Introduction

Parts of this chapter have been adapted from the review article co-authored by Fiona Hamey and Sonia Shaw (Hamey et al. 2016).

1.1. The adult haematopoietic system

The haematopoietic system consists of cells of the blood and immune systems, and is created by haematopoiesis, the process of mature blood cell formation (Orkin and Zon 2008; Laurenti and Göttgens 2018). The adult human has an estimated turnover of around 10^{12} blood cells per day, sustained by the constant production of new cells (Ogawa 1993). All cells of the haematopoietic system arise from haematopoietic stem cells (HSCs), and the mature blood and immune cells are involved in performing critical functions such as oxygen and nutrient transport, immune protection, and wound repair.

The haematopoietic system has been extensively studied due to the accessibility of material and the ability to isolate and study cells at multiple stages of differentiation (N. K. Wilson et al. 2015). It serves as a model system of adult stem cell biology, as well as the cell fate decisions that occur during HSC differentiation (Section 1.2) (Orkin and Zon 2008; Laurenti and Göttgens 2018). The haematopoietic system is an important system to study, as dysregulation during haematopoiesis can lead to serious blood disorders, such as leukaemia (Section 1.3).

1.1.1. The haematopoietic hierarchy

The adult haematopoietic system is classically described as a hierarchy, in which long-term HSCs (LT-HSCs) lie at the apex of the tree and differentiate through progressively more committed progenitors that give rise to all the mature blood cell types (Fig. 1.1). The mature cells produced can be divided into myeloid, erythroid, and lymphoid cells, with specific functions ranging from oxygen transport by erythrocytes and phagocytosis of pathogens by macrophages, to functions in innate and adaptive immunity by lymphocytes such as natural killer cells or B- and T-cells (Bryder, Rossi, and Weissman 2006).

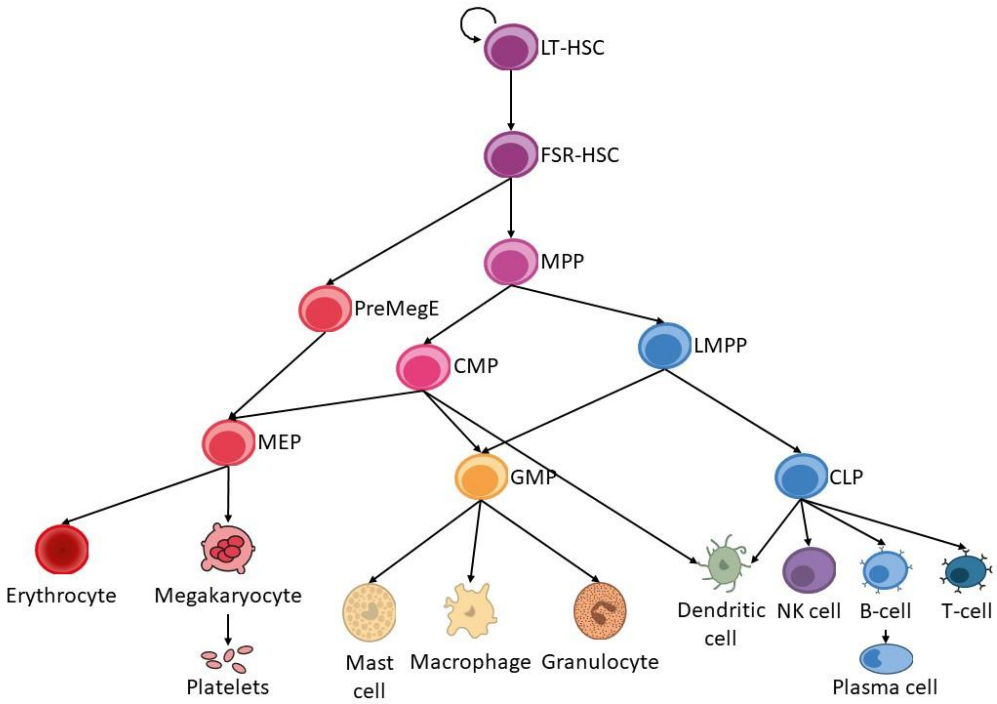


Figure 1.1. The haematopoietic differentiation hierarchy. The classic view of the haematopoietic tree, based on publications by Moignard *et al.* and Wilson *et al.* (Moignard et al. 2013; N. K. Wilson et al. 2015). Long-term HSCs (LT-HSCs) sit at the top of the hierarchy and differentiation through progressively more committed progenitor cells into mature blood cell types. FSR-HSC: finite self-renewal HSC; MPP: multipotent myeloid progenitor; PreMegE: pre-megakaryocyte-erythrocyte progenitor; MEP: megakaryocyte-erythrocyte progenitor; CMP: common myeloid progenitor; GMP: granulocyte-macrophage progenitor; LMPP: lymphoid multipotent progenitor; CLP: common lymphoid progenitor; NK cell: natural killer cell.

Haematopoietic stem and progenitor cells (HSPCs) are found in the bone marrow, the home of haematopoiesis in adult mammals (Eaves 2015). At the top of the hierarchy, HSCs are responsible for maintaining the entire haematopoietic system. **LT-HSCs** are capable of both symmetric and asymmetric division, meaning they can produce two HSCs or two progenitor cells, or one of each, respectively. They generally exist in a quiescent, non-replicative state, and are defined by their ability to reconstitute the haematopoietic system in irradiated mice over several months (Morrison and Weissman 1994). They were first phenotypically defined as Lineage$^-$ Thy1.1^{low} Sca1$^+$, where Lineage (Lin) describes a cocktail of mature blood cell markers (Spangrude, Heimfeld, and Weissman 1988). Since then, HSC isolation protocols have advanced and can isolate enriched HSC populations with increasing functional purity. These strategies are largely based on the LSK (Lin$^-$ Sca1$^+$ c-Kit$^+$) phenotype and isolate HSCs with up to 50% functional purity; however, higher enrichment can be achieved by using the signalling lymphocyte activation molecule (SLAM) family of surface markers, including CD48 and CD150, as well as endothelial protein C receptor (EPCR) (Beerman et al. 2010; Dykstra et al. 2007; Goodell et al. 1996; Morita, Ema, and Nakauchi

2010; Kiel et al. 2005; Balazs et al. 2006; D. G. Kent et al. 2009). In fact, 67% functional purity was achieved by isolating SLAM ($CD48^-$ $CD150^+$) $Sca1^{hi}$ cells (N. K. Wilson et al. 2015).

The LSK compartment, which makes up 0.05-1% of the adult bone marrow, also contains short-term or finite self-renewal HSCs (FSR-HSC), and multipotent progenitors (MPPs). **FSR-HSCs** and **MPPs** are multipotent progenitors that can transiently provide multilineage repopulation in lethally irradiated mice but have limited self-renewal capacity (Spangrude, Heimfeld, and Weissman 1988; Curtis et al. 2004; Morrison et al. 1997; Okada et al. 1992; Harrison and Zhong 1992). MPPs can be subdivided into four subpopulations: MPP1-4 (Pietras et al. 2015). MPP1 cells, or FSR-HSCs, were identified as a metabolically active subset of HSCs (A. Wilson et al. 2008; Cabezas-Wallscheid et al. 2014). MPP2 and MPP3 cells are distinct myeloid-biased subsets, with MPP2 biased towards megakaryocyte production, and MPP3 displaying granulocyte-macrophage potential. Finally, MPP4 cells are a lymphoid-primed MPP subset (Pietras et al. 2015). All four MPP subtypes work together to control blood production. It was originally thought that MPPs give rise to mature blood cells by directly differentiating into common myeloid progenitors (CMPs) and common lymphoid progenitors (CLPs) (Akashi et al. 2000; Kondo, Weissman, and Akashi 1997). However, more recent work showed that MPPs high in Flk2 expression produce both lymphoid and granulocyte-monocyte lineages (Adolfsson et al. 2001, 2005). These cells were named lymphoid-primed multipotent progenitors (LMPPs).

LMPP cells expressed granulocyte-macrophage and lymphoid genes, and can produce granulocytes, macrophages, T- and B-cells, and only very rarely megakaryocyte-erythroid cells (Luc et al. 2008; Adolfsson et al. 2001, 2005). A study by Pronk *et al.* compared transcriptional profiles of CLPs and a pre-granulocyte-macrophage cell type with predominant myeloid potential, and found they were more similar to each other than either cell type was to cells of the erythroid lineage (Pronk et al. 2007). These results suggest that the lymphoid and myeloid lineages branch off earlier than the erythroid lineage, and LMPPs mark a first step towards lymphoid and myeloid lineages. As cells become increasingly lymphoid-restricted, granulocyte-monocyte potential is lost, resulting in **CLPs**. These cells lack myeloid potential, but can rapidly produce natural killer (NK) cells, B- and T-cells (Kondo, Weissman, and Akashi 1997).

CMPs were identified by Akashi *et al*. as a population that can give rise to all myeloid lineages, producing megakaryocyte-erythroid progenitors (MEPs) and granulocyte-monocyte progenitors (GMPs) (Akashi et al. 2000). The functional and cell-fate identity of this population has been

challenged by several studies (Section 1.1.2). As it is possible to isolate both CMPs and LMPPs, it is possible **GMPs** can arise independently from both populations. These cells produce macrophages, granulocytes and mast cells (Dahlin and Hallgren 2015; Akashi et al. 2000). **MEPs**, on the other hand, exclusively produce erythrocytes, megakaryocytes, or mixed colonies (Akashi et al. 2000). Evidence suggests that MEPs may be derived from CMPs or directly from HSCs (H. Iwasaki et al. 2005). More detailed erythroid and myeloid progenitors can be elucidated using CD105, CD150 and CD41, which identified the **PreMegE** (pre-megakaryocyte-erythrocyte) population (Pronk et al. 2007). These cells produced erythrocyte, megakaryocyte, and mixed colonies, indicating PreMegEs have no granulocyte-macrophage potential and include bipotential progenitors as well as cells in early commitment stages. They are likely to occur slightly earlier in development than MEPs as functional and gene expression analyses indicate they are upstream of early unipotent cell populations that specifically give rise to erythroid colonies or megakaryocytes (Pronk et al. 2007).

1.1.2. Challenges to the classic view of the haematopoietic hierarchy

In recent years, questions have been raised that are altering the classical view of the haematopoietic hierarchy, in which all mature blood cells are produced from a single LT-HSC population (Cavazzana-Calvo et al. 2011). In mice, four subtypes of adult bone marrow HSCs were identified (α, β, γ and δ) that vary in their lineage bias and self-renewal activity (Dykstra et al. 2007). Specifically, α and β cells have robust self-renewal activity, but α cells are lymphoid potential-deficient; γ cells are multipotent but have limited self-renewal activity and are derived from β cells, suggesting they represent an intermediate stage of repopulating cells; δ are the most lymphoid-primed, lacking durable myeloid potential, and also have limited self-renewal capacity. Only β cells demonstrate traits typically associated with LT-HSCs (Dykstra et al. 2007). Other studies have also suggested the existence of lineage-primed HSCs alongside balanced HSCs based on repopulation kinetics, differential cytokine responses, and transcription factor expression patterns (Muller-Sieburg et al. 2002, 2004; Sieburg et al. 2006; Adolfsson et al. 2005; Arinobu et al. 2007). Overall, recent evidence suggests that more heterogeneity and lineage-priming exist in HSCs than previously recognised.

Furthermore, the functional and cell-fate identity of CMPs has been questioned by multiple groups (Adolfsson et al. 2005; Arinobu et al. 2007; Bendall et al. 2014; Görgens et al. 2013; J. H. Levine et al. 2015; Murre 2007; Pronk et al. 2007; Laurenti et al. 2008; Yamamoto et al. 2013). A recent

study questioned the existence of the CMP population by using massively parallel RNA-sequencing (MARS-seq) to measure gene expression in 2,730 myeloid progenitors. Paired with transplantation assays, the authors found that CMPs are not heterogeneous cells with undetermined fates, but are instead primed towards erythroid or myeloid fates (Paul et al. 2015). Perié *et al.* reached a similar conclusion with a genetic barcoding experiment, which showed that CMPs produce highly biased myeloid or erythroid outputs after 14 days (Perié et al. 2015). These studies suggest that cells defined as CMPs may actually be at an early commitment stage, and therefore question the usefulness of this classically defined population (Hamey et al. 2016).

While the haematopoietic hierarchy has been well-defined, evidence of heterogeneity in HSPCs, as well as questions raised about cell functions and identities, demonstrate gaps in our understanding of the differentiation pathways towards myeloid, erythroid, and lymphoid cell fates.

1.2. Cell fate decision making

Cell fate decisions are controlled by patterns of gene expression and are essential in maintaining the haematopoietic system and determining haematopoietic differentiation pathways. Gene expression occurs when a DNA sequence is transcribed into RNA, which, for protein-coding genes, is translated into a protein by a ribosome. The process is regulated at every step by extrinsic signalling and epigenetic factors; understanding how gene expression is regulated is critical for understanding how cell fate decisions occur in haematopoiesis.

1.2.1. Chromatin structure and epigenetics

The genomes of higher eukaryotes are primarily made up of non-coding regions, as well as protein-coding and RNA-coding genes, which only make up around one percent of the 3.3Gbp of sequence (Lander et al. 2001). The non-coding regions have important functions such as encoding regulatory DNA elements (section 1.2.2) and structural DNA elements, such as centromeres and telomeres (Alexander et al. 2010).

Due to its large size, DNA must be packaged efficiently within the nucleus of every cell, while still allowing access to DNA for gene expression and DNA replication. This is achieved by supercoiling the DNA around histones to form nucleosomes (Kornberg and Lorch 1992, 1999). Accessibility to DNA is modified by chromatin remodellers, which are ATP-dependent enzymes that move histones along the DNA helix. This produces two chromatin states: euchromatin, which is open

and transcriptionally active, and heterochromatin, which is dense and silent (Felsenfeld and Groudine 2003).

Epigenetic modifications are also able to influence transcription factor binding to DNA and regulation of gene expression by acting on both DNA and histones. DNA methylation at cytosine residues is a stable and heritable repressive mark, causing epigenetic silencing. It is essential for normal development and is involved in processes such as genomic imprinting and X-chromosome inactivation. Epigenetic post-transcriptional modifications of histones include acetylation, methylation, phosphorylation, ubiquitination and sumoylation of the histone tails (Jin, Li, and Robertson 2011). Specific proteins regulate histone modifications. These proteins include histone acetyl and methyl transferases, as well as histone deacetylases (Kouzarides 2007). Histone modifications regulate processes such as gene transcription, DNA replication and DNA repair (Cedar and Bergman 2009).

1.2.2. Gene regulatory elements

Gene regulatory elements (GREs) are categorised by their function and include promoters, enhancers, silencing elements, insulators, and locus control regions. GREs are generally *cis* acting, meaning they act on the same DNA strand (Maston, Evans, and Green 2006).

RNA polymerase binds the promoter region to initiate transcription. The promoter has well characterised features that facilitate pre-initiation complex assembly, including specific DNA sequences to which general transcription factors bind, leading to the recruitment of the RNA polymerase. The most common is the TATA box, which is approximately 30bp upstream of the transcriptional start site. Once bound by the TATA-binding protein, the DNA is partially unwound to facilitate transcription (Lee and Young 2000).

Enhancers are important for activating transcription above basal levels, resulting in tissue-specific gene expression. They are able to do so independent of location and orientation (Maniatis, Goodbourn, and Fischer 1987). Transcription factors bind to distal enhancers and cause the enhancer to loop to the promoter, whereas co-localised enhancer and promoter elements form “chromatin hubs” and integrate transcriptional regulation with other *cis*-regulatory elements (Bulger and Groudine 2011). The Mediator protein complex is important for coordinating promoter-enhancer interactions with the basal transcriptional machinery (Malik and Roeder 2010; Maston, Evans, and Green 2006).

Insulators block the transcriptional effect of neighbouring genes and prevent crosstalk between genomic regions. They are important for preventing inappropriate regulation of adjacent genes, as well as separating the genome into functional and non-functional transcriptional units (Bushey, Dorman, and Corces 2008; Joughin et al. 2010). Silencers also have a negative effect on transcription, and are sequence-specific elements that may exert their effect by directly interfering with the general transcription factor assembly, or by passively inhibiting other elements upstream of the target gene (Ogbourne and Antalis 1998; Maston, Evans, and Green 2006). Locus control regions enhance the expression of linked genes at distal chromatin regions in a tissue-specific and copy number-dependent manner. They are involved in regulating gene expression through chromatin domain-opening activity, as well as increasing transcriptional activity to physiological levels (Q. Li et al. 2002). Overall, different functions are executed by *cis*-acting GREs to contribute to the regulation of cell lineage-specific gene expression.

1.2.3. Transcription factors

Transcription factors are a key class of *trans*-acting transcriptional regulators that exert their effects by binding to DNA elements. Transcription factors rarely act independently, but rather form complexes with other transcription factors, chromatin modifiers and co-factors (Ravasi et al. 2010; Vaquerizas et al. 2009). An enhancer is usually bound by multiple transcription factors, which may have mutually exclusive or stabilising functions (M. Levine and Davidson 2005). Transcription factors may bind directly to specific DNA binding sites or indirectly through other transcription factors (Slattery et al. 2014). They often have cell type-specific expression patterns and play a key role in determining gene expression profiles and cell fates (Vaquerizas et al. 2009).

1.2.3.1. Transcription factors in haematopoiesis

A number of transcription factors are known to play key roles in haematopoiesis and HSC regulation. Indeed, examples of key transcription factors can be identified along the erythroid, myeloid, and lymphoid lineages. *Tal1* encodes the transcription factor SCL, which is key for the development of HSPCs (Bloor et al. 2002). Disruption of *Tal1* also causes defects in the erythroid and megakaryocytic lineages (Hall et al. 2003). Neutrophil differentiation is regulated by the transcription factor encoding gene *Gfi1*, as mice lacking its expression have defects in neutrophil production (Hock et al. 2003; Karsunky et al. 2002). *Ikzf1* encodes the transcription factor IKAROS, which is key in normal B-cell development; mouse models demonstrate that an *Ikaros*-null mutation results in foetal and adult defects in lymphopoiesis (J. H. Wang et al. 1996; Marke,

van Leeuwen, and Scheijen 2018). Therefore, given the key role transcription factors play in determining cell-fates, it is important to investigate their regulation to improve our understanding of gene expression dynamics during HSPC differentiation.

Specific transcription factors important for haematopoiesis will be discussed in more detail in Chapters 4-6.

1.2.4. Gene regulatory networks

Haematopoiesis is determined by the complex regulation of gene expression through gene regulatory networks, which are made up of transcription factors, epigenetic regulators, as well as GREs (Davidson 2009). Transcription factors can regulate their own expression as well as that of other transcription factors, generating feed-back and feed-forward loops and creating complex interconnected circuits (John E. Pimanda and Gottgens 2010; Davidson 2009; Miranda-Saavedra and Göttgens 2008; Davidson 2010). Gene regulatory networks describe highly interconnected relationships between transcription factors and the genes they regulate, in which the effect on gene expression caused by transcription factor binding determines the stability of gene expression (Alon 2007). The network therefore determines which transcription factors are expressed at any moment, defining cell identities. As such, investigating gene regulatory networks can be useful to understand how transcription factors interact to cause cell fate changes. Network reconstruction will be further discussed in Section 1.5.4.3.

1.3. Dysregulation of transcriptional control in haematopoiesis

The haematopoietic system is tightly regulated as any dysregulation in the gene regulatory network, such as transcription factor overexpression, deletion, or abnormal gene fusions, can result in imbalance and malignancy (Sive and Göttgens 2014). The disruption of transcription factor function can, for example, lead to leukaemia. The transcription factor *Runx1* is essential during definitive haematopoiesis in embryonic development, and its absence causes embryonic death in mouse models (Q. Wang et al. 1996; Okuda et al. 1996). In adult human haematopoiesis, however, disruption of RUNX1 results in predisposition to acute myeloid leukaemia (AML) (Sun and Downing 2004; Sakurai et al. 2014; M. Ichikawa et al. 2004). The transcription factor was in fact discovered by an observation that the gene was rearranged in leukaemic cells of t(8;21) AML patients to form the RUNX1-ETO fusion protein (Miyoshi et al. 1991). The transcription factor

SCL was also identified by its role in T-cell acute lymphoblastic leukaemia (ALL), in which it was ectopically expressed due to t(1;14) chromosomal translocations (Begley et al. 1989). *Lmo2* encodes a transcription factor that normally activates an erythroid-specific gene expression program (Warren et al. 1994). It was identified by a t(11;14) translocation in T-cell ALL (Boehm et al. 1991). Mouse models have shown that activation of LMO2 by the translocation causes self-renewal of committed T-cells, which accumulate additional mutations and eventually cause leukaemia (Herblot et al. 2000; Curtis and McCormack 2010).

Understanding the processes regulating HSPC differentiation during normal haematopoiesis is therefore key for uncovering how its dysregulation results in aberrant decision-making, which in turn lead to serious blood disorders (Tenen 2003).

1.4. Mammalian genome editing

Targeted genome editing was first achieved by homologous recombination, in which a DNA sequence is exchanged between two similar sequences; it is also a DNA repair mechanism which replaces damaged DNA (Scherer and Davis 1979; Smithies et al. 1985; X. Li and Heyer 2008). Traditional methods took advantage of endogenous homologous recombination to alter the genome, which was a precise but inefficient approach and was only possible in mice due to the absence of culturable embryonic stem cells in other mammals (D. Carroll 2017).

Advances in genome editing technologies have made it possible to perform genetic manipulations at higher efficiencies in all types of cells and organisms. These high-efficiency methods are programmable to cause a double-stranded break at the desired target, stimulating repair by homologous recombination or non-homologous end-joining (K. Lim 2015). These genome-editing technologies use engineered nucleases composed of a sequence-specific DNA-binding domain, which searches for the target locus, and a non-specific DNA cleavage molecule, which generates the double-stranded break (Gaj et al. 2013; K. Lim 2015). These methods include zinc-finger nucleases (ZFN), transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9. Briefly, ZFNs are a hybrid of a DNA cleavage domain derived from a bacterial protein and zinc fingers identified from sequence-specific eukaryotic transcription factors. TALENs use the same DNA cleavage domain, but its sequence-specific DNA-binding domain is derived from DNA recognition modules from transcription factors of plant pathogenic bacteria (D. Carroll 2017). CRISPR/Cas9, on the other hand, is a prokaryotic acquired immunity system and uses RNA molecules to target a genomic locus (Barrangou et al.

2007; Makarova et al. 2006). It has become the system of choice for genome editing, as it is a more efficacious, less expensive, and less cumbersome technology than ZFNs or TALENs. The CRISPR/Cas9 system was used in this thesis to perturb target genes in HSCs (Chapter 4).

1.4.1. The CRISPR/Cas9 system

The CRISPR/Cas9 system for genome editing was adapted from the acquired immunity mechanism of bacteria and archaea. The microbes capture viral DNA segments and integrate them between CRISPR sequences, which are transcribed into CRISPR RNAs (crRNA). The crRNA then guides the silencing of invading DNA by Cas nucleases. Cas9 in particular participates in crRNA biogenesis and destruction of invading DNA (Jinek et al. 2012; Sorek, Lawrence, and Wiedenheft 2013).

This system has recently been adapted for mammalian genome editing (Cong et al. 2013; Mali et al. 2013). A short guide RNA (gRNA) sequence is designed to bind a target sequence in the genome. The gRNA also binds the Cas9 nuclease and guides it to cut at the target location. This technology allows virtually any genomic location to be effectively targeted (P. D. Hsu, Lander, and Zhang 2014). The CRISPR/Cas9 technology also makes it possible to more easily perform genome-wide knockout screens. Individual cells can be perturbed by one or more gRNA from a pooled library targeting thousands of genes. Cells that have had essential genes knocked out will die, but the surviving cells can be sequenced to uncover the genes causing their perturbation phenotype (Shalem et al. 2014; T. Wang et al. 2014).

1.4.1.1. CRISPR/Cas9 use in investigations of the haematopoietic system

CRISPR/Cas9 has also been used to perform genome editing of HSPCs. For example, lentiviral transduction of gRNAs and Cas9 into a single HSC was used to modify five genes and produce myeloid malignancy, generating an AML mouse model (Heckl et al. 2014). In human HSPCs, transfection of Cas9 and gRNA encoding plasmids was used to target *B2M* and *CCR5* to investigate the possible therapeutic implications of the technology (Mandal et al. 2014). A CRISPR-Cas9 knockin mouse, which expresses Cas9 and can be transfected with target gRNAs, has also been effectively used for genomic research (Platt et al. 2014; Pettitt et al. 2009). Tzelepis *et al.* designed a genome-wide CRISPR screening platform to identify therapeutic targets in AML. Cas9-expressing mice were used to validate their murine lentiviral gRNA library. By transducing Cas9-expressing cancer cell lines with the gRNA library, they identified 492 AML-specific cell-essential

genes as well as many other clinically actionable candidates (Tzelepis et al. 2016). These studies demonstrate the value of using CRISPR/Cas9 genome editing to interrogate HSPC biology with potential therapeutic implications.

1.5. Single cell biology

Analysis of transcriptional regulation has traditionally used bulk cell populations, often due to the limitations of available technologies. However, these analyses can only provide information on cell population averages (Moignard and Göttgens 2014). While HSPCs regulate their functional output at the population level, differentiation is a stochastic process occurring in individual cells (A. M. Klein and Simons 2011; Simons and Clevers 2011). Bulk expression analysis assumes homogeneity within a population and may therefore fail to capture the heterogeneity of the fate decision-making processes of individual cells (Fig. 1.2A). Advances in single-cell profiling technology have now made it possible to profile large numbers of individual cells simultaneously, which has been widely explored in HSPCs to increasingly define the heterogeneity within different cell types (N. K. Wilson et al. 2015; Paul et al. 2015; Hamey et al. 2016).

1.5.1. Heterogeneity in HSPC populations

It is widely accepted that haematopoiesis occurs in a hierarchical fashion, in which HSCs differentiate into mature blood cells, passing through intermediate progenitor states with varying lineage potential (Bryder, Rossi, and Weissman 2006; Pronk et al. 2007; Laurenti and Göttgens 2018). Technological advances have made it possible to study the haematopoietic hierarchy in further detail, revealing heterogeneity within HSPC populations, including the HSC pool. Single-cell clonal output investigations have shown that heterogeneity exists within the HSC compartment in terms of their self-renewal and repopulation potential, as well as their differentiation output, highlighting the need to study individual cells rather than population averages (Dykstra et al. 2007). Furthermore, the MPP compartment has been subdivided into four functionally distinct subgroups (Pietras et al. 2015; Cabezas-Wallscheid et al. 2014).

Indeed, single-cell studies have also challenged the structure of the haematopoietic hierarchy. A study that isolated human cells from myeloid, erythroid, and megakaryocytic fates found that most of the multipotent cells in the bone marrow were in the HSC compartment and there was an absence of intermediate progenitor populations (Notta et al. 2015). Single-cell barcoding can be used to study functional properties of cells by tagging individual cells with unique barcodes that are

noncoding stretches of DNA (Naik, Schumacher, and Perié 2014). Cells are commonly labelled with these barcodes by retroviral transduction, allowing the cells to be tracked *in vivo*. Perié *et al.* used this technique to track the contribution of individual cells to different lineages after transplantation into lethally irradiated mice, sequencing the cells to reveal the barcode identity (Perié et al. 2015). They investigated whether the CLP-CMP divide is the first step of lineage commitment. As previously discussed, they found that the CMP population is composed of heterogeneous cells primed towards myeloid or erythroid outputs, suggesting they are at an early commitment stage. Furthermore, after transplanting barcoded HSCs and MPPs, they found that cell fates toward myeloid and erythroid lineages were mainly determined by the HSCs while most MPPs were restricted to a single fate (Perié et al. 2015). These studies together demonstrate the heterogeneity that exists within HSPCs, as well as the value single-cell analysis offers for resolving the behaviour of individual cells.

1.5.2. Fluorescence-Activated Cell Sorting

An essential first step in many single-cell profiling techniques is the isolation of individual cells. To do so, many studies take advantage of fluorescence-activated cell sorting (FACS), which quantitatively assesses fluorophores at single-cell resolution. Cells are stained with fluorophore-conjugated antibodies against cell surface markers and separated based on multiple parameters, including size, granularity and fluorescent properties correlated to surface marker expression (Lindström 2012). Flow cytometers can detect 30 or more different parameters per cell, limited by the number of distinguishable fluorophores and capabilities of the instrument and analysis software (Nettey, Giles, and Chattopadhyay 2018; Brummelman et al. 2017). The sorted cells can then be used in various applications, such as functional or gene expression analyses. However, the cells can usually only be used for one type of experiment and therefore represent a snapshot of the cell population at the point of collection. If collected at the single-cell level, the snapshots are likely to reveal heterogeneity among the isolated cells.

Index sorting is an important advancement in FACS, as it collects data for all the parameters measured, including well position for each single-cell sorted into a 96- or 384-well plate. By pairing FACS with index sorting, it is possible to obtain the FACS phenotype of every cell, to be retrospectively reviewed (Osborne 2011; Schulte et al. 2015). The index sorting data can then be paired with gene expression analysis to compare populations based on both gene and surface marker expression (Fig. 1.2B). This technique is widely applicable and can be used to characterise

many cellular systems (Schulte et al. 2015; Hayashi et al. 2010). In the haematopoiesis field, Wilson *et al.* paired single-cell gene expression profiling with functional analyses to interrogate heterogeneity within HSC isolation strategies (N. K. Wilson et al. 2015). They defined a refined sorting strategy that separated HSCs from non-HSCs by retrospectively linking gene expression profiles and functional outcome to the index sorting data. Single-cell transplantation and RNA-sequencing of the enriched population identified genes associated with long-term self-renewal. The study highlighted a key difference in the enrichment for cell cycle genes between functional HSCs and non-HSCs, where HSCs expressed higher levels of genes associated with negative regulation of cell proliferation and non-HSCs were primed towards proliferation (N. K. Wilson et al. 2015).

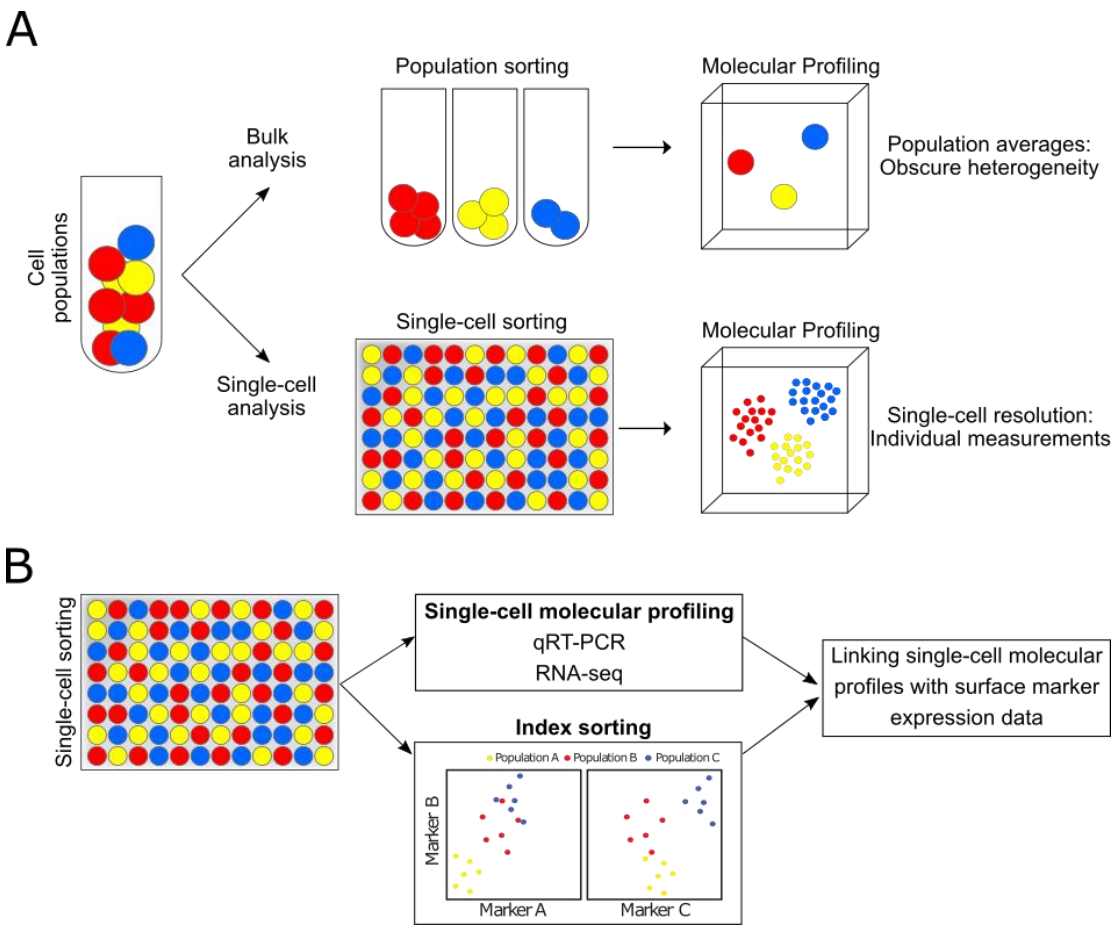


Figure 1.2. Single-cell analysis. (A) Schematic demonstrating the importance of single cell analysis. Profiling a population using bulk analysis will only reveal population averages and therefore will obscure the heterogeneity that exists in individual cells. Single-cell technology makes it possible to collect measurements for individual cells, revealing the variance that exists in heterogeneous populations. (B) Sorting cells using fluorescence-activated cell sorting (FACS) makes it possible to isolate individual cells to be analysed using single-cell molecular profiling techniques. The technique used will vary based on the experimenter's interests and include qRT-PCR or scRNA-seq. FACS paired with index sorting collects data for all parameters measured, including surface marker expression and well position for each cell. The gene expression profiles can therefore be compared to the surface marker expression data to characterise populations and inform future work. qRT-PCR: quantitative real-time polymerase chain reaction; RNA-seq: RNA-sequencing. Figure adapted from Hamey *et al.* (2016).

Another technique for studying single cells is mass cytometry (Cytometry by Time of Flight/CyTOF). While FACS uses fluorochromes to label antibodies, mass cytometry instead requires the antibodies to be labelled with transition element isotopes, which are then quantified by the concentrations of metal-tagged antibodies (Bendall et al. 2012; Behbehani et al. 2012). Mass cytometry can measure 40 or more parameters simultaneously, allowing for in-depth investigations of cell phenotypes, limited only by the choice of antibodies. It can be a useful method for studying signalling cells in a variety of experimental conditions (Behbehani et al. 2012). However, the cells are not available for further molecular and functional analyses after mass cytometry; instead, the results may provide insights for designing FACS strategies for further investigations (Behbehani et al. 2012).

1.5.3. Single-cell gene expression profiling

The heterogeneity of cell populations also affects the analysis of regulatory relationships. Using population expression data to extrapolate to the single-cell level forces assumptions about how individual cells behave. It is necessary to know whether genes are expressed in the same cell to determine whether they are co-regulated or regulate each other; these relationships will be masked at the population level, which suggests all cells express the same level of each gene. Therefore, single-cell gene expression analysis is a useful tool to reveal the complex relationships that exist within a gene regulatory network.

Single-cell technologies are rapidly advancing and expanding, with new technologies frequently being introduced. To study gene expression at the single-cell level, techniques such as quantitative real-time polymerase chain reaction (qRT-PCR) and single-cell RNA-sequencing (scRNA-seq) can be applied. qRT-PCR amplifies specific messenger RNA (mRNA) transcripts to measure the expression of selected genes; when performed on single cells, expression of multiple genes can be measured at the single-cell level (Sanchez-Freire et al. 2012). Fluidigm BioMark™ is a dynamic array integrated microfluidics circuit which allows gene expression studies of up to 96 selected genes in 96 cells. This can be further extended by using multiplexing approaches (G. Guo et al. 2013). The genes of interest are selected by the investigator, meaning the technique is best used for looking at specific questions, targets or systems (Moignard et al. 2013). Single-cell qRT-PCR has been widely used to interrogate heterogeneity and transcriptional regulation in HSPCs (G. Guo et al. 2013; N. K. Wilson et al. 2015; Moignard et al. 2013, 2015; Pina et al. 2012, 2015).

In contrast to qRT-PCR, which is limited to a small, curated set of genes, scRNA-seq is a transcriptome-wide approach for measuring gene expression. The mRNA from single cells is reverse-transcribed and the resulting complementary DNA (cDNA) is amplified and sequenced as a pool, with unique combinations of indexes marking each cell (Picelli et al. 2014; Kolodziejczyk et al. 2015). This method can be used to profile gene expression in individual cells within a population of interest, providing insights into differentiation lineages and the regulatory programmes governing these populations (Nestorowa et al. 2016). Similarly, MARS-seq is an automated method of RNA sequencing, designed to process thousands of multiplexed cells that are barcoded at the molecular, cellular and plate level (Jaitin et al. 2014). This approach allows the processing of thousands of cells as well as the characterisation of multiple cell populations within a single dataset for an in-depth and broad picture of variability and heterogeneity; however, compared to scRNA-seq, it results in much shallower sequencing depth per cell. New technologies are constantly emerging in the single-cell field, aiming to increase throughput while keeping costs low; the technologies relevant to this thesis are scRNA-seq and qRT-PCR, which will be explored further in Chapters 3 and 5, respectively.

1.5.4. Computational analysis of single-cell data

Single-cell expression profiling can produce tens to thousands of gene expression measurements per cell, in which each cell represents part of a heterogeneous population. These datasets are complicated to interpret due to the high-dimensionality nature of the data. Single-cell expression profiling therefore needs to be paired with computational methods to resolve the complex datasets and enable investigations of the underlying biology.

1.5.4.1. Dimensionality reduction

Dimensionality reduction methods make it possible to visualise high-dimensional data in a low-dimensional space, most frequently in two or three dimensions. Doing so enables comparisons of gene expression between groups of cells. Dimensionality reduction methods have been widely applied to population expression data to uncover differences between cells (Fig. 1.3A) (Hamey et al. 2016).

Principal component analysis (PCA) is a widely used linear dimensionality reduction method (Hotelling 1933a, 1933b; Pearson 1901; Jolliffe 2011). PCA applies a linear transformation to the data so that each principal component (PC) explains the maximum variance within the data. They

are ordered so that PC1 has the largest variance, followed by PC2, and so on; as such, plotting the data in the first two or three PCs can reveal separation between different cell states. PCA has been applied to single-cell analysis of HSPCs, for example to investigate aging and differentiation in HSCs or to interrogate the regulation of erythroid and myeloid fates (Kowalczyk et al. 2015; Pina et al. 2015).

However, linear techniques may struggle with capturing complex structures and may not always be the most suitable visualisation for single-cell expression profiles. Recently, non-linear methods such as t-distributed stochastic embedding (t-SNE) and diffusion maps have been applied to single-cell data to uncover more complex relationships in the data (Maaten and Hinton 2008; Haghverdi, Buettner, and Theis 2015). t-SNE embeds the data in a low-dimensional space while conserving the distribution of distances in the high-dimensionality space, meaning that cells with similar expression profiles are nearby on the reduced dimensionality plot. This method has been used on single-cell datasets to interrogate diverse aims, such as visualising the overlap between haematopoietic populations or investigating the molecular mechanisms underlying development in the mouse embryo (N. K. Wilson et al. 2015; Scialdone et al. 2016). These studies demonstrate the usefulness of t-SNE for representing heterogeneous single-cell datasets.

Single-cell expression profiling can be used to capture molecular changes throughout differentiation. These datasets need to be visualised with an appropriate dimensionality reduction method, which can capture the continuous nature of differentiation. Diffusion maps consider lengths of diffusion-like random walks through the data in the high-dimensional space to determine a projection of the cells. This method has been adapted for use with single-cell expression data (Haghverdi, Buettner, and Theis 2015). Moignard *et al.* used the diffusion map method on single-cell qRT-PCR data to visualise cell progression during early blood development, in which diffusion maps successfully separated populations from different time points and illustrated a progression through differentiation (Moignard et al. 2015).

In addition to these methods, researchers have developed web interfaces that are accessible to non-bioinformaticians. The advantage of these methods is the same as their disadvantage—the interfaces are quick and easy to use, producing a plot of a complex dataset within minutes, but the investigator has little to no input on how these plots are constructed. Still, these methods can be highly informative and do not require any computational skill to analyse the dataset. Two such methods are STREAM (Single-cell Trajectories Reconstruction, Examination and Mapping) and

SPRING (H. Chen et al. 2018; Weinreb, Wolock, and Klein 2018). STREAM was developed to reconstruct differentiation trajectories and capture gene expression changes during differentiation using pseudotime ordering, which orders cells in a putative differentiation trajectory based on gene expression (Section 1.5.4.2). STREAM uses a non-linear dimensionality method called modified local linear embedding (MLLE), which uses multiple weight vectors for each point while embedding the low-dimensional data (Z. Zhang and Wang 2006). STREAM is unique in that the visualisation includes density information throughout pseudotime to show how many cells of each investigated population are in which stage of differentiation. The online web-interface can also be used to identify genes important in defining branching points, as well as genes that transition across a given lineage branch (H. Chen et al. 2018).

SPRING, on the other hand, uses a force-directed layout of k-nearest neighbour graphs to capture long-distance relationships between cells (Weinreb, Wolock, and Klein 2018). Cells are connected based on similarities in their expression profiles, and the connections form the edges of the graph. Edges are weighted by the strength of similarities between cells, and the graph is then generated by bringing similar cells together: the edges cause an attracting force between similar cells or a repelling force between cells that differ. Force-directed graphs have previously been used to visualise haematopoietic data collected by mass cytometry. The Klein lab have recently created an interactive web-interface for researchers to explore their scRNA-seq expression data (Spitzer et al. 2015; J. H. Levine et al. 2015; Weinreb, Wolock, and Klein 2018).

Overall, a wide range of dimensionality-reduction and visualisation methods are available; the best choice for a specific project will ultimately depend on the dataset and question at hand. These methods are further explored in Chapter 3.

Dimensionality reduction makes it possible to not only visualise heterogeneity within a dataset, but to also group cells and query differences between populations. Clustering methods separate cells into groups in an unsupervised and unbiased way, based on information such as gene expression profiles (Fig 1.3B). Gene expression specific to individual populations can then be used to identify cell types or novel marker genes (Hamey et al. 2016). Hierarchical clustering is a well-established method that has been used extensively to identify subgroups of cells within single-cell data, and therefore connects or separates cells in a heterogeneous dataset based on their gene expression (N. K. Wilson et al. 2015; Moignard et al. 2013; G. Guo et al. 2013).

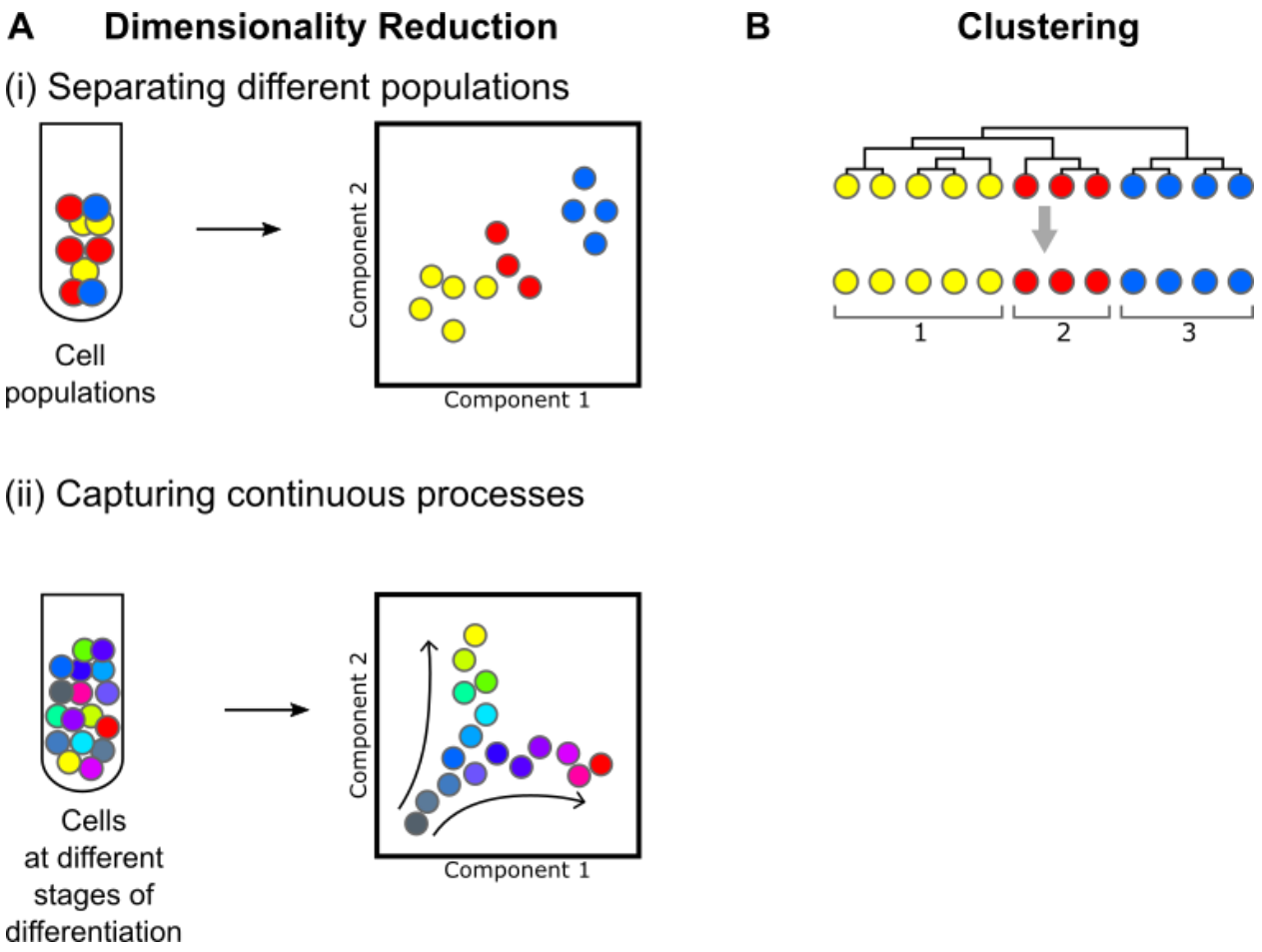


Figure 1.3. Single-cell profiling enables the exploration of cell population heterogeneities. (A) Dimensionality reduction techniques allow visualisation of heterogeneities within a population of cells based on single-cell expression profiles. Plotting cells in this two-dimensional coordinate system can confirm that subpopulations separate based on their expression profiles (i) or can be used to visualise data with a continuous nature, for example describing differentiation processes (ii). (C) Unbiased clustering techniques can be used to explore similarities between cells in single-cell datasets. In hierarchical clustering, as shown here, the most similar groups of cells are closely connected in the dendrogram. This structure then allows exploration of different levels of clustering within the data: for example, the cells can be split into three groups corresponding to their cell type. Figure adapted from Hamey *et al.* (2016).

1.5.4.2. Reconstructing lineage differentiation

During haematopoiesis, a cell increasingly specialises as it commits to one of several cell fates. Isolating populations at different stages of differentiation and profiling them at the single-cell level describes cell differentiation but is limited by time resolution and therefore assumes the cells are synchronised. However, as discussed in section 1.5.1, single-cell analysis reveals the large variation that exists in populations previously thought to be homogeneous in bulk studies (N. K. Wilson et al. 2015; Moignard et al. 2013, 2015; Pina et al. 2015). To resolve the heterogeneity

within data, *in silico* lineage reconstruction uses computational methods to infer lineage differentiation based on single-cell data.

Individual cells undergoing differentiation can be clustered into groups based on single-cell expression profiling data. The most similar groups can be connected into a structure representing a lineage hierarchy (Fig. 1.4A) (Hamey et al. 2016). The spanning-tree progression analysis of density-normalised events (SPADE) algorithm uses this approach to construct lineage hierarchies from flow and mass cytometry data (P. Qiu et al. 2011). SPADE calculates a density-dependent sample of the data to ensure that rare populations are not obscured, after which the cells are clustered based on their expression profiles. The clusters are then linked in a tree structure representing the lineage hierarchy. An advantage of SPADE is that it includes rare cell populations and does not require prior information to infer the lineage structure, but the different random density-dependent samples can lead to different clusters and therefore alternative tree structures, limiting the stability of this approach (Hamey et al. 2016). SPADE was used by Guo *et al.* to construct a lineage tree of the haematopoietic hierarchy (G. Guo et al. 2010). The authors questioned whether commitment occurs at the CMP stage using single-cell qRT-PCR data quantifying 280 commonly used surface markers. The lineage tree constructed by SPADE showed CMPs were found in both megakaryocyte-erythrocyte (MegE) and lympho-myeloid lineages, and that MegE cells were closely connected to LT-HSCs, indicating a very early lineage bias (G. Guo et al. 2013).

Scaffold is another computational method used to reconstruct differentiation hierarchies. The approach involves an initial clustering step, from which a force-directed graph is constructed. Spitzer *et al.* used Scaffold to visualise the hierarchy of the murine immune system based on single-cell mass cytometry data (Spitzer et al. 2015). The authors constructed Scaffold maps for cells from different samples to compare the immune system organisation in different tissues, species, and genetic backgrounds, and demonstrated that circadian rhythm influenced the distribution of immune cells. Scaffold also allows new data to be projected onto an existing structure, making it possible to integrate multiple datasets from various tissues and disease states (Spitzer et al. 2015). These studies demonstrate how constructing differentiation hierarchies *in silico* can provide insights into biological systems.

Ordering individual cell profiles by progress through differentiation is an exciting extension of inferring differentiation hierarchies. Assuming that gene and protein expression continuously

change as cells differentiate, and that a sample contains cells spread at a sufficient density through differentiation, it was hypothesised that single-cell expression profiles could be used to arrange cells in 'pseudotime', where the position of a cell in pseudotime corresponds to its progress through differentiation (Fig. 1.4B) (Hamey et al. 2016). Algorithms have been designed based on these assumptions to solve this computational ordering problem. Trapnell *et al.* describe Monocle, which performs a dimensionality reduction in the data before constructing a graph on this lower dimensional representation and finding the minimum spanning tree (Trapnell et al. 2014). Cells are then ordered in pseudotime based on their position in the minimum spanning tree, making it possible to investigate changes in gene expression patterns throughout pseudotime. The authors used this approach to reconstruct the differentiation of human primary myoblasts and identify branching towards an alternative cell fate present in their data (Trapnell et al. 2014).

Another algorithm, Wanderlust, was applied to single-cell mass cytometry data to capture B-cell development in human bone marrow (Bendall et al. 2014). Wanderlust first considers a k-nearest-neighbour graph on the single-cell expression data. The ordering of cells is based on the path lengths originating from a user-defined starting cell. Wanderlust can cope with very large numbers of cells, and subsamples cells to obtain stable orderings, avoiding the possibility of 'short circuits' through the data. Using Wanderlust, Bendall *et al.* ordered cells encompassing B-cell development with the aim of inferring a developmental trajectory, and confirmed that all the landmarks of B-cell lymphopoiesis were correctly ordered (Bendall et al. 2014).

Both Monocle and Wanderlust have limitations, however. Monocle is unstable and susceptible to short circuits, meaning that multiple applications of Monocle on the same dataset may result in multiple different trajectory inferences. Furthermore, Wanderlust cannot be used to identify branches when a differentiation trajectory separates towards multiple lineages. Updates to both these methods have been released. Monocle was upgraded to Monocle 2, which uses reverse graph embedding instead of minimum spanning tree to reconstruct lineage trajectories. This method doesn't require the user to predetermine the number of branches, allowing for more unbiased investigation of lineage potential (Mao et al. 2017; X. Qiu et al. 2017). Wishbone improved upon Wanderlust to avoid problems that could arise due to outliers in the data, and made it possible to identify cells that are differentiating towards an alternative cell fate (Setty et al. 2016).

Furthermore, alternative algorithms have been developed for constructing pseudotime orderings. Haghverdi *et al.* developed diffusion pseudotime (DPT) to provide a more robust method that can

identify branching towards multiple fates (Haghverdi et al. 2016). DPT measures the distance between cells based on lengths of random walks through single-cell expression data. It then considers the relationships between orderings from the starting point and end point of the main trajectory to identify branching points. The authors used DPT to reconstruct gene expression changes through differentiation in murine developmental haematopoiesis using previously published single-cell qRT-PCR data (Haghverdi et al. 2016; Moignard et al. 2015). These methods therefore make it possible to investigate continuous differentiation processes within single-cell snapshot data and identify lineage trajectories to improve our understanding of differentiation hierarchies within biological systems.

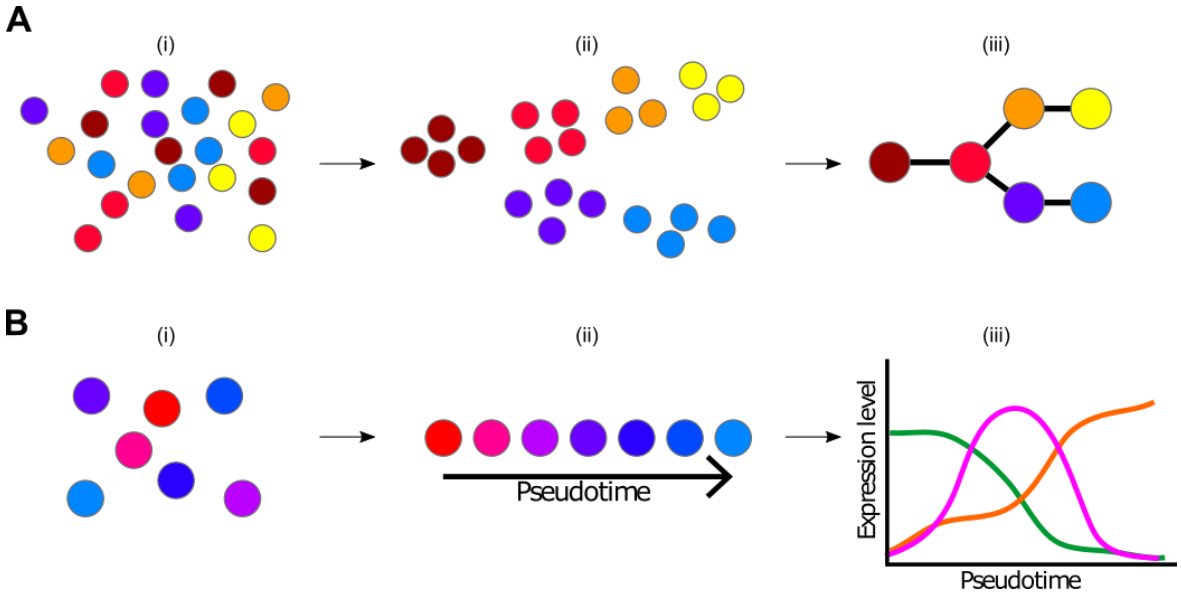


Figure 1.4. Reconstructing lineage differentiation using single-cell expression profiles. Methods have been developed to reconstruct lineage differentiation from single-cell measurements under the assumption that the cells closest in the differentiation process will have the most similar gene or protein expression profiles. (A) A cell population can contain several subpopulations (represented by different colours) from different stages of lineage differentiation (i). Individual cells can be clustered into groups based on gene or protein expression profiles (ii). By assigning similarity scores between groups, a graph can be constructed where each node corresponds to a cell cluster and the edges between nodes are weighted by similarity scores between clusters (iii). This graph can then be used to find a reconstruction of the lineage tree. (B) A seemingly homogenous population can contain cells at multiple stages of differentiation, depicted on a red-to-blue spectrum (i). These cells can be computationally ordered in pseudotime based on similarities in their expression profiles, representing their progress through differentiation (ii). Gene or protein expression patterns can then be explored along pseudotime to identify key biological events or factors linked to the differentiation process (iii). Figure adapted from Hamey *et al.* (2016).

1.5.4.3. Identifying and modelling regulatory relationships

It is important to define the underlying regulatory programmes governing haematopoietic differentiation to better understand how multipotent cells choose between different fates (Gottgens 2015; Peter and Davidson 2015). Transcriptional regulatory networks are composed of

transcription factors and the *cis*-regulatory elements to which they bind (Section 1.2). Network reconstruction directly from experimental evidence has been limited to the simplest organisms due to the number of possible regulations and complex network structures in more complex organisms. Many studies have instead focused on inferring regulatory networks using gene expression data collected from multiple experimental perturbations or conditions. Small sample sizes and masked heterogeneity within cell types constrain network inference from population expression data; single-cell approaches represent a powerful alternative for identifying new regulatory relationships, as each cell represents an observation with its own expression levels, vastly increasing sample sizes (Hamey et al. 2016).

The large sample sizes obtained by measuring single-cell gene expression can be used to identify potential regulatory relationships by considering correlations between genes (Fig. 1.5A). Setting a threshold on correlation strength can identify putative networks consisting of genes linked with the highest positive or negative correlations. Several studies have used this approach to identify and experimentally validate regulatory relationships between highly correlated genes (Moignard et al. 2013; Pina et al. 2015).

Decision-making in cells is governed by complex networks of transcriptional factors with possible combinatorial interactions between network elements. A regulatory relationship between two genes cannot therefore be considered in isolation, as it may depend on the presence or absence of additional transcription factors. Logical relations can be abstracted as Boolean functions, where gene expression is either “on” or “off” (Fig. 1.5B). These abstractions form part of a Boolean network, which make it possible to model and simulate regulatory networks. Single-cell gene expression data can be converted to binary data for each cell, providing a large number of possible Boolean states (Fig. 1.5C). Boolean network modelling has been used to computationally infer networks from single-cell expression data describing embryonic blood development and embryonic stem cells (Moignard et al. 2015; Xu et al. 2014). A drawback of Boolean modelling, however, is that the abstraction of gene expression levels to binary on/off states may discount quantitative expression differences. A recent study addressed this problem, making it possible to consider more than two expression levels for genes using Boolean modelling (Collombet et al. 2017).

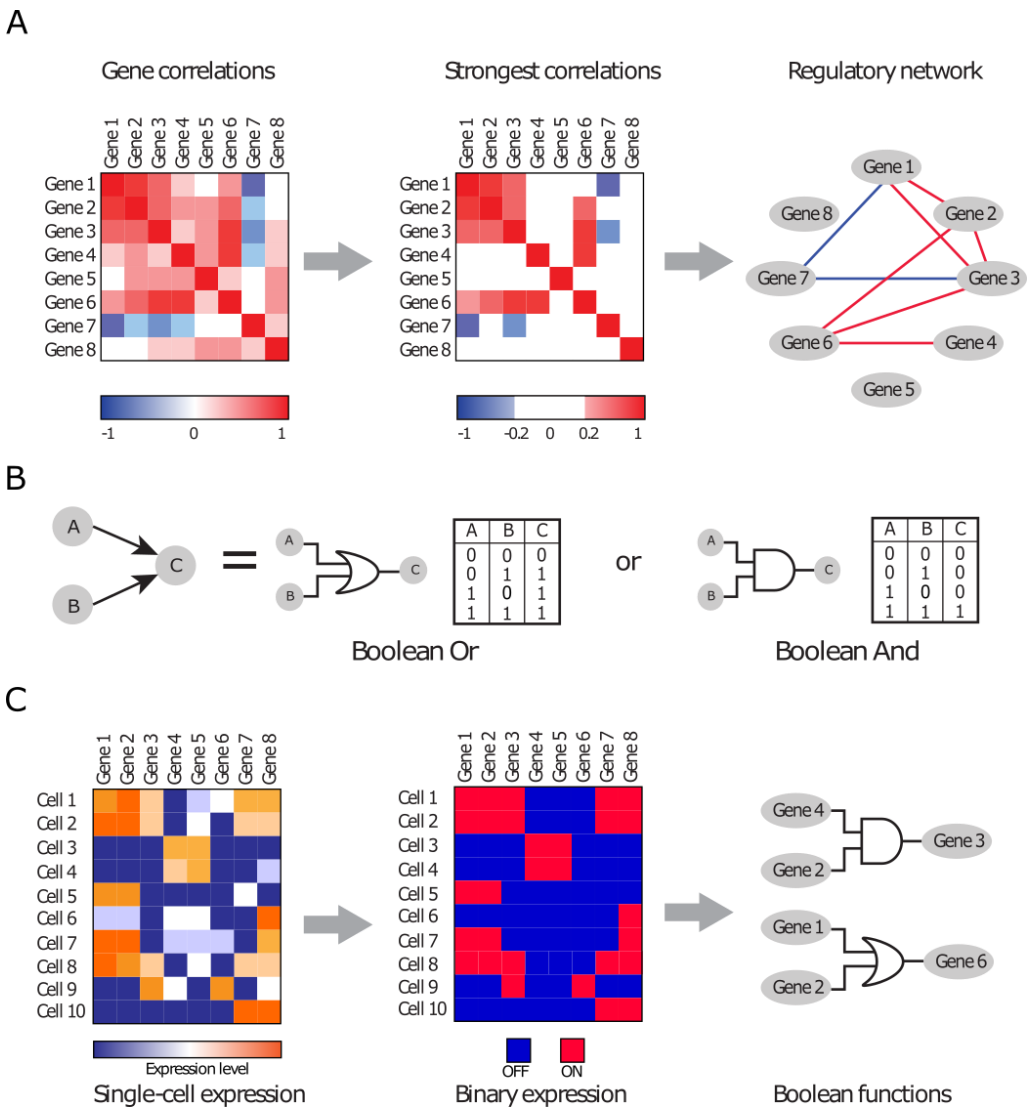


Figure 1.5. Inferring regulatory relationships from single-cell expression data. (A) Correlation between gene pairs can be calculated from single-cell gene expression measurements. Some gene pairs will exhibit positive correlation (red), whereas others will be negatively correlated (blue) (left gene-gene correlation heatmap). The strongly correlated gene pairs can be selected by setting thresholds; here, correlations below the threshold are in white (middle gene-gene strongest correlation heatmap). Connections between the highly correlated pairs can be drawn in a network diagram representing positive or negative correlations with red or blue lines, respectively (right). (B) Transcription factors can be part of combinatorial regulatory relationships. Activation of factor C by both factor A and B could be described by two different scenarios, represented by Boolean logic functions: either A or B alone activate C (Boolean Or), or both A and B need to bind to activate C (Boolean And). The truth tables detail the output of the And/Or functions. (C) Boolean functions can be used to model regulatory networks. Single-cell gene expression measurements (left) are converted into binary (ON/OFF) expression (middle) by setting an expression threshold. Computational methods applied to the binary data allow inference of regulatory relationships, represented by Boolean And/Or functions (right). Image adapted from Hamey *et al.* (2016).

1.5.4.4. HSPC regulatory networks

Single-cell data have been very useful to identify previously unrecognised regulatory networks and important factors in lineage commitment. Pina *et al.* used single-cell qRT-PCR to investigate self-

renewing cells and erythroid- or myeloid-committed progenitors in mice (Pina et al. 2015). Lineage commitment was associated with negative gene regulatory relationships and *Ddit3* was identified as a previously unrecognised key player in lineage commitment. *Ddit3* was positively associated with *Gata2* in self-renewing and committed cells, and negatively associated with *Cebpa*, important for neutrophil commitment. Knockdown of *Ddit3* caused a loss of erythroid function and a switch to myelo-monocytic potential; conversely, enforcing its expression in GMPs increased self-renewal properties and reduced myeloid potential. Analysing wild-type and *Ddit3*-overexpressing GMPs by PCA confirmed that *Ddit3* positively regulates erythroid fates while negatively regulating myeloid fates. The authors found that overexpressing *Ddit3* altered the global transcription network of GMPs, increasing connectivity with *Gata2* and preventing myeloid fate by stabilising primitive megakaryocyte-erythrocyte precursors. Regulatory network modelling therefore demonstrated that conflicting lineage-potential programmes exist at the point of cell commitment and identified a regulatory relationship between *Gata2* and *Ddit3* (Hamey et al. 2016; Pina et al. 2015).

To analyse regulatory relationships between transcription factors during haematopoiesis, Moignard *et al.* used single-cell qRT-PCR to profile the expression of 18 key transcription factors in LT-HSCs, PreMegEs, GMPs and CMPs (Moignard et al. 2013). Correlation analysis showed a reduction in negative correlations between transcription factors for individual populations compared to all cell populations as a whole, suggesting a lack of repression may be important in cell fate transitions. Two new regulatory links were identified: *Gata2*-*Gfi1b* and *Gata2*-*Gfi1*, highlighting a previously unrecognised regulatory triad where mutual inhibition between *Gfi1b* and *Gfi1* is regulated by *Gata2* (Hamey et al. 2016; Moignard et al. 2013).

These studies highlight the utility of single-cell network interrogation in finding regulatory networks obscured in bulk population studies and increasing our understanding of cell-fate decision making during haematopoiesis.

1.6. Aims

This thesis explores the transcriptional landscape of haematopoietic stem and progenitor cell differentiation. The "transcriptional landscape" describes gene expression during haematopoiesis, in which cells differentiate from an HSC to mature cells, moving through intermediate progenitor states (John E. Pimanda and Gottgens 2010; Waddington 1957). To understand how cell fate decisions are regulated and lead towards the different blood cell types, it is important to study gene expression regulation in the context of HSPC heterogeneity. Advances in single-cell technology make it possible to isolate and profile thousands of cells to study relationships between genes through differentiation. Furthermore, advances in genome editing and CRISPR/Cas9 technology make it possible to easily perturb target genes and study the implications on HSPC biology. In the work presented in this thesis, experimental and computational approaches are used to investigate haematopoiesis and address the following aims:

- Profile the transcriptional landscape of haematopoietic stem and progenitor cells at the single-cell level
- Investigate genes implicated in HSC biology using the CRISPR/Cas9 system
- Identify and validate regulatory networks controlling differentiation in haematopoietic stem and progenitor cells

Chapter 2: Materials & Methods

2.1. Cell culture

2.1.1. Mammalian cell lines

2.1.1.1. 416B

$CD34^+$ mouse haematopoietic suspension cell line. This cell line was used for luciferase assays (Section 2.2.3). 416B cells were grown in Roswell Park Memorial Institute Medium (RPMI 1640, Sigma R8758) with 10% foetal calf serum (FCS, Sigma), 1% penicillin-streptomycin (P/S, Sigma, P0781-100ml) and 1% L-Glutamine (Sigma) at 37°C and 5% CO_2. Cells were passaged every 2-3 days to maintain cells between $2x10^5$ cells/ml and $1x10^6$ cells/ml. The cell count should never exceed $1x10^6$ cells/ml.

2.1.1.2. 293T

Human cell line derived from the HEK 293 cell line containing the SV40 T antigen. This cell line was used to produce recombinant retroviruses. 293T cells were grown in Dulbecco's Modified Eagle's Medium (DMEM D6429, Sigma) with 10% FCS, 1% P/S, and 1% L-Glutamine at 37°C and 5% CO_2. Cells were passaged every 2-3 days to maintain 30-80% confluency.

2.1.1.3. HoxB8-Cas9

Mouse cell line generated from the HoxB8-FL cell line (Redecke et al. 2013) and contains the Cas9 protein. This cell line was used for validating guide RNAs for CRISPR screening (Section 2.2.2). HoxB8-Cas9 cells were grown in RPMI 1640, 10% Hyclone foetal bovine serum (FBS, Gibco), 5% Flt3 conditioned media (produced in house; Section 2.1.1.3.1), 1% P/S, 1% L-Glutamine, 50 µM 2-mercaptoethanol (Invitrogen) and 1 µM oestradiol (Sigma). The cells were maintained at 37°C and 5% CO_2. Cells were passaged every 3-4 days to maintain cells between $1x10^5$ cells/ml and $1.5x10^6$ cells/ml.

2.1.1.3.1. Flt3 conditioned media

This medium was used to supplement the HoxB8 complete medium. FLT3 medium was made from B16-FL cells. B16-FL cells were cultured in RPMI 1640, 10% Hyclone FBS, 50 µM 2-mercaptoethanol, 1% P/S and 1% L-Glutamine. Once the cells were confluent, the medium was

harvested and replaced daily for 3 days. After 3 days, the supernatant was pooled, centrifuged, filtered through a 0.2 µM filter and stored at -20°C.

2.1.2. HSC maintenance

Primary HSCs were isolated from wild-type (C57BL/6) and Cas9 transgenic mice. The cells were grown in StemSpan SFEM (StemCellTechnologies), 1% P/S, 1% L-Glutamine, 100 µM 2-mercaptoethanol, 20 ng/ml IL-11 (R&D Systems), 300 ng/ml SCF (R&D Systems) and 10% FBS. FBS was batch-tested by the Kent lab. Cells were maintained in flat- or U-bottom 96-well plates at 37°C and 5% CO_2. Cell medium was changed every 4 days and cells were only passaged if they were kept in culture for more than 7 days.

2.1.2.1. Cas9 transgenic mice

Dr George Vassiliou generously gifted our lab with the Cas9 transgenic mouse line used in this thesis (Tzelepis et al. 2016). The mouse line was kept in a C57BL/6 background and was created by inserting a human *EF1a* promoter-driven Cas9 expression cassette into the *Rosa26* locus in the JM8 mouse embryonic stem cell line (Pettitt et al. 2009). CBS maintained the mouse stock.

2.2. Cell Biology

2.2.1. Flow cytometry

BD Falcon 5 ml polypropylene tubes were used for all sorting procedures on the BD Influx4, Influx5 and BD FACSMelody (BD Biosciences). BD Falcon 5ml polystyrene tubes were used for all flow cytometry analysis on the BD LSRFortessa™ (BD Biosciences) unless otherwise stated.

2.2.1.1. Fluorescence Activated Cell Sorting (FACS): bone marrow HSPCs

The vertebrae, sternum, femurs, tibiae, and iliac crests were collected into 15 ml FACS buffer (2% heat-inactivated FBS in phosphate-buffered saline (PBS) without calcium and magnesium) from mice (Table 2.1). All procedures were performed in compliance with the United Kingdom Home Office regulations for animal work.

All centrifugation steps were performed at 300 g for 5 minutes.

The bones were crushed with a pestle and mortar and bone marrow tissue was harvested and filtered through a 70 µm cell strainer (BD Falcon). Cells were centrifuged and resuspended in 3 ml FACS buffer. 5 ml of ammonium chloride (StemCellTechnologies) was added for red cell lysis. The cells were incubated on ice for 10 minutes after which 7 ml of FACS buffer was added to neutralize the ammonium chloride, and the cells were centrifuged again. The cells were resuspended in 500 µl FACS buffer in a 5 ml polystyrene tube and lineage depleted using the EasySep™ Mouse Haematopoietic Progenitor Cell Enrichment Kit (StemCellTechnologies, used 2015-2017) or the EasySep™ Mouse Haematopoietic Progenitor Cell Isolation Kit (StemCellTechnologies, used 2017 onwards). The protocol for both kits was altered from the manufacturer's protocol.

EasySep™ Mouse Haematopoietic Progenitor Cell Enrichment Kit modified protocol: 5 µl/ml of EasySep™ Mouse Haematopoeitic Progenitor Cell Isolation Cocktail was added to the resuspended cells and the cells were incubated on ice for 15 minutes. 80 µl/ml of EasySep™ Biotin Selection Cocktail was then added, and the cells were incubated on ice for another 15 minutes. The EasySep™ Mouse Progenitor Magnetic Microparticles were vortexed for 30 seconds and 50 µl/ml were added to the cells. The cells were incubated on ice for a further 10 minutes, after which the tube was topped up with FACS buffer up to 2.5 ml.

EasySep™ Mouse Haematopoietic Progenitor Cell Isolation Kit modified protocol: 1:100 of EasySep™ Mouse Hematopoietic Progenitor Isolation Cocktail was added to the resuspended cells and the cells were incubated on ice for 15 minutes. The EasySep™ Streptavidin RapidSpheres™ were vortexed for 30 seconds and added to the cells in a 1:25 ratio. The cells were incubated on ice for a further 10 minutes, after which the tube was topped up with FACS buffer up to 2.5 ml.

The cells were run through the EasySep™ Magnet (StemCellTechnologies) twice, each time for 3 minutes. After lineage depletion, cells were stained for 30 minutes with the appropriate panel of antibodies (Section 2.2.1.4). Cells were then centrifuged and resuspended in 500 µl FACS buffer or stained with a secondary antibody when necessary. Prior to FACS, 7-amino-actinomycin (7-AAD, ThermoFisher) or 4',6-Diamidino-2-Phenylindole (DAPI, Invitrogen) was added to each sample as a viability dye. The BD Influx4 and Influx5 were used for all cell sorting. Cell sorting was performed by the CIMR Flow Cytometry Core.

Single-stain controls and an all-stain control were made using non-lineage depleted cells and were used as gate-setting controls. The all-stain control was made by staining mouse bone marrow cells before lineage depletion and was made for every mouse used.

Table 2.1. Mouse strain, age, sex, and number used for FACS.

Chapter	Experiment	Strain	Age (weeks)	Sex (F/M)	Number
5	single-cell qRT-PCR	C57BL/6	8-12	F	3
3	single-cell RNA-seq	C57BL/6	8-12	F	5
4	CRISPR screen	Cas9	19-22	F/M	3-8

For single-cell gene expression analysis, single cells were sorted into 96-well plates into either the mixture required for Fluidigm BioMarkTM HD (Section 2.4.1) or for Smart-Seq2 (Section 2.4.2) (Picelli et al. 2014). For CRISPR screening, bulk HSCs (250 cells) were sorted into 96-well U-bottom plates into HSC base medium (without cytokines).

2.2.1.2. HoxB8-Cas9 retroviral transduction analysis

HoxB8-Cas9 cells were transduced with guide RNA (gRNA) retroviruses (Section 2.2.3.4) and analysed on day 3 post-transduction. The cells were analysed by flow cytometry using the BD LSRFortessa^{TM}. Cells were washed with 1 ml FACS buffer and centrifuged at 300 g for 5 minutes. Once spun down, the cells were resuspended in 500 µl FACS buffer + 7-AAD, except for unstained and green fluorescent protein (GFP) controls, which were resuspended in 500 µl FACS buffer only. The cells were then analysed for GFP expression.

2.2.1.3. HSC retroviral transduction analysis: flow cytometry

2.2.1.3.1. Day 7 analysis by flow cytometry

E-SLAM HSCs (Lin^- $CD48^+$ $CD150^-$ $CD45^+$ $EPCR^+$) from Cas9 transgenic mice were transduced with gRNA retroviruses (Section 2.2.3.5). On day 7 post-transduction, the cells were analysed by flow cytometry using the BD LSRFortessa^{TM}. Cells were pipetted up and down and 100 µl was transferred to a new U-bottom 96-well plate (BD Falcon). The cells were stained in the wells using the E-SLAM2 panel (Table 2.3), which did not require any secondary staining. After 30 minutes on ice, cells from each well were transferred into individual 1.4 ml polypropylene tubes (ThermoFisher). The empty wells were washed with 200µl FACS buffer + 7-AAD, which was then added to the corresponding polypropylene tube (final volume: 300 µl). Single-stain and all-stain controls were prepared in FACS buffer without the addition of 7-AAD. A small fraction of cells taken from an empty vector sample were used as a GFP single-stain control.

Results were analysed using FlowJo_V10 and statistical analysis was performed in GraphPad Prism 8. Both the percentage of GFP^+ Lin^- $EPCR^+$ cells and the median EPCR expression in GFP^+

Lin^- cells was recorded. Each set of gRNAs was compared to the empty vector using a one-way analysis of variance (ANOVA) and Dunnett's multiple comparisons test. In the analysis of median EPCR expression, the results for the three gRNAs for each gene were pooled and reanalysed together using the same statistical significance tests.

2.2.1.3.2. Annexin V apoptosis analysis by flow cytometry

Annexin V BUV395 (BD Biosciences) was added to the E-SLAM2 panel towards the end of this investigation to see whether the CRISPR perturbation phenotypes affected apoptosis. A 10X Annexin V Binding Buffer (BD Biosciences) was required to see the Annexin V staining. It was freshly diluted in distilled water for each use.

Cells were stained in 100 µl HSC medium as described above using the E-SLAM2 panel with the addition of Annexin V BUV395. After 30 minutes on ice, cells from each well were transferred into individual 1.4 ml polypropylene tubes. The empty wells were washed with 200 µl Binding Buffer + 7-AAD, which was then added to the corresponding polypropylene tube (total volume: 300 µl). The 1.4 ml polypropylene tubes were placed inside 5 ml polystyrene tubes for flow cytometry analysis. Single-stain and all-stain controls were prepared in Binding Buffer only. A small fraction of cells taken from an empty vector sample were used as a GFP single-stain control.

Cells that were positive in both Annexin V and 7-AAD were already dead, whereas cells that were Annexin V positive and 7-AAD negative were in early apoptosis. Viable cells were negative for both viability markers.

Results were analysed using FlowJo_V10 and statistical analysis was performed in GraphPad Prism 8. Each set of gRNAs was compared to the empty vector using a one-way analysis of variance (ANOVA) and Dunnett's multiple comparisons test.

2.2.1.3.3. FACS: GFP^+ cells for genotyping and differentiation assays

On day 7 post-transduction, the HSCs that were analysed by flow cytometry were pooled and spun down at 300 g for 5 minutes. Once centrifuged, 200 µl of medium was removed and replaced with 100 µl of fresh complete medium.

On day 10 post-transduction, the HSCs were sorted for genotyping and differentiation assays. Only the $EPCR^+$ samples were used for these investigations. Although the cells were sorted on 7-AAD and GFP only, they were additionally stained with the GFP sorting panel (Table 2.3). 100 µl of

medium was removed for staining (total staining volume: 100 µl). After 30 minutes on ice, cells from each well were transferred into individual 5 ml polypropylene tubes. The empty wells were washed with 200µl FACS buffer + 7-AAD, which was then added to the corresponding tube (total volume: 300µl). The cells were sorted on the BD FACSMelody for genotyping and differentiation assays as detailed below:

- Differentiation assays: cells were sorted into 1.5 ml Eppendorf tubes filled with 300 µl HSC base medium. Each sample was sorted for GFP^+ and GFP^- cells and for three cell counts: 100, 200 and 400 cells. When there were too few cells, only the 100 and 200 cell counts were sorted.
- Genotyping: cells were sorted into 1.5 ml Eppendorf tubes filled with 100 µl FACS buffer. Each sample was sorted for GFP^+ cells only. Cells were sorted for the maximum number of cells possible, i.e. until the sample was empty.

2.2.1.4. Antibody Panels

Table 2.2. List of antibodies used for FACS and flow cytometry.

Antigen	Fluorophore	Manufacturer	Clone
Annexin V	BUV395	BD Biosciences	---
B220	Pe-Cy7	Biolegend	RA3-6B2
CD105	APC	Biolegend	MJ7/18
CD150	Pe-Cy7	Biolegend	TC15/12F12.2
CD150	PB	Biolegend	TC15/12F12.2
CD16/32	PE	Biolegend	93
CD16/32	Al647	Biolegend	93
CD3	Pe-Cy7	Biolegend	17A2
CD34	FITC	BD Pharmingen	RAM34
CD41	FITC	BD Pharmingen	Mw/Reg30
CD45	BV711	Biolegend	30-F11
CD45	FITC	Biolegend	30-F11
CD48	APC	Biolegend	HM48-1
CD48	PB	Biolegend	HM48-1
c-Kit	APC-Cy7	Biolegend	2B8
EPCR	PE	Stem Cell Tech.	RMEPCR1560
Flk2 (CD135)	Pe-Cy5	e-Bioscience	A2F10
Gr-1	Pe-Cy7	Biolegend	RB6-8C5
IL-7Ra	Biotin	Biolegend	A7R34
Lineage Cocktail (Lin)		Stem Cell Tech.	*
Mac1	Pe-Cy7	Biolegend	M1/70
Nk1.1	Pe-Cy7	Biolegend	PK136
Sca1	PB	Biolegend	E13-161.7
Sca1	BV421	Biolegend	108133
Sca1	BV605	Biolegend	108133
Streptavidin	BV510	Biolegend	*

Antigen	Fluorophore	Manufacturer	Clone
Ter119	Pe-Cy7	Biolegend	Ter-119

*Lineage Cocktail (StemCellTechnologies) was a pre-mixed combination of biotinylated antibodies targeting non-haematopoietic cells and non-progenitor cells (CD5, CD11b, CD19, CD45R/B220, Ly6G/C(Gr-1), TER119, 7-4). The lineage cocktail was used with Streptavidin BV510 for FACS and flow cytometry analysis.

Table 2.3. Antibody panels used for FACS and flow cytometry.

Staining Panel	Antibodies	Chapter
PreMegE	Lin BV510, Sca1 PB, CD105 APC, CD150 Pe-Cy7, CD16/32 PE, CD41 FITC, 7-AAD	5
MPP	Lin BV510, IL-7Ra Biotin, c-Kit APC-Cy7, Sca1 PB, Flk2 Pe-Cy5, CD34 FITC, 7-AAD	5
ST-HSC	Lin BV510, IL-7Ra Biotin, c-Kit APC-Cy7, Sca1 PB, CD48 APC, CD150 Pe-Cy7, CD34 FITC, CD16/32 PE, CD135 Pe-Cy5, 7-AAD	5
HSPC	Lin BV510, EPCR PE, CD48 PB, Sca1 BV605, CD34 FITC, Flk2 Pe-Cy5, CD150 Pe-Cy7, CD16/32 Al647, c-Kit APC-Cy7, DAPI	3
E-SLAM 1	Lin BV510, EPCR PE, CD48 APC, CD150 Pe-Cy7, c-Kit APC-Cy7, Sca1 BV421, CD45 FITC, 7AAD	4
E-SLAM 2	Gr1 Pe-Cy7*, Mac1 Pe-Cy7*, B220 Pe-Cy7*, CD3 Pe-Cy7*, Ter119 Pe-Cy7*, Nk1.1 Pe-Cy7*, Sca1 BV605, CD45 BV711, CD150 PB, EPCR PE, CD48 APC, c-Kit APC-Cy7, 7AAD º Annexin V BUV395	4
GFP Sorting Panel	Gr1 Pe-Cy7*, Mac1 Pe-Cy7*, B220 Pe-Cy7*, CD3 Pe-Cy7*, Ter119 Pe-Cy7*, Nk1.1 Pe-Cy7*, EPCR PE, 7AAD	4

* Antibodies used to generate a lineage cocktail were conjugated to the same fluorophore.
º Annexin V BUV395 was added to the E-SLAM2 panel for apoptosis analysis (Section 2.2.1.3.2)

2.2.2. HSC retroviral transduction analysis: differentiation assays

Differentiation assays were performed on day 10 post-transduction. After flow analysis, the remaining cells representing each perturbation phenotype were pooled. Initially, fractions of cells equalling to roughly 400, 200 and 100 cells were taken and placed into a methylcellulose-based medium (MethoCult™ GF M3434, StemCellTechnologies). However, it was not possible to differentiate GFP positive and negative cells microscopically. Therefore, an additional FACS step was included. Cells were sorted based on GFP expression on either the BD FACSMelody (BD Biosciences) or the BD Influx 4 sorter. Cells were sorted directly into 300 µl HSC medium without cytokines. Specific numbers of both GFP^+ and GFP^- cells were sorted separately (400, 200, and 100 cells). If there were too few cells, only the 100 and 200 cell counts were sorted. The cells were then placed into 2.7 ml methylcellulose-based medium and vortexed. Once all air bubbles had dissipated from the medium (usually after 5-10 minutes), the mixture was split between 2 wells of

a 6-well SmartDish™ (StemCellTechnologies) using a 5 ml syringe and 16-gauge needle. The plates were incubated at 37°C and 5% CO_2. After 10 days, the plates were imaged and analysed using STEMvision™ Acquisition and Analyzer software (StemCellTechnologies) on the "Human 14-Day CFU" setting. Burst forming unit-erythroid (BFU-E), colony forming unit-granulocyte-macrophage (CFU-GM), and colony forming unit-granulocyte-erythrocyte-macrophage-megakaryocyte (CFU-GEMM) colonies were manually assigned in the STEMvision™ Marker software.

2.2.3. Retrovirus Production and Transduction

2.2.3.1. Gene selection

Gene selection for the HSC CRISPR screen (Chapter 4) was performed based on the MolO and SuMO genes that were previously identified, which are listed in Table 2.4 (N. K. Wilson et al. 2015). The expression of each gene was visualized on the mouse single-cell gene expression atlas (Nestorowa et al. 2016). Genes that did not have expression specific to the LT-HSC region were not considered for further analysis. Violin plots were produced for the remaining candidate genes to select for genes that have highest expression in the E-SLAM and LT-HSC populations. By doing so, the list of 44 genes was narrowed down to 16 candidate genes.

Table 2.4. MolO and SuMO genes.

MolO Genes				SuMO Genes		
Cd82	*Gimap6*	***Pdzk1ip1***	*Vwf*	*Ablim1*	***Inhba***	***Wfdc2***
Cdkn1c	*Gstm1*	*Ptpn14*		*Cd74*	*Ly6e*	*Ifitm1**
Cldn10	*Limd2*	***Smtnl1***		*Cyp27a1*	*Mapk12*	*Ly6a**
Ctsf	*Ltb*	***Sox18***		*Gbp6*	***Ndnf***	*Mllt3**
Fads3	*Mettl7a1*	*Sqrdl*		***Gbp8***	*Ralgapa1*	***Procr****
Fgfr3	***Neo1***	***Trim47***		***Gm4951***	*St8sia4*	***Ramp2****
Gimap1	***Pde1b***	*Ubl3*		*Ifitm3*	*Tftp2*	***Sult1a1****

Genes were identified by Wilson *et al.* (2015). Genes in **bold** are genes that were chosen for the CRISPR screen.
* represent genes that are both MolO and SuMO genes.

2.2.3.2. Oligo Generation

Once the candidate genes were selected, gRNA sequences were chosen. The Brie gRNA library (Doench et al. 2016) provided sequences for gRNAs for most of the candidate genes; however, where unavailable, gRNAs were designed using the "Broad Institute sgRNA Designer: CRISPRko" tool, selecting *S. pyogenes (NGG)* as the CRISPR enzyme and mouse as the target

taxon (Broad Institute 2018). The chosen gRNA sequences were checked using UCSC Genome Browser to make sure that each sequence was present only within the gene of interest. If there was overlap with other genes, a new sequence was generated and verified. Three gRNA sequences were chosen per gene of interest.

The gRNAs were modified to meet the criteria of the published protocol for cloning into the pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W CRISPR gRNA expression vector (Tzelepis et al. 2016). A G nucleotide was appended at the 5' end of each sequence. The 5' overhang of the top oligo was CACC and the 5' overhang of the complement bottom oligo was CAAA. All oligos (Table 2.5) were ordered through Sigma at 100 mM in TE buffer, purified by desalt purification.

2.2.3.3. Production of retroviruses in 293Ts

293T cells at 70% confluency (10 cm plates) were used to produce retroviruses. On day 0, 293Ts were transfected using TransIT-LT1 Transfection Reagent (Mirus). A transfection reagent mix (500 µl DMEM + 45 µl TransIT-LT1) and DNA mix (5 µg construct + 5 µg pMD2.G + 5 µg ΔR8.9 in 500 µl DMEM) were prepared for each construct of interest. The DNA mix was slowly and drop-wise added to the transfection reagent mix, mixed gently, and incubated at room temperature for 30 minutes. The mixture was then added slowly and drop-wise to 10 cm plates containing 293T cells in 10 ml 293T medium. The plates were incubated at 37°C. On day 1, the medium was changed from 293T medium to 6 ml base HSC medium (no cytokines and 2-mercaptoethanol). On day 2, the supernatant was recovered from each plate using a 5ml syringe, filtered through a 0.45µm filter and aliquoted into 2ml cryogenic tubes (ThermoFisher Nunc™ Cryovials). These tubes were then stored at -80°C.

This protocol was also adapted to be performed in 6 well plates, using 100 µl DMEM + 9 µl TransIT-LT1 for the transfection reagent mix and 1 µg of the construct, pMD2.G and ΔR8.9 in 100 µl DMEM for the DNA mix. On day 1, the medium was changed to 1.5 ml base HSC medium. All other steps are as previously described.

2.2.3.4. Batch-testing of retroviruses in HoxB8-Cas9 cells

Retroviruses made for the HSC CRISPR screen (Chapter 4) were tested for transduction efficiency in the HoxB8-Cas9 cell line before being used in primary cells. On the day of transduction (Day 0), HoxB8-Cas9 cells were plated at 3.3 x 10^5 cells in 440 µl HoxB8 complete medium in a 24-well plate. 5.33 µl polybrene (1 mg/ml, Sigma) was added to each well for a final concentration of

8 µg/ml. For batch testing, between 12 and 22 of the viruses made at one time (Section 2.2.3.3) were tested, along with an empty vector and untreated control. 100 µl of viral supernatant was added to the appropriate well and topped up with 120 µl of HoxB8 medium. The plate was then centrifuged for 90 min at 779 g at 32°C (maximum acceleration, no brake). After centrifugation, the plate was incubated for 90 min at 32°C. 330 µl of medium was then removed and replaced with 560 µl of fresh complete medium. The plate was then incubated at 37°C and 5% CO_2. The following day (Day 1), 440 µl of fresh medium was added to each well. On Day 2, the cells were split 1:2 to prepare for analysis the next day. On Day 3, the cells were analysed for GFP expression by flow cytometry (Section 2.2.1.2).

2.2.3.5. Retroviral transduction of HSCs

To prepare for retroviral transduction of HSCs, 96-well U-bottom plates were first coated with Retronectin (Clontech) to improve transduction efficiency. The stock Retronectin (1 µg/µl) was diluted 1:100 in PBS and 50 µl was added to the each well. The plates were incubated for 2 hours at room temperature before the Retronectin was discarded. 50 µl of 2% bovine serum albumin (BSA, Sigma)/PBS was then added to the wells and incubated for 30 minutes at room temperature. After 30 minutes, the 2% BSA/PBS was removed and replaced with 110 µl base HSC medium. The plate was then kept in the fridge until required later that day.

HSCs were isolated from the bone marrow of Cas9 transgenic mice as previously described (Section 2.2.1.1) and were first sorted into two bulk populations using the E-SLAM1 panel: Lin^- $CD48^+$ $CD150^-$ $EPCR^+$ (referred to as $EPCR^+$ cells), and Lin^- $CD48^+$ $CD150^-$ $EPCR^-$ (referred to as $EPCR^-$ cells), where $EPCR^-$ cells were used as controls. All sorting was done on the BD Influx4 and Influx5. The first sort was performed in 2.0 Drop Enrich mode. The $EPCR^-$ and $EPCR^+$ cells were then sorted in bulk (250 cells) directly into 110 µl base HSC medium in a 96-well U-bottom plate, using the 1.0 Drop Pure mode.

Once the cells were sorted, each well was topped up with 10 µl of a 12X Cytokine Cocktail (5 µg/ml IL-11, 100µg/ml SCF), 0.2 µl 50mM 2-mercaptoethanol, and 1 µl of 1 mg/ml polybrene. 1 µl of virus was added to the wells as appropriate. Empty vector and untreated controls were used for each FACS experiment and 3 replicates for each virus were collected when possible (depending on the number of cells sorted).

After the addition of virus, the cells were spun down at 600 g for 30 minutes at 32°C (maximum acceleration, no brake). The cells were then incubated for 30 minutes at 32°C, after which they were maintained in a 37°C and 5% CO_2 incubator. The next day (Day 1), each well containing cells was topped up with 80 µl of complete HSC medium. The cell medium was changed on Day 4 by carefully removing 120 µl of medium from each well and replacing it with 120 µl of fresh complete medium. On Day 7, cells were analysed for changes in EPCR expression by flow cytometry (Section 2.2.1.3). On Day 10, cells were sorted for differentiation assays (Section 2.2.2) and genotyping (Section 2.4).

2.2.3.6. Oligo sequence list

Table 2.5. Sequences of gRNAs used for the HSC CRISPR screen.

Oligo Name	Sequence (5' - 3')
Cdkn1c_sg1_pKLV2_L	CACCGCGGGTCGGAGGTCGCGACCA
Cdkn1c_sg2_pKLV2_L	CACCGAGACGACCAGGGCCTCGAAG
Cdkn1c_sg3_pKLV2_L	CACCGCGTGGCGACTCGGGACGGCG
Cdkn1c_sg1_pKLV2_R	AAACTGGTCGCGACCTCCGACCCGC
Cdkn1c_sg2_pKLV2_R	AAACCTTCGAGGCCCTGGTCGTCTC
Cdkn1c_sg3_pKLV2_R	AAACCGCCGTCCCGAGTCGCCACGC
Fgfr3_sg1_pKLV2_L	CACCGGTATAGTTGCCACGATCGGA
Fgfr3_sg2_pKLV2_L	CACCGGAGGCTGGCAGCGTGTACGC
Fgfr3_sg3_pKLV2_L	CACCGTGACAAGGACCTGTCGGACC
Fgfr3_sg1_pKLV2_R	AAACTCCGATCGTGGCAACTATACC
Fgfr3_sg2_pKLV2_R	AAACGCGTACACGCTGCCAGCCTCC
Fgfr3_sg3_pKLV2_R	AAACGGTCCGACAGGTCCTTGTCAC
Neo1_sg1_pKLV2_L	CACCGCGTAACCGATGGCATAACCT
Neo1_sg2_pKLV2_L	CACCGGGTTCCAAGATTATCCACAG
Neo1_sg3_pKLV2_L	CACCGAACACCGTTATCTGGCAATG
Neo1_sg1_pKLV2_R	AAACAGGTTATGCCATCGGTTACGC
Neo1_sg2_pKLV2_R	AAACCTGTGGATAATCTTGGAACCC
Neo1_sg3_pKLV2_R	AAACCATTGCCAGATAACGGTGTTC
Pde1b_sg1_pKLV2_L	CACCGCAACACCATCTCGATAACCA
Pde1b_sg2_pKLV2_L	CACCGAAAACTCATCAGAAACACTG
Pde1b_sg3_pKLV2_L	CACCGGGACTGCAGTAGAGTATGTG
Pde1b_sg1_pKLV2_R	AAACTGGTTATCGAGATGGTGTTGC
Pde1b_sg2_pKLV2_R	AAACCAGTGTTTCTGATGAGTTTTC
Pde1b_sg3_pKLV2_R	AAACCACATACTCTACTGCAGTCCC
Ramp2_sg1_pKLV2_L	CACCGGAATCAATCTCATCCCACTG
Ramp2_sg2_pKLV2_L	CACCGACCAAGCCGAGATCCACCCG
Ramp2_sg3_pKLV2_L	CACCGCTCTTGTACTCATACCAGCA
Ramp2_sg1_pKLV2_R	AAACCAGTGGGATGAGATTGATTCC
Ramp2_sg2_pKLV2_R	AAACCGGGTGGATCTCGGCTTGGTC

Oligo Name	Sequence (5' - 3')
Ramp2_sg3_pKLV2_R	AAACTGCTGGTATGAGTACAAGAGC
Smtnl1_sg1_pKLV2_L	CACCGCTAAGAGTGGCGAATCAGGG
Smtnl1_sg2_pKLV2_L	CACCGAGAGACCGGAAGTGACACAA
Smtnl1_sg3_pKLV2_L	CACCGAGCAGAGGTTACTGTCAACG
Smtnl1_sg1_pKLV2_R	AAACCCCTGATTCGCCACTCTTAGC
Smtnl1_sg2_pKLV2_R	AAACTTGTGTCACTTCCGGTCTCTC
Smtnl1_sg3_pKLV2_R	AAACCGTTGACAGTAACCTCTGCTC
Sox18_sg1_pKLV2_L	CACCGAGCAGCGGCCCCGATTCCAG
Sox18_sg2_pKLV2_L	CACCGCAGAGTGGGTAGCTCGCGGA
Sox18_sg3_pKLV2_L	CACCGCCGACGAGTTGCGCATTCGG
Sox18_sg1_pKLV2_R	AAACCTGGAATCGGGGCCGCTGCTC
Sox18_sg2_pKLV2_R	AAACTCCGCGAGCTACCCACTCTGC
Sox18_sg3_pKLV2_R	AAACCCGAATGCGCAACTCGTCGGC
Inhba_sg1_pKLV2_L	CACCGACAAGCAATCCGCACGTCCA
Inhba_sg2_pKLV2_L	CACCGCGAGGAAATGGGCTTAAAGG
Inhba_sg3_pKLV2_L	CACCGCTGCTGCTGAAATAGACGGA
Inhba_sg1_pKLV2_R	AAACTGGACGTGCGGATTGCTTGTC
Inhba_sg2_pKLV2_R	AAACCCTTTAAGCCCATTTCCTCGC
Inhba_sg3_pKLV2_R	AAACTCCGTCTATTTCAGCAGCAGC
Ndnf_sg1_pKLV2_L	CACCGGACTGTGGTACGTCCAAAGG
Ndnf_sg2_pKLV2_L	CACCGAAGGGGTTAAAGTCTAGACC
Ndnf_sg3_pKLV2_L	CACCGCAGTGGAACTCAAAGACGGG
Ndnf_sg1_pKLV2_R	AAACCCTTTGGACGTACCACAGTCC
Ndnf_sg2_pKLV2_R	AAACGGTCTAGACTTTAACCCCTTC
Ndnf_sg3_pKLV2_R	AAACCCCGTCTTTGAGTTCCACTGC
Wfdc2_sg1_pKLV2_L	CACCGCACACTACTAAACCACCGGG
Wfdc2_sg2_pKLV2_L	CACCGACAGTAGCAACCCTAGTAGG
Wfdc2_sg3_pKLV2_L	CACCGTGGGACTACTACTCAATCAG
Wfdc2_sg1_pKLV2_R	AAACCCCGGTGGTTTAGTAGTGTGC
Wfdc2_sg2_pKLV2_R	AAACCCTACTAGGGTTGCTACTGTC
Wfdc2_sg3_pKLV2_R	AAACCTGATTGAGTAGTAGTCCCAC
Pdzk1ip1_sg1_pKLV2_L	CACCGCGGCGAAGACGATTGCAACA
Pdzk1ip1_sg2_pKLV2_L	CACCGAGAACACAGCGACAGCAATG
Pdzk1ip1_sg3_pKLV2_L	CACCGCAACCACTTCTGGTGCCAGG
Pdzk1ip1_sg1_pKLV2_R	AAACTGTTGCAATCGTCTTCGCCGC
Pdzk1ip1_sg2_pKLV2_R	AAACCATTGCTGTCGCTGTGTTCTC
Pdzk1ip1_sg3_pKLV2_R	AAACCCTGGCACCAGAAGTGGTTGC
Procr_sg1_pKLV2_L	CACCGTCCAAGACAACCATCATGTG
Procr_sg2_pKLV2_L	CACCGTGCGCCCTTTGTAACTCCGA
Procr_sg3_pKLV2_L	CACCGGCCACATCGAAGAAGACATG
Procr_sg1_pKLV2_R	AAACCACATGATGGTTGTCTTGGAC
Procr_sg2_pKLV2_R	AAACTCGGAGTTACAAAGGGCGCAC
Procr_sg3_pKLV2_R	AAACCATGTCTTCTTCGATGTGGCC

Oligo Name	Sequence (5' - 3')
Sult1a1_sg1_pKLV2_L	CACCGGGTGGCAAGCTAGATAAGTG
Sult1a1_sg2_pKLV2_L	CACCGATGTGTCTTAATGATCCGTG
Sult1a1_sg3_pKLV2_L	CACCGTCCGCAAAGTATTTGATGAG
Sult1a1_sg1_pKLV2_R	AAACCACTTATCTAGCTTGCCACCC
Sult1a1_sg2_pKLV2_R	AAACCACGGATCATTAAGACACATC
Sult1a1_sg3_pKLV2_R	AAACCTCATCAAATACTTTGCGGAC
Trim47_sg1_pKLV2_L	CACCGTGATGAGGGCCACAGTACGG
Trim47_sg2_pKLV2_L	CACCGACGCGACAGTAGCGCTCCAG
Trim47_sg3_pKLV2_L	CACCGCTGGGACCGGCCCAACATTG
Trim47_sg1_pKLV2_R	AAACCCGTACTGTGGCCCTCATCAC
Trim47_sg2_pKLV2_R	AAACCTGGAGCGCTACTGTCGCGTC
Trim47_sg3_pKLV2_R	AAACCAATGTTGGGCCGGTCCCAGC
Gbp8_sg1_pKLV2_L	CACCGCACTAAACCAGAGCACACCC
Gbp8_sg2_pKLV2_L	CACCGCGTCTGGCAGGACAGAATCA
Gbp8_sg3_pKLV2_L	CACCGGCTAGAGCTGAAGTTAAATG
Gbp8_sg1_pKLV2_R	AAACGGGTGTGCTCTGGTTTAGTGC
Gbp8_sg2_pKLV2_R	AAACTGATTCTGTCCTGCCAGACGC
Gbp8_sg3_pKLV2_R	AAACCATTTAACTTCAGCTCTAGCC
Gm4951_sg1_pKLV2_L	CACCGTGGGGTGACGGACAAAACCA
Gm4951_sg2_pKLV2_L	CACCGGGTGAGAGCAACATTGAGCG
Gm4951_sg3_pKLV2_L	CACCGATTCTACTTCGTGAGAACAC
Gm4951_sg1_pKLV2_R	AAACTGGTTTTGTCCGTCACCCCAC
Gm4951_sg2_pKLV2_R	AAACCGCTCAATGTTGCTCTCACCC
Gm4951_sg3_pKLV2_R	AAACGTGTTCTCACGAAGTAGAATC

Table 2.6. Primer sequences used to sequence the pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W vector.

Primer Name	Primer Sequence (5' - 3')
SN037_Seq_F1	AGATAATTAGAATTAATTTGACTG
SN037_Seq_F2_LKO.15	GACTATCATATGCTTACCG
SN037_Seq_R1	CATGCTCCAGACTGCCTTG

2.2.4. Luciferase Assays

Luciferase assays were used to validate regulatory relationships between genes, inferred from in silico regulatory network models (Chapters 5 and 6). The two relationships interrogated were between *Gata2* and *Nfe2*, and *Gata2* and *Cbfa2t3h*. The *Cbfa2t3h* minimal and full promoter sequences, as well as the *Nfe2* enhancer sequence, were cloned into pGL2-Basic and pGL2-Promoter vectors (Section 2.3.4.4). The mm10 mouse genome coordinates of chromosomal regions tested are as follows: chr8:122699004-122701098 for the *Cbfa2t3h* full promoter,

chr8:122699111-122699377 for the *Cbfa2t3h* minimal promoter, and chr15:103258245-103258850 for the *Nfe2* enhancer. Both wild-type and mutant GATA2 constructs were generated, in which all GATA2 binding sites were mutated to prevent binding activity.

Luciferase assays were performed with the help of Sarah Kinston as previously described (Bockamp et al. 1995). 416B cells were transfected with the linearized *Nfe2* and *Cbfa2t3h* constructs and pPGK-Neo by electroporation. The transfected cells were harvested, washed, and centrifuged. A luciferase buffer was made up with LB buffer (25 mmol/L Tris-phosphate buffer pH 7.8, 8 mmol/L MgCI2, 1 mmol/L 1,4-Dithio-DL-threitol [DTT], 1% Triton-X 100, 1% BSA and 15% glycerol), ATP and the luciferase substrate D-Luciferin (Promega). 100 µl of sample was mixed with the luciferase buffer and incubated for 5 minutes in the dark at room temperature. The samples were then assayed for light emission in a Berthold LB 953 luminometer (Berthold).

In analysing the results from the luciferase assay, the luciferase activity was normalised against empty vector and the fold change was used to show differences in luciferase activity.

2.3. Molecular Biology

2.3.1. Agarose gel electrophoresis

DNA fragments were separated by size using agarose gel electrophoresis. The amount of agarose (Biogene) used ranged between 0.8-3.0% depending on the application. Agarose was mixed with 1xTBE (Tris/Borate/EDTA; produced on site) and dissolved in a microwave. Once the solution cooled, 0.5 µg/ml ethidium bromide solution (Sigma) was added to stain DNA.

Samples were mixed with a 6x loading dye (NEB). DNA reference ladders (50 bp-1 kb, NEB) were used according to the expected size of the DNA. Gels were run on an electric field of 80-120 V and the DNA was then visualized using a UV transilluminator.

2.3.2. DNA purification

2.3.2.1. Gel extraction

DNA was visualized using a UV transilluminator and the bands of interest were excised using a scalpel blade. DNA was then extracted using the QIAquick Gel Extraction Kit (Qiagen) as per manufacturer's instructions and eluted in 30 µl elution buffer.

2.3.2.2. PCR purification

PCR fragments were purified using the QIAquick PCR Purification Kit (Qiagen) as per manufacturer's instructions and eluted in 30 µl elution buffer.

2.3.3. Plasmid purification

Plasmids were purified from starting cultures of various volumes. Starting cultures consisted of lysogeny broth (LB) + 100 µg/ml ampicillin (Sigma) inoculated with a single bacterial colony. Bacteria were cultured for 16 hours at 37°C, shaking at 220 rpm. Culture volumes used are detailed in the appropriate subsections.

2.3.3.1. Mini Preps

Mini preps were performed following the QIAprep Spin Miniprep Kit (Qiagen) as per manufacturer's instructions, from a 2 ml starting culture. DNA was eluted in 30 µl ultra-pure water. The DNA was then sequenced (Source BioScience) and run on an agarose gel to confirm the presence of correctly sized DNA fragments.

2.3.3.2. Maxi Preps

Maxi preps were performed to purify large quantities of plasmid DNA. The volume of the starting culture was 100 ml. Maxi preps were performed following the Plasmid DNA Purification Kit (Macherey-Nagel) following the "Maxi" instructions (AX 500). DNA was resuspended in 200-300 µl ultra-pure water.

2.3.3.2. Xtra Maxi Prep

Xtra maxi preps were performed when a very large quantity of plasmid DNA was required, for example to purify a vector used as a backbone or for retrovirus production. The volume of the starting culture was 300 ml. Xtra maxi preps were performed following the NucleoBond Xtra® Maxi Kit (Macherey-Nagel) as per manufacturer's instructions. DNA was resuspended in 200-300 µl ultra-pure water.

2.3.4. Cloning strategies

2.3.4.1. Restriction Digests

Restriction enzymes and buffers were purchased from New England Biolabs. Digests were performed in 50-100 µl volumes depending on the application and quantity required. Restriction

enzymes made up no more than 10% of the reaction volume. Reactions consisted of the restriction enzyme, the appropriate 10X buffer, DNA, and ultra-pure H_2O. The reactions were incubated at 37°C for 1 hour to overnight, and then purified by gel extraction or PCR purification.

2.3.4.2. Weissman protocol

This protocol was used for cloning gRNA into the pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W plasmid (Addgene), which was digested with BbsI. The protocol was originally described by the Weissman lab (Adamson et al. 2016) and then adapted as described below.

Annealing: Oligos were annealed by mixing 1µl of the forward and reverse oligos (10 µM) with 23 µl dH_2O and 25 µl 2X Annealing Buffer (200mM potassium acetate, 60 mM HEPES-KOH pH 7.4, 4 mM Magnesium Acetate). The mixture was incubated at 95°C for 5 minutes, then left to anneal by gradually cooling to room temperature (on the benchtop, not in a PCR machine) for 10-20 minutes. The annealed oligos were diluted 20-fold for ligation.

Ligation: 500ng of the digested vector backbone was mixed with 10 µl of 1:20 diluted annealed oligos, 2 µl of fresh 10X T4 ligase buffer (NEB), 1 µl T4 ligase (NEB). The volume was made up to a 20 µl total reaction volume. A negative control was made by using dH_2O instead of annealed oligos. The ligation mixture was incubated at room temperature for 1-4 hours, or overnight at 16°C.

Transformations were done using MegaX DH10B™ T1® Electrocomp™ Cells (ThermoFisher) following the manufacturer's protocol. All of the transformed bacteria were plated onto a LB+ Ampicillin plate (made on site) and incubated at 37°C overnight.

2.3.4.2.1. Modified Weissman protocol

The Weissman protocol was modified to incorporate a Rapid DNA Ligation Kit (Roche), which contained a 5X DNA dilution buffer, T4 DNA ligation buffer, and ligase. The oligos were annealed as previously described and diluted 1:20 in water. The 5X DNA dilution buffer was diluted in sterile H_2O. 100 ng of the digested vector backbone was mixed with 2 µl of the diluted annealed oligos in up to 10 µl of 1X DNA dilution buffer. 10 µl of T4 DNA ligation buffer and 1 µl ligase were added to the mix and incubated at room temperature for 5 minutes. The ligation mixture was then used for transformations in DH5α cells (Invitrogen).

2.3.4.3. Yusa Lab protocol

This protocol was used for cloning gRNA into the pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W and pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W-ccdB plasmids (provided by Oliver Dovey, Vassiliou lab), which were digested with BbsI. The protocol was described by the Yusa lab (Tzelepis et al. 2016).

Oligo phosphorylation and annealing: 1 µl each of the forward and reverse strand oligos (10 mM) were mixed with 1 µl of 10X T4 ligase buffer (NEB), 0.5 µl of T4 PNK (NEB) and 6.5 µl dH_2O. The mixture was placed in a PCR machine and the following program was run: 30 min at 37°C → 5 min at 95°C → ramp down to 25°C at 0.1°C/sec → 4°C (hold). The annealed oligos (ds-oligo) were diluted to 7.1 pmol/µl in EB buffer (Qiagen) by diluting 2 µl ds-oligo in 139 µl EB, 3µl of which was then diluted in 57 µl EB.

Ligation: 1 µl of 20ng/µl digested vector backbone was mixed with 2 µl diluted ds-oligo, 1 µl 10X ligase buffer, 1 µl T4 ligase and 5 µl dH_2O. A negative control was made by using dH_2O instead of ds-oligo. The mixture was incubated at 16°C for 2 hours to overnight.

Transformations were done using Library Efficiency™ DH5α™ Competent Cells (ThermoFisher). 5 µl of the ligation mixture was mixed with 50 µl bacterial cells and incubated for 30 minutes on ice. The cells were heat-shocked at 42°C for 45 seconds, and then incubated on ice for 5 minutes. 250 µl SOC medium was added to the cells and the transformed bacteria were incubated at 37°C for 30 min (shaking at 300 rpm). The transformed bacteria were plated onto a LB+ Ampicillin plate and incubated at 37°C overnight.

2.3.4.4. Cloning for luciferase assays

GeneArt Strings (ThermoFisher) were ordered for the *Cbfa2t3h* minimal promoter and full promoter, as well as the *Nfe2* enhancer region. These GeneArt String products had both wildtype and mutated GATA2 sites. 20 bp of the pBluescript KS vector were inserted at either end of the sequence to enable Gibson Assembly (Section 2.3.4.4.1). The mutated sequences were cloned into the pGL2-Basic Vector or pGL2-Promoter Vector by Gibson Assembly.

For Gibson Assembly, the digested backbone and GeneArt String fragments were used in a 1:1 ratio. They were mixed with deionized H_2O and 2X Gibson Assembly Master Mix (NEB) in a 20 µl reaction volume. The mixture was incubated at 50°C for 15 minutes, after which it could be used for transformations.

2.3.4.4.1. GeneArt String Sequences

Restriction enzyme site	**GATA2 binding site**	**Mutated base pair**

Cbfa2t3h minimal promoter (restriction enzymes XhoI/HindIII): chr8:122699111-122699377

GTACCGAGCTCTTACGCGTGCTAG*CTCGAG*GTGGGAGGTCTCAGGGCTACAGGCGG**GATA**GGAGGAAGTTG
TTGGGAAGTCAGACCGGAATGGCATGGTGGAGGGAGAACCGGCAACCAGGCAGATGGTTCCTGACGAGGAA
GCTCTGGGCACAGCTGCAGGCCCCCGACCCCCACCGCAT**TATC**ACTGTGACACAGCTGGCTGCCTCACCCC
TGAAGGCTGCAGGAGGACCTCCCCCATGCTGTCCCCAAGCCCGCCCCGTGTCACATGAGGCCCTGCAGACT
CCCACCCTCCGTC*AAGCTT*GGCATTCCGGTACTGTTGGTAAAA

Cbfa2t3h minimal promoter - mutated (restriction enzymes XhoI/HindIII)

GTACCGAGCTCTTACGCGTGCTAG*CTCGAG*GTGGGAGGTCTCAGGGCTACAGGCGG**GCGA**GGAGGAAGTTG
TTGGGAAGTCAGACCGGAATGGCATGGTGGAGGGAGAACCGGCAACCAGGCAGATGGTTCCTGACGAGGAA
GCTCTGGGCACAGCTGCAGGCCCCCGACCCCCACCGCAT**TGGC**ACTGTGACACAGCTGGCTGCCTCACCCC
TGAAGGCTGCAGGAGGACCTCCCCCATGCTGTCCCCAAGCCCGCCCCGTGTCACATGAGGCCCTGCAGACT
CCCACCCTCCGTC*AAGCTT*GGCATTCCGGTACTGTTGGTAAAA

Cbfa2t3h full promoter (restriction enzymes Nhe1/HindIII): chr8:122699004-122701098

GGGAGGTACCGAGCTCTTACGCGT*GCTAGC*ACCCTGGCATGGGAAGAGTGTGAGGGAACAGGAGGGAGCCC
CCAATCCCTGGCATAAAGCTGGGTGCAGACTGCAGACGGCCATGGTTTCTGCAGGGAACCGCGGCCCCTCA
AGCTCTCTGCAGCCACTTCCTTCCCCACTCCTTCCT**TATCTATC**GGACCACCCAGCGCAGAAAGCCACAAC
CAACGTCTACTTCCCCACAAACACCAACTGCCCTCCTGGGGGCATAGGGAGCGGATGATCACCCCCAGGTG
CTGGGGCAGTCCCCCACTTGTGAGCAACTTTCTTCTCTTTCAAACCCCACCTCCGTGAGACCTGCTGATTG
GAAAAGCATGCGGACAGGGCAGTAAGCAACGCCAGGCCTCCTCTTGAAAGCGGAAGTGTCCCATCTGTGGG
CCCCACAGGTGCAAGCCAGGAACATCTGCCTCGGAGCTGGTGGGGAGCCACCCACTCCCACCAGCCGCACC
CAGGGTCCAGCCCTGCTGCCTCTTGGGCTAGGGGCTGTTTTCAAAAACCCCTCACTGCCCCTGATCCAAAC
CCTTCCTTGGCTCACTCACTCACTCT**GATA**TATTTGGGGGGAGGGGCTGTGGCTCAGCTGCACTGGTTCTG
GGGACCATGACACTGCTCTGTTCTTCAGAGGTCAAGGAGCCTGGCCCCTGCAGGACTGCTGGCACTGAAAA
TAAACAGACACAACCCAGTGCCCTCGTTGATCCCATGCTCCACCCTCATCAGAACACAGGGAACACAGTGA
CACTGGTGGACTGTCC**TATC**TCACAGCAGGGTCCTCTCTCCTCAAACCCTGAAGAGAACAGTCAGCCAGCA
GGGTAGAAGGGGCAAGCAGGACCCCATGTTTTGGAACTCAAGACCTAAGCTTGTAGCTTGTTGTATTACTC
TTAATTTCAAGACACTCAAGGCCAAGACAGTGACATCCAGGAAGACATGAAACAAGGGGTTTTTAGTAGCC
AAC**TATC**TACATGAACAAGCTAAACGTTAAATCAAAGCAAAGCATAGTGGCCCTATTAGCATCCTAGGACC
ATGTCTGGGAACCCTTACCAGGTTTAAGAGAACATTATTCCTTCCCATGAAACCACCCTATAGTTTCATCA
CCGTTTAGTAAGTGGCTTTCTCTGGAACCTGAAGCTATAATAGCATTATTTCTACTGGGATTCACTTGGAA
AGGGGCCTGGAGTTTCTCCCAAGTAGACAGCCTTTCTAGCAAGCTCCAGGTCACTAGGCAGAGAGGCCAAC
TGCCTATGTGTGGCCTGTTTGGCTCAAGCTTCCTTTAAAGAAGGTAATATAGGATGGGGTTTCCACAGCTG
GGAGACTAGACTGTGACCCTGATCC**TATC**TGAGGCCAGGCACAACATGACAATGGAGATTCTTCCAGGCAA
GAGGTGACCAGAGAACCACTGATGGTGATGGAGAGACCACCTGAGCCATGGGTGGGCAGGAGGCTGGTCAG
AGTCCTGACCATGCTCACTCCTGACTCTGCCAGCACCCCCAATAGGTCGTTCATTGGAGGGCAAATAGAAA
GGGTTTCTTGGTCTTGGGCAAAGTGTGGCTGAAGGCAGGCAAAGTGGGTGGAGTCTCTGGACCTCCCAGGC
CACAGGAAGGGCTATGCCCCATGGGAA**GATA**GACAGATGAGCCAGAGGCTCTGGCCAGGATGGACATTACT
CCAGGGTGGAGGAAACACACTCCATGAGCCACACCAGATTGGGCCCAGTGGGAGGTCTCAGGGCTACAGGC
GG**GATA**GGAGGAAGTTGTTGGGAAGTCAGACCGGAATGGCATGGTGGAGGGAGAACCGGCAACCAGGCAGA
TGGTTCCTGACGAGGAAGCTCTGGGCACAGCTGCAGGCCCCCGACCCCCACCGCAT**TATC**ACTGTGACACA
GCTGGCTGCCTCACCCCTGAAGGCTGCAGGAGGACCTCCCCCATGCTGTCCCCAAGCCCGCCCCGTGTCAC
ATGAGGCCCTGCAGACTCCCACCCTCCGTCCAGGGCCACAACCCAGCTCTGCCGGCTGTAGTGACTAGAAA

GGCCTGGAGCCTCCAAGGAACAGAGGCACGGGCTCCGAGACGCCAAAGCTCCTCCAGCCCTCCTGT*AGATC*
*T*AAGTAAGCTTGGCATTCCGGTACT

Cbfa2t3h full promoter - mutated (restriction enzymes Nhe1/HindIII)

GGGAGGTACCGAGCTCTTACGCGT*GCTAGC*ACCCTGGCATGGGAAGAGTGTGAGGGAACAGGAGGGAGCCC
CCAATCCCTGGCATAAAGCTGGGTGCAGACTGCAGACGGCCATGGTTTCTGCAGGGAACCGCGGCCCCTCA
AGCTCTCTGCAGCCACTTCCTTCCCCACTCCTTCCT**TGTCTGGC**GGACCACCCAGCGCAGAAAGCCACAAC
CAACGTCTACTTCCCCACAAACACCAACTGCCCTCCTGGGGGCATAGGGAGCGGATGATCACCCCCAGGTG
CTGGGGCAGTCCCCCACTTGTGAGCAACTTTCTTCTCTTTCAAACCCCACCTCCGTGAGACCTGCTGATTG
GAAAAGCATGCGGACAGGGCAGTAAGCAACGCCAGGCCTCCTCTTGAAAGCGGAAGTGTCCCATCTGTGGG
CCCCACAGGTGCAAGCCAGGAACATCTGCCTCGGAGCTGGTGGGGAGCCACCCACTCCCACCAGCCGCACC
CAGGGTCCAGCCCTGCTGCCTCTTGGGCTAGGGGCTGTTTTCAAAAACCCCTCACTGCCCCTGATCCAAAC
CCTTCCTTGGCTCACTCACTCACTCT**GCAA**TATTTGGGGGGAGGGGCTGTGGCTCAGCTGCACTGGTTCTG
GGGACCATGACACTGCTCTGTTCTTCAGAGGTCAAGGAGCCTGGCCCCTGCAGGACTGCTGGCACTGAAAA
TAAACAGACACAACCCAGTGCCCTCGTTGATCCCATGCTCCACCCTCATCAGAACACAGGGAACACAGTGA
CACTGGTGGACTGTCC**TTTC**TCACAGCAGGGTCCTCTCTCCTCAAACCCTGAAGAGAACAGTCAGCCAGCA
GGGTAGAAGGGGCAAGCAGGACCCCATGTTTTGGAACTCAAGACCTAAGCTTGTAGCTTGTTGTATTACTC
TTAATTTCAAGACACTCAAGGCCAAGACAGTGACATCCAGGAAGACATGAAACAAGGGGTTTTTAGTAGCC
AAC**TCGC**TACATGAACAAGCTAAACGTTAAATCAAAGCAAAGCATAGTGGCCCTATTAGCATCCTAGGACC
ATGTCTGGGAACCCTTACCAGGTTTAAGAGAACATTATTCCTTCCCATGAAACCACCCTATAGTTTCATCA
CCGTTTAGTAAGTGGCTTTCTCTGGAACCTGAAGCTATAATAGCATTATTTCTACTGGGATTCACTTGGAA
AGGGGCCTGGAGTTTCTCCCAAGTAGACAGCCTTTCTAGCAAGCTCCAGGTCACTAGGCAGAGAGGCCAAC
TGCCTATGTGTGGCCTGTTTGGCTCAAGCTTCCTTTAAAGAAGGTAATATAGGATGGGGTTTCCACAGCTG
GGAGACTAGACTGTGACCCTGATCC**TCTC**TGAGGCCAGGCACAACATGACAATGGAGATTCTTCCAGGCAA
GAGGTGACCAGAGAACCACTGATGGTGATGGAGAGACCACCTGAGCCATGGGTGGGCAGGAGGCTGGTCAG
AGTCCTGACCATGCTCACTCCTGACTCTGCCAGCACCCCCAATAGGTCGTTCATTGGAGGGCAAATAGAAA
GGGTTTCTTGGTCTTGGGCAAAGTGTGGCTGAAGGCAGGCAAAGTGGGTGGAGTCTCTGGACCTCCCAGGC
CACAGGAAGGGCTATGCCCCATGGGAA**GACA**GACAGATGAGCCAGAGGCTCTGGCCAGGATGGACATTACT
CCAGGGTGGAGGAAACACACTCCATGAGCCACACCAGATTGGGCCCAGTGGGAGGTCTCAGGGCTACAGGC
GG**GCGA**GGAGGAAGTTGTTGGGAAGTCAGACCGGAATGGCATGGTGGAGGGAGAACCGGCAACCAGGCAGA
TGGTTCCTGACGAGGAAGCTCTGGGCACAGCTGCAGGCCCCCGACCCCCACCGCAT**TGGC**ACTGTGACACA
GCTGGCTGCCTCACCCCTGAAGGCTGCAGGAGGACCTCCCCCATGCTGTCCCCAAGCCCGCCCCGTGTCAC
ATGAGGCCCTGCAGACTCCCACCCTCCGTCCAGGGCCACAACCCAGCTCTGCCGGCTGTAGTGACTAGAAA
GGCCTGGAGCCTCCAAGGAACAGAGGCACGGGCTCCGAGACGCCAAAGCTCCTCCAGCCCTCCTGT*AGATC*
*T*AAGTAAGCTTGGCATTCCGGTACT

Nfe2 enhancer (restriction enzymes SacI/XhoI): chr15:103258245-103258850

AGCTAACATAACCCGGGAGGTACC*GAGCTC*TGGGAAAATCC**TATC**CACATGTAAACTGGAACACAAGGAAA
ATAACCGATGACTCTGGAGATCTGACTCCAC**TATC**TAGCAAAGTTTTTACTTTATACTACCCCCACTCCCG
ACTTCATCAGGGGAGCGTGAGTATTCCTGGGTCCCAGGCGTCCTTCACCACCACCCCCATCCGGGGCAACC
GCCCTGCTCTGCTGCTTTG**GATA**ACACCGGGCCCTCCCCCTATTCCCCCTGTGGCTGCCTCCCCCTTCCGT
CTGTTGAGAGAGGAAGCCAGGGGGTGGCGGGTGCAATGCTGTGGGGCACT**GATA**AAAGGCCAGTAC**TATC**C
CCGCCCTCTGGGGCCACTGCGGTCACACCAGTAGGCAATCCAGCAAGGCAGCCAGTTCCCTGTGGGACCCA
TGGCCCTCCCCTGGTTCCACCTCTAGCCACCCCGCCCTGCTCACCCTTCTCGGGAAGCTGGTTGCATAACC
CAGTGGGGTGTTTGGCAACAATGCTTGTGGCTTGACCTGATGCTGCTGGTGGTGTGCACATACGTAGTGGA
GGTGGGTTGGACTGGGGTGACAGGTTAACTATTTAGGGGTTGGGTGAGCAGCAAAAGTGGAAAATGTT*CTC*
*GAG*ATCTGCATCTCAATTAGTCAGCAA

Nfe2 enhancer - mutated (restriction enzymes SacI/XhoI)

AGCTAACATAACCCGGGAGGTACC*GAGCTC*TGGGAAAATCCGGTCCACATGTAAACTGGAACACAAGGAAA
ATAACCGATGACTCTGGAGATCTGACTCCACTGTCTAGCAAAGTTTTTACTTTATACTACCCCCACTCCCG
ACTTCATCAGGGGAGCGTGAGTATTCCTGGGTCCCAGGCGTCCTTCACCACCACCCCCATCCGGGGCAACC
GCCCTGCTCTGCTGCTTTGGACAACACCGGGCCCTCCCCCTATTCCCCCTGTGGCTGCCTCCCCCTTCCGT
CTGTTGAGAGAGGAAGCCAGGGGGTGGCGGGTGCAATGCTGTGGGGCACTGACGAAAGGCCAGTACTGTCC
CCGCCCTCTGGGGCCACTGCGGTCACACCAGTAGGCAATCCAGCAAGGCAGCCAGTTCCCTGTGGGACCCA
TGGCCCTCCCCTGGTTCCACCTCTAGCCACCCCGCCCTGCTCACCCTTCTCGGGAAGCTGGTTGCATAACC
CAGTGGGGTGTTTGGCAACAATGCTTGTGGCTTGACCTGATGCTGCTGGTGGTGTGCACATACGTAGTGGA
GGTGGGTTGGACTGGGGTGACAGGTTAACTATTTAGGGGTTGGGTGAGCAGCAAAAGTGGAAAATGTT*CTC
GAG*ATCTGCATCTCAATTAGTCAGCAA

2.3.4.5. Plasmid List

Table 2.7. List of plasmids used as backbones for various applications.

Plasmid Name	Source	Chapter
pGL2-Basic Vector	Göttgens Lab	6 (Luciferase assays)
pGL2-Promoter Vector	Göttgens Lab	6 (Luciferase assays)
pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W	Addgene #67976	4 (gRNA cloning)
pMD2.G	Göttgens Lab	4 (retrovirus production)
ΔR.89	Göttgens Lab	4 (retrovirus production)

2.4. HSC retroviral transduction analysis: genotyping

2.4.1. Genotyping primer design

Primers were designed using Primer3 (Koressaar and Remm 2007; Untergasser et al. 2012). The gene sequence was looked up on Ensembl (Zerbino et al. 2018) and visualized in ApE Plasmid Editor. The guide sequences were identified in the gene and enclosed in square brackets in Primer3, to ensure they were included in the primer sequence. Table 2.8 shows the overhangs used in the primer design. The order in which the primers were used was determined by their proximity to the cleavage site. Primers were designed for all guides but only the ones listed in Table 2.9 were used. Specifications for primer design included:

- The primer closest to the cleavage site must be between 70-200bp away from the cleavage site
- Total product size including overhangs should be between 400-600bp
- Avoid big differences in primer T_m within primer pairs by setting it between 56-63°C in Primer3

- Avoid long stretches of single bases
- Avoid big differences in GC% within primer pairs

Table 2.8. Primer overhang sequences

Proximity to cleavage site	Overhang sequence (5' - 3')
Closest	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Furthest	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Table 2.9. Genotyping primers

Primer Name	Sequence (5' - 3')
Procr_sg1_geno1_L	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGTCCGATTGCAGACC TCAGTT
Procr_sg1_geno1_R	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGGGAGGATGGTGACGT TTTGG
Trim47_sg3_geno1_L	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGAGGCCTCTGAAATCACCACA
Trim47_sg3_geno1_R	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGTGCGCTGTTCCTCTTGTCTA
Wfdc2_sg3_geno1_L	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGTTTAGGACCGAGCGAAGGAG
Wfdc2_sg3_geno1_R	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGTTCCTGGTCCCTCTTTGTCC
Gbp8_sg1_geno1_L	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGCTTCACAGGCATAGCTCCCT
Gbp8_sg1_geno1_R	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGCCTCCTCCTACCTTTTCCACA

2.4.2. Isolating gDNA

Genomic DNA was isolated from sorted GFP^+ cells using the QIAamp DNA Micro Kit (Qiagen) following the suggested protocol. The carrier RNA diluted in Buffer AE was always added to Buffer AL to improve yield. The gDNA was eluted into 30µl water and stored at -4°C.

2.4.3. Testing genotyping primers

All genotyping primers used were first tested using HoxB8-FL gDNA, which was isolated by Iwo Kucinski.

To test the primers, the following PCR reaction mixture was prepared on ice: 1.25 µl of the forward and reverse primers (10mM), 5 µl of 5X Phusion® High-Fidelity Reaction Buffer (NEB), 0.5 µl of 10 mM dNTP solution (NEB), 0.25 µl of Phusion ® High-Fidelity Polymerase (NEB), 10-20 ng/µl of the gDNA template and water up to a 25 µl total reaction volume. The following PCR program was used: 98℃ for 30s → 30 cycles of 98℃ for 10s, 59℃ for 20 s, 72℃ for 20 s → 72℃ for 7 min. The product was then checked on an 0.9% agarose gel.

A band corresponding to the expected product size was seen for all primers, indicating that all the primers worked at an annealing temperature of 59℃.

2.4.4. Genotyping protocol

The same PCR as described in Section 2.4.3 was used for genotyping the gDNA of interest. After the product was checked on a 0.9% agarose gel, a second PCR reaction was set up to anneal the indexing primers (Nextera XT Index Kit v2 Set A, Illumina) to the previous PCR product. 1 µl of the forward and reverse primers were mixed with 10 µl of KAPA HiFi HotStart PCR Mix (KAPA Biosystems, Roche), 2 µl of template from the previous PCR, and 6 µl of water for a total reaction volume of 20 µl. The following PCR program was then used: 98℃ for 3 min → 20 cycles of 98℃ for 20s, 55℃ for 15s, 72℃ for 1min → 72℃ for 5 min. The product was again checked on an 0.9% agarose gel. After confirming that the product was as expected, it was cleaned up using Agencourt AMPure XP Beads (Beckman Coulter). 4 µl of each sample was pooled in a 1.5 ml Eppendorf tube. The beads were added to the sample in a 1:0.7 ratio (sample: beads). The mixture was incubated at room temperature for 5 minutes and then placed on a magnetic stand (Invitrogen) for 2 minutes. The supernatant was carefully removed, and the beads were washed twice with freshly prepared 80% ethanol. Once the beads were dry, 20 µl of EB was added to each tube. The samples were vortexed and incubated at room temperature for 2 minutes before being placed on the magnetic stand for 2 minutes. The entire volume of supernatant was then transferred to another 1.5 ml Eppendorf tube without disturbing the remaining beads. The product was run on the BioAnalyzer system (Agilent Genomics) using the Agilent High Sensitivity DNA Kit (Agilent Genomics) to visualize the product size and quality. Finally, the product was quantified using the KAPA Library Quantification Kit (KAPA Biosystems, Roche) following the suggested protocol. The quantified sample was diluted to the appropriate concentration (10 nM) and sent for sequencing at the Genomics Core, CRUK CI (University of Cambridge) on the MiSeq Nano System (Illumina).

2.5. Single Cell Gene Expression Analysis

2.5.1. Single cell gene expression analysis (Fluidigm BioMark™ HD)

The protocol for isolating single bone marrow HSPCs is detailed above (Section 2.2.1). Cell processing for single cell gene expression analysis was performed as previously described (Moignard et al. 2013). TaqMan® assays (Table 2.11, Applied Biosystems) in TE Buffer (Life Technologies) were pooled in a 1:100 dilution of each assay, constituting a 0.2X TaqMan® assay mix. This mix was aliquoted and stored at -20°C.

Single cells were sorted directly into a 96-well PCR plate by FACS. Each well contained 5 µl of 2X Reaction Mix (CellsDirect One-Step qRT-PCR kit, Life Technologies), 0.1 µl SUPERase RNase Inhibitor (Ambion), 2.5 µl of the 0.2X assay mix, 1.2 µl TE buffer and 1.2 µl Superscript III/Platinum Taq (CellsDirect One-Step qRT-PCR kit, Life Technologies) for a total volume of 10 µl per well. After sorting, the plates were vortexed then centrifuged at 700 g for 2 minutes at 8°C. The plates were stored at -80°C.

2.5.1.1. Specific Target Amplification

Reverse transcription and preamplification were performed using the conditions listed in Table 2.10. Victoria Moignard previously determined the optimum number of preamplification cycles for haematopoietic cells to be 22 cycles, which brings the gene expression within the dynamic range of the Fluidigm BioMark™ HD platform. The cDNA was then stored at -20°C.

Table 2.10. Thermocycler conditions for synthesis and specific target amplification of cDNA from single cells.

Step	Temperature (°C)	Time	Cycles
cDNA synthesis	50	15 min	---
Inactivation of SuperScript III/ Activation of Platinum Taq	95	2 min	---
Specific target amplification	95 60	15 s 4 min	22
Hold	4	∞	---

2.5.1.2. qRT-PCR on the Fluidigm BioMark™ HD platform

After preamplification, qRT-PCR was performed on the Fluidigm BioMark™ HD platform using a 48:48 Dynamic Array integrated fluidics chip (Fluidigm). The cDNA was diluted 1:5 in TE buffer.

A 96-well plate was used to prepare the reagents for loading onto the 48:48 Dynamic Array. On one half of the plate, 2.7 µl of the diluted cDNA was mixed with 3 µl of a TaqMan® Universal Mastermix (Applied Biosystems) and 0.3 µl Gene Expression Sample Loading Reagent (Fluidigm). Each well contained a different sample. On the other half of the plate, 3 µl of each FAM-labelled TaqMan® assay was mixed with 3 µl Gene Expression Assay Loading Reagent (Fluidigm). Each well contained a different assay. 4.5 µl of each assay or sample was loaded into individual assay or sample inlets on the 48:48 Dynamic Array. Samples and assays were loaded into integrated fluidics chip using the IFC Controller MX (Fluidigm). The 48:48 Dynamic Array was then transferred to the Fluidigm BioMark™ HD for qRT-PCR. The following qRT-PCR program was used: 95°C for 10 minutes → 40 cycles of 95°C for 15 seconds, 60°C for 1 minute.

Table 2.11. List of TaqMan® assays used for single cell gene expression analysis.

Gene name	Assay ID	Gene name	Assay ID
Bptf	Mm01251151_m1	*Ldb1*	Mm00440156_m1
Cbfa2t3h	Mm00486780_m1	*Lmo2*	Mm01281680_m1
Cdkn2a	Mm00494449_m1	*Lyl1*	Mm01247198_m1
Csf1r	Mm01266652_m1	*Mecom*	Mm01289155_m1
Dnmt3a	Mm00432881_m1	*Meis1*	Mm00487659_m1
Egfl7	Mm00618004_m1	*Mitf*	Mm01182480_m1
Eif2b1	Mm01199614_m1	*Mpl*	Mm00440310_m1
Erg	Mm01214246_m1	*Myb*	Mm00501741_m1
Ets1	Mm01175819_m1	*Nfe2*	Mm00801891_m1
Ets2	Mm00468977_m1	*Nkx2-3*	Mm01199403_m1
Etv6	Mm01261325_m1	*Notch1*	Mm00435249_m1
Fli1	Mm00484409_m1	*Pbx1*	Mm04207617_m1
Gata1	Mm00484678_m1	*Polr2a*	Mm00839493_m1
Gata2	Mm00492300_m1	*Prdm16*	Mm00712556_m1
Gata3	Mm00484683_m1	*Procr*	Mm00440992_m1
Gfi1	Mm00515855_m1	*Runx1*	Mm01213405_m1
Gfi1b	Mm00492318_m1	*Spi1*	Mm00488142_m1
Hhex	Mm00433954_m1	*Sh2b3*	Mm00493156_m1
HoxA5	Mm00439362_m1	*Smarcc1*	Mm00486224_m1
HoxA9	Mm00439364_m1	*Tal1*	Mm01187033_m1
HoxB4	Mm00657964_m1	*Tcf7*	Mm00493445_m1

Gene name	Assay ID	Gene name	Assay ID
Ikzf1	Mm01187882_m1	*Tet2*	Mm00524395_m1
Itga2b	Mm00439768_m1	*Ubc*	Mm01201237_m1
Kit	Mm00445212_m1	*Vwf*	Mm00550376_m1

2.5.2. Single cell gene expression analysis (single-cell RNA sequencing)

The protocol for isolating single bone marrow HSPCs is detailed above (Section 2.2.1). scRNA-seq analysis was performed as described previously (Picelli et al. 2014). All centrifugation steps occurred at 8℃ for 1 minute at 700 g. Volumes for all mixtures mentioned are for a single 96-well plate, unless otherwise stated.

2.5.2.1. Single cell lysis

Single cells were sorted by FACS directly into 96-well PCR plates. Each well contained 2.3 µl lysis buffer. The lysis buffer contained 1 µl of SUPERase RNase Inhibitor (Invitrogen) to 19 µl of 0.2% (vol/vol) Triton X-100 (Sigma). Once the cells were sorted, the plate was vortexed and spun down. The plates were stored at -80℃ for up to 6 months.

2.5.2.2. Reverse transcription and preamplification

ERCC (External RNA Controls Consortium) RNA Spike-In Mix (Invitrogen) was diluted 1:300,000 in water containing SUPERase RNAse Inhibitor. 10 µl of the diluted ERCCs were mixed with 10 µl of 100 µM oligo-dT (Table 2.14), 100 µl of 10 mM dNTP mix (ThermoFisher), and 80 µl of water to make up the annealing mixture. Once the plate of sorted cells had thawed on ice, 2 µl of the annealing mixture was added to each well. The plate was centrifuged before being incubated for 3 minutes at 72℃.

A reverse transcription (RT) mix was prepared with the following reagents: 50 µl of Superscript II RT (Invitrogen), 200 µl of 5X Superscript II First Strand Buffer (Invitrogen), 50 µl of 100 mM DTT (Invitrogen), 25 µl of SUPERase RNAse Inhibitor, 200 µl of 5M Betaine (Sigma), 6 µl of 1 M $MgCl_2$ (Ambion), 10 µl of 100 µM TSO (Table 2.14) and 29 µl of water. 5.6 µl of the reverse transcription mixture was added to each well and the plate was centrifuged before being transferred into the thermocycler. RT was performed using the conditions detailed in Table 2.12.

Table 2.12. Thermocycler conditions for reverse transcription (Smart-seq2* protocol).

Step	Temperature (°C)	Time (min)	Cycles
RT and template switching	42	90	---
Unfolding of RNA secondary structures	50	2	10
Completion/continuation of RT and template switching	42	2	
Enzyme inactivation	70	15	---
Hold	4	∞	---

*(Picelli et al. 2014)

A PCR preamplification mix was made with the following reagents: 1250 µl of 2X KAPA HiFi HotStart Ready Mix (KAPA Biosystems), 25 µl of 10 µM IS PCR primer (Table 2.14) and 225 µl of water. After reverse transcription, 15 µl of the PCR preamplification mix was added to each well. The plate was centrifuged and preamplification was performed using the conditions detailed in Table 2.13.

Table 2.13. Thermocycler conditions for preamplification (Smart-seq2* protocol).

Step	Temperature (°C)	Time	Cycles
Denature	98	3 min	---
	98	20 sec	21
Anneal	67	15 sec	
Extend	72	6 min	
	72	5 min	---
Hold	4	∞	---

*(Picelli et al. 2014)

Table 2.14. Oligo sequences.

Oligo	Source	Sequence
TSO (LNA oligo)	Exiqon	AAGCAGTGGTATCAACGCAGAGTACATrGrG+G
Oligo-dT30VN	Biomers.net	AAGCAGTGGTATCAACGCAGAGTAC(T30)VN
IS PCR	Biomers.net	AAGCAGTGGTATCAACGCAGAGT

All oligos are HPLC purified.

After preamplification, the plates of cDNA were stored at -20°C. The cDNA was cleaned up using the Beckman Coulter Biomek FX^P (Beckman Coulter). A 1:0.8 ratio of sample to Agencourt AMPure XP beads was used for PCR purification. After being washed with 80% ethanol, the samples were eluted in 22 µl of EB buffer. 20 µl of the supernatant was collected and transferred to a new 96-well plate. The cDNA library was then checked for quality and size distribution on the BioAnalyzer system using an Agilent High Sensitivity DNA Kit.

2.5.2.3. Library preparation

The following steps were performed using the Nextera XT DNA Library Preparation Kit (Illumina) and the Nextera XT 96-Index Kit (384 samples, Illumina). All reagents, unless otherwise mentioned, are contained in this kit. The protocol is based on the Tagmentation protocol from Fluidigm.

A pre-mix was made using 264 µl of Tagmentation DNA Buffer (warmed to room temperature) and 132 µl Amplicon Tagment Mix. 3.75 µl of the pre-mix was added to each well of a fresh 96-well plate ("library prep" plate). 1.25 µl of each sample from the cDNA library plate was then added to individual wells of the "library prep" plate. The plate was then centrifuged and incubated at 55°C for 10 minutes, after which 1.25 µl of NT buffer was immediately added to each well to neutralize the samples.

Once the plate was centrifuged again, 3.75 µl of the Nextera PCR Master Mix was added to each well. 1.25 µl each of Index Primer 1 (N701-N712) and Index Primer 2 (S517, S502-S508) were added to each well, creating unique combinations of the indexes in each well. It was essential to know the order of the indexes for data analysis.

After the addition of the indexes, the plate was centrifuged before being transferred into the thermal cycler for PCR amplification (Table 2.15). After amplification, the plates could be stored at -20°C long-term.

Table 2.15. Thermocycler conditions for amplification of cDNA libraries.

Temperature (°C)	Time	Cycles
72	3 min	---
95	30 sec	---
95	10 sec	
55	30 sec	12
72	60 sec	
72	5 min	---
10	∞	---

For clean-up, 2 µl of sample from each well was pooled together into one 1.5 ml Eppendorf tube and the total pooled volume was measured. Agencourt AMPure XP beads were added at 0.9% of the total pooled library volume. The mixture was incubated at room temperature for 5 minutes and then placed on a magnetic stand (Invitrogen) for 2 minutes. The supernatant was then carefully

removed, and the beads were washed twice with freshly prepared 80% ethanol. Once the beads were dry, 50 µl of EB was added to each tube. The samples were vortexed and incubated at room temperature for 2 minutes before being placed on the magnetic stand for 2 minutes. The entire volume of supernatant was then transferred to another 1.5 ml Eppendorf tube without disturbing the remaining beads.

The library size distribution was checked using the BioAnalyzer system with an Agilent High Sensitivity DNA Kit. The library was then quantified using the KAPA Library Quantification Kit and diluted to the appropriate concentration (10 nM). Libraries were sequenced at the Genomics Core, CRUK CI (University of Cambridge) using the Illumina HiSeq 2500 or Illumina HiSeq 4000 system (single-end 125bp reads).

2.6. Computational Analysis

2.6.1. Single cell gene expression data analysis (Fluidigm BioMark™ HD)

All analysis of qPCR data from the Fluidigm BioMark™ was performed using R (www.r-project.org). Most scripts used were written by Fiona Hamey or Victoria Moignard and Fernando Calero-Nieto, and Fiona Hamey and Wajid Jawaid helped with some coding aspects. All analyses were carried out by Sonia Shaw unless otherwise stated. Fiona Hamey performed the pseudotime inference, network construction and stable state analyses. Specific details are provided in Chapters 5 and 6.

2.6.1.1. Data processing and filtering

Single-cell gene expression data was collected using Fluidigm Data Collection software and analysis was performed as previously described (Moignard et al. 2013; N. K. Wilson et al. 2015). ΔCt values were calculated by normalising mean expression levels to housekeeping genes *Ubc* and *Polr2a* (G. Guo et al. 2010). Where a gene could not be detected, the maximum ΔCt value for a gene/assay was calculated and 3.5 was added.

2.6.1.2. Downstream analyses

All housekeeper genes (*Ubc*, *Polr2a*, *Eif2b1*), *Cdkn2a*, *Egfl7*, *Gfi1*, and *Spi1* were removed from the dataset for downstream analysis. *Cdkn2a* was not expressed in any of the cell types, and *Egfl7*, *Gfi1* and *Spi1* were removed due to technical issues.

The data collected for FSR-HSC2, MPP and PreMegE cells in this investigation were projected onto a principal component analysis (PCA) plot together with data collected by Wilson *et al.* for the following populations: LMPPs, CMPs, GMPs, MEPs, FSR-HSCs, and HSCs. The data were also re-analysed with data from Wilson *et al.* (N. K. Wilson et al. 2015). Since the projected PCA plot and the re-analysed PCA plot showed similar correlations between the cell populations, analysis was continued using the re-analysed data set containing all 12 populations.

Hierarchical clustering was performed using the *hclust* function and heatmap.2 from the *gplots* package. Spearman rank correlations and ward linkage were used. PCA was performed using the default settings for the *prcomp()* function. T-distributed Stochastic Neighbour Embedding (t-SNE) analysis was performed using the *tsne* package. Diffusion maps dimensionality reductions were calculated using the *destiny* package using centred cosine distance and $\sigma = 0.3$ (Angerer et al. 2016). Cells were retrospectively coloured based on clusters or the population to which they belonged. Subsequent analyses in which MolO cells were projected onto the data were performed using the *roots* package.

2.6.1.3. Additional information

Single cell gene expression data were also collected for HoxB8-FL cells. These data were processed by Fiona Hamey. All the single cell gene expression data from the Fluidigm BioMark™ platform can be downloaded from:

http://blood.stemcells.cam.ac.uk/single_cell_qpcr.html

2.6.2. Single cell gene expression data analysis (scRNA-seq)

All analysis of the scRNA-seq data was performed using R (www.r-project.org) unless otherwise stated. The script used for the analysis was written by Fiona Hamey. All analyses were carried out by Sonia Shaw unless otherwise stated.

2.6.2.1. Aligning reads and quality control

Reads were aligned using G-SNAP (T. D. Wu and Nacu 2010) and the mapped reads were assigned to Ensembl genes (release 81) (Zerbino et al. 2018) by HTSeq (Anders, Pyl, and Huber 2015). This was done by Evangelia Diamanti.

Quality control and data normalization was performed by Fiona Hamey. To pass quality control, cells had to meet the following requirements:

- Cells need to have at least 200,000 reads mapping to nuclear genes
- Cells need to have at least 4,000 genes detects
- Less than 10% of mapped reads should map to mitochondrial genes
- Less than 50% of mapped reads should map to ERCC spike-ins.

The reads were normalised using the method of Lun *et al.* (Lun, Bach, and Marioni 2016). Technical variance was estimated using ERCC spike-ins, as described by Brennecke *et al*. (Brennecke et al. 2013). The data were normalized in R using *flowCore* (Hahne et al. 2009) and *ComBat* (Johnson, Li, and Rabinovic 2007).

2.6.2.2. Assigning population thresholds

Population thresholds were assigned retrospectively by comparing normalised index data with published literature (A. Wilson et al. 2008; Pronk et al. 2007; Pietras et al. 2015; Cabezas-Wallscheid et al. 2014). The index data was plotted in FlowJo (Treestar) and gated to define HSPC, MPP and progenitor populations. CD45 was not available in the index data; therefore, E-SLAM cells were gated using the following strategy: $EPCR^+$ $CD48^-$ $CD150^+$. The set gates either covered all cells (broad gating) or left unclassified cells in between populations to prevent overlap between gates (narrow gating).

2.6.2.3. Downstream analyses

Hierarchical clustering was performed using the *hclust* function with average linkage and (1 – Spearman's correlation)/2 distance. Clusters were identified using the *cutreeDynamic* function from the *dynamicTreeCut* package using a minimum cluster size of 10 and the deepSplit parameter set at 1. Wilcoxon rank sum tests with Benjamini-Hochberg correction tested for differential expression in genes expressed in at least half of the cells in a cluster. Diffusion maps dimensionality reductions were calculated using the *destiny* package with cosine distance and Gaussian kernel width = 0.3 (Angerer et al. 2016). Pseudotime analysis was performed by Fiona Hamey.

2.6.2.4. Haematopoietic differentiation landscape – Online resource

An interactive website was designed by Blanca Pijuan-Sala, which allows other researchers to view the expression of their genes of interest on the HSPC differentiation landscape. The website also contains the surface marker and cell phenotype visualisations. The raw data was also made available for others to use the HSPC differentiation atlas in their research. This interactive website can be found following this link:

http://blood.stemcells.cam.ac.uk/single_cell_atlas.html

2.6.2.5. STREAM analysis

STREAM analysis (Single-cell Trajectories Reconstruction, Exploration and Mapping) was performed on the scRNA-seq dataset by Huidong Chen from the Pinello lab at Harvard University (H. Chen et al. 2018). The data was made available at:

http://stream.pinellolab.org/

Within this online resource, Sonia Shaw visualized the expression of various genes and made observations about genes involved in branching and transitioning points in the pseudotime ordering.

2.6.2.6. SPRING analysis

SPRING analysis was performed on the scRNA-seq dataset by Caleb Weinreb from the Klein lab at Harvard University (Weinreb, Wolock, and Klein 2018). The data was made available at:

https://kleintools.hms.harvard.edu/tools/springViewer.html?cgi-bin/client_datasets/gottgens_prenorm

Within this online resource, Sonia Shaw visualized the expression of various genes and proteins and made observations about the discernible cell populations.

2.6.3. Analysis of genotyping data

Genotyping results were analysed in R studio. The genotyping protocol is described in Section 2.4. Evangelia Diamanti mapped the reads to the mouse genome and generated .fastq files for each library sequenced using MiSeq Nano. The first 10,000 reads of each .fastq file were aligned to the appropriate reference sequence using custom functions created by Iwo Kucinski. The output of the script included the fraction of indels and frameshift mutations for each sample, which were compared to the empty vector controls to determine whether the CRISPR gRNAs successfully targeted the genes of interest.

Chapter 3: A single-cell atlas of adult murine haematopoiesis

Parts of this section have been modified from Nestorowa *et al.* (2016). Isolation of primary mouse bone marrow cells and scRNA-seq analysis were carried out by Sonia Shaw. Blanca Pijuan-Sala performed index data normalisation and designed the interactive website associated with the publication (http://blood.stemcells.cam.ac.uk/single_cell_atlas.html). Sonia Shaw performed the computational analysis described in this chapter, except for the pseudotime ordering on the diffusion map, which was performed by Fiona Hamey.

3.1. Background

The haematopoietic system is formed through a series of cell-fate decisions, in which haematopoietic stem cells (HSCs) differentiate towards multiple mature cell types through dynamic gene expression. It is necessary to study populations at various stages of maturation together to gain insights into cell-fate decision making. Researchers have developed increasingly pure isolation strategies for haematopoietic stem and progenitor cells (HSPCs), providing many significant advances for the haematopoietic community (Beerman et al. 2010; D. G. Kent et al. 2009; Kiel et al. 2005; Morita, Ema, and Nakauchi 2010; Challen et al. 2010). Purification protocols require the use of increasingly restrictive gates to optimise purity. These strict gates exclude 'contaminating cells,' which may in fact be transitional cells moving from one cellular state to another.

Bulk-expression studies have helped improve our understanding of the haematopoietic system; however, it is a very heterogeneous system and bulk-expression profiling will only capture average expression states within cells, inadequately representing individual cells. Single-cell analysis is therefore required to resolve heterogeneity within HSPC populations. Advances in flow cytometry have allowed for the collection of surface marker expression data for single-cells during fluorescence-activated cell sorting (FACS) (Osborne 2011). The surface marker expression can be coupled with single-cell gene expression profiling, making it possible to use broader sorting strategies to capture a wider range of cell types while retaining the ability to classify cells into well-defined haematopoietic populations (Schulte et al. 2015).

Recently, over 2,700 cells were profiled using massively parallel single-cell RNA sequencing (MARS-seq) coupled with index sorting to investigate heterogeneity within the common myeloid progenitor (CMP) population. The investigation spanned CMPs, granulocyte-monocyte progenitors (GMPs), and megakaryocyte-erythroid progenitors (MEPs), and collected surface marker and gene expression data to create a reference mouse model of haematopoiesis (Paul et al. 2015). While this study demonstrates the power of index-sorting coupled with FACS, it is restricted by sequencing depth, limiting the information gained from the dataset. In fact, although single-cell gene expression studies of HSPCs were published prior to this investigation, they were limited by inadequate coverage of the haematopoietic hierarchy, or were restricted in the number of genes profiled when using quantitative real-time PCR (qRT-PCR) (Grover et al. 2016; Moignard et al. 2013; Paul et al. 2015; Kowalczyk et al. 2015; Wilson et al. 2011, 2015; Hamey et al. 2016). Where the number of genes quantified was restricted, bias would be introduced as the investigator selected the genes to profile, potentially missing dynamic genes and limiting the opportunity for discovering novel gene expression patterns. As such, at the time of this investigation, a single-cell HSPC transcriptomic atlas that covered all cell types in early haematopoiesis was not available, and the creation of such an atlas would be a powerful resource for investigating gene expression changes during differentiation.

3.1.1. Aims

The aims of this chapter were to:

- Generate a comprehensive atlas of murine HSPC differentiation at the single-cell level
- Interrogate the HSPC atlas to better understand differentiation trajectories

These aims were addressed by profiling over 1,600 HSPCs from mouse bone marrow using single-cell RNA sequencing (scRNA-seq). Cells were isolated using broad sorting strategies to capture cells that may represent important transitional stages in haematopoietic differentiation, normally excluded by narrow gating strategies. The scRNA-seq profiles were used to visualise differentiation from HSCs to more mature haematopoietic progenitor cells. Using various analysis methods, the data was interrogated to show gene and surface marker expression trends, as well as the breadth of information that can be obtained from transcriptomic profiling.

3.2. Isolation of haematopoietic stem and progenitor cells for single-cell analysis

To comprehensively sample cells across the HSPC transcriptional landscape, cells were isolated from mouse bone marrow. The cells were sorted by single-cell FACS coupled with index sorting, which collects the surface marker expression data for every cell sorted (Fig. 3.1A). Two broad gates based on c-Kit and Sca1 expression were used to capture cells from the HSPC gate (Lin^- $Sca1^+$ $c\text{-}Kit^+$) and Progenitor (Prog) gate (Lin^- $Sca1^-$ $c\text{-}Kit^+$). The HSPC gate encompassed long-term HSCs (LT-HSCs), finite self-renewal HSCs (FSR-HSCs), multipotent progenitors (MPPs), and lymphoid multipotent progenitors (LMPPs), whereas the Prog gate included CMPs, MEPs and GMPs (Fig. 3.1B). LT-HSCs are much less frequent than the other cell types in the HSPC gate (Fig. 3.1C). LT-HSCs were sorted separately to ensure sufficient coverage of the population, using the Lin^- $Sca1^+$ $c\text{-}Kit^+$ $Flk2^-$ $CD34^-$ sorting strategy.

The cells were sorted into twenty 96-well PCR plates with HSPC, Prog and LT-HSC cells represented on each plate (Fig 3.1A) for a total of 1,920 cells (840 HSPC cells, 840 Prog cells, and 240 LT-HSCs). This layout was used to account for possible batch effects that may arise from only one cell type being represented per plate, and to decrease the number of cells of a particular cell type lost if there was a plate issue. The cells were processed for scRNA-seq as previously described (Picelli et al. 2014). Quality control (QC) removed empty wells and low-quality profiles, leaving a total of 1,654 cells suitable for further analysis (701 HSPC cells, 798 Prog cells, and 155 LT-HSCs). A high sequencing depth resulted in a median of over 8,600 genes per cell detected for each cell passing QC (Fig 3.1A). Technical variance analysis was performed to remove genes that were affected by technical noise or showed low variation across cell types (Brennecke et al. 2013). ERCC spike-ins were used to estimate technical variance, finding 4,290 genes with variance exceeding the estimated threshold. These highly variable genes were used for downstream analysis.

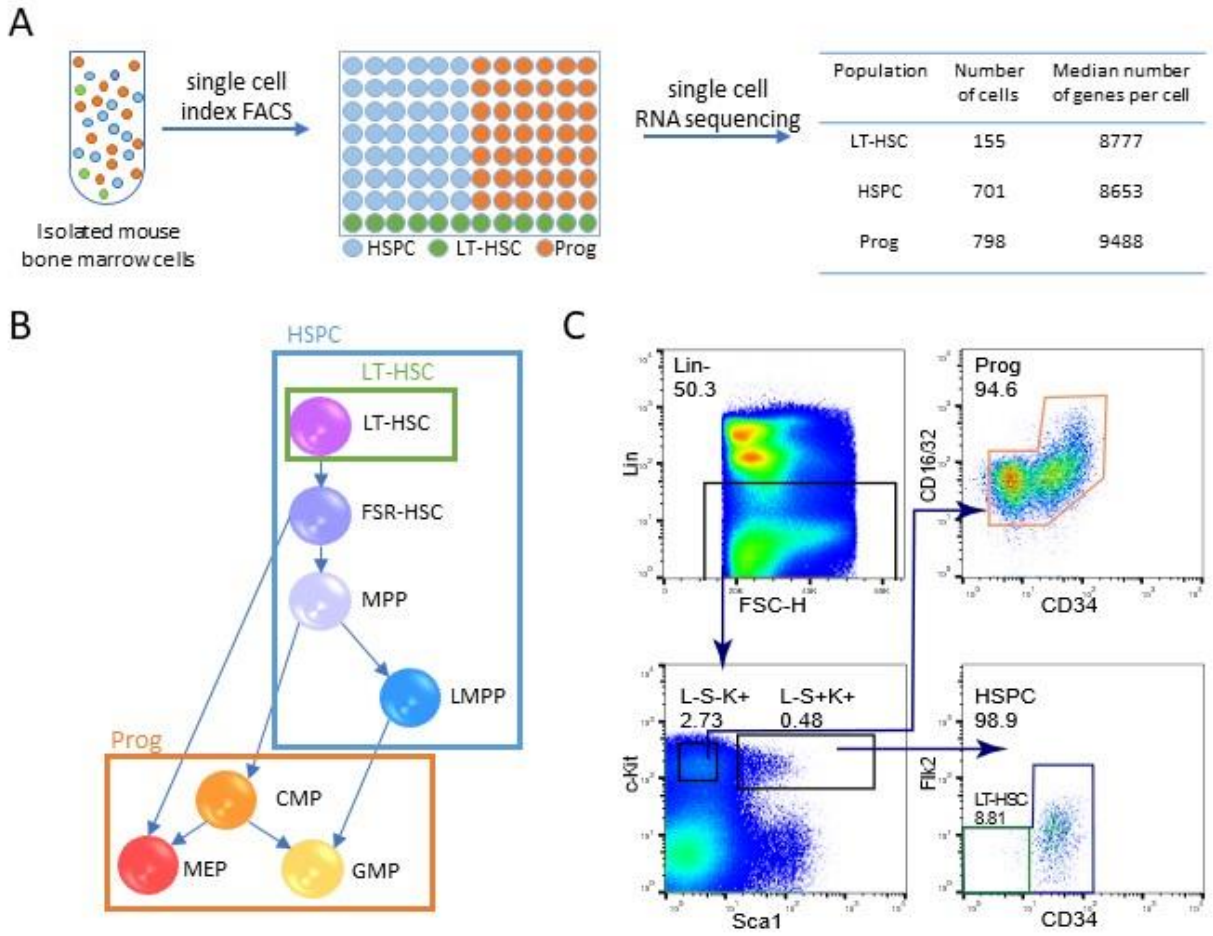


Figure 3.1. Isolation and profiling of HSPCs at the single cell level. (a) Schematic of the experimental design. Bone marrow cells isolated from mice were sorted using three broad gates based on surface marker expression. Index sorting data was collected during FACS. The cells were processed for scRNA-seq; the table summarises the number of cells and median number of genes detected for the three sorted populations after quality control. HSPC: Lin- Sca1+ c-Kit+; LT-HSC: Lin- Sca1+ c-Kit+ CD34- Flk2-; Prog: Lin- Sca1- c-Kit+. (b) Breakdown of the populations encompassed in the broad sorting gates, indicated by coloured boxes. (c) Flow cytometry plots showing the gating strategy used for FACS. Numbers above each gate indicate the percentage of those cells present in its parent population. Gate colours correspond to cell/well colour in Fig 3.1A, and box colour in Fig 3.1B. L-S-K+: Lin- Sca1-c-Kit+ (Prog parent gate); L-S+K+: Lin- Sca1+ c-Kit+ (LT-HSC and HSPC parent gate).

3.3. Single-cell gene expression analysis reveals distinct HSPC clusters

Unsupervised hierarchical clustering was performed using the expression of the 4,290 highly variable genes to investigate heterogeneity between the cell populations. Clustering allows the data to be grouped based solely on gene expression and can be used to gain insights about similarities and differences between groups of cells. Unsupervised learning does not require the input of any classifications by the investigator, and therefore finds commonalities within the data without the investigator's bias.

The clustering partitioned the 1,654 cells into 5 broad clusters (Fig. 3.2). Cell populations were retrospectively assigned based on traditional sorting strategies using the index data collected during sorting, which is further explained in Section 3.6. The cell types making up each cluster and their gene expression patterns can therefore be distinguished on the heatmap. Cluster 1 is made up of mainly LT-HSCs and is represented by genes such as *Procr*, *Trim47* and *F11r* (Sugano et al. 2008; Schulte et al. 2015; Gerrits et al. 2009). Cluster 2 contains mostly LT-HSCs and LMPPs, as well as a small proportion of FSR-HSCs, MPP3, and CMPs. Clusters 3 and 4 are both composed of all the cell types investigated and share expression of many of the representative genes. However, they are differentiated by higher expression of certain genes, such as *Ccl9*, *Clec12a*, and *Tyrobp* in Cluster 3, or *Apoe* in Cluster 4. Cluster 5 is mainly made up of MEPs and is characterized by the expression of erythroid genes such as alpha-haemoglobin (*Hba-a1)* (Stadhouders et al. 2015).

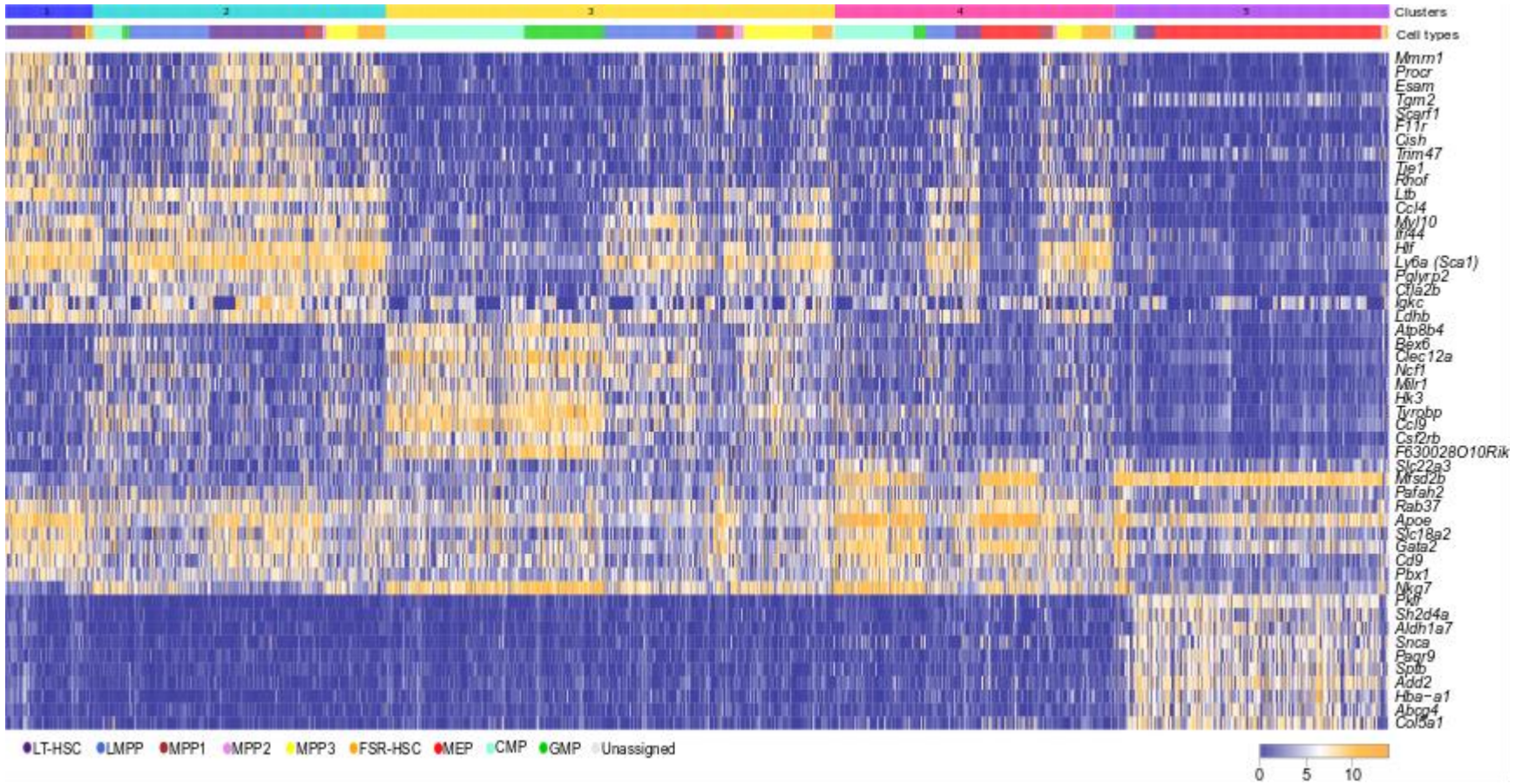


Figure 3.2. Unbiased hierarchical clustering reveals heterogeneity in the gene expression of HSPC clusters. Heatmap showing gene expression across clustered scRNA-seq data. Differential expression ranked genes by fold-change between cells within a cluster versus other clusters. Only the top 10 genes most specific for each cluster are displayed. Colour bars indicate the cluster identity and population identity for each cell. Cell types are coloured based on retrospective gating; the legend explains the colour-coding of cell populations: LT-HSC – purple; LMPP – blue; MPP1 – brown; MPP2 – pink; MPP3 – yellow; FSR-HSC – orange; MEP – red; CMP – light green; GMP – dark green. Grey cells represent cells unassigned to any cell population. Cluster colours: 1 – blue; 2 – turquoise; 3 – yellow; 4 – pink; 5 – purple

3.4. Transcriptional profiling reveals a differentiation continuum

Clustering is a useful tool to characterise gene expression patterns within a dataset but forces the data into discrete groups, which does not adequately represent the differentiation continuum occurring in haematopoiesis. To visualise the data as a continuum, dimensionality-reduction methods were implemented. These methods condense the high-dimensionality data into a low-dimensionality plot and emphasise key differences in the dataset.

Principal component analysis (PCA) was performed to visualise the distribution of the clusters in relation to each other based on their gene expression (Fig. 3.3A). These clusters were previously determined by unsupervised hierarchical clustering (Section 3.3). The PCA grouped Clusters 1, 2 and 4 together and separated Cluster 5 from the rest of the data as a more distinct cluster. The cells in Cluster 3 overlapped with Clusters 2 and 4, but also formed a projection in the data, indicating some differentiating gene expression patterns, consistent with the pattern seen on the heatmap (Fig. 3.2). While PCA is an informative method of dimensionality-reduction and pulls out variations in the data, it assumes the data has a linear structure and may therefore miss any non-linear patterns (Lever, Krzywinski, and Altman 2017). Furthermore, PCA can only take into consideration two or three principal components at a time, whereas other methods consider all components and try to plot them in two dimensions. As this scRNA-seq dataset has many dimensions and has a complex structure, PCA may be unable to capture all the lineage differentiation patterns in the dataset.

Diffusion maps have been successful as a non-linear dimensionality-reduction method to capture continuous differentiation processes from single-cell snapshot data (Coifman et al. 2005; Haghverdi, Buettner, and Theis 2015). The diffusion map method was applied to this dataset and the plot was visualised in the first three diffusion components after determining these components were optimal for showing the continuous nature of the data (Fig. 3.3B). The PCA and diffusion map both separated Clusters 3 and 5 from the remaining clusters. However, in the diffusion map, Clusters 1 and 2 also form a separate projection together. This gives a better resolution of the haematopoietic hierarchy, as Cluster 1 represents the most immature cells, the LT-HSCs, and Cluster 2 consists of LT-HSCs and early progenitors. Furthermore, the structure of the diffusion map suggests a continuum in which the cells of Cluster 1 give rise to three main trajectories, represented by Clusters 2, 3 and 5. The dataset was visualised using the diffusion map for all downstream analyses, unless otherwise stated.

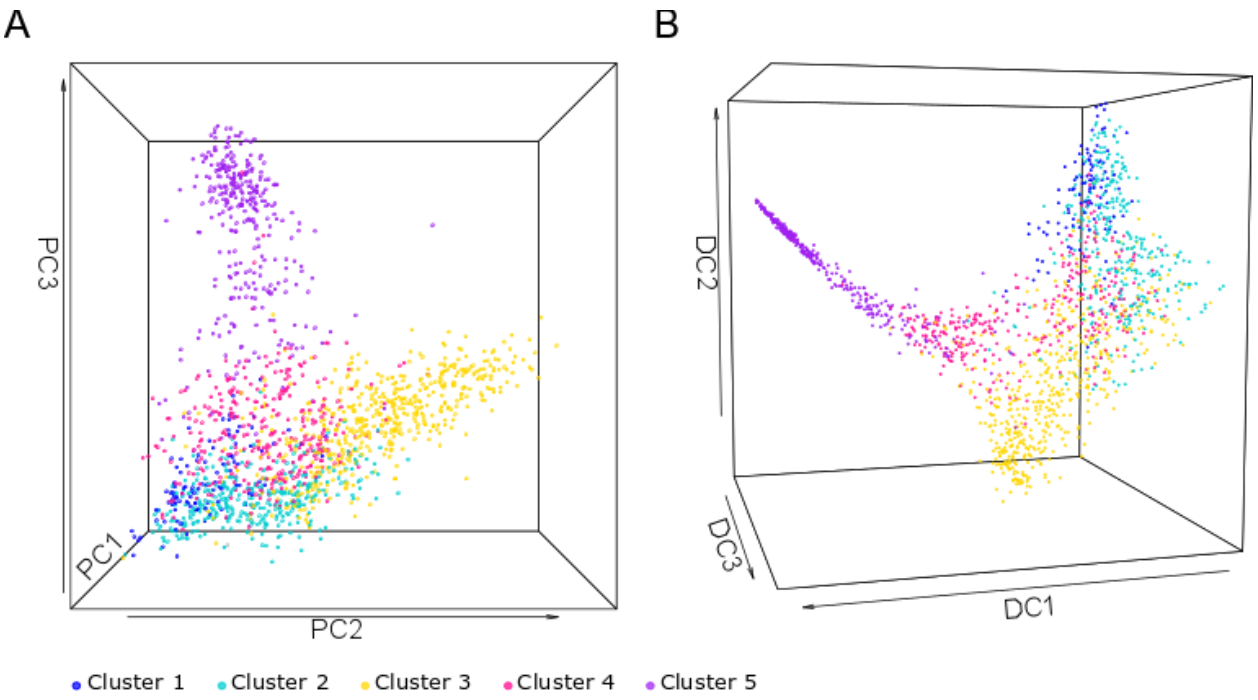


Figure 3.3. Multidimensional analysis shows a differentiation continuum towards different blood lineages. (a) PCA calculated on the expression of 4,290 variable genes measured by scRNA-seq. PC: Principal Component. The plot is coloured by the clusters identified in Section 3.3. (b) Diffusion map calculated on the expression of 4,290 variable genes measured by scRNA-seq. DC: Diffusion Component. The plot is coloured by the clusters identified in Section 3.3. Cluster colours: 1 – blue; 2 – turquoise; 3 – yellow; 4 – pink; 5 – purple.

3.5. Visualising gene expression on the HSPC differentiation continuum

Expression levels of individual genes can be plotted on the diffusion map to show their expression across the HSPC differentiation landscape (Fig. 3.4) This can be a useful tool to not only identify characteristics of the projections seen on the diffusion map, but also to visualise a gene of interest and understand how its expression differs among the HSPC populations.

Procr and *Mpl* are known to be important in HSCs and are mainly expressed at the top of the diffusion map, corresponding with Clusters 1 and 2 (Balazs et al. 2006; Kimura et al. 1998). Additionally, the expression of recently reported LT-HSC markers *HoxB5* and *Fgd5* is concentrated at the top of the diffusion map (J. Y. Chen et al. 2016; Gazit et al. 2014). These genes highlight the HSC region. The lymphoid genes *Ccl3* and *Dntt* are highly expressed in Cluster 2, highlighting cells of the early lymphoid trajectory (Rothenberg 2014). Cluster 3 consists of cells expressing the myeloid marker genes *Mpo* and *Ctsg*, representing the myeloid trajectory (Olsson et al. 2016). Similarly, the expression of *Cebpa*, which is involved in cell-fate decisions during

myeloid differentiation, is mostly concentrated to Cluster 3 (L. Scott et al. 1992). *Gata1*, a transcription factor important in erythroid and megakaryocytic differentiation, has concentrated expression in Cluster 5 (Evans and Felsenfeld 1989). *Klf1* and *Gypa* are also highly expressed in cluster 5 but further along in the projection, marking the erythroid trajectory (Dzierzak and Philipsen 2013). Therefore, scRNA-seq analysis of cells isolated using broad sorting gates captured the transcriptional profiles of cells from HSCs through to erythroid, myeloid, and lymphoid trajectories.

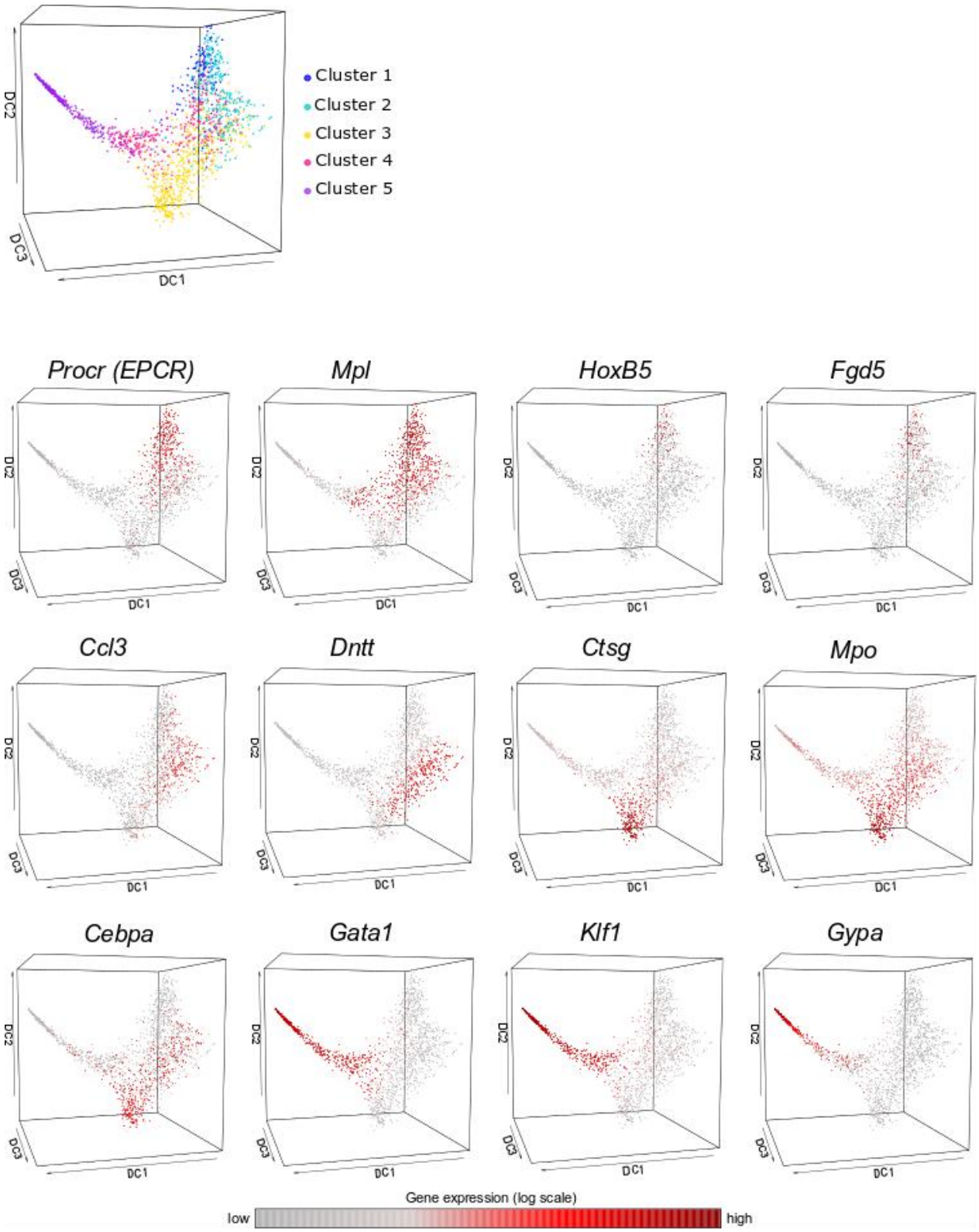


Figure 3.4. Gene expression in the HSPC differentiation atlas. Diffusion map of all cells coloured based on the expression of selected genes. The genes were chosen based on published literature or were identified computationally as highly expressed in specific cell populations. The colour corresponds to a $\log_2$ scale of expression ranging between 0 and the maximum value for each gene. The diffusion map coloured by clusters from Figure 3.3B is included for reference. Cluster colours: 1 – blue; 2 – turquoise; 3 – yellow; 4 – pink; 5 – purple. DC: Diffusion Component

3.6. Linking cell phenotypes with the transcriptome

Index-sorting data was collected during FACS, meaning that surface marker expression data was available for all cells along with the transcriptional profiles made available by scRNA-seq. Surface marker expression data could therefore be visualised on the diffusion map (Fig. 3.5). These visualisations were performed to confirm that the transcriptome recapitulates the well-characterised structure of the cell surface phenotypes of haematopoietic differentiation.

Along with the antibodies required to isolate the HSPC, Prog and LT-HSC populations, all cells were also stained with antibodies against CD48, CD150 and EPCR. The surface markers Sca1, Flk2, CD34, CD16/32, CD48, CD150, and EPCR marked coherent territories on the diffusion map (Fig. 3.5A). These territories were consistent with the separation of HSCs and more mature progenitors and matched the gene expression patterns described in Section 3.5. EPCR expression was highest at the top of the diffusion map, consistent with the expression pattern of *Procr*, the gene encoding EPCR (Fig. 3.4). The lineage cocktail (Lin) is made up of antibodies against CD5, CD11b, CD19, CD45R, Gr-1, TER119, and 7-4, which are markers of mature haematopoietic cells. Lin, c-Kit and Sca1 were the main surface markers used for sorting and their expression, as visualised on the diffusion map, reflects the sorting strategies used. All cells were sorted as c-Kit positive and Lin negative; the colours in Figure 3.5 are normalised between the minimum and maximum expression value for each marker, and therefore regions of the c-Kit and Lin expression plots may appear varied based on cell marker expression within a defined gate. Overall, c-Kit and Lin expression were consistent throughout the landscape (Fig. 3.5B), whereas Sca1 is not expressed in the erythroid and myeloid projections (Fig. 3.5A). Forward-scattered light-height (FSC-H), which represents cell size, was higher in myeloid and erythroid cells than in the HSCs and early lymphoid cells. This is consistent with the larger size of more mature cells compared to immature HSCs.

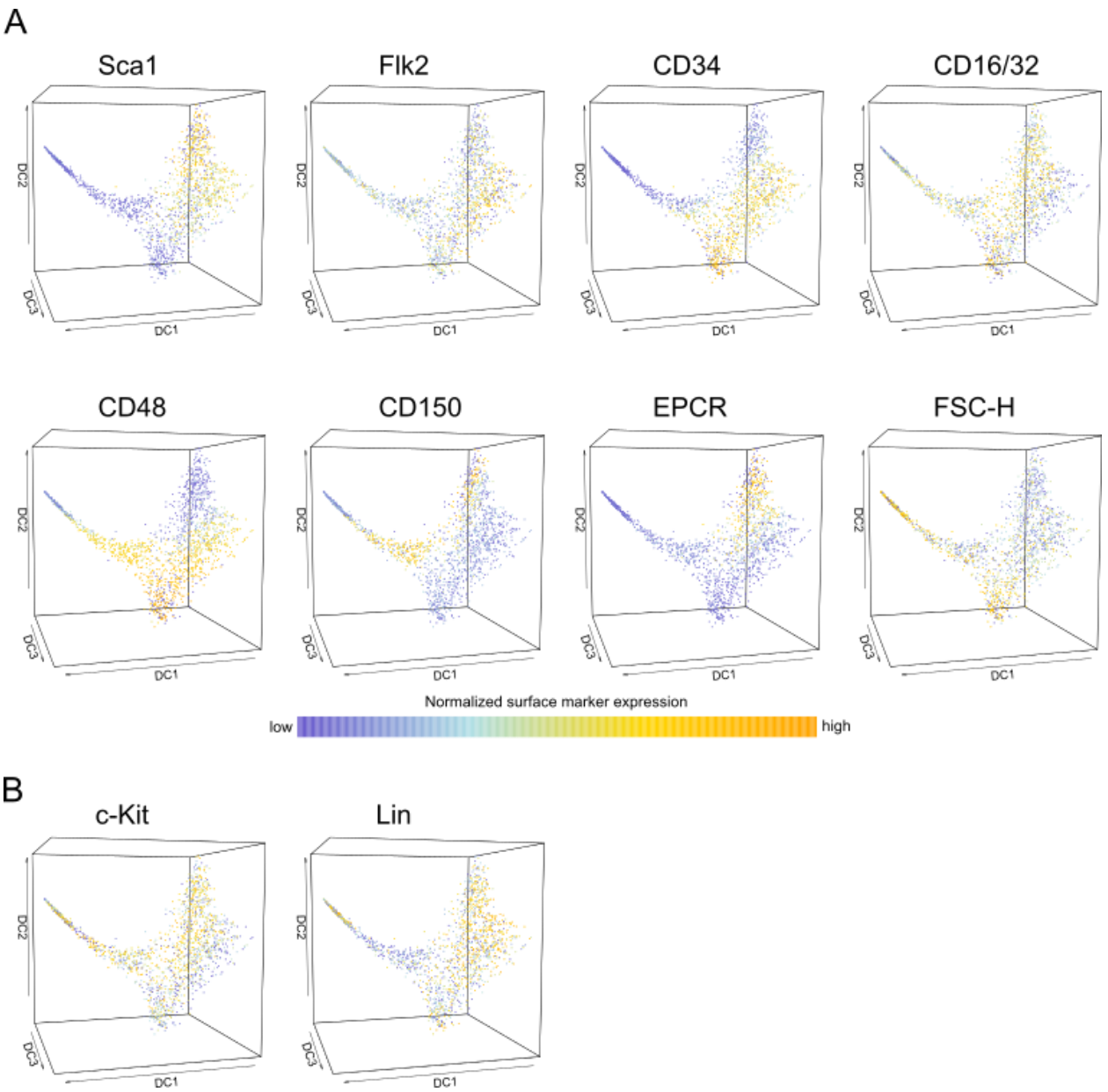


Figure 3.5. Surface marker expression on the HSPC differentiation atlas. Diffusion map of all cells coloured based on surface marker expression data collected using index-sorting. Surface markers in (a) mark territories on the diffusion map, whereas surface markers in (b) do not mark any territories due to their global expression. The colour corresponds to normalised expression values ranging between the minimum and maximum value for each marker. Flow cytometry data were normalised across two sort days. FSC-H: forward-scattered light-height; Lin: Lineage cocktail (CD5, CD11b, CD19, CD45R, Gr-1, TER119, 7-4); DC: Diffusion Component

Three broad sorting strategies were used to isolate the HSPC, Progenitor and LT-HSC populations. Using strict gates could limit the information gained for creating a differentiation continuum; they are limited by the investigator's knowledge of existing populations and may miss stepwise processes occurring between known HSC and progenitor populations. By using a liberal gating strategy, it could be possible to capture cells in the differentiation continuum that would normally

be missed by using strict sorting gates. The index sorting data collected during FACS was used to retrospectively assign cells into the cell populations outlined in Fig. 3.1B using common sorting strategies (Fig. 3.6A). By additionally staining for CD150, CD48 and EPCR, cells could be retrospectively assigned to other populations, such as E-SLAM (D. G. Kent et al. 2009) or the MPP subpopulations (A. Wilson et al. 2008; Pietras et al. 2015). The index data was visualised on the diffusion map to show the cell type distributions (Fig. 3.6B). The expression of marker genes correlated with the unsupervised clustering and transcriptional profiling of cells.

Three sorting strategies were visualised for LT-HSCs using retrospective gating, ordered in Fig. 3.6B by decreasing stringency: E-SLAM ($CD48^-$ $CD150^+$ $EPCR^+$), Lin^- $Sca1^+$ $c\text{-}Kit^+$ $CD34^-$ $Flk2^-$ $CD48^-$ $CD150^+$, and the LT-HSC sorting strategy (Lin^- $Sca1^+$ $c\text{-}Kit^+$ $CD34^-$ $Flk2^-$). The sorting strategy stringency has a reverse correlation with heterogeneity, as E-SLAM cells occupy the least heterogeneous space at the top of the diffusion map.

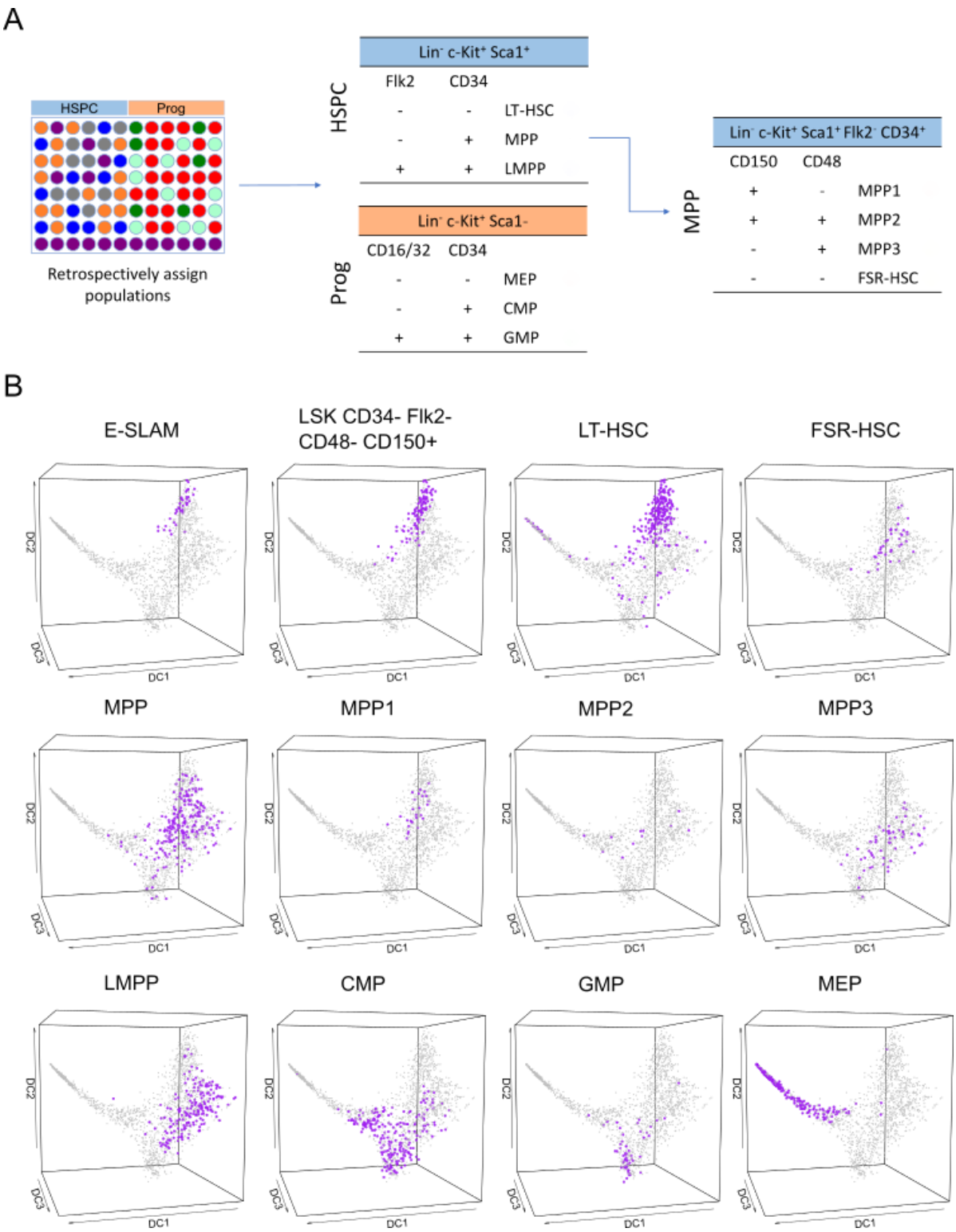


Figure 3.6. Visualising HSPC populations on the HSPC differentiation atlas. (a) Schematic of the strategies used for retrospectively gating the cells using the index-sorting data. The 96-well plate is coloured by a possible distribution of cell types sorted. The colours are consistent with those shown in the gating strategy diagram. Purple – LT-HSC; grey – MPP; blue – LMPP; red – MEP; light green – CMP; dark green – GMP. (b) Diffusion maps of all cells coloured based on cell types assigned by retrospective gating. The population of interest is coloured in purple; all other cells are in grey. DC: Diffusion Component.

3.7. Capturing changes in gene and protein expression using pseudotime ordering

Pseudotime ordering links cells together to order each transcriptional profile by its progress through differentiation (Bendall et al. 2014; Trapnell et al. 2014). Pseudotime analysis assumes gene and protein expression constantly change throughout differentiation, and that the dataset includes a large sample of cells broadly covering the differentiation process. In pseudotime ordering, the position of a cell corresponds to its progress in differentiation.

Fiona Hamey performed pseudotime ordering on the diffusion map to capture changes in both gene and surface marker protein expression during differentiation, and to identify differentiation lineages towards mature haematopoietic cells. Differentiation trajectories were identified from HSCs towards erythroid (E), granulocyte-macrophage (GM), and lymphoid (L) lineages (Fig. 3.7A). Changes in index-sorting parameters along these pseudotime trajectories were visualised (Fig. 3.7B). The surface markers used for isolating cells showed changes in expression consistent with the sorting strategies used. FSC-H increased along all three trajectories, indicating an increase in cell size, as seen previously when the index sorting data was visualised on the diffusion map (Fig. 3.5). The increase in FSC-H occurred more gradually along the L lineage than along either the E or GM lineage. EPCR, which was not used for sorting but was included in the antibody staining panel, showed decreased expression along all three trajectories, consistent with EPCR being a marker for HSCs (Balazs et al. 2006).

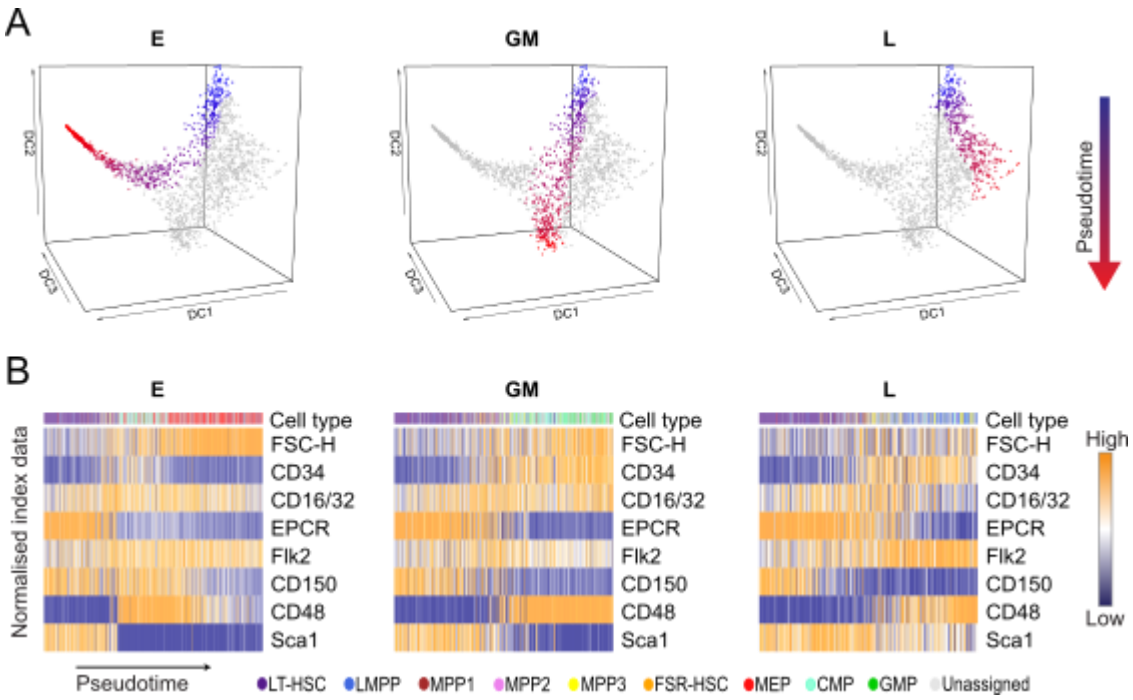
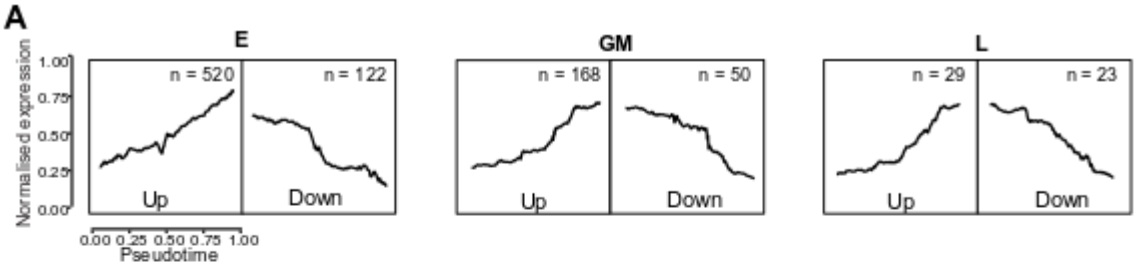


Figure 3.7. scRNA-seq profiles can be computationally ordered along pseudotime trajectories. (A) Diffusion map highlighting cells in three differentiation trajectories. Cells were ordered from HSCs along erythroid (E), granulocyte-macrophage (GM) and lymphoid (L) lineages. The pseudotime value for each cell indicates its position in the differentiation trajectory, moving from blue to red. DC: Diffusion Component. (B) Surface marker and FSC-H dynamics along the pseudotime trajectories. Index data is scaled so that each variable ranges from 0 (low) to 1 (high) in each trajectory. The colour bar at the top of each heatmap indicates the cell type. LT-HSC – purple; LMPP – blue; MPP1 – brown; MPP2 – pink; MPP3 – yellow; FSR-HSC – orange; MEP – red; CMP – light green; GMP – dark green. Grey cells represent cells unassigned to any cell population. Figure was created by Fiona Hamey and modified by Sonia Shaw.

Gene set enrichment analysis (GSEA) interrogates a given dataset for biological processes, molecular functions, or phenotypes in which the genes are known to be involved. GSEA can be used to gain a better understanding of statistically significant biological processes occurring along a differentiation trajectory. Gene sets that were up- or down-regulated throughout pseudotime were identified for each trajectory (Fig. 3.8A). GSEA revealed enrichment terms for these gene sets that corresponded to the relevant trajectory (Fig. 3.8B). The L trajectory showed an increase in genes involved in B-cell lineage commitment, genes involved in megakaryocyte-erythrocyte progenitor phenotypes were upregulated along the E trajectory, and neutrophil degranulation genes were upregulated along the GM trajectory. Both the E and GM trajectories had significant terms related to cell-cycle, such as “mitotic cell cycle” and “DNA replication.” However, the L trajectory did not show any significant upregulation of genes related to cell cycle, which may be due to the low number of genes up- (n=29) and down-regulated (n=23) in this 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^{-5}	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)

Figure 3.8. Pseudotime ordering reveals changes in gene expression during differentiation. (a) Normalised expression of genes up- or down-regulated along pseudotime for erythroid (E), granulocyte-macrophage (GM), and lymphoid (L) trajectories. Gene expression was smoothed by a sliding window of size 20 along pseudotime. Mean normalised expression is shown as a black line ± standard deviation (grey shading). n indicates the number of genes in each subset. (b) Gene set enrichment analysis terms for the above gene sets. Adjusted p-value is shown for significant terms (Benjamini-Hochberg). Figure was created by Fiona Hamey and modified by Sonia Shaw.

3.8. Ordering cells along differentiation trajectories using STREAM

The diffusion map is a useful non-linear dimensionality reduction method to visualise continuous differentiation processes within the scRNA-seq dataset. The analysis informed on gene and surface marker expression throughout the HSPC atlas and was used to identify three differentiation trajectories within the data using pseudotime ordering. However, the visualisation lacks information about cell density and cell type composition.

Recently, a new trajectory inference tool called STREAM (Single-cell Trajectories Reconstruction, Examination and Mapping) was developed to reconstruct differentiation trajectories and capture gene expression changes during differentiation using pseudotime ordering (Trapnell et al. 2014; H. Chen et al. 2018). STREAM uses a non-linear dimensionality reduction method called Modified Local Linear Embedding (MLLE) and infers trajectories using a novel method called ElPiGraph

(Lever, Krzywinski, and Altman 2017; Z. Zhang and Wang 2006). ElPiGraph differs from other methods as it does not require drastic dimensionality reduction or pre-clustering to infer trajectories. In addition to capturing trajectories within a dataset, STREAM uses a visualisation method which includes density information throughout pseudotime. This is a useful tool to track cell population composition changes along a trajectory.

The Pinello lab used our scRNA-seq dataset to demonstrate STREAM and their interactive web-tool. The online interface allows the user to manipulate the data to interrogate branching and gene expression patterns. STREAM is available at the following link:

http://stream.pinellolab.org/MNoPZ/

HSCs were selected as the start of the branching structure and the scRNA-seq dataset was visualised on a "subway plot" (Fig. 3.9A) and a "stream" plot (Fig. 3.9B). The subway plot orders cells according to their pseudotime score and distance from their assigned branch. The purpose of the subway plot is to visualise the branching structure of the data to understand the pseudotime progression. The stream plot also orders cells based on their pseudotime score but incorporates information on the density and composition of cell types along the different trajectories. The stream plot visualisation requires the user to input cell type information and is made from the subway plot using a sliding window approach. The length of the plot represents a cell's location along pseudotime, whereas the width of the plot is proportional to the number of cells.

The subway plot showed that STREAM analysis identified three lineages in the data: erythroid, myeloid, and lymphoid. The analysis suggests the lymphoid cells entered their trajectory before the myeloid and erythroid cells. The stream plot showed that the lymphoid branch was composed mostly of LMPPs, the erythroid branch of MEPs, and the myeloid branch of CMPs and GMPs. The lymphoid trajectory stopped before the other lineages, which is due to there being fewer lymphoid cells in the analysis as more mature lymphoid cells were excluded from the sorting gates. The expression of the genes *Procr, Klf1, Ctsg* and *Dntt* were visualised on stream plots to represent the HSCs, the erythroid, myeloid, and lymphoid lineages, respectively (Fig. 3.9C). The branching and gene expression patterns were consistent with the diffusion map visualisation. *Dntt* expression was observed in the lymphoid branch as well as in cells heading towards the erythroid and myeloid lineages. The stream plot (Fig. 3.9B) showed LMPPs are present in the trajectory at this stage (S1-S3 on the subway plot), accounting for the observed *Dntt* expression pattern.

STREAM analysis also detects genes important in defining branching points in the data (Fig. 3.9D). The user can identify which branch they want to investigate using the annotations marked on the subway plot, and STREAM identifies genes differentially expressed between the diverging branches. *Cd63* and *Hlf* were highly expressed on the HSC branch compared to cells after the first bifurcation event. *Cd63* encodes for an endosome-associated protein that has previously been identified as a marker of HSCs in cultured human $CD34^+$ HSCs, and *Hlf* has recently been shown to be a key regulator in HSC quiescence (Komorowska et al. 2017; Beckmann et al. 2007). Conversely, *Il12a* and *Cst7* were more highly expressed after the bifurcation event than in HSCs. *Il12a* encodes for a subunit of the IL-12 cytokine, a main activator of natural killer cells, and *Cst7* is involved in normal eosinophil function (Seaman 2000; Halfon et al. 1998). *Ltb* and *Uhrf1* were identified as differentially expressed between the lymphoid and erythroid/myeloid lineages, respectively. *Ltb* is involved in the development of normal lymphoid tissue, whereas *Uhrf1* is an epigenetic regulator required for establishing DNA methylation patterns of erythroid genes (Koni et al. 1997; J. Zhao et al. 2017). STREAM also found genes marking the second bifurcation event. *Gimap6*, which encodes a protein required for T-cell maintenance, was more highly expressed in cells before the differentiation point (Pascall et al. 2018). Conversely, *Sdsl* and *Rab44*, which are associated with the erythroid and myeloid lineages, respectively, were more highly expressed in their respective lineages (Poczobutt et al. 2016; Khoramian Tusi et al. 2018). Finally, when the erythroid and myeloid lineages were directly compared, *Mfsd2b* and *Hk3* were differentially expressed in the two trajectories. *Mfsd2b* is involved in red cell morphology, whereas *Hk3* is involved with neutrophil differentiation, supporting the notion that these genes may mark a branching point between the lineages (Vu et al. 2017; Federzoni et al. 2012).

STREAM analysis was also used to look for transition genes, defined as genes for which the expression correlated with the pseudotime ordering on a given branch (Fig. 3.9E) (H. Chen et al. 2018). These genes were selected by the STREAM interface based on their differential expression across the stream plot. This analysis can also give insight into cell-fate decision making and has the potential to discover novel genes. *Tgm2* had increasing expression towards the tip of the HSC branch, whereas *Tespa1* showed increased expression moving away from HSCs. *Tgm2* is an extracellular matrix protein previously suggested to be a regulator of LT-HSCs (Forsberg et al. 2005). *Tespa1*, on the other hand, is a signalling molecule that plays a wide range of roles in more differentiated cells, including T-cells and mast cells (Liang et al. 2017; D. Wang et al. 2012). *Igsf6*, which is involved in myeloid differentiation, and *Ctla2a* were inversely correlated with the myeloid

branch (Stein and Baldwin 2013). The expression of *Smim1*, which encodes an erythroid transmembrane protein, marked the tip of the erythroid trajectory, while *Coro1a* expression was higher in the less differentiated cells (Storry et al. 2013). Finally, *Tyms*, a gene involved in DNA replication and repair, was identified as a transition gene moving away from the lymphoid trajectory (Ozer et al. 2015). *Ltb* (Fig. 3.9D), involved in normal lymphoid organogenesis, was the transition gene identified towards the lymphoid trajectory (Koni et al. 1997). Except for *Ltb*, these genes were not identified in previous analyses, such as hierarchical clustering, demonstrating that STREAM can be used to find novel genes of interest.

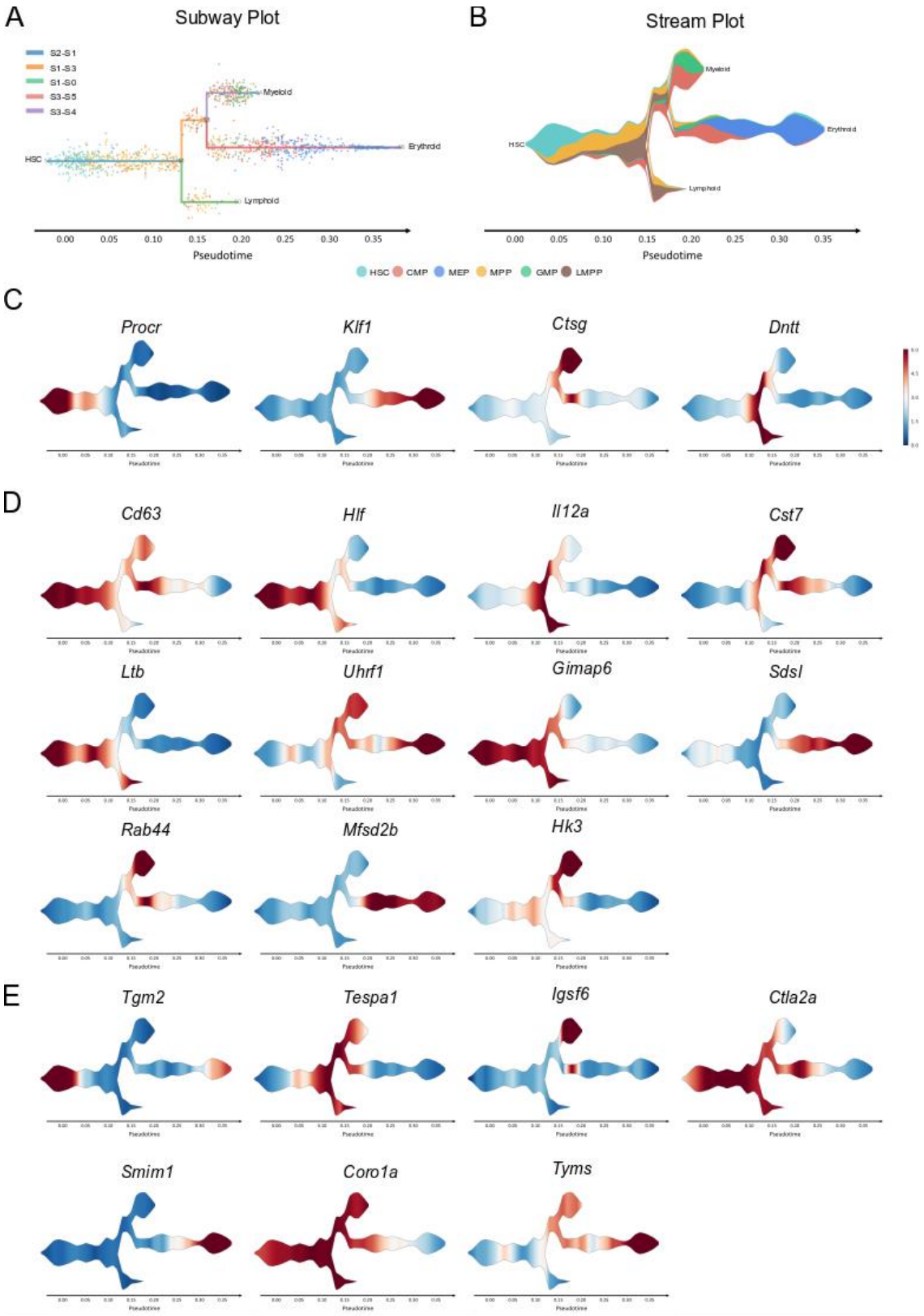


Figure 3.9. STREAM analysis reveals information about pseudotime ordering in the HSPC differentiation landscape. (a) Subway plot of the scRNA-seq data visualised based on its pseudotime ordering. Cells are coloured based on cell type. The trajectories are ordered and coloured for the user to easily manipulate the data. S2-S1 (blue) – HSC to first branching point; S1-S0 (green) –lymphoid trajectory; S1-S3 (orange) – branch into myeloid and erythroid trajectories; S3-S4 (purple) – myeloid trajectory; S3-S5 (red) – erythroid trajectory (b) Stream plot of the scRNA-seq

data visualised based on its pseudotime ordering. The width of each branch is proportional to the total number of cells. Branches are coloured based on cell type composition. HSCs – turquoise; CMP – red; MEP – blue; MPP – yellow; GMP – green; LMPP – brown. (c) Stream plots of all cells coloured based on the expression of selected genes. The genes chosen were previously used in Figure 3.4 to mark branches within the diffusion map. (d) Stream plots of genes identified to be differentially expressed between erythroid and myeloid branches. (e) Stream plots of genes identified to be correlated with the pseudotime ordering on the HSC branch. Genes are ordered based on which branch they are associated with (moving from HSCs to erythroid).

3.9. Visualising HSPCs and capturing rare populations using SPRING

The diffusion map and STREAM analysis both revealed three differentiation trajectories within the scRNA-seq dataset. However, neither method is able to capture the full complexity of the data in a two-dimensional plot. A new visualisation method, SPRING, uses a force-directed layout of k-nearest neighbour graphs to capture long-distance relationships between cells (Weinreb, Wolock, and Klein 2018). SPRING brings similar cells together and repels cells which differ based on their gene expression while ensuring all cells remain connected via nodes. SPRING can visualise continuous expression topologies and could be a useful tool for uncovering biological processes that were not captured by previous approaches.

The Klein lab used our scRNA-seq dataset to demonstrate SPRING using their interactive web-tool. The online interface allows the user to compare gene expression profiles, focus on subpopulations within the data, and potentially discover marker genes. SPRING is available at the following link: https://kleintools.hms.harvard.edu/tools/spring.html

Cell populations, determined by retrospective gating using index-sorting data, were visualised using the SPRING web-tool (Fig. 3.10). The SPRING plot recapitulated a similar pattern compared to the diffusion map, with LT-HSCs, LMPPs, GMPs and MEPs forming distinct branches. LT-HSCs and LMPPs were found close together and MEPs formed a long protrusion from the dataset. GMPs, however, formed a branch that is more separated and elongated than in the diffusion map. Furthermore, the presence of five additional small branches on the SPRING plot suggests that there are other lineages represented in the data that were not captured by the diffusion map or STREAM (Fig. 3.11).

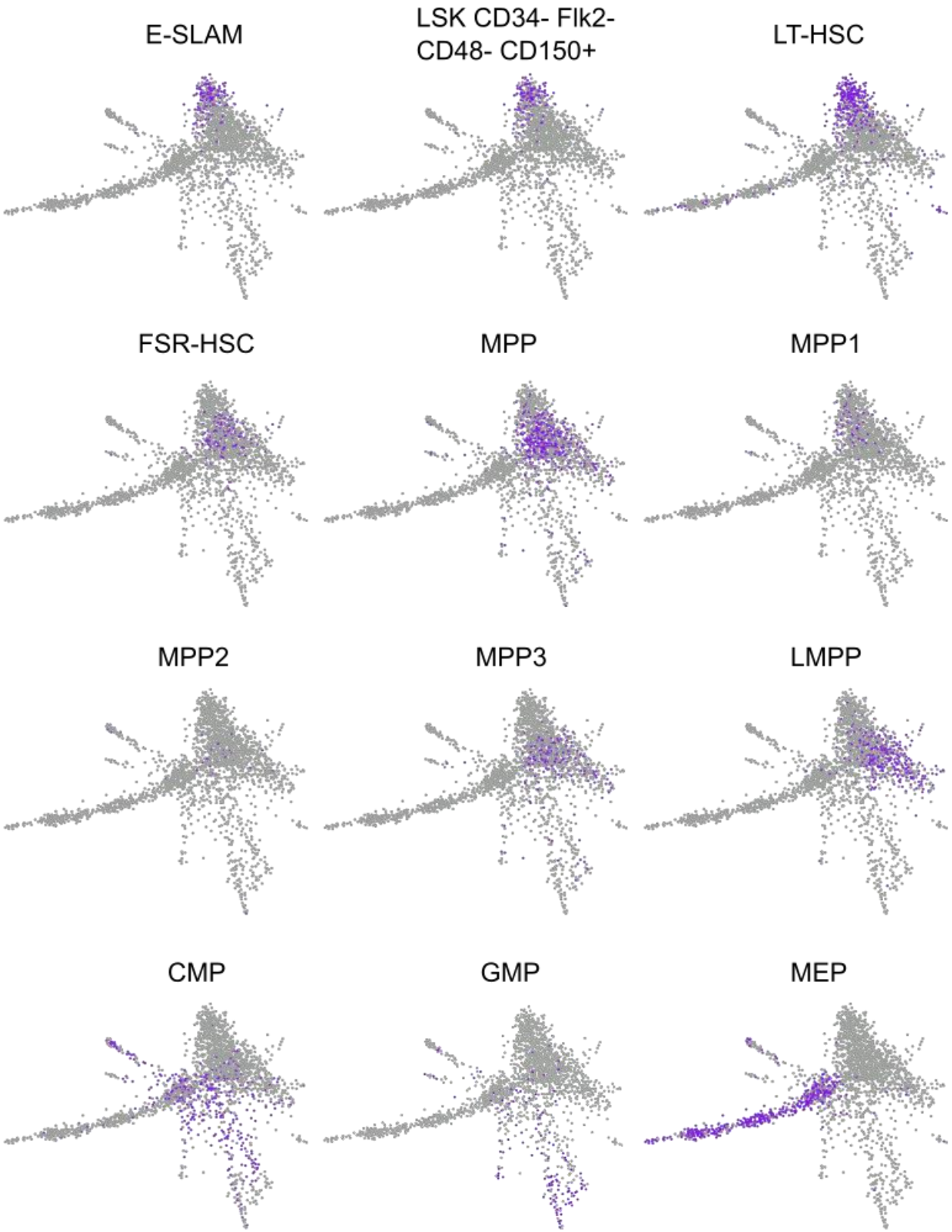


Figure 3.10. Visualising HSPC populations using the SPRING interface. SPRING plots of all cells coloured based on cell types assigned by retrospective gating of the index-data. The population of interest is coloured in purple; all other cells are in grey.

The expression of the genes *Procr*, *Klf1, Ctsg* and *Dntt* was visualised on SPRING plots to represent HSCs, erythroid, myeloid, and lymphoid lineages, respectively (Fig. 3.11A). The cell surface marker data and gene expression profiles separate the HSPC populations well. The SPRING interface allows the user to select cells of interest to reveal their expression profiles. This tool was used to identify the additional branches and to better determine genes involved in differentiation towards these lineages. Differential expression analysis was performed on the remaining branches and the top differentially expressed gene for each branch was visualised (Fig. 3.11B). *Gb1pp* is a gene encoding for part of a platelet complex and may mark the megakaryocyte lineage (Savoia et al. 2011). *Ltf*, which encodes for an iron-binding protein, has been previously identified as a late-stage differentiation marker of neutrophils, macrophages, and subtypes of dendritic cells (Kovacic et al. 2014). *Ltf* may therefore be marking a myeloid lineage. *Ms4a2* is expressed in mast cells and basophils and its presence in a distinct branch on the SPRING plot could indicate an early point in the mast cell/basophil lineage (Dwyer et al. 2016). *Ly6c2* was expressed at the tip of the GMP branch and encodes for Ly6C, a marker for inflammatory monocytes (J. Yang et al. 2014). *Cd19* is a reliable B-cell marker, which is expressed from pre-B cells until differentiation into plasma cells (K. Wang, Wei, and Liu 2012). Its expression was close to the LMPP population, and its presence in a distinct branch of the SPRING plot could indicate cells differentiating into B-cells. *Il7r* marks a cell's entry into the lymphoid lineage and is expressed in common lymphoid progenitors (J. Wang et al. 2012). Finally, *Ifit3b* is an uncharacterised gene within the IFIT family, which are induced by interferon and have anti-viral functions (Vladimer, Górna, and Superti-Furga 2014). IFIT3 expression is induced in human dendritic cells in response to viral infections (Y.-L. Hsu et al. 2013). Therefore, the branch of cells expressing *Ifit3b* could indicate an early point in differentiation towards dendritic cells.

The genes identified by SPRING analysis to be the top differentially expressed genes in each branch were also plotted on the diffusion map for comparison (Fig. 3.11C). Except for *Il7r* and *Ly6c2*, the expression of the selected genes is more disperse. *Gb1pp* has concentrated expression at the start of the erythroid lineage branch and *Ms4a2* is expressed in the CMP/MEP regions. *Ltf* and *Cd19* are expressed by few cells and are difficult to identify on the diffusion map. SPRING analysis therefore appears to capture more of the biological relationships within the data and reveals lineages previously obscured within the bulk of the landscape, allowing for deeper interrogation of trajectories along the HSPC differentiation landscape.

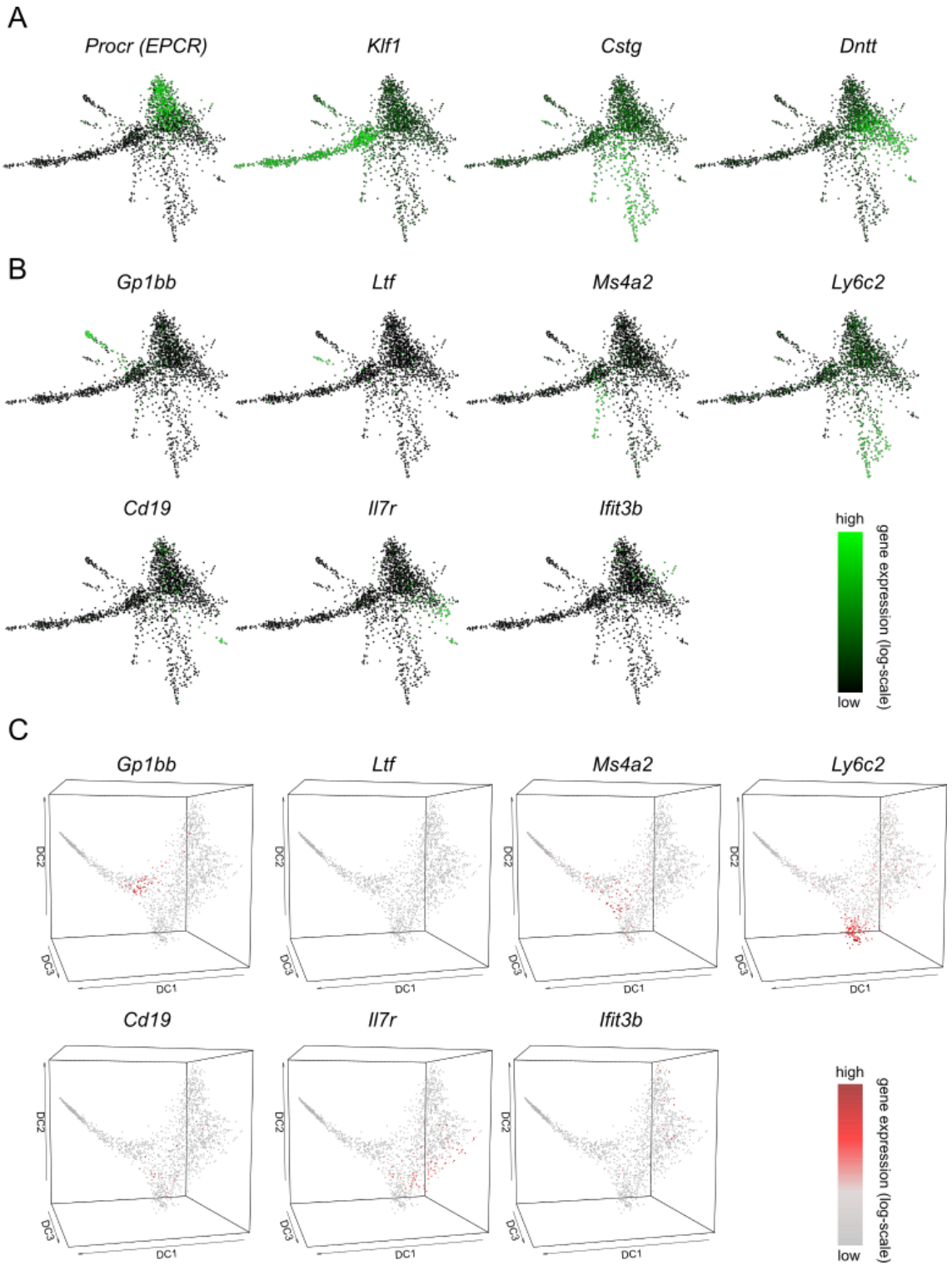


Figure 3.11. SPRING analysis captures rare populations in the HSPC differentiation atlas. (a) SPRING plot of all cells coloured based on the expression of selected genes. The genes chosen were previously used in Figure 3.4 to mark branches within the diffusion map. (b) SPRING plot of all cells coloured based on the expression of selected genes. The genes chosen were identified by SPRING analysis to be the top differentially expressed gene for each branch. (c) Diffusion map of all cells coloured based on the expression of selected genes. The genes chosen were identified by SPRING analysis to be the top differentially expressed gene for each branch. The colour corresponds to a $\log_2$ scale of expression ranging between 0 and the maximum value for each gene. DC: Diffusion Component.

3.10. Conclusions

The main aim of this chapter was to generate a comprehensive atlas of HSPC differentiation. This was achieved by isolating individual cells from mouse bone marrow and profiling them by scRNA-seq. The isolated cells encompassed HSCs and early progenitor populations to create a thorough depiction of the transcriptional landscape of early HSPC differentiation. The resulting landscape was then interrogated by pseudotime ordering and differential expression analysis to better understand the differentiation trajectories captured in the data.

The cells were isolated using two broad sorting gates and an additional LT-HSC sorting gate. The gates used captured HSC and progenitor populations from known sorting strategies without specifying narrow restrictions. In doing so, the aim was to capture stepwise processes that may exist outside of narrowly defined population gates and may fill in gaps within the differentiation continuum. Surface marker and gene expression showed that retrospective gating using the index sorting data correctly placed all the HSPC populations. The data was visualised using a diffusion map, a recently developed non-linear dimensionality reduction method that was particularly designed to capture the branching structure in single-cell data (Haghverdi, Buettner, and Theis 2015). Previously defined HSPC populations were restricted to distinct regions on the diffusion map, with the exception of the CMP population. Unsupervised clustering grouped the data into five broad clusters. CMPs were represented mostly in clusters 3 and 4 which consisted of cells of all populations, and in total were present in four of the five clusters. The clustering and diffusion map visualisation show CMPs are a highly heterogeneous population, and in fact previous investigations have also suggested that CMPs are mainly erythroid- or myeloid-committed cells (Paul et al. 2015; Perié et al. 2015).

Pseudotime ordering was performed on the diffusion map, ordering cells during differentiation based on their gene expression profiles. The ordering revealed three trajectories through the data from stem cells to erythroid, myeloid/granulocyte-macrophage, and lymphoid lineages. Surface marker expression was ordered along pseudotime and showed cells increased in cell size as they differentiated, particularly in the erythroid and granulocyte-macrophage lineages. This was consistent with MEP, CMP and GMP cells representing more mature cells compared to LMPPs, which are earlier progenitors along the lymphoid trajectory. GSEA was performed on genes up- and down-regulated along the lineages and showed that the erythroid and myeloid lineages had an upregulation in cell-cycle related terms, whereas the lymphoid trajectory did not. This suggests that

cell-fate specification occurred independently of cell cycling. However, there were few genes associated with the lymphoid trajectory. Cell cycle analysis was performed and reported in the Nestorowa *et al.* (2016) publication, in which cells were assigned to G_0/G_1, S and G_2/M cell cycle categories (Scialdone et al. 2015). The distribution of cells across these cell cycle categories correlated well with the enrichment of cell cycle terms in genes upregulated in the E and GM trajectories. The analysis also showed that the large-scale transitioning of cells to S and G_2/M phases occurred after the divergence of the L trajectory from the E and GM trajectories, supporting the evidence presented in this thesis suggesting that the transition to rapid cell cycling is secondary to transcriptional diversification (Nestorowa et al. 2016). Overall, the diffusion map was a useful tool for visualising the branching structure of the data and identifying differentiation lineages, which differ in their cell type composition as well as surface marker and gene expression patterns.

Since the publication of the paper associated with this chapter, a number of interactive web-tools were developed for analysing and visualising scRNA-seq data. One of the great advantages of these web-tools is their accessibility for researchers with limited bioinformatics knowledge; on the flipside, these tools limit the investigator's input into the analysis of the data and can be difficult to use if issues arise. STREAM and SPRING are two recently developed web-tools which were used in this chapter to elucidate more information about the differentiation trajectories (Weinreb, Wolock, and Klein 2018; H. Chen et al. 2018). STREAM analysis defined three trajectories within the data but suggested that the bifurcation event towards myeloid and erythroid trajectories occurs after the lymphoid branch is formed, consistent with the classic view of the haematopoeitic hierarchy. Uniquely, STREAM inputs cell composition data into its pseudotime ordering visualisation. This clearly represents the cell types present and their density at any given point in the pseudotime ordering, which is useful for understanding the structure of the lineages. This is, however, limited to the data provided by the user. The data analysed in this chapter was precomputed by Chen *et al.* to demonstrate the use of STREAM (H. Chen et al. 2018). They classified the cells into HSCs, LMPPs, CMPs, MEPs and GMPs, leaving out FSR-HSCs and the subpopulations of MPPs. It would be useful to see how the density of cell types would be affected if the data was fully annotated. Using STREAM, genes involved in branching points within the data were identified. Furthermore, genes that correlated with pseudotime along the myeloid, erythroid, and lymphoid branches were identified, which may be useful for recognising patterns of gene up- and down-regulation that result in particular cell phenotypes. The genes found by

STREAM were not identified using the hierarchical clustering or diffusion map, suggesting that STREAM has the potential to identify novel markers for HSPC populations.

STREAM and diffusion map analysis are both informative methods for analysing scRNA-seq data but are limited in the dimensions used for capturing the structure of the data. Although using more dimensions may increase the power to decipher which genes may be driving lineage decisions, it can also introduce noise into the analysis as the high-dimensional space is sparser. As such, neither STREAM nor diffusion map analyses are able to fully capture the structure within a dataset. In contrast, SPRING captures long-distance relationships in the data. Using a force-directed layout algorithm, SPRING can visualise continuous expression topologies within an interactive web interface (Weinreb, Wolock, and Klein 2018). SPRING analysis separated the HSPC atlas into eight branches. The main erythroid, myeloid and lymphoid trajectories were represented in the dataset. Additionally, SPRING identified branches into megakaryocyte, neutrophil, mast cell/basophil, and B-cell lineages. These lineages were obscured in the bulk of the HSPC atlas in both STREAM and diffusion map visualisations but were captured using SPRING. This tool allows for more specificity when looking at the differential expression of genes in a specific lineage, and therefore may be superior for interrogating the HSPC differentiation landscape.

3.10.1. Limitations & future work

As previously mentioned, the diffusion map and STREAM analyses were unable to fully represent the differentiation lineages in the HSPC landscape. This shortcoming was addressed by SPRING, which captured five additional trajectories in the data. Future work should include exploring pseudotime ordering on the SPRING plot to capture gene expression changes occurring towards more specific blood lineages.

In every visualisation method used to show the HSPC landscape, the lymphoid branch was shorter than the myeloid and erythroid branches. This is due to the sorting gates being focused on capturing cells of the myeloid and erythroid lineages, only including LMPPs as representatives of the lymphoid branch. A lack of lymphoid cells resulted in a limited analysis of transition genes in the lymphoid branch using STREAM, where only *Ltb* was identified as transitioning towards the lymphoid trajectory. To full recapitulate the landscape of HSC differentiation into early blood progenitors, more cells of the lymphoid lineage would need to be represented.

Limitations also exist in the transcriptional profiling method itself. scRNA-seq profiling is associated with a substantial cost of sequencing, limiting the number of cells profiled based on budget. Furthermore, scRNA-seq has a limited quantitative range and can suffer from dropout events which cause technical and biological noise. To investigate the regulatory relationships controlling cell-fate decisions, a more directed approach could be taken. High-throughput quantitative real-time PCR could be used to further interrogate single cell data and identify dynamic gene expression through differentiation within a specific subset of genes. This will be further explored in Chapters 5 and 6.

The HSPC atlas is a great tool for visualising the expression of specific genes to guide *in vitro* investigations. It could be used to interrogate genes important to characteristics of populations of interest, such as HSCs. This will be further explored in Chapter 4.

3.10.2. Summary

In summary, scRNA-seq analysis of 1,654 HSPCs resulted in a comprehensive atlas of haematopoietic differentiation. The investigation showed how single-cell profiling can give insight into many aspects of differentiation, identifying genes involved in branch points and transitions through pseudotime. Furthermore, SPRING analysis demonstrated the breadth of biological relationships present in the HSPC differentiation landscape.

Chapter 4: CRISPR screening of HSC-enriched genes

4.1. Background

Haematopoietic stem cells (HSCs) sit at the apex of the haematopoietic differentiation hierarchy and have the capacity to self-renew as well as differentiate into all adult blood cells (Morrison, Uchida, and Weissman 1995). For three decades, there have been intense efforts to molecularly and functionally characterise HSCs, and researchers have developed protocols to isolate increasingly refined haematopoietic stem and progenitor cell (HSPC) populations (Beerman et al. 2010; Challen et al. 2010; D. G. Kent et al. 2009; Kiel et al. 2005; Morita, Ema, and Nakauchi 2010). The HSPC atlas described in Chapter 3 contributed to these efforts by providing a comprehensive web interface of single-cell transcriptomic data, which can be used to analyse gene and surface marker expression in HSPC populations (Nestorowa et al. 2016).

The HSPC atlas is a great resource to visualise population-specific gene expression trends. Chapter 3 focused on identifying an appropriate visualisation of the high-dimensionality dataset, as well as categorising the differentiation lineages that are captured by the HSPC atlas. The focus of this chapter is to identify which genes are important in HSC biology, using the HSPC atlas to identify candidate targets specific to long-term HSCs (LT-HSCs).

In their paper focused on resolving HSC heterogeneity, Wilson *et al.* identified a subpopulation of HSCs that were molecularly similar and included cells isolated by four commonly used sorting strategies (N. K. Wilson et al. 2015). This subpopulation was termed the molecularly overlapping (MolO) population. Single-cell RNA-sequencing of Lin^- $Sca1^+$ $c\text{-}Kit^+$ $CD34^-$ $Flt3^-$ $CD48^-$ $CD150^+$ HSCs led to the identification of 29 genes that were positively associated with the MolO population (Table 4.1). The gene set included *Cdkn1c*, *Ptpn14*, and *Ifitm1*, which are negative regulators of cell proliferation, suggesting that the MolO genes are involved in maintaining the HSC state. Wilson *et al.* also performed single-cell transplantation experiments of $CD48^+$ $CD150^-$ (SLAM) $Sca1^{hi}$ HSCs and recorded the surface marker expression during FACS paired with index sorting. This analysis permitted the integration of functional and transcriptional information from cells that were sorted on the same day using the same parameters. These cells were compared based on their surface phenotype. Cells with similar phenotypes were called SuMO cells (surface marker overlap) and analysis of this population revealed 21 genes that were positively associated with repopulating

HSCs (Table 4.1) (N. K. Wilson et al. 2015). There was a large degree of overlap between the MolO and SuMO gene lists, which gave a high degree of confidence; therefore, the lists were combined, resulting in a list of 44 genes associated with functional HSCs.

Table 4.1. List of all MolO and SuMO genes identified by Wilson *et al.* (2015).

MolO Genes			SuMO Genes		Shared Genes
Cd82	*Cdkn1c*	*Cldn10*	*Ablim1*	*Cd74*	*Ifitm1*
Ctsf	*Fads3*	*Fgfr3*	*Cyp27a1*	*Gbp6*	*Ly6a*
Gimap1	*Gimap6*	*Gstm1*	*Gbp8*	*Gm4951*	*Mllt3*
Limd2	*Ltb*	*Mettl7a1*	*Ifitm3*	*Inhba*	*Procr*
Neil2	*Neo1*	*Pde1b*	*Ly6e*	*Mapk12*	*Ramp2*
Pdzk1ip1	*Ptpn14*	*Smtnl1*	*Ndnf*	*Ralgapa1*	*Sult1a1*
Sox18	*Sqrdl*	*Trim47*	*St8sia4*	*Tgtp2*	
Ubl3	*Vwf*		*Wfdc2*		

The *Streptococcus pyogenes*-derived type II clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 system has become a very successful tool for genome editing (Mali et al. 2013; Cong et al. 2013). CRISPR sequences, or guide RNAs (gRNA), locate matching target DNA, which is cut by Cas9, effectively perturbing the target gene. This tool has been widely used for genetic research in many organisms, including genetic screens for essential genes or therapeutic targets (T. Wang et al. 2014; Shi et al. 2015; Shalem et al. 2014; Koike-Yusa et al. 2014). Recently, Tzelepis *et al*. used a Cas9 transgenic mouse to interrogate the genomes of acute myeloid leukaemia cells for potential therapeutic targets (Tzelepis et al. 2016). The investigation described in this chapter uses CRISPR-Cas9 technology to examine whether perturbing the candidate genes implicated in HSC biology influences HSC characteristics or function.

4.1.1. Aims

The aims of this chapter were to:

- Identify MolO and SuMO genes uniquely expressed in HSCs using the HSPC atlas (Chapter 3)
- Interrogate the effect of perturbing candidate genes using CRISPR-Cas9 by analysing EPCR expression, apoptosis, and differentiation

These aims were addressed by designing a study that took advantage of CRISPR-Cas9 technology and gene profiling previously performed by Wilson *et al.* (2015). Using flow cytometry and colony-forming unit (CFU) assays, the impact of perturbing candidate genes on HSPC biology was interrogated.

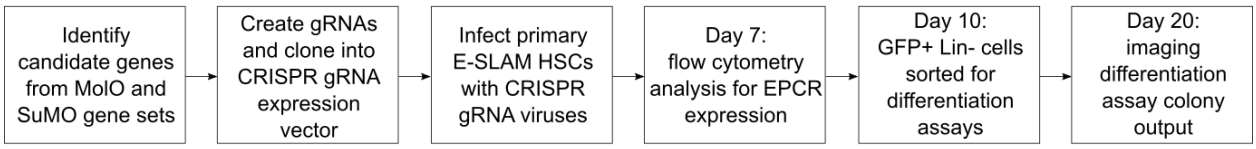
4.2. Study design to investigate genes implicated in HSC biology

To investigate genes implicated in HSC biology, a CRISPR perturbation study was designed (Fig. 4.1A). Candidate genes were identified from the MolO and SuMO gene sets previously described (N. K. Wilson et al. 2015). The list of 44 genes was refined by selecting genes that were more highly expressed in LT-HSCs (Section 4.3). First, the expression of each gene was visualised on the HSPC scRNA-seq atlas to determine whether it had widespread expression or was concentrated in HSCs. Violin plots were drawn for each gene to quantify its expression across all the cell populations, only selecting the genes that were most highly expressed in LT-HSCs. Three gRNA constructs were designed for each candidate gene using gRNAs from the Brie library or the Broad Institute sgRNA designer (Doench et al. 2016). All guides were cloned into a CRISPR gRNA expression vector with a GFP marker (pKLV2-U6gRNA5(BbsI)-PGKpuro2AmAG-W). E-SLAM HSCs (Lin^- $CD48^-$ $CD150^+$ $CD45^+$ $EPCR^+$) were isolated from 18-20 week old Cas9 transgenic mice, kindly gifted to the Göttgens lab by Dr George Vassiliou (Tzelepis et al. 2016). The E-SLAM HSCs were isolated by FACS and aliquots of 250 cells were sorted directly into individual wells of a U-bottom 96-well plate containing HSC medium. Cells were transduced with the CRISPR gRNA viruses and kept in culture for 10 days; on day 7, approximately half of the cells of each well were analysed by flow cytometry for EPCR expression changes (Sections 4.5-4.7). On day 10, the remaining cells were pooled and sorted for GFP^+ Lin^- cells, which were put into M3434 methylcellulose to study colony outputs (Section 4.8). In later experiments, cells were also sorted to be used for genotyping analysis (Section 4.9).

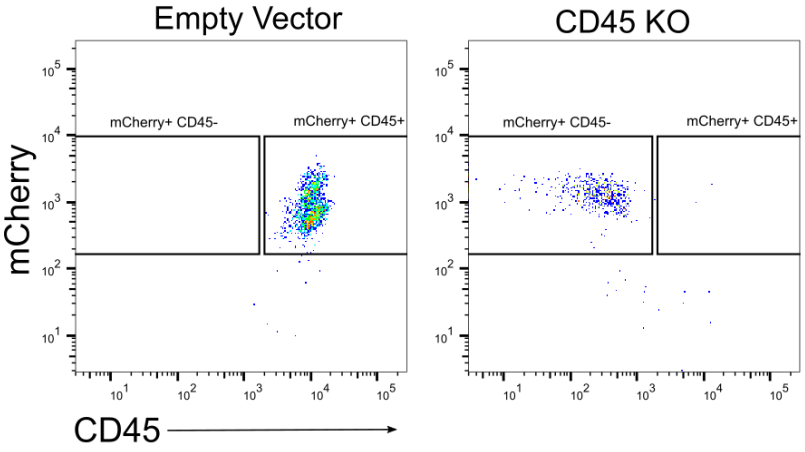
To confirm whether the CRISPR study design was valid, a trial experiment was performed in the same conditions as planned for the screen but looking at CD45 expression. CD45 is a haematopoietic cell-specific antigen and is expressed in most haematopoietic cell types (Ogata et al. 2005). A CRISPR gRNA expression vector with gRNA targeting the *Ptprc* gene (encoding the CD45 antigen) was provided by Iwo Kucinski. The gRNA expression vector was identical to that used for the remainder of the study but contained an alternative reporter gene, mCherry, instead of GFP. E-SLAM cells were sorted into HSC medium in a 96-well U-bottom plate and kept in culture for 7 days, after which they were analysed by flow cytometry (Fig. 4.1C). In the empty vector control, all $mCherry^+$ cells were $CD45^+$ (100.00±0.00% of $mCherry^+$ cells). In contrast, cells that were $mCherry^+$ after treatment with the *Ptprc* gRNA expression vector were predominantly $CD45^-$ (94.70±8.67% of $mCherry^+$ cells). These results demonstrated that the CRISPR gRNA successfully

perturbed the *Ptprc* gene in E-SLAM cells; furthermore, it indicated that the culture conditions were suitable for the CRISPR screen.

A



B



	CD45-	SD	CD45+	SD
Empty Vector	0.00	0.00	100	0
CD45 KO	94.70	8.67	0.32	0.45

Figure 4.1. Study design to investigate genes implicated in HSC biology. (A) Schematic of study design. Candidate genes were identified from the MolO and SuMO gene sets based of their expression patterns. gRNAs for each candidate gene were cloned into a CRISPR gRNA expression construct and transduced into E-SLAM HSCs from Cas9 transgenic mice. On day 7, cells were analysed for EPCR expression by flow cytometry. On day 10, cells were placed into CFU assays, which were analysed on day 20. (B) CD45 expression in E-SLAM cells treated with empty vector (control) or a *Ptprc* gRNA expression vector, analysed by flow cytometry after 7 days. CD45 is plotted against mCherry to distinguish cells that were not successfully infected with the mCherry vectors. A representative plot is shown. Table below the flow cytometry plots denotes the percentage of mCherry$^+$ cells expressing CD45 in both samples (n = 3; mean ± SD).

4.3. Identifying candidate genes important to HSC characteristics

To identify potential regulators of the HSC state, candidate genes were selected from the MolO and SuMO gene sets identified by Wilson *et al.*, as these represented genes associated with a molecularly similar HSC population and repopulating HSCs (N. K. Wilson et al. 2015). Initially, the expression of each gene was plotted on the HSPC atlas described in Chapter 3 (Fig. 4.2-4.4). The HSPC atlas was structured based on the gene expression of 1,654 single cells describing HSPC populations. The LT-HSCs (Lin$^-$ Sca1$^+$ c-Kit$^+$ Flk2$^-$ CD34$^-$) were found at the top of the branching

structure, and the E-SLAM HSCs (Lin^- $CD48^-$ $CD150^+$ $CD45^+$ $EPCR^+$) formed a more homogenous sub-cluster of the LT-HSC population (Fig. 4.2A). Twelve of the 44 genes were specifically expressed in the LT-HSCs only (Fig. 4.2B) including the recently suggested HSC marker *Neo1* (de Haan et al. 2017; Balazs et al. 2006). Seven more genes were most highly expressed in LT-HSCs but also expressed in neighbouring populations (Fig. 4.2C). This group of genes included the HSC marker *Procr*, the expression of which minimally extended outside the LT-HSCs, and *Pdzk1ip1*, which was also expressed in MEPs.

The genes which were also expressed in early progenitor cells were not outright eliminated as candidates; instead, their expression in the individual HSPC populations was quantified and assessed using violin plots (Fig. 4.3/4.4). All genes had varying expression in the HSPC populations; however, all the selected candidate genes had highest expression in LT-HSCs, and more specifically, E-SLAM cells (Fig. 4.3/4.4A). Some genes were also expressed in MPPs, including *Pdzk1ip1* and *Trim47*, but as their expression was still highest in the HSC populations, they were not excluded from the study. *Gbp6*, *Mapk12*, and *Neil2*, on the other hand, appeared to have relatively LT-HSC specific expression, yet the violin plots revealed their expression was similar across all populations, most likely due to very few cells expressing these genes at high levels (Fig. 4.4B). These genes were excluded from the study, along with 25 genes that were eliminated outright based on their expression outside the LT-HSC population (Fig. 4.5).

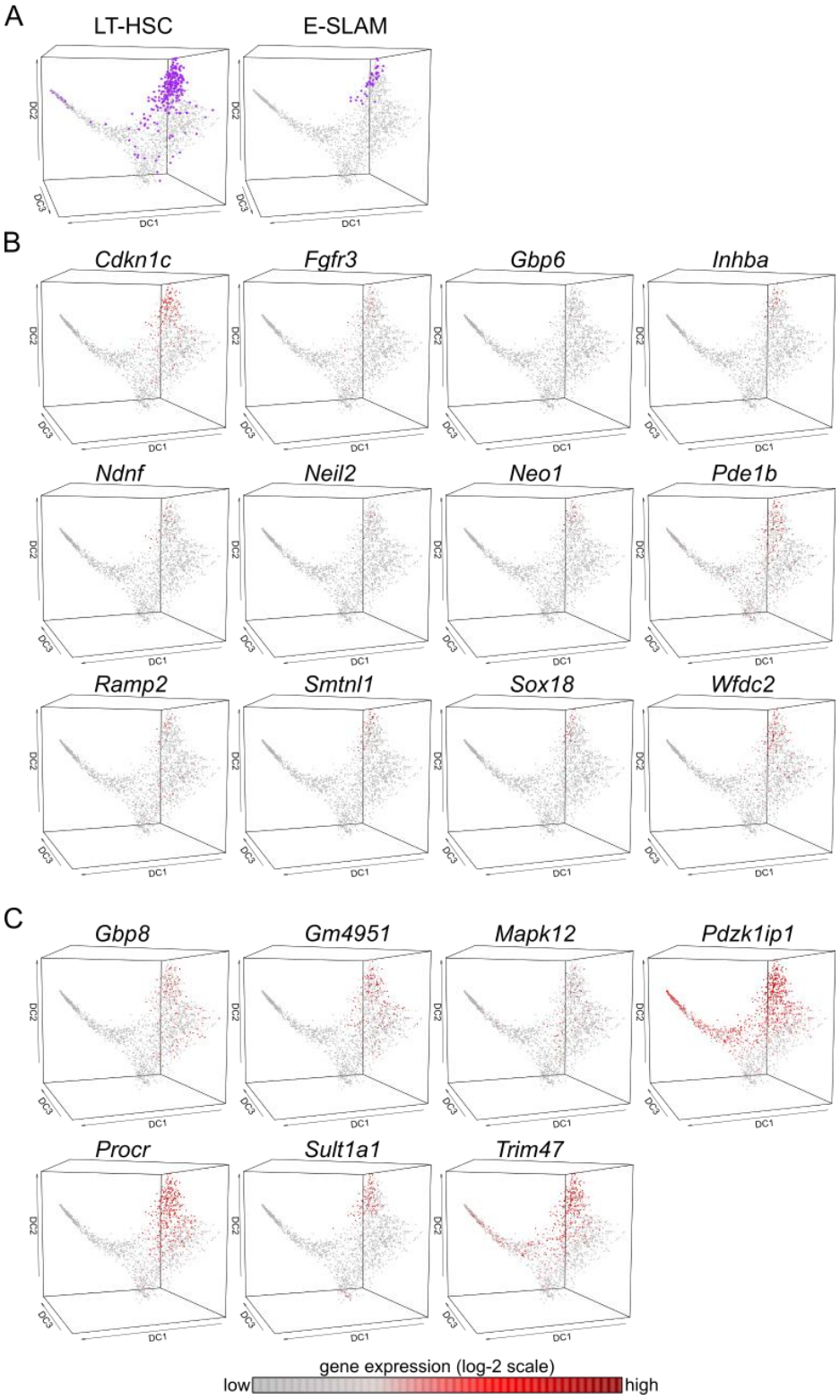


Figure 4.2. Expression plots of potential candidate genes for the CRISPR study. (A) Diffusion map of the HSPC atlas described in Chapter 3. LT-HSCs (Lin^- $Sca1^+$ $c\text{-}Kit^+$ $Flk2^-$ $CD34^-$) and E-SLAM cells (Lin^- $CD48^-$ $CD150^+$ $CD45^+$ $EPCR^+$) are coloured in purple and were assigned by retrospective gating. (B-C) Diffusion map of all cells coloured based on the expression of selected genes from the MolO and SuMO gene sets. Genes in (B) are specifically expressed in LT-HSCs, whereas genes in (C) are also expressed outside of the LT-HSC population to a varying extent. The colour corresponds to a $\log_2$ scale of expression ranging between 0 and the maximum value for each gene. Genes are ordered alphabetically. DC: Diffusion Component.

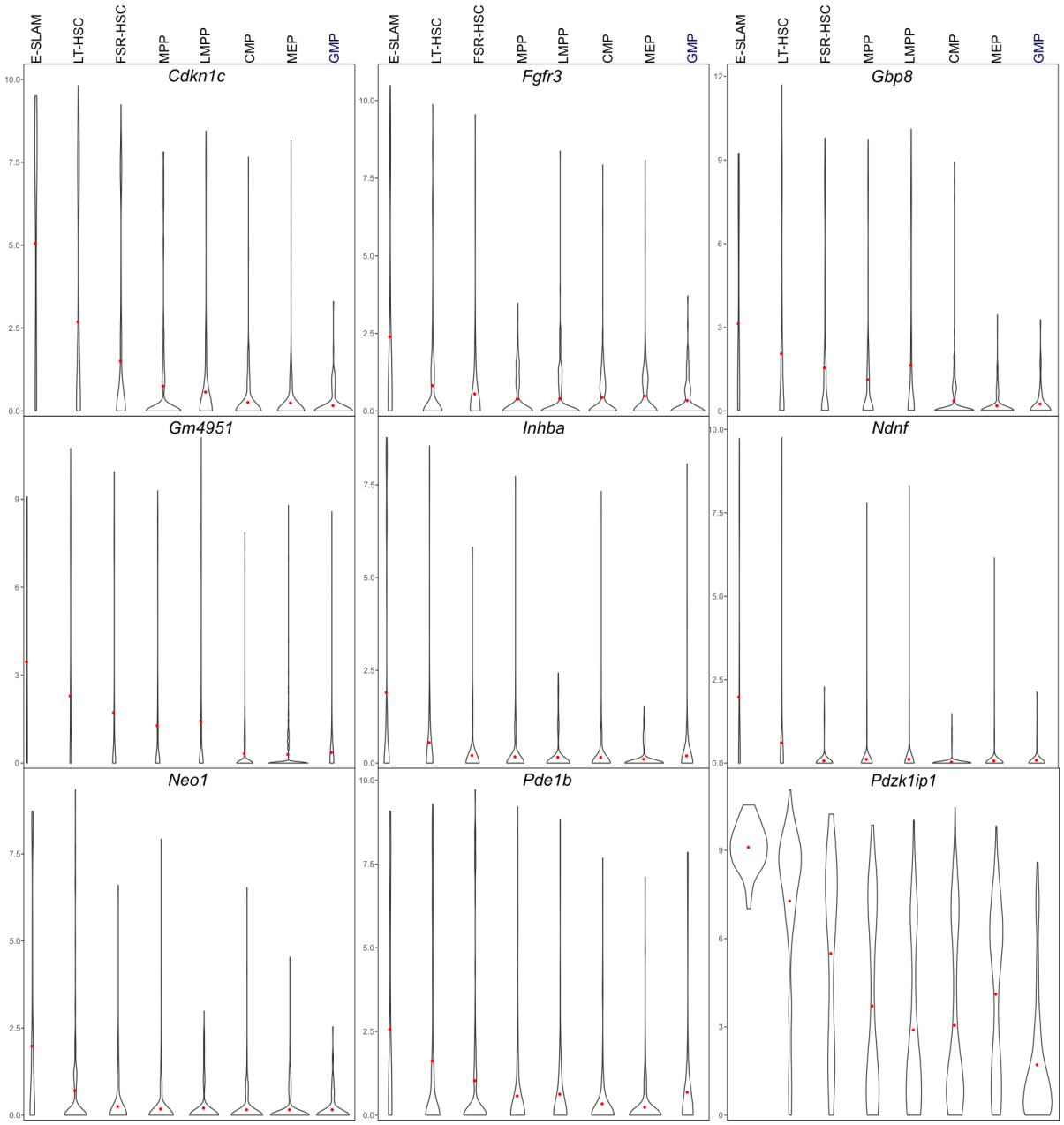


Figure 4.3. Violin plots for selected candidate genes. Violin plots detailing the expression of selected genes in specific HSPC populations. The width of each bar corresponds to the number of cells expressing the gene at the corresponding expression level; the red dot denotes the median expression of the gene in each population. Genes are ordered alphabetically.

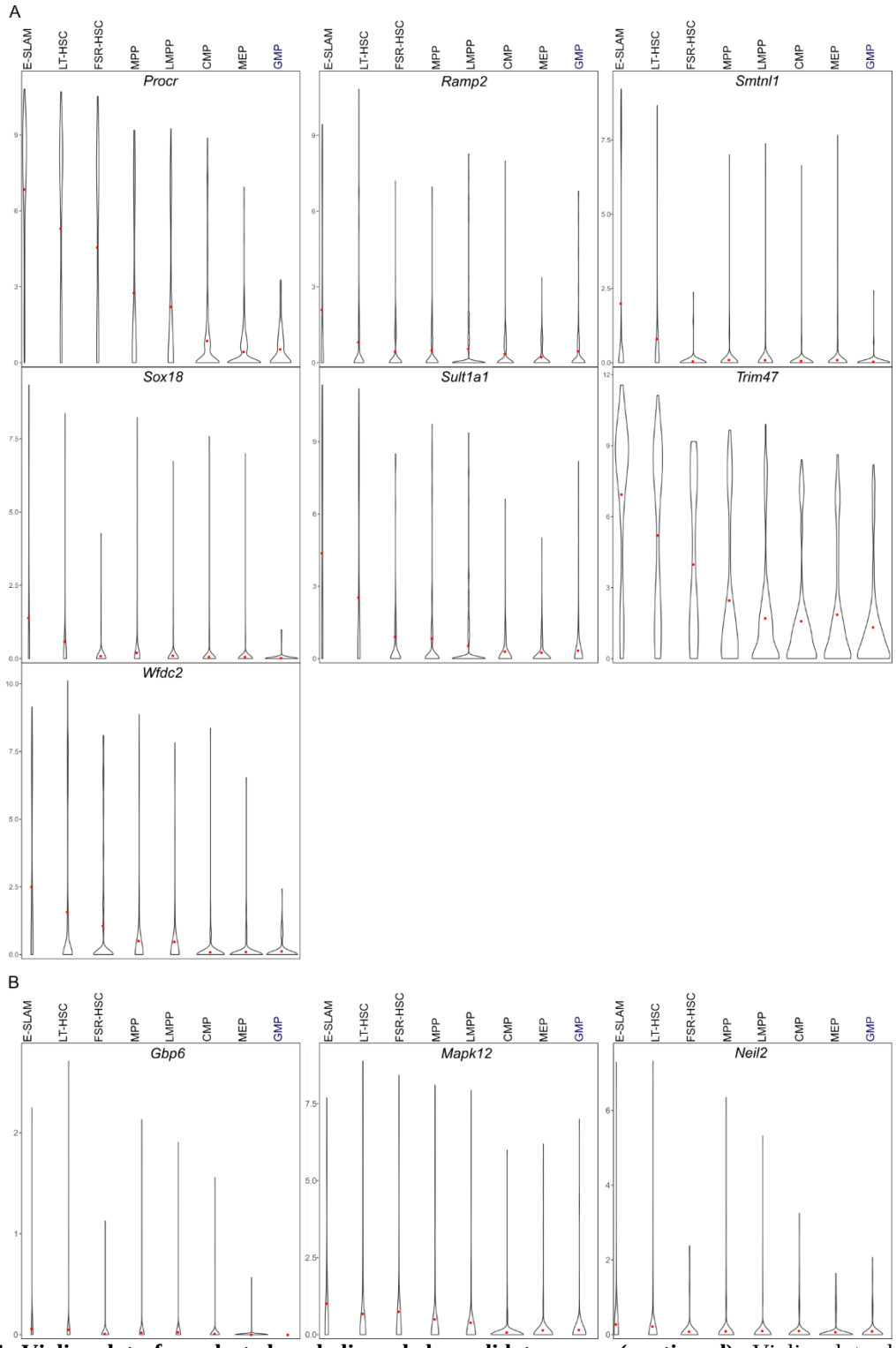


Figure 4.4. Violin plots for selected and discarded candidate genes (continued). Violin plots detailing the expression of selected genes in specific HSPC populations. The width of each bar corresponds to the number of cells expressing the gene at the corresponding expression level; the red dot denotes the median expression of the gene in each population. Genes are ordered alphabetically. (A) Continuation of candidate genes selected for the CRISPR study. (B) Expression of *Gbp6*, *Mapk12* and *Neil2* in HSPC populations. Although these genes appeared relatively HSC-specific on the HSPC atlas, they were excluded from further analysis based on their relatively constant expression across the atlas.

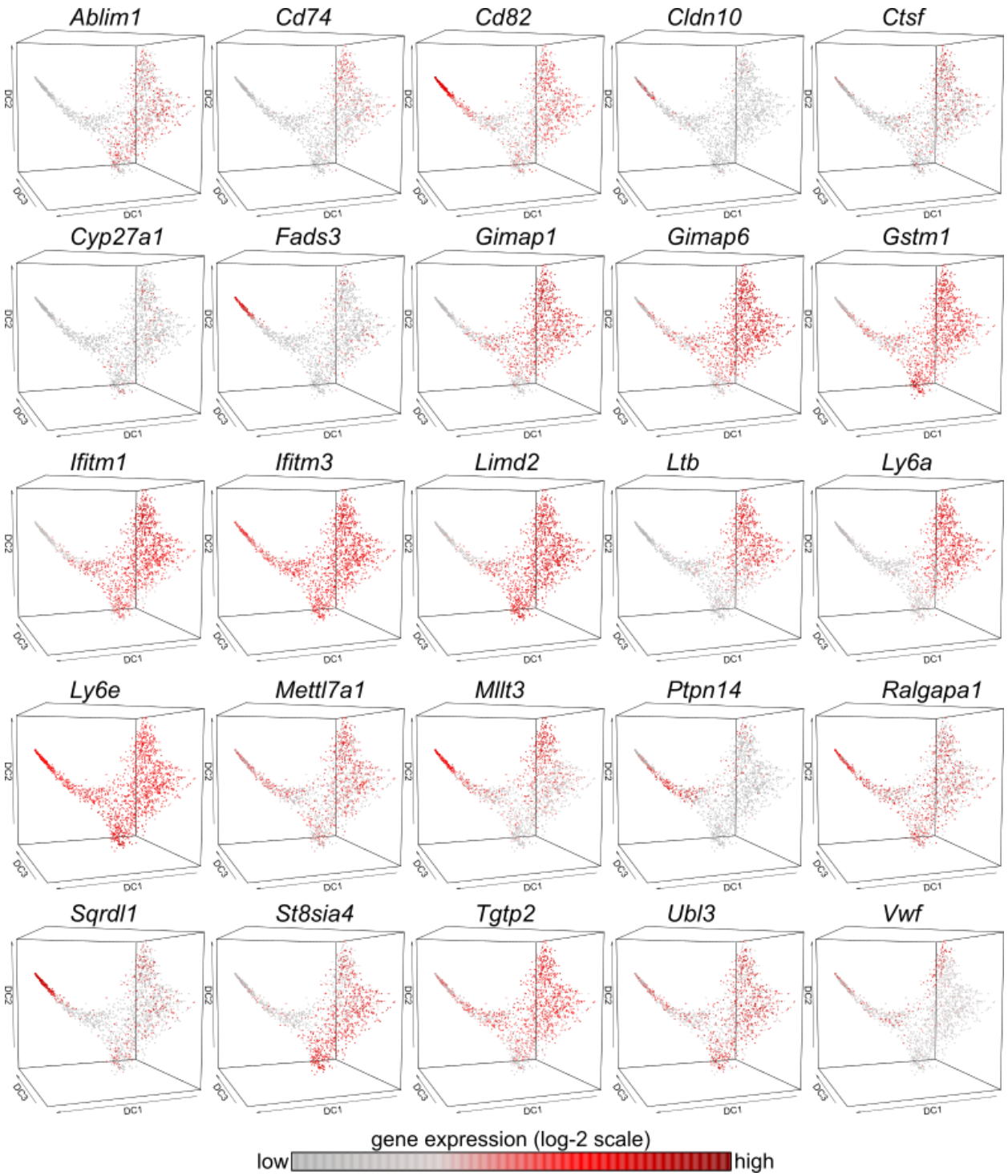


Figure 4.5. Expression plots of MolO and SuMO genes eliminated as candidates for the CRISPR study. Diffusion map of all cells coloured based on the expression of selected genes from the MolO and SuMO gene sets. These genes had broader expression across the HSPC atlas and were therefore excluded from the study. The colour corresponds to a $\log_2$ scale of expression ranging between 0 and the maximum value for each gene. DC: Diffusion Component.

4.4. CRISPR screen candidate genes

The gene list was narrowed down to 16 candidates based on their expression across the HSPC atlas. These 16 remaining candidates were more highly expressed in LT-HSCs and were therefore a good starting point for investigating the effect of perturbing HSC-enriched genes on HSC biology. These candidates are described in Table 4.2. As these genes were identified from a scRNA-seq experiment, rather than a curated haematopoiesis-specific list, many of the genes are poorly characterised and currently do not have a reported functional role in HSC biology. It may therefore be possible to discover novel HSC regulator candidates from this gene list.

Table 4.2. Properties of candidate genes selected for the CRISPR screen. Genes are ordered alphabetically.

Gene	Summary
Cdkn1c	• Cyclin dependent kinase inhibitor (p57) • Implicated in HSC quiescence • HSCs deficient in p57 have deficits in self-renewal and maintenance (Matsumoto et al. 2011; Zou et al. 2011)
Fgfr3	• Fibroblast growth factor (FGF) receptor • Involved in the negative regulation of definitive haematopoiesis during embryonic development (Pouget et al. 2014) • FGF signalling positively regulates adult HSCs (de Haan et al. 2003; Yeoh et al. 2006)
Gbp8	• Guanylate binding protein • Induced by interferon (Olszewski, Gray, and Vestal 2006) • Highly expressed in HSCs compared to the progenitor compartment (Ali et al. 2017)
Gm4951	• Interferon-gamma inducible GTPase • Higher expression in HSCs than progenitors • Induced by macrophages (Gautier et al. 2012)
Inhba	• Inhibin beta A • Differentially expressed in megakaryopoiesis during human HSC differentiation (Komor et al. 2005)
Ndnf	• Neuron-derived neurotrophic factor • Downregulated during differentiation towards MPP1 (Cabezas-Wallscheid et al. 2014)
Neo1	• Neogenin-1 • Recently identified as a marker and key regulator of HSC function (de Haan et al. 2017) • Regulates HSC quiescence and maintenance (Renders et al. 2017)
Pde1b	• Phosphodiesterase in the PDE1 family • Downregulated during differentiation towards MPP1 (Cabezas-Wallscheid et al. 2014) • The splice isoform PDE1B2 is upregulated after monocyte to macrophage differentiation (Lerner and Epstein 2006)
Pdzk1ip1	• PDZK1-interacting protein 1/Map17 • Expressed in HSCs but reduced in more differentiated cells (Cabezas-Wallscheid et al. 2014) • Preferentially expressed in HSCs with the most undifferentiated phenotype and lowest proliferation rate (Sawai et al. 2016)

Gene	Summary
Procr	• Endothelial protein C receptor (EPCR) • Marker of haematopoietic stem cells (Balazs et al. 2006)
Ramp2	• Receptor activity modifying protein • Essential in vasculogenesis (Ichikawa-Shindo et al. 2008)
Smtnl1	• Smoothelin-like 1 • Modulates smooth muscle contraction and relaxation (Borman, MacDonald, and Haystead 2004)
Sox18	• SRY-box 18 transcription factor • Upregulated in LT-HSCs compared to MPPs (Forsberg et al. 2005) • Implicated in vascular development (Hosking et al. 2004)
Sult1a1	• Sulfotransferase • Knockout mice are viable and have normal HSC function (Gazit et al. 2014)
Trim47	• Tripartite motif containing protein (GOA) • Trim47 is not functionally described in HSCs but has been identified as a candidate HSC gene (Gerrits et al. 2009; N. K. Wilson et al. 2015)
Wfdc2	• WAP four-disulphide core domain protein (HE4) • Protease inhibitor that may function in innate immunity (Chhikara et al. 2012) • Increased expression in ovarian carcinomas (Drapkin et al. 2005)

4.5. Isolation and analysis of E-SLAM HSCs using flow cytometry

Flow cytometry was used throughout this investigation to both isolate and analyse HSCs before and after treatment with gRNA expression vectors. Cells were isolated directly into a 96-well U-bottom plate using FACS (Fig. 4.6A). E-SLAM HSCs (Lin^- $CD48^-$ $CD150^+$ $CD45^+$ $EPCR^+$) were sorted as this isolation strategy enriches for cells that have multilineage and self-renewal potential at a high purity (D. G. Kent et al. 2009). Cells were isolated from both female and male Cas9 transgenic mice between 18 and 20 weeks old; 250 cells were sorted per well directly into HSC medium. Three replicate wells for each construct, including empty vector and cell-only controls, were obtained for each experiment. Once the cells were transduced with the individual gRNA expression vectors, they were kept in culture for 10 days with regular medium changes.

On day 7, approximately half of the cells in each well were stained with antibodies for flow cytometric analysis (Fig. 4.6B). Many surface makers change their expression in *in vitro* culture due to the culturing process as well as biological changes occurring to the cells; therefore, the full panel of stem cell and progenitor markers could not be used. The study focused on EPCR expression as it is an established marker of functional HSCs; a change in EPCR expression would suggest than the gene perturbations affected the proportion of cells in a primitive HSC state in culture. To determine whether treatment with the various gRNA expression vectors influenced EPCR expression, cells were first gated based on GFP expression. The percentage of $EPCR^+$

Sca1^{+} cells (EPCR^{+}) in the GFP^{+} Lin^{-} population was initially recorded. However, gating on the EPCR^{+} population is subjective as the investigator decides where to set the gates. Furthermore, there is only a small number of EPCR^{+} cells in culture, and integration of the different experiments across different days is challenging due to the differences in the intensity of fluorescent markers. Therefore, median EPCR expression in GFP^{+} Lin^{-} cells was also recorded for a more objective measure of the effect of perturbing genes implicated in HSC biology.

Later in the experiment, Annexin V BUV395 was added to the day 7 staining panel. Annexin V is a Ca^{2+}-dependent phospholipid binding protein that binds membrane phosphatidylserine that has translocated to the cell surface, an early event occurring during cell apoptosis (Hingorani et al. 2011). Annexin V was included to determine whether the CRISPR gRNAs caused apoptosis in the perturbed cells. To determine whether the cells were dying or already dead, 7-AAD was also used: apoptosing cells were Annexin V positive but 7-AAD negative, whereas dead cells were positive for both markers (Fig. 4.6C).

The cells that were not analysed on day 7 were pooled for each replicate and kept in culture for three more days to be sorted for CFU assays and genotyping (Section 4.8-4.9). On day 10, cells were sorted based on GFP expression only. For CFU assays, 100, 200 or 400 cells were sorted into HSC medium, transferred into M3434 methylcellulose, and plated in 6-well SmartDish^{TM} plates, to be kept in culture for an additional 10 days. Where possible, the remaining cells were sorted into FACS buffer (2%FBS/PBS) to be used for genotyping analysis.

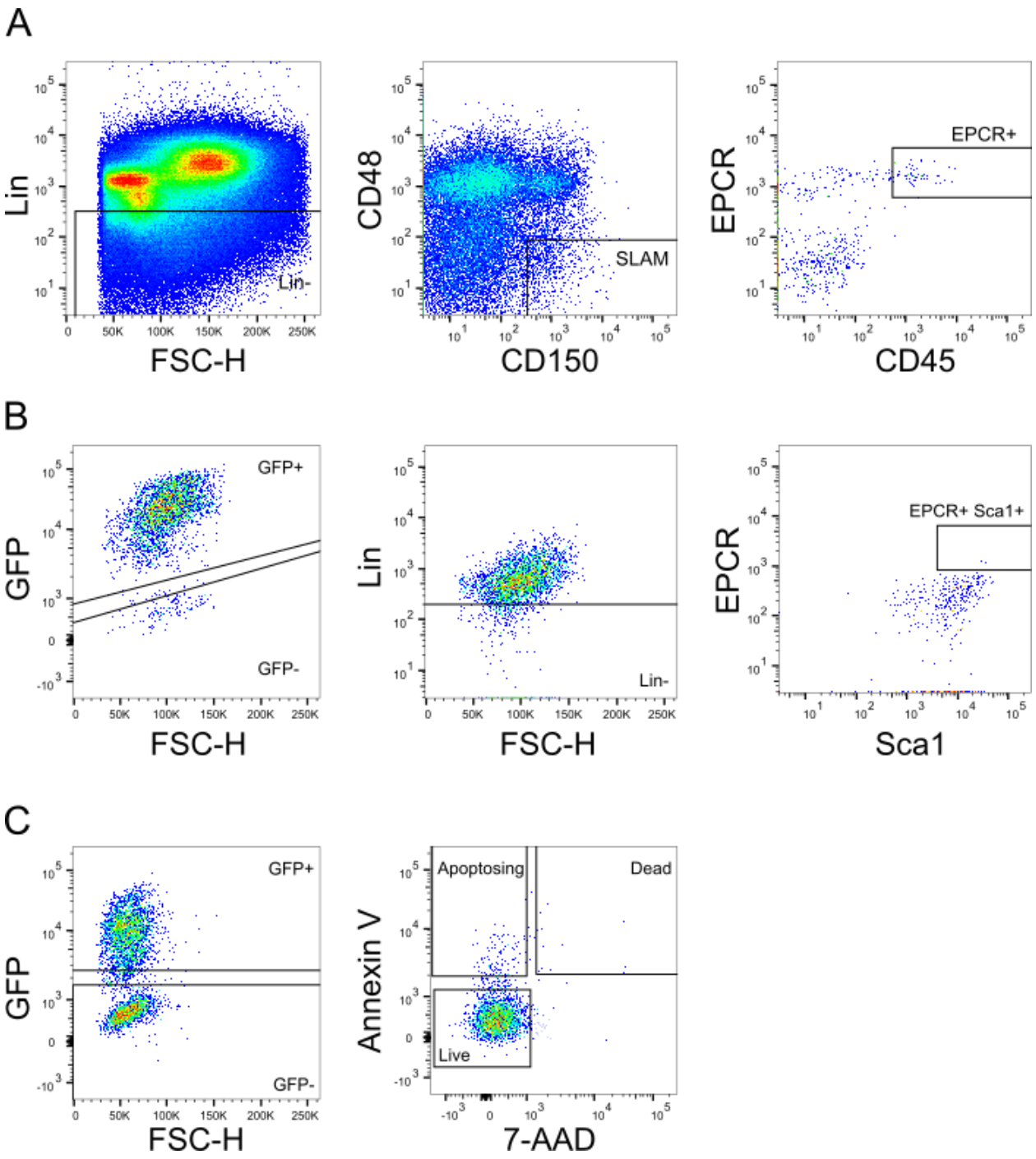


Figure 4.6. Flow cytometry sorting and analysis strategies for E-SLAM cells. (A) Sorting strategy for isolating E-SLAM cells (Lin^- $CD48^-$ $CD150^+$ $CD45^+$ $EPCR^+$) (B) Gating strategy for analysing EPCR expression on day 7 after treatment of E-SLAM cells with gRNA expression vectors. Cells were gated based on GFP, Lineage marker (Lin), EPCR and Sca1 expression. The percentage of $EPCR^+$ $Sca1^+$ cells in the Lin^- and GFP^+ gates was recorded, along with the median EPCR expression of Lin^- cells. (C) Gating strategy for analysing apoptosis on day 7 after treatment of E-SLAM cells with gRNA expression vectors. Cells were gated on GFP, 7-AAD and Annexin V expression. Apoptosing cells were negative for 7-AAD and positive for Annexin V.

4.6. Effect of candidate gene perturbations on EPCR expression in E-SLAM cells

To investigate the effect of perturbing the candidate genes on HSC characteristics, EPCR expression was analysed by flow cytometry after culturing the cells for seven days. EPCR has been shown to be representative of HSC activity even after *in vitro* culture, and was therefore the focus of this investigation (Balazs et al. 2006).

The investigation was performed over twelve separate screens due to the rarity of E-SLAM cells, as well as limitations in the numbers of mice and FACS time available. The percentage of $EPCR^+$ cells in the GFP^+ Lin^- population was recorded for each gRNA used and normalised to the empty vector control (Fig. 4.7A). The percentage of $EPCR^+$ cells does not significantly change compared to the empty vector for the majority of candidate genes. Two *Ramp2* gRNAs, Ramp2_g1 and Ramp2_g2, appeared to cause a significant increase in $EPCR^+$ cells ($p<0.0001$). In one screen, there were technical issues with the empty vector, but the cell-only control, which was a sample of $EPCR^+$ cells that was not treated with any gRNAs, was consistent with the cell-only controls and empty vector controls in the other eleven screens. The samples from this screen were therefore normalised against the cell-only control (denoted with a ^ in Fig. 4.6A). All *Pdzk1ip1* gRNAs and one *Wfdc2* gRNA caused a significant increase in the percentage of $EPCR^+$ cells.

However, it is difficult to have confidence in these percentages when gating on small numbers of cells. Visualising the flow cytometry results showed that large differences in percentages could actually be caused by very small changes in cell numbers (Fig. 4.7B). All gRNAs targeting *Pdzk1ip1* caused a significant increase in the percentage of $EPCR^+$ cells compared to the cell-only control. However, the figures clearly show that the cell-only control had more cells than any of the *Pdzk1ip1* samples. Furthermore, the numbers of cells that contribute to the $EPCR^+$ population varies greatly between replicates for each individual gRNA: for the Pdzk1ip1_g1 samples, the $EPCR^+$ population includes 10, 2 and 4 cells in the individual replicates (an average of 5.33 ± 4.16 $EPCR^+$ cells overall). Therefore, it is difficult to draw conclusions from these percentages and be confident that the statistical significance reflects biologically significant results.

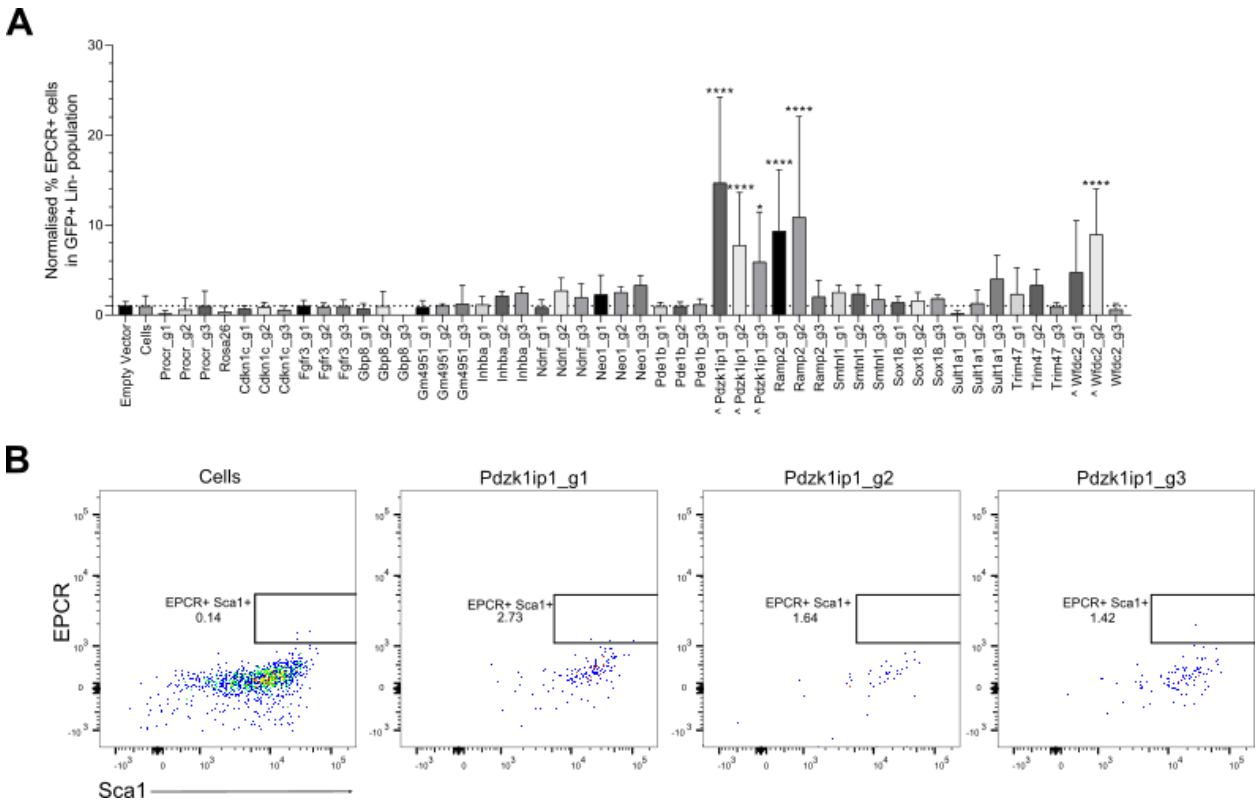


Figure 4.7. Changes in the percentage of EPCR+ cells after candidate gene perturbation. (A) EPCR expression after candidate gene perturbation, represented by the normalised percentage of EPCR$^+$ cells. All results are normalised to the empty vector, except for those marked with a ^, which were normalised against the cell-only control due to a technical issue with the empty vector. Samples are shown according to the gRNA with which they were treated. The dotted line marks the normalised EPCR expression of the empty vector for easy comparison. Flow analysis was performed seven days after E-SLAM cells were transduced with CRISPR gRNAs targeting HSC candidate genes. Significance was determined using a one-way ANOVA and Dunnett's multiple comparisons test. Mean ± SD (n = 3); * p<0.05; ** p<0.01; *** p<0.001; ****p<0.0001 (B) Example flow cytometry plots of the percentage of EPCR$^+$ cells present in culture seven days after candidate gene perturbation. The gating strategy for these cells was: GFP$^+$ Lin$^-$ EPCR$^+$ Sca1$^+$, as described in Fig. 4.6. The numbers in each plot denote the percentage of EPCR$^+$ Sca1$^+$ cells in the GFP$^+$ Lin$^-$ population.

Compared to measuring expression changes using percentages, median fluorescence intensity measurement is a more objective approach to visualising changes in surface marker expression that is also less affected by fluctuations in cell numbers. Within a selected gate, the median fluorescence of any fluorophore used in the antibody-staining panel can be calculated. The results can then be compared to an empty vector or untreated control to determine whether there was a shift in surface marker expression.

Median EPCR expression was measured for GFP$^+$ Lin$^-$ cells in the perturbed samples. Normalising the median EPCR expression to the empty vector showed that the perturbations did not significantly affect EPCR expression for most candidates, and previous candidates, identified from the percentages of EPCR$^+$ cells, are no longer valid (Fig. 4.8A). *Procr* encodes the EPCR protein and

was included as this gene is highly expressed in HSCs and serves as a control as perturbation of *Procr* should significantly reduce EPCR expression. Two *Procr* guides, Procr_g1($p<0.0001$) and Procr_g2 ($p<0.01$) showed decreased median expression. The *Ramp2* guide Ramp2_g1 caused a significant increase in median expression ($p<0.01$), as did the *Wfdc2* guide, Wfdc2_g2 ($p<0.0001$). Wfdc2_g3 ($p<0.01$), on the other hand, significantly decreased EPCR expression in GFP^+ Lin^- cells. The significant increase in $EPCR^+$ cells caused by perturbing *Pdzk1ip1* was not reflected in the median EPCR expression.

The significant results observed were caused by only one or two gRNAs for each gene. In the case of *Wfdc2*, the gRNAs caused opposing results. The median EPCR expression caused by each gRNA was pooled for each gene to determine whether their effects remained significant (Fig. 4.8B). Only *Procr* caused a significant perturbation and reduced the median EPCR expression by 33% compared to the empty vector (0.67 ± 0.28, $p<0.0001$). The results obtained by pooling the individual gRNA results demonstrate that overall, the candidate gene perturbations did not impact EPCR expression and that the effect caused by one gRNA is not enough to draw a conclusion, as it could be caused by an off-target effect and may not accurately represent the perturbed phenotype.

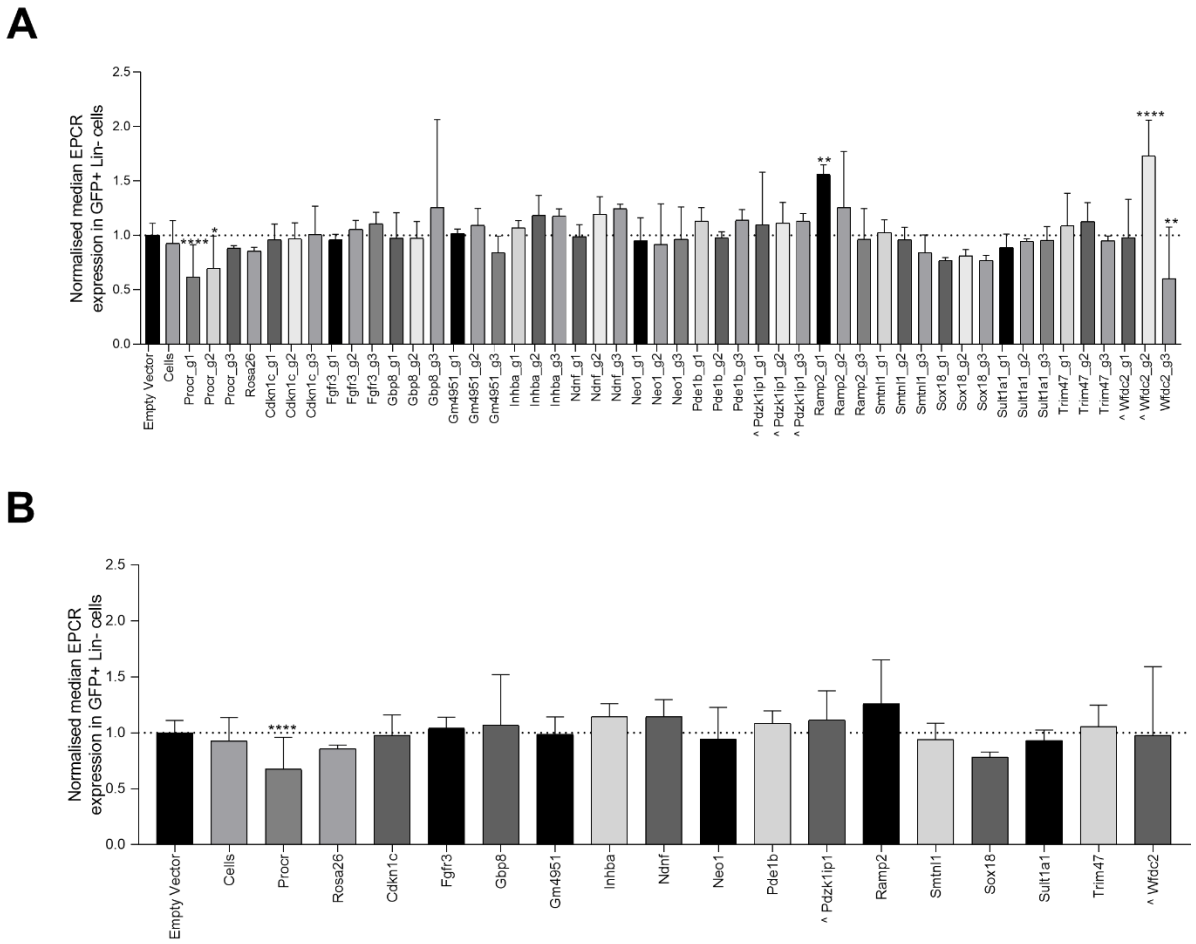


Figure 4.8. Changes in median EPCR expression after candidate gene perturbation. (A) EPCR expression after candidate gene perturbation, represented by the median EPCR expression in GFP^+ Lin^- cells. All results are normalised to the empty vector, except for those marked with a ^, which were normalised against the cell-only control due to a technical issue with the empty vector. Samples are shown according to the gRNA with which they were treated. The dotted line marks the normalised EPCR expression of the empty vector for easy comparison. Flow analysis was performed seven days after E-SLAM cells were transduced with gRNAs targeting HSC candidate genes. Mean ± SD (n = 3) (B) Median EPCR expression after candidate gene perturbation, grouped by target gene. The results from perturbing cells using three gRNAs for each target gene were pooled. The dotted line marks the normalised EPCR expression of the empty vector for easy comparison. Significance was determined using a one-way ANOVA and Dunnett's multiple comparisons test. Mean ± SD (n = 9); * $p<0.05$; ** $p<0.01$; *** $p<0.001$; ****$p<0.0001$

4.7. Candidate gene perturbation does not influence apoptosis in E-SLAM cells

All wells started with 250 sorted E-SLAM HSCs; however, on day 7, it was often observed that the cells treated with CRISPR gRNAs had variable cell counts. To determine whether the candidate gene perturbations were causing apoptosis, which could also contribute to the lack of significant change in EPCR expression, Annexin V was added to the staining panel. Annexin V staining detects cells in the early stages of apoptosis by binding to membrane phosphatidylserines that have

translocated to the cell surface (Hingorani et al. 2011). As it was added to the panel in later screens, not all candidate gene perturbations have recorded results for apoptosis (Fig. 4.9).

Apoptosing cells were marked by a negative 7-AAD and positive Annexin V profile. The percentage of live, dead, or apoptosing cells were normalised to the empty vector control in each case. No significant differences were observed in all categories (Fig. 4.9A). The number of cells that were dead or dying in all samples was variable and contributed to by small cell numbers, resulting in large standard deviations for each sample.

The cells that were perturbed with gRNAs targeting *Gbp8* are shown separately due to the great amount of variation in the percentage of dead cells in these samples (Fig. 4.9B). Gbp8_g3 appeared to increase cell death 284 times more than the empty vector (283.91±491.74); however, visualising the data reveals that very few cells are contributing to these results (Fig. 4C). The Gbp8_g3 gRNA only had 0.53% efficiency, effectively transducing 77 cells, most of which were live cells. The dead cell counts for the Gbp8_g3 replicates show that only one repeat is contributing to the large number of dead cells observed (Fig. 4D).

Overall, there is a great degree of variation in the number of dead and dying cells in the perturbed samples; however, this variation appears to be the result of very few cells contributing to these populations. Based on these observations, perturbing the candidate genes did not cause a reliably detectable change in apoptosis, and therefore an increase in cell death or apoptosis is not the main reason for a lack of significant changes in EPCR expression.

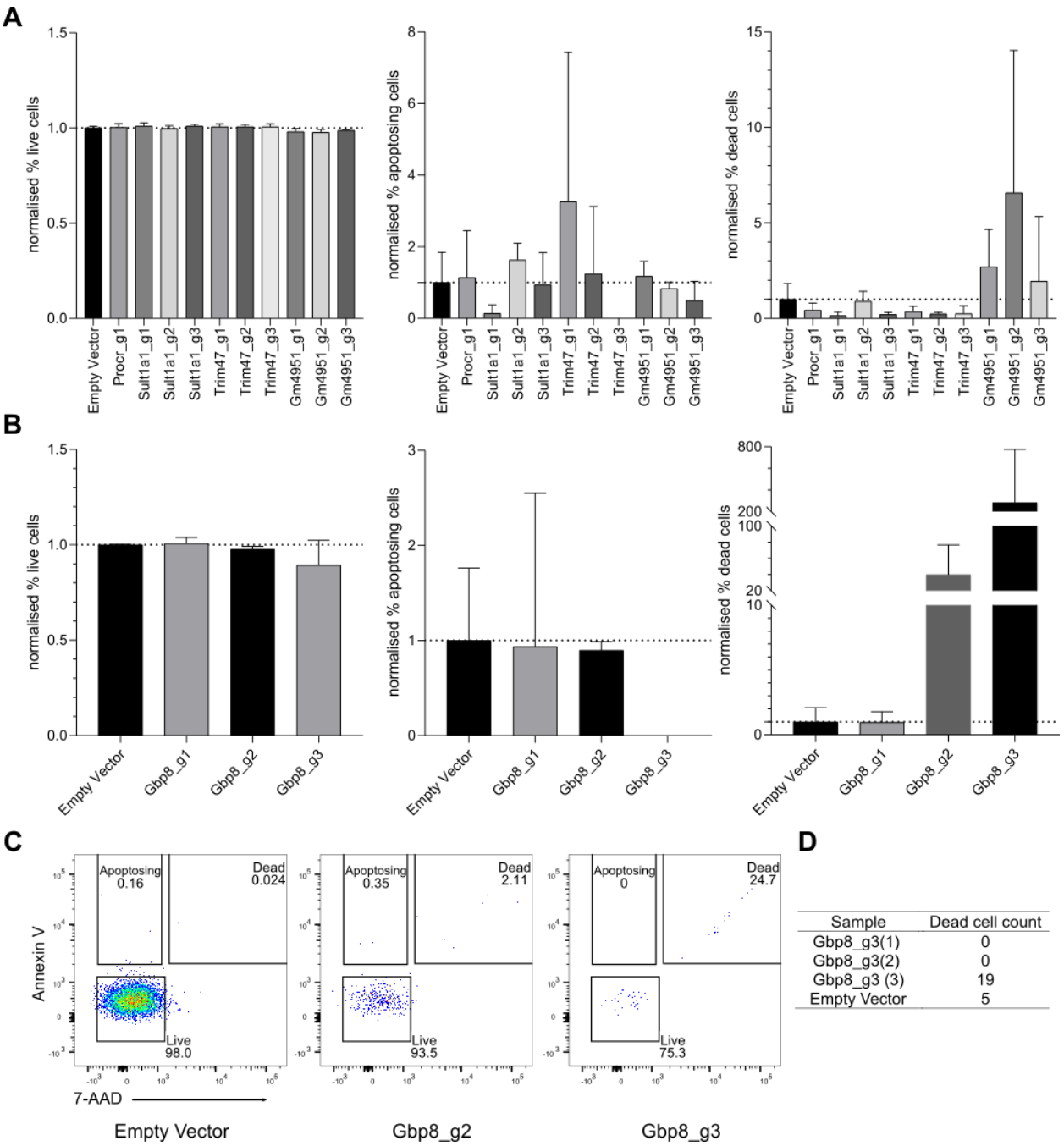


Figure 4.9. Candidate gene perturbation does not influence apoptosis in E-SLAM cells. (A) Normalised percentage of live (left), apoptosing (middle), and dead (right) cells after candidate gene perturbation in E-SLAM cells, recorded after seven days in culture. All results are normalised to the empty vector. The dotted line marks the percentage of each cell population in the empty vector sample for easy comparison. Mean ± SD (n = 3). (B) Normalised percentage of live (left), apoptosing (middle), and dead (right) cells after *Gbp8* perturbation in E-SLAM cells, recorded after seven days in culture. The percentage of cells in each gate is normalised to the empty vector. These results are pictured separately due to the variation in dead cells after *Gbp8* perturbation. Mean ± SD (n = 3). (C) Flow cytometry plots of the percentage of GFP^+ live/apoptosing/dead cells after *Gbp8* perturbation with Gbp8_g1 and Gbp8_g2. These results are compared to the empty vector. (D) Table showing the number of cells contributing to the dead cell gate in the three Gbp8_g3 replicates compared to the empty vector.

4.8. Changes in lineage output after candidate gene perturbation in E-SLAM cells

Although significant changes in EPCR expression were not observed for most candidate genes, it was still possible that the perturbations may cause a change in colony output due to changes in transcriptional networks governing differentiation lineages.

After E-SLAM HSCs were sorted into 96-well U-bottom plates, they were transduced with CRISPR gRNAs targeting the candidate genes and kept in culture for 10 days. On day 10, GFP^+ cells were sorted from the perturbed samples and cultured in MethoCult™ GF M3434 (StemCellTechnologies). Either 100, 200 or 400 cells were sorted, resuspended in MethoCult, and split between two wells for culturing. On average, these starting cell counts represented 5.5%, 11%, and 22% of all GFP^+ cells in each sample, respectively. The methylcellulose is supplemented with cytokines which support the growth of primitive erythroid progenitors (BFU-E), granulocyte-macrophage progenitor cells (CFU-G/M/GM), and multi-potential granulocyte, erythroid, macrophage, megakaryocyte progenitor cells (CFU-GEMM) (C. L. Miller and Lai 2005). The cultures were analysed after 10 days using the STEMvision™ system, which automatically images and counts colonies in haematopoietic CFU assays.

As CFU assays were added to the protocol for later screens, not all candidate gene perturbations have recorded results for colony output (Fig. 4.10). Low cell numbers in the GFP^+ Lin^- gate and limitations in the initial sort (i.e. number of mice available and length of sorting time) meant that the assays could not be repeated. The empty vector and cell-only controls were included in every screen (n=3). Furthermore, Procr_g1 and Gbp8_g1 were included in two screens (n=2) and give an indication of whether the colony output results are reliable. Due to low cell numbers, a 400 starting cell culture was not included for *Sult1a1* and two *Trim47* gRNAs.

Although all three colony types were measured, there were very few CFU-GEMM and BFU-E colonies in all samples, with little variation observed between the perturbations (Fig. 4.10A). The predominant colony type was CFU-G/M/GM. The remaining analyses looked at the total colony counts only. Visualising the total number of colonies demonstrated a proportional increase in colonies produced from increasing numbers of starting cells in all perturbation samples (Fig 4.10B). This was consistent among the individual gRNAs targeting each candidate gene. Therefore, the colony counts for the gRNAs were pooled to compare the lineage output in perturbed samples

versus the empty vector control (Fig. 4.10C). These results were collected in three separate experiments and the colony counts were normalised against the relevant empty vector for each individual screen. Pooling the results for each gRNA made it possible to perform statistical analyses, which showed that there were no statistically significant differences in colony output in all perturbed samples compared to the empty vector. Therefore, it appears that perturbing the candidate genes had no effect on lineage output.

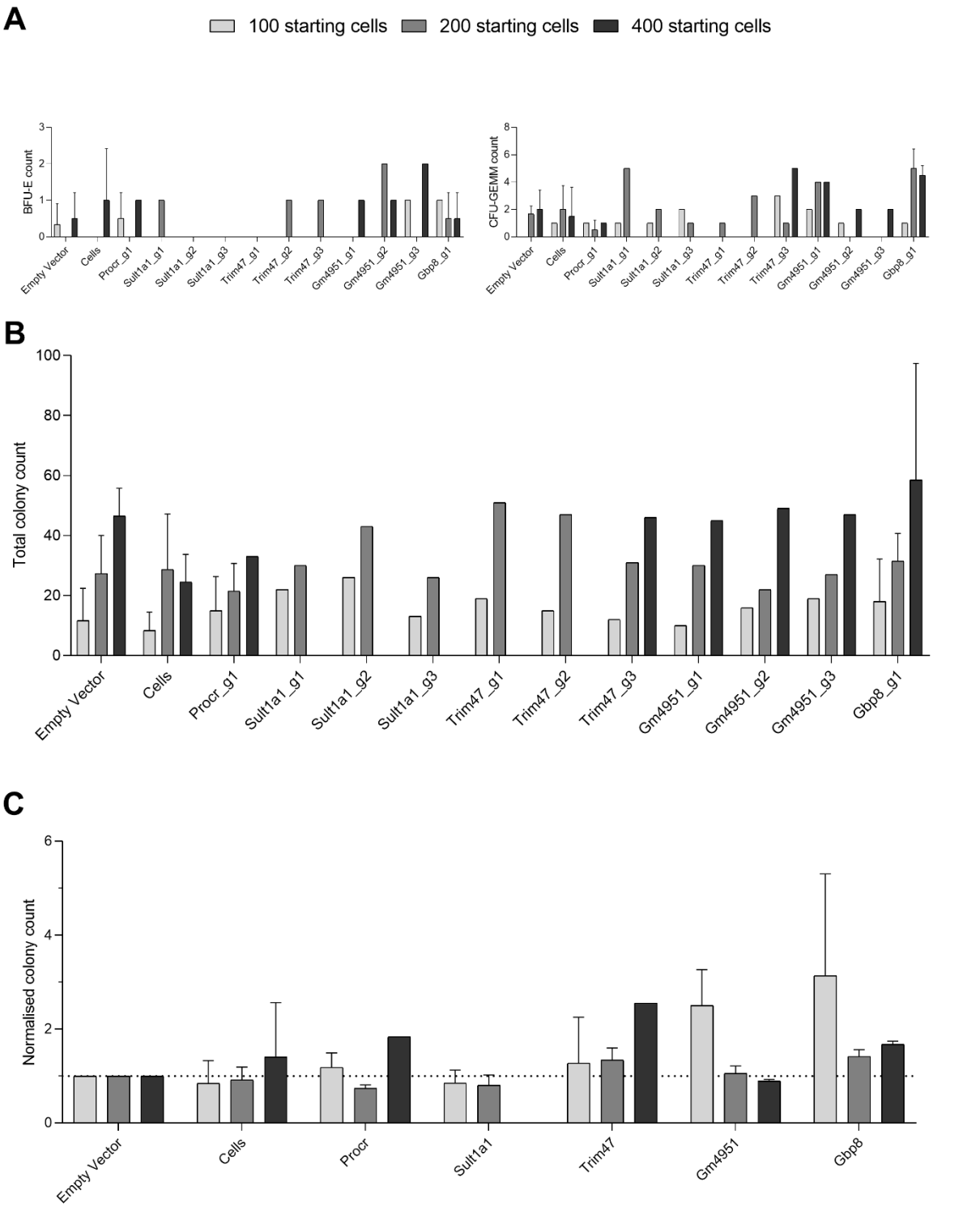


Figure 4.10. Changes in lineage output after candidate gene perturbation. GFP$^+$ cells were sorted 10 days after infection of E-SLAM HSCs with CRISPR gRNA expression vectors against candidate genes. The cells were sorted at 100, 200 or 400 starting cell densities and cultured for 10 days in MethoCult^{TM} GF M3434. Only Gbp8_g1 was repeated (n = 2), as well as cell-only and empty vector controls, which were included in each individual experiment (n = 3). (A) Colony count of BFU-E (left) and CFU-GEMM (right) for all perturbations, resulting from 100, 200 or 400 starting cell densities. (B) Total colony output for all perturbations from 100, 200 or 400 starting cell densities. (C) Total colony counts normalised to empty vector from 100, 200, and 400 starting cell densities. A dotted line shows the normalised colony count for the empty vector sample for easy comparison. The results from each individual gRNA perturbations were pooled for each candidate gene (n=3; mean ± SD). Statistical analysis was performed using a one-way ANOVA and Dunnett's multiple comparison test.

4.9. Genotyping confirms CRISPR gRNAs are correctly targeting the candidate genes

As the flow analysis and colony assays did not yield significant results, genotyping analysis was performed on a sample of perturbed candidate genes to confirm whether the CRISPR gRNAs are correctly targeting the genes of interest.

On day 10 post-infection with CRISPR gRNA expression vectors, GFP^+ cells were sorted from cells treated with *Procr*, *Trim47* and *Wfdc2* gRNAs, as well as an empty vector (WT) control. *Procr* was included because it was used as a control for flow analysis, and therefore it was important to confirm whether it was correctly perturbed, whereas *Trim47* and *Wfdc2* were analysed as a random subsample of the candidate genes. The genomic DNA (gDNA) was isolated from each sample and prepared for genotyping. Briefly, each target sequence was amplified with specifically designed primers, one of which was designed to be 60-100bp away from the 20bp cut site, and the other was up to 320bp away, for a total PCR product length of 300-500bp. The sequencing libraries were then prepared by adding a unique combination of index primers to each sample. The libraries were then pooled and cleaned up, after which they were sent for sequencing on the MiSeq Nano.

The sequencing results were analysed against a reference sequence for the correct starting sequence, perfect matches with the reference sequence, indels and frameshift mutations (Fig. 4.11). All samples had at least 60% reads with the correct starting sequence. In the WT samples, at least 47% of the reads starting with the correct sequence were a perfect match to the reference sequence, whereas less than 10% of reads in all CRISPR perturbations (CRISPR KO) matched the reference. The CRISPR KOs all had a higher fraction of indels and frameshift mutations than the WT, indicating that the genes were successfully targeted by the gRNAs. It was expected that roughly two-thirds of the indels would cause frameshift mutations, which is reflected in the results. The WT *Wfdc2* sample had a relatively high fraction of indels and frameshift mutations, at 31% and 21%, respectively, which may be due to PCR clean-up issues as well as a high GC content in the sequence that may have caused sporadic PCR amplification errors. Overall, it appears that the gRNAs designed to perturb the HSC candidate genes were correctly targeting the genes of interest.

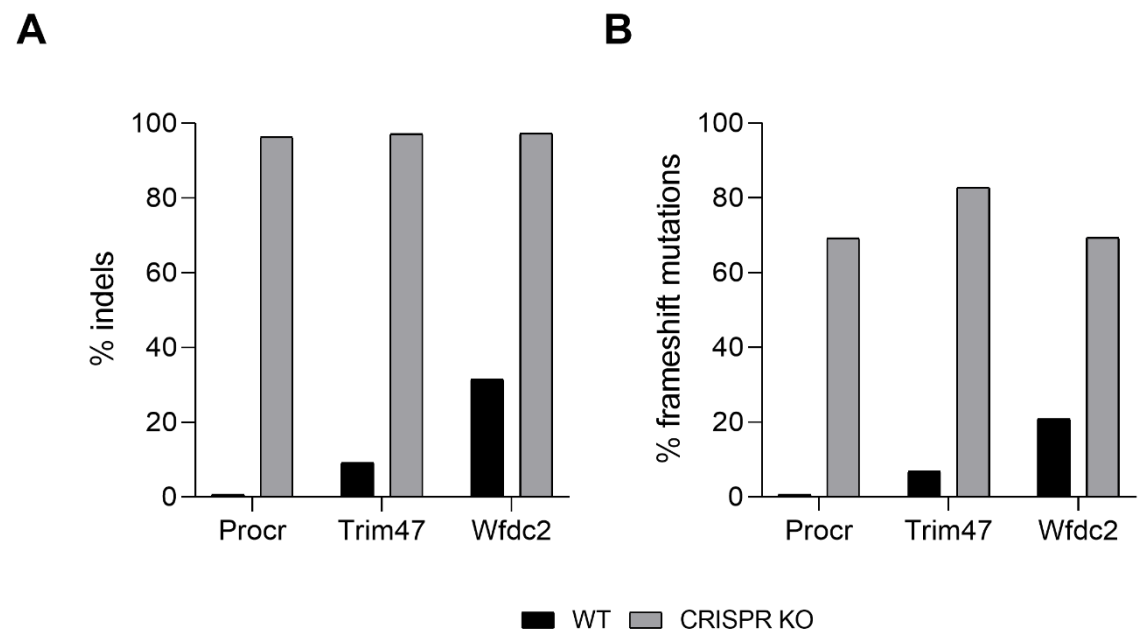


Figure 4.11. Genotyping analysis indicates the CRISPR gRNAs are correctly targeting the candidate genes. Results from genotype sequencing of three candidate gene perturbations samples (CRISPR KO/grey) and a control sample (WT/black). (A) Percentage of indels in sequenced samples. (B) Percentage of frameshift mutations in sequenced samples. 10,000 reads were analysed for each sample.

4.10. Conclusions

The main aim of this chapter was to investigate the effect of perturbing genes that were implicated in HSC biology. This was attempted by designing a CRISPR screen for a curated gene list and assessing changes in HSC characteristics using flow cytometry and CFU-assay analysis.

Chapter 3, in which the single-cell HSPC atlas was introduced, focused on using different dimensionality-reduction methods to best represent the data, as well as ordering cells through pseudotime to elucidate lineage trajectories. However, the HSPC atlas could be further used to guide *in vitro* investigations into genes of interest. Gene expression can be visualised across the entire HSPC trajectory to determine whether a gene is specific to an HSPC population or expressed throughout the entire HSPC compartment; doing so can identify genes from an existing list that may be interesting to investigate further.

To study HSC-specific genes, the gene list used for this investigation was defined from the MolO and SuMO gene lists previously described (N. K. Wilson et al. 2015). These genes were positively associated with an HSC subpopulation with similar gene expression and in vivo repopulation activity, respectively, and were therefore an ideal dataset for investigating genes involved in HSC

characteristics. The 16 candidate genes were perturbed in primary E-SLAM cells from Cas9 transgenic mice using CRISPR gRNA expression vectors with GFP markers. The treated cells were cultured and analysed by flow cytometry for changes in EPCR expression and induction of apoptosis. CFU-assays were performed to determine whether the gene perturbations influenced colony output.

EPCR is a well-established marker of HSCs and its expression in HSC marks cells with long-term bone marrow reconstitution potential (Balazs et al. 2006; Hiroko Iwasaki et al. 2010). To determine whether the gene perturbations influenced HSC function, the study focused on changes in EPCR expression. EPCR retention in culture is indicative of a more primitive cell that should not express lineage markers and should be less proliferate; therefore, a decrease in EPCR positive cells could indicate a loss of cells that are in a primitive HSC state and a potential skewing towards more mature progenitor populations. Changes in EPCR expression were assessed as both a percentage of GFP^+ Lin^- cells, as well as the median EPCR expression in the same population. The number of $EPCR^+$ cells after seven days in culture was variable but generally very low, and therefore percentages were determined by a small number of cells and difficult to compare between experiments. Median EPCR expression was considered a more objective measure, as it did not require any gate setting by the investigator and was a more accurate representation of EPCR expression in the perturbed cells. Interestingly, no significant decreases in EPCR expression were noted, except for in the *Procr* treated cells. *Procr* encodes the EPCR protein; therefore, it was reassuring to see that perturbing this gene significantly reduced EPCR expression in this study design.

The effect of candidate gene perturbation on apoptosis was also interrogated to determine whether an increase in apoptosis influenced the lack of significant results observed. Apoptosing cells were marked by positive Annexin V and negative 7-AAD expression; however, none of the perturbations caused a significant change in the percentage of live, dead, or apoptosing cells. Overall, it appears that after 10 days in culture, the gene perturbations did not impact EPCR expression nor apoptosis.

Although there was a lack of significant changes in surface marker expression, it was possible that the candidate gene perturbations do not influence EPCR expression but may impact the overall colony output of GFP^+ cells. CFU assays cultured over 10 days were used to investigate whether there was a shift in differentiation potential after perturbing the genes, demonstrated by changes in total colony count. The CFU assays were performed using 100, 200, and 400 starting cell densities,

which showed a proportional increase in the number of colonies. The results from the gRNAs targeting each gene were pooled, allowing for statistical analysis, which showed that the gene perturbations did not influence lineage output.

Due to the lack of significant results observed during flow cytometry and CFU-assay analysis, genotyping analysis was performed on a small sample of perturbed candidate genes and confirmed that the designed gRNAs were successfully causing indels and frameshift mutations in the treated cells.

The results suggest that while the CRISPR gRNAs were effectively targeting the candidate genes, their perturbation did not significantly impact HSC biology when assayed with the *in vitro* tests used in this investigation. Loss of function of a single gene may not necessarily influence the overall function of a cell due to genetic compensation or gene redundancies (El-Brolosy and Stainier 2017; Velten et al. 2017). However, the lack of significant results also suggests that there are weaknesses in the study design, which are discussed below.

4.10.1. Limitations

Flow analysis and CFU assays were performed seven and ten days after the initial perturbation, respectively. Due to low cell numbers in the *Sult1a1* and *Trim47* samples, a 400 starting cell culture was not included. The reduced cell count may suggest that the gRNAs were killing the cells, which may have occurred during an earlier time point and therefore the cells that managed to survive to day 10 were less effected. Furthermore, it is possible that analysing the cells at day 7 or day 10 is too late and the perturbations may have exerted their effect at an earlier point in the culture. However, earlier time points were not measured, and therefore it is not possible to determine whether the analyses were performed on the optimal day.

In analysing the data, it was determined that gating on the $EPCR^+$ population was too subjective, as the investigator decided where to set the gates. A Fluorescence Minus One (FMO) control for EPCR should have been included to accurately determine the $EPCR^+$ gate. This would have made the analysis of the gated population more reliable.

In each well, 250 cells were transduced with the CRISPR gRNA expression vectors to analyse the different perturbations using flow cytometry and CFU assays. Doing bulk studies with HSCs makes analysis of changes in surface marker protein expression and colony outputs difficult, as it is unclear which HSC gives rise to which cells, and only records the effect of candidate gene

perturbations on the population average. Redesigning the study to instead culture single-cells and optimise flow cytometry protocols for low cell number analysis may reveal trends in the data that were indiscernible at bulk level. Smaller bulk cultures of 10-50 cells may have also helped to make out differences in the perturbations. However, a problem with both of these approaches is that the perturbation efficiencies would have been greatly reduced. An alternative approach could be to re-sort the cells into individual wells after the infection, making it possible to efficiently perturb the cells and study the effect on HSC function in single cells or small bulk populations.

The study focused on the loss of EPCR expression, a protein that is already rarely expressed. Investigating a negative outcome is always challenging, whereas looking for retention of EPCR expression or the gain of a different HSPC marker may have been easier to score. A previous study showed that $EPCR^+$ cells lose their long-term repopulation potential and have significantly reduced *Procr* expression after two days in culture (Hiroko Iwasaki et al. 2010). The cells used in this study were cultured in different medium conditions, but still call into question whether the study needs to be redesigned to interrogate genes important in HSC biology differently.

Possible suggestions for redesigning the study will be further explored in the thesis discussion (Chapter 7).

4.10.2. Summary

This chapter aimed to study genes implicated in HSC biology by perturbing them in primary E-SLAM cells using CRISPR/Cas9 technology. Flow cytometry was used to analyse changes in EPCR expression and apoptosis, and CFU-assays interrogated changes in differentiation output. Genotyping analysis confirmed that the CRISPR gRNAs were successfully targeting the candidate genes and causing indels and frameshift mutations; however, EPCR expression and lineage output were not significantly affected by the perturbations. These results suggest that a single gene perturbation may not influence a cell's phenotype due to compensatory mechanisms, or that the study design was flawed and requires improvement.

Chapter 5: Resolving heterogeneity in HSPC populations

Parts of this chapter have been modified from Hamey *et al.* (2017). Isolation of HSPCs was performed by Sonia Shaw, David Kent and Nicola Wilson. Sonia Shaw and Nicola Wilson processed the single cells using the Fluidigm BioMark™ platform. Sonia Shaw performed the quality control and normalisation of single-cell data, as well as all analyses described in this chapter unless otherwise stated. Fiona Hamey carried out the pseudotime ordering analysis.

5.1. Background

In the adult mammalian blood system, haematopoietic stem cells (HSCs) differentiate into all mature blood cell types and self-renew to maintain the HSC pool. Individual cells make fate choices, but the overall balance of cell types is regulated at the population level. An imbalance in the regulation processes can cause biased production of cell types and result in severe blood disorders. It is therefore important to understand cell fate decision making and its regulation during normal blood cell development.

Transcriptional regulation is a key process in cell fate decision making, in which the primary players are transcription factors that function in complex networks of interactions to regulate gene expression (Gottgens 2015; Peter and Davidson 2015). Many existing studies use bulk expression data to study transcriptional regulation within the haematopoietic system; however, it is a highly heterogeneous system and the intricacies of cell fate regulation may be missed at the population level. Recent developments in high-throughput single-cell technologies make it possible to investigate how heterogeneity in haematopoietic stem and progenitor cell (HSPC) populations is related to fate choices (Paul et al. 2015; N. K. Wilson et al. 2015). Gene expression profiles can be obtained at the single-cell level using methods such as scRNA-seq and quantitative real-time PCR (qRT-PCR) (Hamey et al. 2016).

This chapter investigates how HSPC fate decisions are controlled by exploring HSPC heterogeneity and using pseudotime ordering to order cells through differentiation (Bendall et al. 2014; Trapnell et al. 2014). To provide a sufficient number of cells for this investigation and to ensure that the haematopoietic hierarchy is represented comprehensively, a previously published dataset was

extended. Wilson *et al*. used the Fluidigm BioMark™ platform to profile single HSPCs using qRT-PCR (N. K. Wilson et al. 2015). They aimed to design an unbiased sorting strategy that enriched for functional HSCs at a higher purity, which at the time hovered around 50% for existing strategies (Beerman et al. 2010; Morita, Ema, and Nakauchi 2010; Dykstra et al. 2007; Goodell et al. 1996). They isolated long-term HSCs using four common sorting strategies, as well as a finite self-renewal HSC (FSR-HSC) fraction, lymphoid-primed multipotent progenitors (LMPP), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), and granulocyte-monocyte progenitors (GMP) (Adolfsson et al. 2005, 2001; D. G. Kent et al. 2009; Kiel, Radice, and Morrison 2007; Weksberg et al. 2008; Akashi et al. 2000). The expression of 48 genes involved in HSPC biology was quantified using single-cell qRT-PCR. The analysis was extended in this investigation to include three intermediate progenitor populations to obtain a comprehensive coverage of the haematopoietic hierarchy. FSR-HSCs were isolated from murine bone marrow using an alternative sorting strategy by fluorescence activated cell sorting (FACS), in addition to multipotent progenitors (MPP) and pre-megakaryocyte-erythroid progenitors (PreMegE) (Pronk et al. 2007). These populations were profiled using the same single-cell qRT-PCR assays and combined with the earlier profiles to provide extensive coverage of murine HSC populations.

5.1.1. Aims

The aims of this chapter were to:

- Isolate intermediate progenitor populations at the single-cell level to compliment the qRT-PCR data generated by Wilson *et al.* (2015) and extend coverage of the haematopoietic hierarchy
- Interrogate the qRT-PCR data to better understand heterogeneity and regulatory relationships in HSPC populations

These aims were addressed by profiling FSR-HSCs, PreMegEs, and MPPs from mouse bone marrow using qRT-PCR on the Fluidigm BioMark™ platform. The gene expression profiles generated were analysed together with qRT-PCR HSPC data previously generated and published by the lab. Using various dimensionality reduction methods and computational analyses, regulatory relationships and heterogeneity were explored in the HSPC dataset.

5.2. Isolation of haematopoietic stem and progenitor cell populations

To study the transcriptional control of HSPC differentiation, a dataset previously published by the Göttgens lab was used as a starting point for this investigation. Wilson *et al*. used single-cell FACS

paired with index sorting to resolve heterogeneity within four LT-HSC populations (N. K. Wilson et al. 2015). They isolated 1,975 cells from HSC, FSR-HSC, LMPP, CMP, MEP and GMP populations and profiled their gene expression using the Fluidigm BioMark^{TM} platform (Fig. 5.1A). The focus of this study was to identify an enriched HSC population by pairing single-cell functional assays with gene expression analysis. As such, most cells collected were part of different HSC isolation strategies and the dataset did not provide comprehensive coverage of intermediate progenitor populations.

The results of Chapter 3 demonstrated the value of including intermediate progenitor populations to study differentiation trajectories. Additional populations were isolated in this chapter to supplement the Wilson *et al*. dataset and to be able to explore regulatory networks within the haematopoietic hierarchy, specifically focused on known regulators of the stem cell compartment.

205 cells from each of the following populations were isolated using single-cell FACS: FSR-HSC2 (Lin^- $Sca1^+$ $c\text{-}Kit^+$ $Il7Ra^-$ $CD34^+$ $Flk2^-$), MPP (Lin^- $Sca1^+$ $c\text{-}Kit^+$ $Il7Ra^-$ $CD34^+$ $Flk2^+$) (Fig. 5.1B), PreMegE (Lin^- $Sca1^-$ $c\text{-}Kit^+$ $CD41^-$ $CD150^+$ $Fc\gamma RII/III^{low}$) (Fig. 5.1C). Each population was isolated during separate FACS experiments due to time restrictions and complexity of the antibody panels (Pronk et al. 2007). Single cells were sorted directly into a 96-well PCR plate containing 10 µl of reaction mixture (Section 2.5) and profiled by Fluidigm BioMark^{TM} analysis (Moignard et al. 2013).

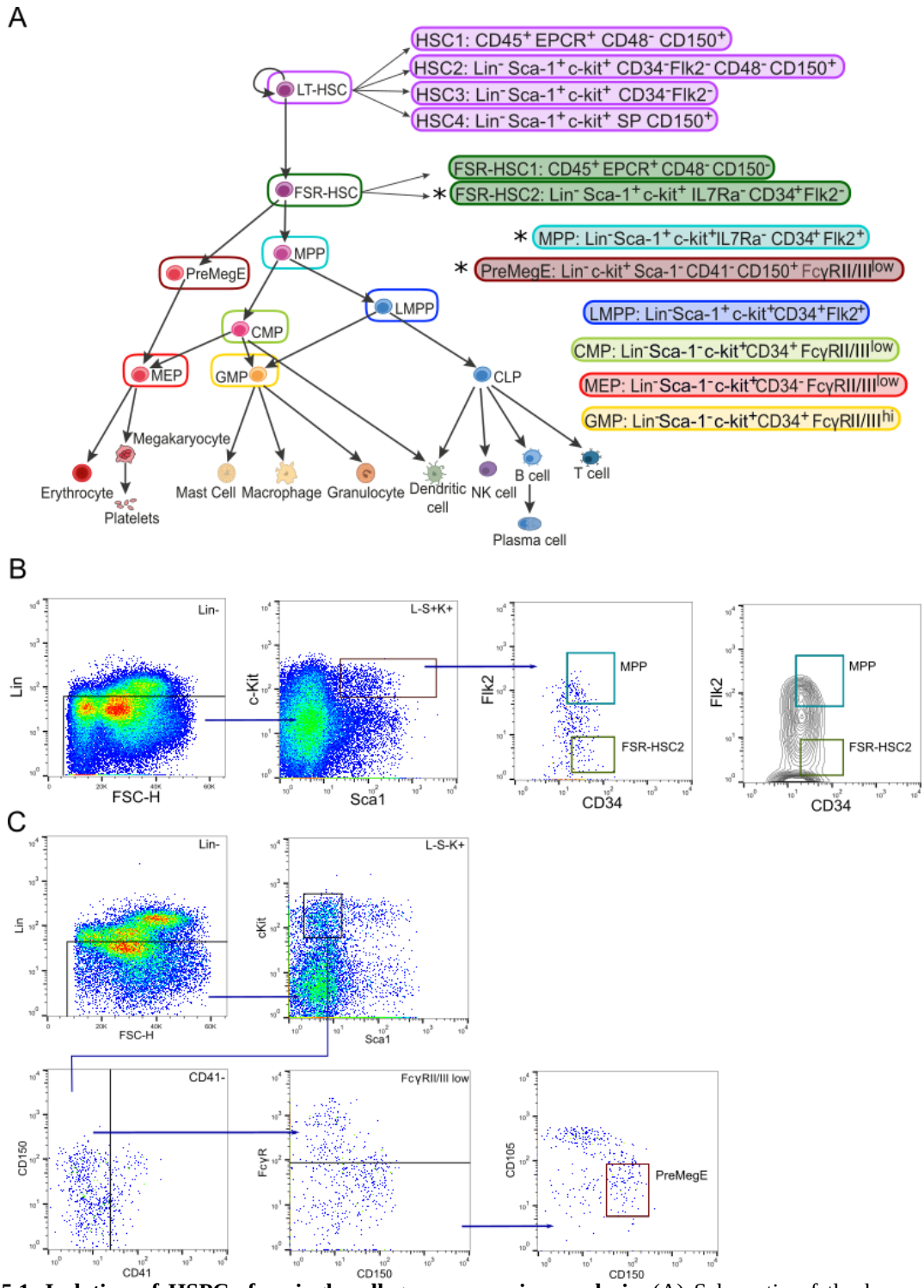


Figure 5.1. Isolation of HSPCs for single-cell gene expression analysis. (A) Schematic of the haematopoietic hierarchy and the different sorting strategies used to isolate HSPCs. Nine populations were previously isolated and published (N. K. Wilson et al. 2015). * indicates the populations sorted specifically for this investigation: FSR-HSC2, MPP, and PreMegE. (B) Flow cytometry plots of the sorting strategies used to isolate MPP (Lin^- $Sca1^+$ c-Kit^+ $Il7Ra^-$ $CD34^+$ $Flk2^+$) and FSR-HSC2 (Lin^- $Sca1^+$ c-Kit^+ $Il7Ra^-$ $CD34^+$ $Flk2^-$) populations. Both pseudocolour and contour plots are shown for the final isolated populations; the gates are coloured according to the colour scheme in Fig. 5.1A. Green – FSR-HSC2; turquoise – MPP. (C) Flow cytometry plots of the sorting strategy used to isolate PreMegEs (Lin^- $Sca1^-$ c-Kit^+ $CD41^-$ $CD150^+$ $Fc\gamma RII/III^{low}$). The gate is coloured according to the colour scheme in Fig. 5.1A (brown).

5.3. Selection of a gene set for HSPC analysis

As Fluidigm BioMark™ is a qRT-PCR approach for single-cell gene expression analysis, it was necessary to select genes to investigate. The gene-set includes 48 genes and is identical to that used by Wilson *et al*. in their published analysis, and builds upon the gene-set used by Moignard *et al.*, which included 18 transcription factors (N. K. Wilson et al. 2015; Moignard et al. 2013). The Wilson *et al.* gene set was designed to investigate HSC heterogeneity and transcriptional networks. It included 33 transcription factors important for HSCs and haematopoiesis, 12 additional genes that are involved in HSC biology, and three housekeeping genes (Table 5.1). The housekeeping genes (*Eif2b1, Polr2a* and *Ubc*) were included for normalisation purposes and are used as a reference to control for differences in cell size and RNA quantities (G. Guo et al. 2010; Pina et al. 2012). The cell surface marker c-Kit, which was used to isolate all the populations, was also included.

Table 5.1. Properties of genes selected for single-cell gene expression analysis. Genes are ordered alphabetically.

Gene	Summary
Bptf	• Bromodomain PHD finger transcription factor • Regulates thymocyte maturation (Landry et al. 2011) • Involved in the maintenance and function of regulatory T-cells (B. Wu et al. 2016)
Cbfa2t3h	• Encodes the repressor ETO2 • Functions in protein complexes to recruit histone deacetylases and repress gene expression • Regulates cell proliferation and differentiation to determine terminal erythroid maturation (Goardon et al. 2006) • Involved in repressing megakaryocyte differentiation (Hamlett et al. 2008)
Cdkn2a	• Tumour suppressor gene encoding for $p16^{INK4a}$ • Cell cycle gene • Inactivation of $p16^{INK4a}$ has been observed in various cancers (R. Zhao et al. 2016), including B-cell lymphomas (Jardin et al. 2010) and childhood acute lymphoblastic leukaemia (ALL) (Sulong et al. 2009)
Csf1r	• Receptor for colony-stimulating factor 1 • Regulates the development of macrophages • Expressed at low levels in HSCs (Sarrazin et al. 2009; Mossadegh-Keller et al. 2013) • Expressed at higher levels in monocytes and macrophages (Guilbert and Stanley 1980; Byrne, Guilbert, and Stanley 1981) • High levels of CSF1R are associated with poor survival in acute myeloid leukaemia (AML) (Rashid et al. 2016)
Dnmt3a	• Encodes DNA methyltransferase 3 alpha • Important for HSC differentiation (Challen et al. 2011)

Gene	Summary
	• Loss of murine *Dnmt3a* is associated with haematological malignancies, including myelodysplastic syndromes (Walter et al. 2011) and AML (Ley et al. 2010)
Egfl7	• Epidermal growth factor-domain gene • Expressed in endothelial cells (Fitch et al. 2004) • Regulates the migration of endothelial cells and is involved in vasculogenesis (Schmidt et al. 2007) • EGFL7 protein and mRNA expression is increased in AML and is associated with worse outcome (Rashid et al. 2016)
Eif2b1	• Encodes a subunit of the translation initiation factor eIF2B • Housekeeping gene (Pfister, Tatabiga, and Roser 2011)
Erg	• Transcription factor belonging to the ETS family • Potent oncogene through chromosomal translocation (Salek-Ardakani et al. 2009) • Required for HSC function and maintenance • Involved in late-stage megakaryocyte maturation (Loughran et al. 2008) • Causally related to Down syndrome-associated megakaryocytic leukaemia (Rainis et al. 2005) • Involved in oncogenesis in AML and Ewing sarcoma (H. Ichikawa et al. 1994; Sorensen et al. 1994)
Ets1 and *Ets2*	• Transcription factors belonging to the ETS family • Reciprocal activity during T-cell activation (Bhat et al. 1990) • ETS1: maintains T-cell quiescence • ETS2: T-cell activation and proliferation • ETS1 is upregulated in chronic myeloid leukaemia (CML) (Rashid et al. 2016) • High ETS2 expression is associated with a worse outcome in AML (L. Fu et al. 2017)
Etv6	• Transcription factor belonging to the ETS family (Tel) • Transcriptional repressor • Controls the survival of HSCs and is required for late-stage megakaryopoiesis (Hock, Meade, et al. 2004) • *ETV6-RUNX1* fusion is common in childhood ALL and enhances the self-renewal of B-lineage progenitor cells (Rashid et al. 2016)
Fli1	• Transcription factor in the ETS family • Highly expressed in HSCs and endothelial cells (Klemszs et al. 1993; Melet et al. 1996; Ben-David et al. 1991) • Abnormal expression of FLI1 is associated with poor prognosis in AML (Kornblau et al. 2011)
Gata1	• Transcription factor in the GATA family • Essential for normal erythropoiesis • Expressed in erythroid cells, megakaryocytes, eosinophils, and mast cells (Leonard et al. 1993; Martin et al. 1990; Zon et al. 1993) • *GATA1* mutations are seen in Down's syndrome-associated AML and transient abnormal myelopoiesis (Hitzler et al. 2003)
Gata2	• Transcription factor in the GATA family • Required for the expansion of multipotent haematopoietic progenitors

Gene	Summary
	• Required for the generation of mast cells (Tsai and Orkin 1997) • Mutated in paediatric AML and a subset of adult CML (S.-J. Zhang et al. 2008; Luesink et al. 2012)
Gata3	• Transcription factor in the GATA family • Contributes to the development of T-lymphocytes (Pandolfi et al. 1995; Vicente et al. 2011) • GATA3 polymorphisms are significantly associated with susceptibility to B-lineage ALL (Q. Hou et al. 2017)
Gfi1	• Transcriptional repressor • Involved in the development of HSCs, B- and T-cells, dendritic cells, granulocytes, and macrophages (Karsunky et al. 2002; Zeng et al. 2004; Rathinam et al. 2005) • Preserves HSC quiescence (Zeng et al. 2004; Hock, Hamblen, et al. 2004) • *Gfi1* is frequently mutated in T-cell lymphomas and T-cell ALL (Q. Hou et al. 2017)
Gfi1b	• Transcriptional repressor • Essential in megakaryocytic and erythroid development (Saleque, Cameron, and Orkin 2002; Osawa et al. 2002; Vassen, Okayama, and Möröy 2007; Vassen et al. 2014)
Hhex	• Haematopoietically expressed homeobox • Expressed in B-lymphocyte and myeloid lineages and MPPs (Bedford et al. 1993) • Overexpressed in AML cells and essential for their propagation (Shields et al. 2016)
Hoxa5	• Transcription factor in the HOX family • Preferentially expressed in expanding HSCs; key regulator of HSC cell cycle (D. Yang et al. 2015) • Constitutive expression inhibits erythropoiesis and causes a shift towards myeloid differentiation (Crooks et al. 1999; Fuller et al. 1999) • Hypermethylated in AML; associated with progression to blast crisis in CML (Kim et al. 2010)
Hoxa9	• Transcription factor in the HOX family • Preferentially expressed in early HSCs (Sauvageau et al. 1994; Pineault et al. 2002) • Important in differentiation of myeloid, lymphoid, and erythroid lineages (Lawrence et al. 1997) • Loss of function perturbs early T-cell development (Izon et al. 1998) • Overexpressed in AML; associated with poor prognosis (Collins and Hess 2016)
Hoxb4	• Transcription factor in the HOX family • Predominantly expressed in HSCs • Essential to HSC expansion (Sauvageau et al. 1995) • Significantly overexpressed in bone marrow of de novo AML patients and inversely correlated with the expression of the multidrug-resistance gene *ABCB1* (Umeda et al. 2012)
Ikzf1	• Transcription factor in the IKAROS family (Ikaros) • Regulator of lymphoid differentiation (Georgopoulos et al. 1994) • Frequently deleted or mutated in B-cell precursor ALL (Mullighan et al. 2009)
Itga2b	• Encodes CD41, part of the CD41/CD61 integrin complex • Upregulated during megakaryopoiesis

Gene	Summary
	• Required for platelet aggregation and clotting (Ginsberg et al. 1995; Emambokus and Frampton 2003) • Its expression marks myeloid-biased HSCs (Gekas and Graf 2013)
Kit	• Encodes a receptor tyrosine kinase (Yarden et al. 1987) • Key in regulating HSC function • Critical for proliferation, survival and differentiation (Edling and Hallberg 2007) • Aberrant activity is associated with AML pathogenesis (Malaise, Steinbach, and Corbacioglu 2009)
Ldb1	• Encodes the LIM-domain binding protein LDB1; closely related to SCL (Visvader et al. 1997) • Expressed in HSCs, with lower levels in myeloid and megakaryocyte progenitor cells • Expressed at all stages of erythroid development (L. Li et al. 2010) • Interacts with LMO1 and LMO2 in human T-cell ALL (Valge-Archer, Forster, and Rabbitts 1998; Layer et al. 2016)
Lmo2	• Encodes the LIM-domain protein LMO2 • Discovered by its association with chromosomal translocations in T-cell leukaemia (Boehm et al. 1991). • Activates an erythroid-specific gene expression program • *Lmo2* downregulation is essential for terminal erythroid differentiation (Kornblau et al. 2011)
Lyl1	• Basic helix-loop-helix transcription factor • A *Lyl1*-deficient mouse is viable but has impaired LT-HSC reconstitution and a reduced number of B-cells (Capron et al. 2006) • Required for HSC function and maintenance of early T-cell lineage progenitors (Zohren et al. 2012; Souroullas et al. 2009) • Involved in adult angiogenesis (Pirot et al. 2010) • Overexpressed in AML (Meng et al. 2005)
Mecom	• Zinc finger transcription factor (Evi1) • Associated with myeloid leukaemia (Russell et al. 1993; Morishita et al. 1992) • Predominantly expressed in HSCs and regulates HSC proliferation (Goyama et al. 2008)
Meis1	• Encodes a homeobox protein • Loss of *Meis1* is embryonic lethal (Azcoitia et al. 2005) • Required for HSC maintenance and self-renewal (Ariki et al. 2014; Kocabas et al. 2012; Unnisa et al. 2012; M. E. Miller et al. 2016) • Co-expressed with HOXA9 in human myeloid leukaemia (Lawrence et al. 1999)
Mitf	• Basic helix-loop-helix transcription factor • Expressed in mast cells • Mutations in *Mitf* result in phenotypic abnormalities and a decrease in mast cells (Kitamura et al. 2002)
Mpl	• Encodes the receptor for thrombopoietin (Kaushansky 1995) • Involved in regulation of HSC production and function (Kimura et al. 1998; Solar et al. 1998) • Involved in maintenance of HSC quiescence (Yoshihara et al. 2007)

Gene	Summary
	• Important in differentiation towards megakaryocytes and platelets (Plo et al. 2017) • Mutations occur in myeloproliferative neoplasms (MPN) (Plo et al. 2017)
Myb	• Transcription factor in the MYB family • Required for B-cell and T-cell development (Thomas et al. 2005; Lieu et al. 2004) • Regulator of HSC quiescence and self-renewal (Cooke, Sutton, and Parker 2010; Lieu and Reddy 2009) • Driver of leukemogenesis in birds and mice; expressed at high levels in AML and ALL (Pattabiraman and Gonda 2013)
Nfe2	• Basic-leucine zipper transcription factor • Expressed in HSCs and erythroid and megakaryocytic lineages (Andrews et al. 1993) • Important for megakaryocyte maturation and platelet production (Shivdasani et al. 1995)
Nkx2-3	• Transcription factor in the NKX family • Important in lymphoid tissue development (Pabst, Zweigerdt, and Arnold 1999; Pabst et al. 2000)
Notch1	• Encodes a transmembrane receptor • Important in cell fate decision making in haematopoiesis (Radtke et al. 2004) • Influences B versus T-cell differentiation (Pui et al. 1999) • Mutations causing continuous activation of Notch signalling occur frequently in T-cell ALL (Liu, Zhang, and Ji 2013)
Pbx1	• Transcription factor in the PBX family • Important in B-cell lineage commitment (Sanyal et al. 2007) • Discovered in pre-B cell ALL in a reciprocal translocation between chromosomes 1 (*Pbx1)* and 19 (*E2A*); causes AML in mice (A. J. Carroll et al. 1984; Kamps and Baltimore 1993)
Polr2a	• Encodes a subunit of RNA polymerase II • Housekeeping gene (Radonić et al. 2004)
Prdm16	• Zinc finger transcription factor • Involved in HSC renewal, quiescence, apoptosis, and differentiation (Aguilo et al. 2011) • Frequently involved in translocations in AML (Corrigan et al. 2018)
Procr	• Encodes endothelial protein C receptor (EPCR) • Marker of haematopoietic stem cells (Balazs et al. 2006)
Runx1	• Transcription factor in the RUNX family • Expressed in HSCs, myeloid, and lymphoid cells • Downregulated during erythroid differentiation (North et al. 2004) • Involved in translocations in AML and paediatric ALL (Sood, Kamikubo, and Liu 2017)
Spi1	• Encodes the transcription factor PU.1 (ETS family transcription factor) • Expressed in HSCs, CLPs and CMPs (Nutt et al. 2005) • Required for the generation of both myeloid and lymphoid lineages (E. W. Scott et al. 1994; McKercher et al. 1996) • Decreased PU.1 expression is linked to AML in mice and humans (Verbiest et al. 2015)
Sh2b3	• Encodes lymphocyte adaptor protein (LNK) • Negative regulator of normal haematopoiesis

Gene	Summary
	• Expressed in HSCs, myeloid and lymphoid progenitors (Takaki et al. 2000; Velazquez et al. 2002) • Controls HSC quiescence and self-renewal through *Mpl* (Bersenev et al. 2008) • Recently identified as a possible genetic predisposition gene to B-precursor ALL • Genetic variations identified in MPN and lymphoid leukaemia (Perez-Garcia et al. 2013; Maslah et al. 2017)
Smarcc1	• Encodes a subunit of the SWI/SNF remodelling complex • Involved in HSC activity through chromatic modification (Deneault et al. 2009) • Important in murine embryonic stem cell differentiation (Schaniel et al. 2009)
Tal1	• Encodes the transcription factor stem cell leukaemia (SCL) • Important in HSC specification (Porcher et al. 1996; Robb et al. 1996) • Expressed in normal HSCs and along erythroid, mast cell and megakaryocytic lineages (Gottgens et al. 1997) • Gain of function mutation in T-cell ALL (O'Neil et al. 2004)
Tcf7	• Essential for normal T-cell development (Weber et al. 2011) • Expressed in naïve T-cells and memory T-cells, but downregulated in effector T-cells (D. M. Zhao et al. 2010) • Key regulator of the switch between self-renewal and differentiation in HSCs (Choi et al. 2017)
Tet2	• Enzyme in the TET family • Involved in DNA methylation: converts 5-methylcytosine to 5-hydroxymethylcytosine (Pastor, Aravind, and Rao 2013) • Regulator of HSC homeostasis (Nakajima and Kunimoto 2014) • Frequently mutated in haematopoietic malignancies; loss of *TET2* leads to induction of leukemogenesis by DNA hypermethylation of active enhancers (Rasmussen et al. 2015)
Ubc	• Encodes polyubiquitin-C • Housekeeping gene (Silver et al. 2008; Chua et al. 2011)
Vwf	• Encodes Von Willebrand factor (vWF) • Mediates platelet adhesion and platelet aggregation (Peyvandi, Garagiola, and Baronciani 2011) • vWF expression marks HSCs with durable self-renewal capacity (D. G. Kent et al. 2009) • vWF expression selects for platelet-biased HSCs (Sanjuan-Pla et al. 2013)

5.4. Processing single cells using the Fluidigm BioMark™ platform

The Fluidigm BioMark™ platform was used to investigate gene expression in HSPC populations. Fluidigm BioMark™ is a qRT-PCR approach that is sensitive at the single-cell level. In addition to single cells, 10-cell controls and empty well controls were collected for each experiment. Cells were processed as previously described and the single-cell expression data was collected using the Fluidigm Data Collection software. Cells were normalised against the housekeeper genes *Ubc* and

Pol2ra. The expression of all three housekeeper genes was analysed and showed that most cells expressed *Ubc* and *Polr2a*, whereas *Eif2b1* expression was more variable and therefore excluded from normalisation. After filtering and normalising the data, only retaining cells that expressed at least 23 genes as well as both housekeeper genes, a total of 2,167 cells remained that were representative of the 8 cell types sorted (Table 5.2). All housekeepers (*Ubc, Polr2a* and *Eif2b1*) were removed from the dataset for downstream analysis. Wilson *et al.* also removed *Cdkn2a* and *Egfl7* from their downstream analysis due to a lack of expression and technical issues, respectively (N. K. Wilson et al. 2015). This was carried forward in this investigation for consistency. In addition, *Gfi1* and *Spi1* were expressed in the empty well controls for the PreMegE, FSR-HSC2 and MPP populations. As this expression was visible on all integrated fluidics circuits (IFCs) run through the Fluidigm BioMark^{TM} for this experiment, it indicated a technical issue and these genes were also removed from downstream analysis. After quality control, the expression of 41 genes, including 31 transcription factor genes, was retained.

Table 5.2. Table indicating number of cells included in this study.

Cell Type	Cell Count		Cells Lost	% Cells Used
	Before filtering	After filtering		
HSC1	210	198	12	94
HSC2	210	166	44	79
HSC3	210	197	13	94
HSC4	210	198	12	94
FSR-HSC1	294	233	61	79
FSR-HSC2*	205	199	6	97
MPP*	205	188	17	92
PreMegE*	205	154	51	75
LMPP	210	178	32	85
CMP	210	147	63	70
MEP	211	124	87	59
GMP	210	185	25	88
Total	2590	2167	423	84

* indicates the populations sorted specifically for this investigation: FSR-HSC2, MPP, PreMegE.

5.5. Resolving populations using multidimensionality analysis

Dimensionality reduction was required to investigate how the cell populations related to each other based on their gene expression. Dimensionality reduction methods are useful for visualising large datasets in a lower dimensionality space. In this investigation, they were used to evaluate the

heterogeneity and structure of the haematopoietic bone marrow compartment in an unsupervised fashion.

Principal component analysis (PCA) was used to visualise relationships between cell populations (Fig. 5.2). PCA is a linear dimensionality method in which principal component (PC) 1 has the largest variance, followed by PC2. Therefore, the data was plotted in the first two components to demonstrate the variance between the cell populations. The new data collected for this investigation was integrated into the Wilson *et al.* dataset and analysed together (Fig. 5.2A). PreMegEs, MPPs and FSR-HSC2 cells are intermediate populations in the haematopoietic hierarchy, which is recapitulated by their location on the PCA plot.

The four HSC populations were clustered together at the top of the graph (Fig. 5.2B). HSC1 showed the most dispersed expression; however, the four strategies enriched for cells with an overall similar expression profile. The FSR-HSC populations were located between the HSCs, MPPs and LMPPs, consistent with the classical view of the haematopoietic hierarchy. For greater visual clarity, the four HSC populations and the two FSR-HSC populations were coloured together (Fig. 5.2C). Although there were no clear projections in the PCA visualisation, the HSCs, MEPs and LMPPs were found at distinct edges of the structure, indicating these populations were the most different from one another. PreMegEs clustered closely to the MEPs whereas the GMPs were in between MEP and LMPP populations, albeit more concentrated near the LMPPs. The CMPs were disperse among the progenitor populations, consistent with previous observations about their heterogeneity (Paul et al. 2015).

The PCA loadings show which genes contributed to the separation of the data (Fig. 5.2D). At the top of the PCA plot, *Mpl*, *Mecom* and *Procr* contributed to the variance that separated HSCs from the other populations, consistent with these genes being important to HSC characteristics (Table 5.1). *Gata1* and *Gfi1b* contributed to the separation on the left side of the PCA plot, and *Notch* and *Csf1r* contributed to the variance on the right, consistent with these regions of the PCA plot being made up of MEPs, LMPPs and GMPs, respectively.

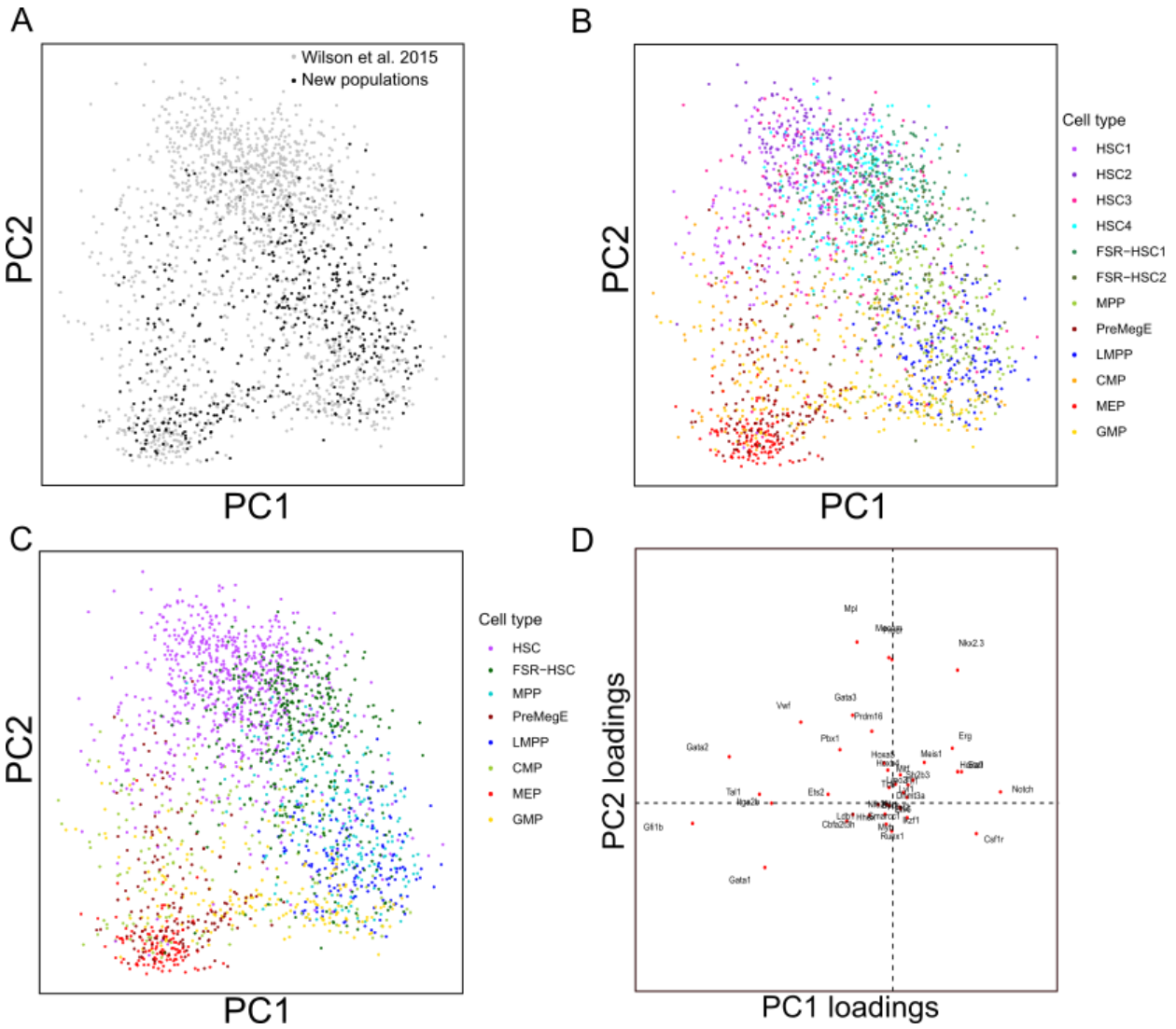


Figure 5.2. Visualisation of single-cell qRT-PCR data using principal component analysis. (A) PCA plot showing the integration of the new data with the data from Wilson *et al.* (2015). New data – black; Wilson *et al.* data – grey. (B) PCA plot of all populations, calculated on the expression of 41 genes measured by qRT-PCR. The plot is coloured by sorting gate. HSC1 – purple, HSC2 – dark purple, HSC3 – pink, HSC4 – cyan, FSR-HSC1 – forest green, FSR-HSC2 – olive green; MPP – yellow green; PreMegE – dark brown; LMPP – blue; CMP – orange; MEP – red; GMP – yellow. (C) PCA plot of all populations, coloured by cell type. The four HSC populations are grouped together (purple) and the two FSR-HSC populations are grouped together (olive green). MPP – light blue, PreMegE – dark brown, LMPP – blue, CMP – yellow green; MEP – red; GMP – orange. (D) PCA loading plots, showing genes that contribute to the variance in PC1 and PC2. PC: Principal Component.

Although PCA is an informative dimensionality-reduction method, it can only capture linear structures in the data. More recently, non-linear dimensionality reduction methods such as t-distributed stochastic neighbour embedding (t-SNE) and diffusion maps have been applied to single-cell data (Maaten and Hinton 2008; Haghverdi, Buettner, and Theis 2015). These methods are able to capture more complex structures in the data. t-SNE aims to conserve the local distances

of the high-dimensionality data in a low-dimensionality structure, so that cells with similar gene expression are nearby on the plot.

The qRT-PCR data was visualised using t-SNE (Fig. 5.3), which recapitulated the structure seen using PCA. The HSCs were located at the top of the landscape. The HSC1 cells showed the most heterogeneity and the HSC4 population appeared more molecularly different from the other HSC sorting strategies than seen in the PCA plot (Fig. 5.3A). As the sorting strategies were different, it was assumed that the functional HSCs would be similar, and each strategy would differ in the phenotype of contaminating cells that it captured. The t-SNE separated the data into two distinct branches, separating MEPs and LMPPs, which is clearly shown when the four HSC populations and two FSR-HSC populations are coloured together (Fig. 5.3B). The CMPs and GMPs were both dispersed among the progenitor cells; on the LMPP branch, GMPs were at the tip of the branch, but in between MEPs and HSCs on the MEP branch.

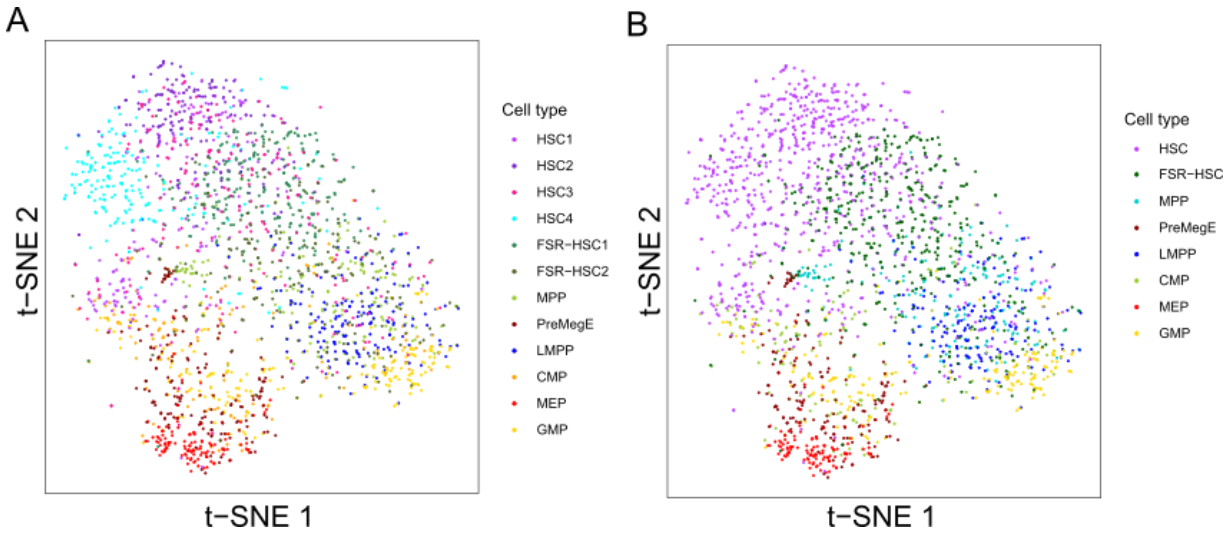


Figure 5.3. Visualisation of single-cell qRT-PCR data using t-distributed stochastic neighbour embedding. (A) t-SNE plot of all populations, calculated on the expression of 41 genes measured by qRT-PCR. The plot is coloured by sorting gate. HSC1 – purple, HSC2 – dark purple, HSC3 – pink, HSC4 – cyan, FSR-HSC1 – forest green, FSR-HSC2 – olive green; MPP – yellow green; PreMegE – dark brown; LMPP – blue; CMP – orange; MEP – red; GMP – yellow. (B) t-SNE plot of all populations, coloured by cell type. The four HSC populations are grouped together (purple) and the two FSR-HSC populations are grouped together (olive green). MPP – light blue, PreMegE – dark brown, LMPP – blue, CMP – yellow green; MEP – red; GMP – orange.

A disadvantage of t-SNE analysis is that it is a stochastic model, which means that while the overall conclusions from the analysis do not change, the t-SNE visualisation will be altered every time it is generated. It is therefore necessary to generate t-SNE plots multiple times to confirm that structure of the dataset is reproducible, and then set the seed parameter to be able to reproduce the same figure every time. Furthermore, both PCA and t-SNE dimensionality reduction methods are

designed to detect differences in the data rather than continuous relationships (Haghverdi, Buettner, and Theis 2015). As haematopoiesis involves the differentiation of an HSC towards a mature cell fate while passing through intermediate progenitor phenotypes, it would be beneficial to visualise the data using a dimensionality-reduction method that is better able to determine more complex structures in the data. Diffusion maps use the length of diffusion-like random walks through the data in high-dimensional space to determine a projection of the cells, and have been adapted to successfully display single-cell data (Coifman et al. 2005).

The qRT-PCR data was visualised on a diffusion map (Fig. 5.4). As in the PCA and t-SNE plots, the HSCs sat at the top of the structure and HSC1 showed the most disperse expression pattern of the four HSC sorting strategies. Furthermore, the diffusion map recapitulated the pattern seen in the t-SNE plot, in which HSC4 was most distinct from the four HSC sorting strategies (Fig. 5.4A). When the HSC and FSR-HSC populations are coloured together, it is easier to visualise that the structure roughly segregated the cells into two projections, separating MEPs and LMPPs (Fig. 5.4B). The LMPPs were located closer to the FSR-HSCs than the MEPs, suggesting their gene expression was closer to that of the early progenitors. The distinct gene expression of PreMegEs and MEPs was more clearly visualised in the diffusion map than using the other methods.

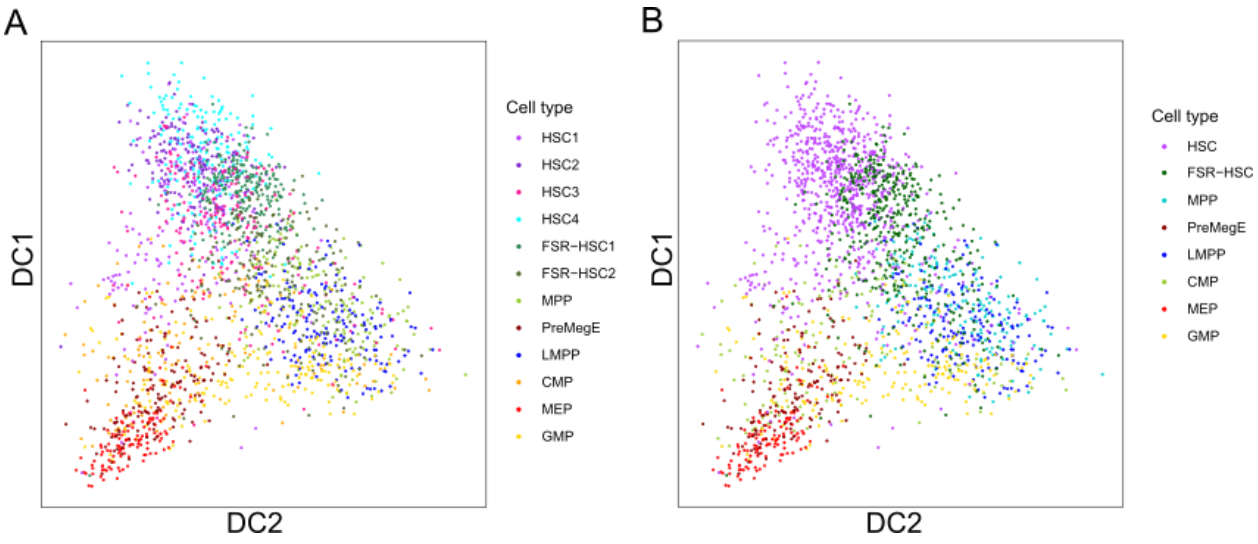


Figure 5.4. Visualisation of single-cell qRT-PCR data using diffusion maps. (A) Diffusion map of all populations calculated on the expression of 41 genes measured by qRT-PCR. The plot is coloured by sorting gate. HSC1 – purple, HSC2 – dark purple, HSC3 – pink, HSC4 – cyan, FSR-HSC1 – forest green, FSR-HSC2 – olive green; MPP – yellow green; PreMegE – dark brown; LMPP – blue; CMP – orange; MEP – red; GMP – yellow. (B) Diffusion map of all populations coloured by cell type. The four HSC populations are grouped together (purple) and the two FSR-HSC populations are grouped together (olive green). MPP – light blue, PreMegE – dark brown, LMPP – blue, CMP – yellow green; MEP – red; GMP – orange. DC: Diffusion Component.

5.6. Single-cell gene expression analysis reveals cell population clusters

Unsupervised hierarchical clustering was performed to investigate heterogeneity between cell populations using the expression of the 41 genes that passed quality control. Clustering analysis can be used to gain insights about groups of cells, which are clustered based on their gene expression patterns without the input of any classifications by the investigator.

The clustering partitioned the cells into two broad clusters, both of which had multiple sub-clusters (Fig. 5.5). Roughly, Cluster 1 included HSCs and the earliest progenitor populations, whereas Cluster 2 was made up of the later progenitor populations. These clusters were differentiated by the expression of HSC-specific genes such as *Mpl, Mecom,* and *Procr* (Yoshihara et al. 2007; Goyama et al. 2008; Balazs et al. 2006). Further sub-clustering Cluster 1 divided it based on the four sorting strategies used to isolate HSCs. HSC4 made up the majority of Cluster 1F and was characterised by the expression of *Gata3*, indicating these cells may contribute to a lymphoid lineage (Pandolfi et al. 1995; Vicente et al. 2011). Cluster 1E was a mixture of cell types, but predominantly contained HSC2 and HSC3 cells, similar to Cluster 1C. However, Cluster 1C also included CMP and GMP populations, whereas Cluster 1E contained FSR-HSCs, MPPs, and PreMegEs. *Procr* was expressed in more cells of Cluster 1E, whereas Cluster 1C contained cells expressing *Itga2b* and *Gata1*, possibly indicating a myeloid-erythroid bias (Gekas and Graf 2013; Leonard et al. 1993). Cluster 1D was made up of mostly LMPPs, as well as a mixture of FSR-HSCs, GMPs, and MPPs, whereas Clusters 1A and 1B were the most heterogeneous and included all cell types in varying proportions. The separations in the sub-clusters of Cluster 2 were clearer. Cluster 2A was made up of predominantly PreMegEs and MEPs and was characterised by the expression of *Gata1, Gata2, Tal1* and *Gf1b*, which are all genes involved in myeloid and erythroid differentiation (Leonard et al. 1993; Tsai and Orkin 1997; Osawa et al. 2002; Gottgens et al. 1997). Cluster 2B predominantly contained GMPs, but also included LMPPs and FSR-HSCs. *Csf1r* was most highly expressed in this cluster, consistent with it having a major GMP component (Guilbert and Stanley 1980). As expected, *Kit* was expressed in all cells as it was used to sort all populations. The clustering analysis shows that the 2,167 cells can be subdivided into distinct clusters that correspond with their sorted identities, based solely on their gene expression. Furthermore, the four HSC populations generally overlap with each other, although individual patterns can be observed, which may indicate lineage biases and/or the presence of "contaminating" cells in the sorting strategy.

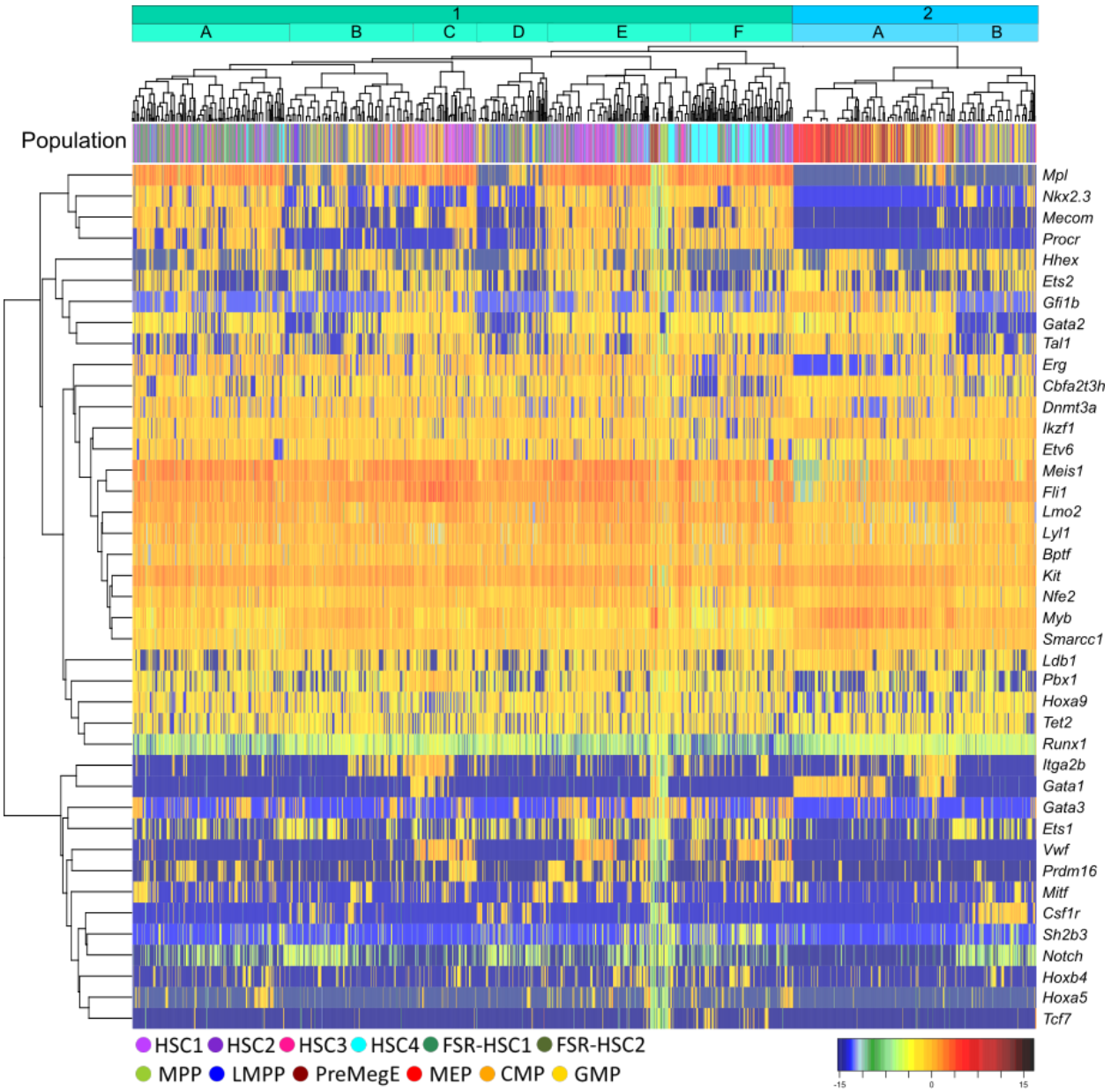


Figure 5.5. Unsupervised clustering of HSPC populations reveals clusters in single-cell gene expression data. Heatmap showing unsupervised clustering of 2,167 cells profiled by qRT-PCR according to the expression of 41 genes. The distances of the population dendrogram are not proportional to the dissimilarity. The bar above the dendrogram indicates the clusters to which the cells belong. Cluster 1 – green; Cluster 2 – blue. The colour bar indicates the population/sorting gates of origin. Sorting gate colours: HSC1 – purple; HSC2 – dark purple; HSC3 – pink; HSC4 – cyan; FSR-HSC1 – forest green; FSR-HSC2 – olive green; MPP – yellow green; LMPP – blue; PreMegE – dark brown; MEP – red; CMP – orange; GMP – yellow.

5.7. Pairwise correlation analysis reveals putative relationships between genes

The dimensionality reduction and clustering analyses indicate that the HSPC populations are characterised by different gene expression patterns, with some overlap between populations. Genes that share regulatory mechanisms may have similar expression patterns, whereas genes that have

very different expression profiles may have unrelated mechanisms governing their expression (Ståhlberg and Bengtsson 2010). Correlation analysis can be used to gain insight into these mechanisms, where a positive correlation suggests a factor may activate another, and negative correlation suggests antagonism.

Pairwise correlation analysis was performed across all 2,167 HSPCs by hierarchical clustering of Spearman Rank correlations between pairs of transcription factors. The analysis revealed both positive and negative correlations between pairs of transcription factors (Fig. 5.6). A positive correlation was observed between *Gata2, Gf1b, Tal1,* and *Gata1*. This is consistent with literature as positive interactions have previously been described between *Scl* and *Gata2*, as well as between *Gata2* and *Gf1b* (Moignard et al. 2013; J. E. Pimanda et al. 2007). A negative correlation was observed between *Gata1*, a key regulator of erythropoiesis, and *Nkx2.3* and *Notch*, both of which are important in lymphoid tissue development (Pabst et al. 2000; Pabst, Zweigerdt, and Arnold 1999; Martin et al. 1990). A strong positive correlation was observed between *Myb* and *Runx1*. A recent study observed that RUNX1 regulates *Myb* in mouse T-cell ALL, in which a RUNX1 deficiency reduced transcription factor binding at the *Myb* +15kb enhancer (Choi et al. 2017). In normal haematopoiesis, *Myb* and *Runx1* cooperate to induce expression of lineage-specific genes in HSPCs; however, a direct regulatory relationship between these two genes has not been reported, suggesting their interaction may be mediated by coactivating proteins (Friedman 2002).

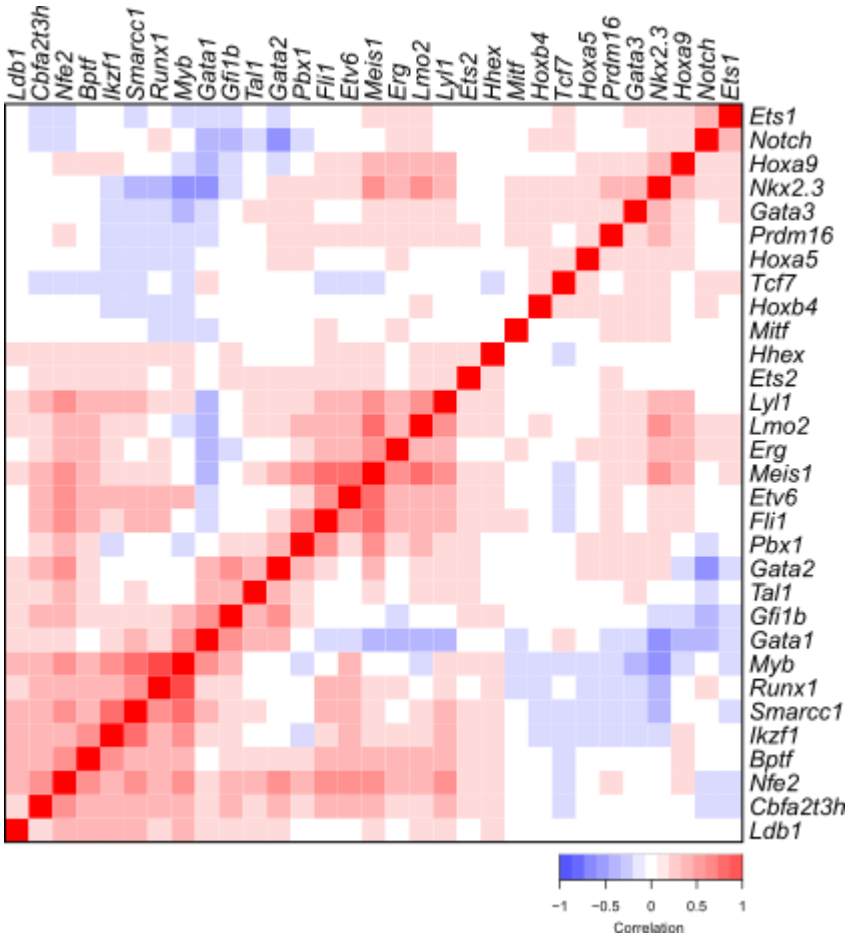


Figure 5.6. Pairwise correlations reveal regulatory relationships between transcription factors. Hierarchical clustering of Spearman Rank correlations between pairs of transcription factors for all 2,167 HSPCs. Positive correlations are shown in red and negative correlations are shown in blue.

To establish whether the identified regulatory relationships changed during differentiation, the correlation analysis was repeated for the eight cell types separately (Fig. 5.7). The HSC and FSR-HSC cells from different sorting strategies were analysed together to see which relationships were strongly shared between the strategies, revealing the correlations of true HSCs or FSR-HSCs rather than the contaminating cells. The correlations observed in HSC and FSR-HSC populations were most consistent to those seen when analysing all cells together (Fig. 5.6). This reflects the greater number of these cells included in the dataset, as well as the bias of the gene set towards genes involved in HSC biology. Many of the strong positive correlations identified in the whole dataset remained stable; in particular, a positive correlation between *Fli1*, *Etv6*, *Meis1*, *Erg*, and *Lmo2* was observed in all populations. More negative correlations were seen in progenitor populations compared to the whole dataset, suggesting that repressive relationships may be particularly important in more differentiated populations.

The HSC and FSR-HSC populations generally had similar pairwise correlation patterns. A key difference between the two cell types was that *Tcf7* was involved in many negative relationships in HSCs, but mainly had a positive role, if any, in FSR-HSCs. *Tcf7* has been implicated in HSC quiescence and self-renewal; therefore, it may attribute to the long-term self-renewal potential that is specific to the stem cell compartment (J. Q. Wu et al. 2012). Interestingly, *Gata2* and *Runx1* were positively correlated in HSCs, but negatively correlated in FSR-HSCs. MPPs had similar relationships to FSR-HSCs, but had a positive correlation cluster between *Hoxb4, Tcf7, Hoxa5, Prdm16,* and *Gata3.* This cluster also included *Nkx2*.3, *Hoxa9, Notch,* and *Ets1* in the PreMegE population only. Compared to the other populations, MEPs, GMPs, and CMPs displayed many negative correlations. *Hoxa5* was generally involved in negative relationships in MEPs, and positive in CMPs and GMPs. Conversely, *Notch* was negatively correlated with erythroid genes such as *Gata1* and *Gf1b* in CMPs and GMPs; *Ets1* was also negatively correlated with these erythroid genes in CMPs, but not in GMPs. Finally, LMPPs displayed the same relationships that were shared among all populations, but lacked correlations with most genes, reflecting the myeloid-erythroid bias of the gene set. While *Ets1* and *Notch,* genes important in lymphoid lineage, were positively correlated with each other, they were negatively correlated with megakaryopoiesis genes *Etv6* and *Nfe2* (Andrews et al. 1993; Hock, Meade, et al. 2004). *Ets1* and *Ets2* were negatively correlated, as previously described (Bhat et al. 1990).

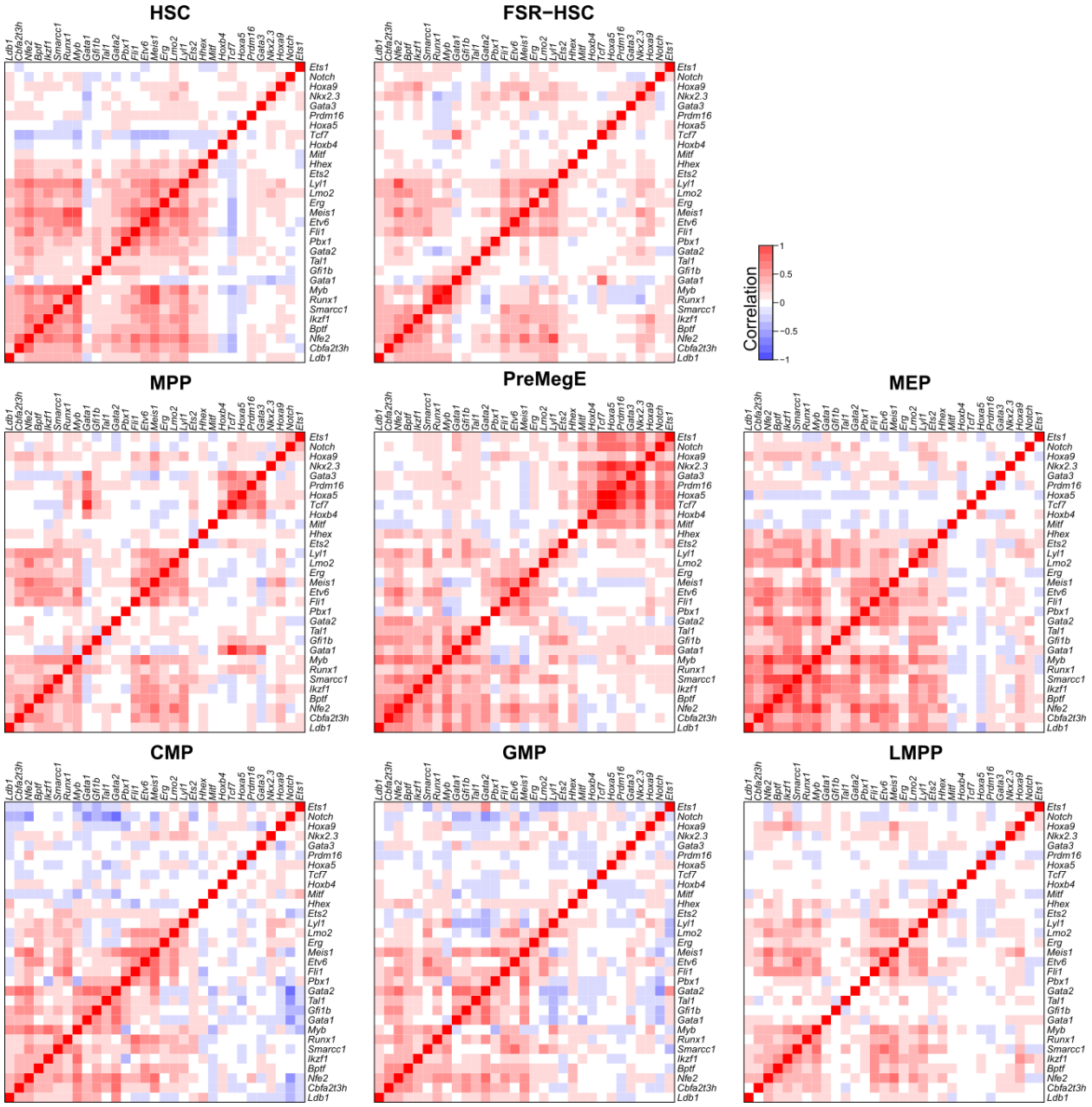


Figure 5.7. Pairwise correlations for each HSPC population. Hierarchical clustering of Spearman Rank correlations between pairs of transcription factors for individual HSPC populations. Cells are grouped based on cell type. Positive correlations are shown in red and negative correlations are shown in blue. Genes are ordered according to the clustering in Fig 5.6.

5.8. Reconstructing differentiation trajectories from single-cell gene expression profiles

The diffusion map representation described in Section 5.6. demonstrates that the HSPC dataset recapitulates the structure of the haematopoietic hierarchy. To confirm the apex of the hierarchy, molecular overlap cells (MolO cells) were projected onto the atlas (Fig. 5.8A). Wilson *et al.*

identified the MolO cells during their investigation for a sorting strategy that enriches for functional LT-HSCs with high purity (N. K. Wilson et al. 2015). The MolO cells are HSCs that share a transcriptional profile and have increased probability of long-term multilineage repopulation potential upon single-cell transplantation. When projected onto the landscape, MolO cells sit at the top of the structure with the most primitive cells, as expected. Cells belonging to MEP and LMPP populations are at the end of the landscape, and intermediate populations such as MPPs and PreMegEs were present between the highlighted cell types.

In Chapter 3, the diffusion map was used to capture cells on differentiation trajectories towards mature cell types. Motivated by the structure of the hierarchy described by the qRT-PCR data, pseudotime analysis was performed to better understand the transcriptional changes occurring throughout differentiation. Pseudotime orders cells based on their gene expression profiles to infer their position in differentiation, which can be used to construct differentiation trajectories through single-cell expression data (Ocone et al. 2015). Coordinates on the diffusion map were used to identify cells on trajectories from HSCs to MEPs and LMPPs (Fig. 5.8B). Cells were assigned to two broad branches and were ordered in pseudotime using the Wanderlust algorithm (Bendall et al. 2014).

The expression of transcription factors was visualised through the pseudotime progression for both MEP and LMPP trajectories (Fig. 5.8C). Distinct expression patterns were noted between the two trajectories. On the MEP trajectory, *Nkx2.3, Meis1* and *Pbx1* expression decreased as *Gata1, Gfi1b* and *Ikzf1* expression increased, consistent with the negative correlation seen between these genes in Fig. 5.6. Along the LMPP trajectory, genes important in lymphoid development, such as *Notch* and *Ets1*, increased, whereas genes important in HSC characteristics, such as *Prdm16* and *Hoxb4*, decreased early in the trajectory and the key erythroid gene *Gata1* was not expressed at all. Overall, fewer genes showed increased expression compared to the MEP trajectory, once again reflecting the myeloid-erythroid bias of the gene set. Pseudotime ordering demonstrates the gene expression dynamics occurring in haematopoiesis and suggests that the data could be further used to investigate regulatory networks along the trajectories.

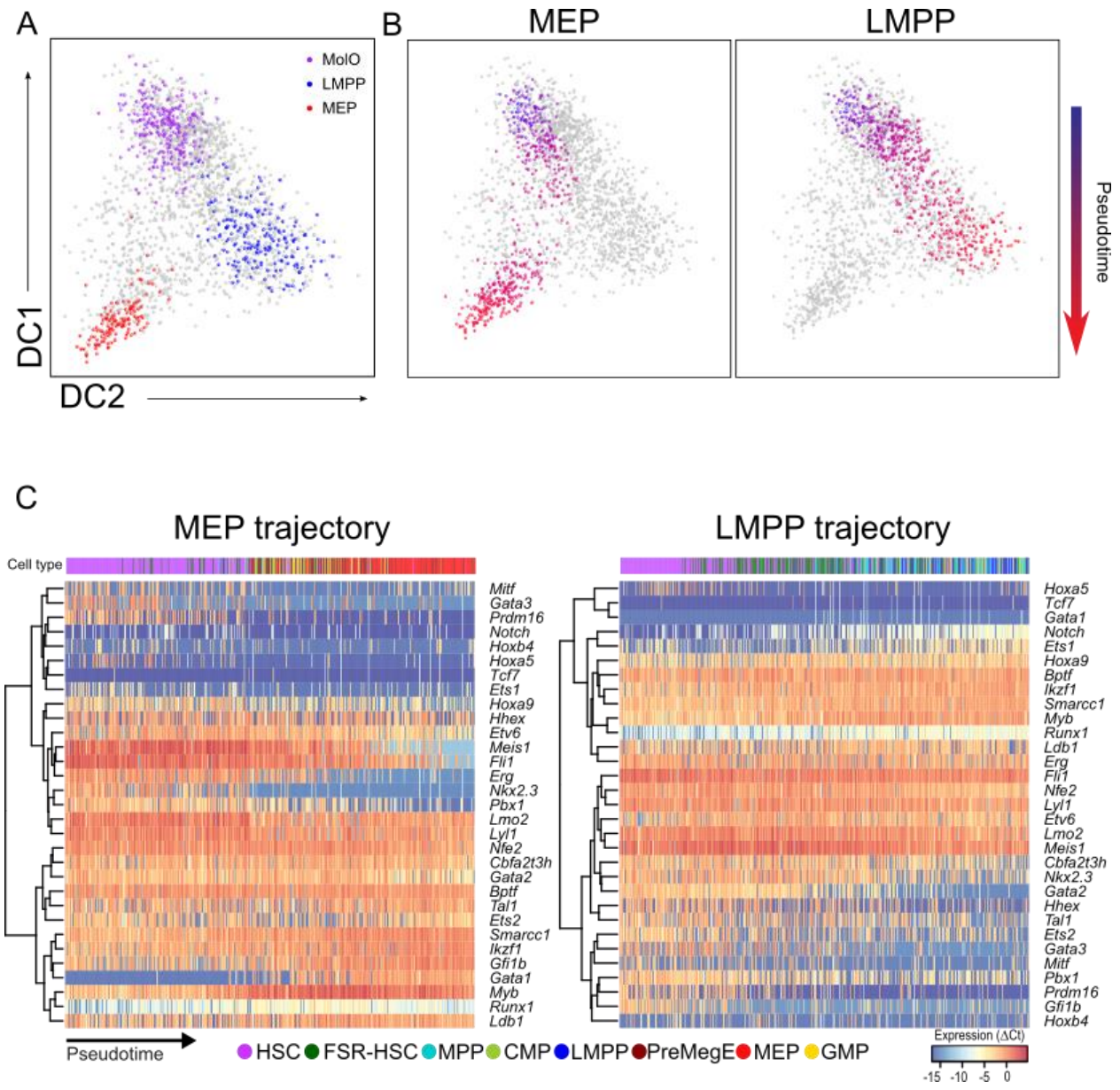


Figure 5.8. Pseudotime ordering reveals two differentiation trajectories in the single-cell HSPC data. (A) Projection of MolO cells onto the qRT-PCR dataset using a diffusion plot visualisation. MEPs and LMPPs are highlighted. MolO cells – purple; MEP – red; LMPP – blue. (B) Differentiation trajectories from stem cells to MEPs or LMPPs. Cells are coloured by their pseudotime value, moving from blue (early in pseudotime) to red (late in pseudotime). Cells not in the trajectory are grey. (C) Heatmaps showing the expression of transcription factor encoding genes. Cells are ordered along the pseudotime trajectories towards MEP or LMPP fates. The colour bar at the top of each heatmap indicated the cell types along each trajectory. HSC – purple; FSR-HSC – forest green; MPP – light blue; CMP – yellow green; LMPP – blue; PreMegE – brown, MEP – red, GMP – yellow. Figure was generated by Fiona Hamey and modified by Sonia Shaw.

5.9. Conclusions

In this chapter, single cell gene expression profiles were generated using the Fluidigm BioMark™ platform to explore heterogeneity and regulatory relationships within the haematopoietic hierarchy. A qRT-PCR dataset of 2,167 cells was generated, spanning HSCs and early progenitors. The dataset was then explored using dimensionality-reduction methods, correlation analysis and pseudotime ordering.

This investigation built on a pre-existing dataset which included HSCs, FSR-HSCs, and four additional progenitor populations: CMPs, GMPs, MEPs and LMPPs (N. K. Wilson et al. 2015). The gene set was handpicked to include 33 transcription factor encoding genes important in HSPC biology, as well as 12 other genes implicated in HSC function. While this was already a large dataset with good coverage of early haematopoiesis, it did not include intermediate progenitor populations that occur during the differentiation process. Including these intermediate populations gave a more complete picture of early haematopoietic differentiation, which proved to be useful for inferring differentiation trajectories and regulatory networks. The three additional populations isolated were FSR-HSCs, MPPs, and PreMegEs. These populations were chosen because FSR-HSCs and MPPs should have multi-lineage potential without the capability of reconstituting a mouse long-term, and PreMegEs are an early precursor of megakaryocytic, erythroid, or mixed colonies (Pronk et al. 2007; Cabezas-Wallscheid et al. 2014).

The HSPCs were visualised using three dimensionality-reduction methods: PCA, t-SNE, and diffusion maps. All three methods recapitulated the haematopoietic hierarchy but gave varying levels of resolution in terms of heterogeneity and gene expression relationships. The PCA plots showed little separation of the four HSC isolation strategies, whereas HSC4 was more separated from the other strategies in both the t-SNE and diffusion map plots. As PCA only recognises linear relationships in the data, it will miss any non-linear relationships and therefore may not provide the most suitable visualisation for more complex structures, such as single-cell data. t-SNE is useful for visualising highly heterogeneous data and positions cells with similar expression profiles close together. In this HSPC dataset, the t-SNE plot positioned the MEPs and LMPPs further away from each other. However, while t-SNEs are often used to represent heterogeneous datasets, they are stochastic and may struggle to display continuous processes such as differentiation. Diffusion maps have been adapted to specifically deal with single-cell expression data (Coifman et al. 2005; Haghverdi, Buettner, and Theis 2015). When this dataset was visualised using diffusion maps, the

LMPPs were clearly separated from PreMegEs and MEPs. Based on the structure of the data, the diffusion map was the best method for recapitulating the structure of the haematopoietic hierarchy.

Unsupervised hierarchical clustering was used to group cells based on their gene expression and clearly separated the more mature progenitors and HSC populations into two distinct clusters. However, within the HSC cluster, there was a great amount of heterogeneity and overlap with FSR-HSC, LMPP, CMP and GMP populations. The clustering showed that the four isolation strategies used for HSCs do overlap but vary in their functional purity. The four strategies capture different "contaminating" cells, i.e. non-HSC cells, and the frequency and nature of these contaminating cells depends on the sorting strategy used. Correlation analysis was performed on all 2,167 cells together as well as the individual populations to examine regulatory relationships between the different populations. Previously published positive interactions were observed between *Scl* and *Gata2* and between *Gata2* and *Gfi1b*, corroborating the accuracy of this dataset (Moignard et al. 2013; J. E. Pimanda et al. 2007). Furthermore, negative correlations were observed between genes involved in opposing branches of haematopoietic differentiation, such as *Notch* and *Gata1*, which are genes involved in the lymphoid and erythroid lineages, respectively (Pui et al. 1999; Hamlett et al. 2008).

The correlation analysis for individual populations showed that many positive correlations were stable among the HSPC populations, but key differences were observed in their negative correlations. MEPs, GMPs, and CMPs in particular had the most negative correlations between genes. A multipotent phenotype may therefore be more associated with positive relationships, while repression, or lack of it, becomes more important in increasingly differentiated populations. CMPs and GMPs had similar regulatory relationships, in which *Notch* was negatively correlated with many erythroid genes. Current research suggests that CMPs may actually be a heterogeneous population primed towards erythroid and myeloid fates (Perié et al. 2015; Jaitin et al. 2014). Indeed, this investigation suggests that CMPs and GMPs are very similar, based on the visualisation of their transcriptional structures, clustering of the cells, and correlation analyses.

The cells were ordered along differentiation using pseudotime ordering, which identified two trajectories in the data from HSCs towards MEPs and LMPPs. Visualising the expression of transcription factor encoding genes along the trajectories showed that the genes had both static and dynamic expression patterns. *Gata1* was differentially expressed along the two trajectories and was associated with the MEP trajectory, while *Notch* and *Ets1* increased along the LMPP trajectory but

were only expressed in the HSC component of the MEP trajectory. Conversely, *Bptf, Smarcc1* and *Myb,* which were positively correlated in the analysis of all 2,167 cells, were constitutively expressed in both trajectories. The complexity of correlations observed, as well as the gene expression changes occurring along pseudotime, imply that it may be a useful dataset for inferring regulatory network models along the two trajectories, explored in Chapter 6.

5.9.1. Limitations

Although pairwise correlations were identified for all populations, fewer relationships could be identified for LMPPs as the gene set was biased towards myeloid-erythroid genes and focused on HSCs rather than the progenitor populations. A limitation of qRT-PCR is that the number of genes profiled is limited and chosen by the investigator, which may hinder discovery of novel regulators, and, in the context of this work, may miss key regulators of haematopoietic differentiation. Furthermore, the limited gene set fails to capture the full transcriptional heterogeneity of the different cell types. Visualisations of this dataset have suggested a significant overlap between GMP and LMPP populations; however, the work in Chapter 3 shows these populations can be separated based on the full transcriptome, where LMPPs and GMPs occupy separate territories on the transcriptional landscape. These populations could potentially be separated better if specific lymphoid genes were included in the gene set, such as *Dntt, Il7r,* or *Cd19*.

Another limitation of this work arose due to technical issues in the processing of samples, which resulted in key regulators such as *Gfi1* and *Spi1* being excluded from the analysis. *Spi1* would have been a valuable addition to the analysis due to its proposed antagonistic relationship with *Gata1* in megakaryocytic-erythroid versus granulocytic-monocytic lineage decision making (Burda, Laslo, and Stopka 2010). Recent research from continuous live cell imaging and reporter mouse lines suggest that these two genes do not initiate the megakaryocytic-erythroid versus granulocytic-monocytic lineage switch, but rather reinforce the lineage choices once made (Hoppe et al. 2016). It would have been interesting to see if these recent findings could be seen in this single cell qRT-PCR dataset. *Gfi1* was previously identified to be part of a regulatory triad with *Gata2* and *Gfi1b,* and would have been a useful addition to compare the work to previous literature (Moignard et al. 2013). Missing these key regulators renders the findings from hierarchical clustering and pairwise correlation analysis incomplete, although the dataset does accomplish its original goal of distinguishing between HSC sorting strategies.

5.9.2. Further work

In this chapter, a qRT-PCR dataset of HSPC expression profiles was generated and interrogated for heterogeneity and regulatory relationships. The results indicate that there are complex relationships occurring between genes during differentiation, and that HSC regulators are not only involved in HSC maintenance but play a role in differentiation decisions as well. Genes that have similar expression profiles may also share regulatory mechanisms, whereas genes that don't have similar expression profiles are likely to have unrelated regulatory mechanisms (Ståhlberg and Bengtsson 2010). It would be interesting to try to infer regulatory networks of transcription factors along the MEP and LMPP trajectories to further explore their unique regulatory mechanisms as well as those shared between them. Furthermore, these networks can be validated using functional assays. The regulatory networks and their validation will be explored in Chapter 6.

5.9.3. Summary

Single-cell expression profiling of HSPC populations using qRT-PCR demonstrated the heterogeneity present within populations of the haematopoietic hierarchy. Pairwise correlations of the different haematopoietic lineages identified regulatory relationships in individual populations and across the HSPC compartment. Pseudotime analysis ordered the cells in two trajectories from HSC to MEP or LMPP fates and was used to compare the dynamics of transcription factor expression along these trajectories.

Chapter 6: Validating regulatory networks models

Parts of this chapter have been modified from Hamey *et al.* (2017). Fiona Hamey developed the network inference method and carried out the *in-silico* modelling of networks. Nicola Wilson profiled HoxB8-FL cells by qRT-PCR and generated the ChIP-seq data. Rebecca Hannah aligned the ChIP-seq data. Sonia Shaw analysed the ChIP-seq data and performed luciferase assays.

6.1. Background

The haematopoietic differentiation system consists of individual cells making cell fate decisions that influence the balance of mature cell type output at the population level; any dysregulation in fate choices can cause an imbalance that leads to serious blood disorders, such as leukaemia. Therefore, it is important to study the regulation of fate decisions during blood development.

Cell fate decision making is heavily influenced by transcription factors, which work together in a transcriptional regulatory network (Peter and Davidson 2015). It is challenging to identify functional relationships between genes as experimental validation does not readily scale to a system-wide approach; therefore, computational network inference methods are used to predict these functional relationships. Attempts at modelling transcriptional regulation in blood have included using literature-curated regulatory relationships to construct networks, which limit discovery as they rely on prior knowledge of the system (Krumsiek et al. 2011; Narula et al. 2010; Chickarmane, Enver, and Peterson 2009; Swiers, Patient, and Loose 2006). Most network construction methods have been restricted to bulk expression data. Using single-cell data is becoming an increasingly recognised method for uncovering regulatory relationships (Pina et al. 2015; Moignard et al. 2013).

In Chapter 5, a single-cell HSPC dataset was established by combining quantitative real-time PCR (qRT-PCR) data from Wilson *et al.* with three intermediate progenitor populations to extend the coverage of the murine bone marrow HSPC compartment (N. K. Wilson et al. 2015). The chapter showed that haematopoietic stem and progenitor cell (HSPC) populations are highly heterogeneous and gene expression is dynamic across differentiation trajectories towards megakaryocytic-erythroid progenitor (MEP) and lymphoid multipotent progenitor (LMPP) cell fates. Fiona Hamey

used this qRT-PCR dataset of 2,167 HSPCs to identify transcriptional regulatory networks as this dataset profiles a large number of cells at different stages across haematopoietic differentiation. A hybrid network inference method was developed based on information about continuous gene expression levels obtained through pseudotime ordering (Ocone et al. 2015). The hybrid method took advantage of Boolean abstraction, which has previously been used to model transcriptional regulatory networks in HSCs, embryonic stem cells, and embryonic blood development (Bonzanni et al. 2013; Dunn et al. 2014; Xu et al. 2014; Moignard et al. 2015). Boolean abstraction converts gene expression to binary functions, where gene expression is either "on" or "off". These functions then form part of a Boolean network to describe gene interactions (Fig. 6.1A). The pseudotime trajectories were used to identify the most suitable Boolean functions. Cells were ordered based on continuous expression data and converted to binary expression profiles. Pairs of cells were treated as input-output pairs for Boolean function to identify which functions best fitted the data (Fig. 6.1B) (Hamey et al. 2017).

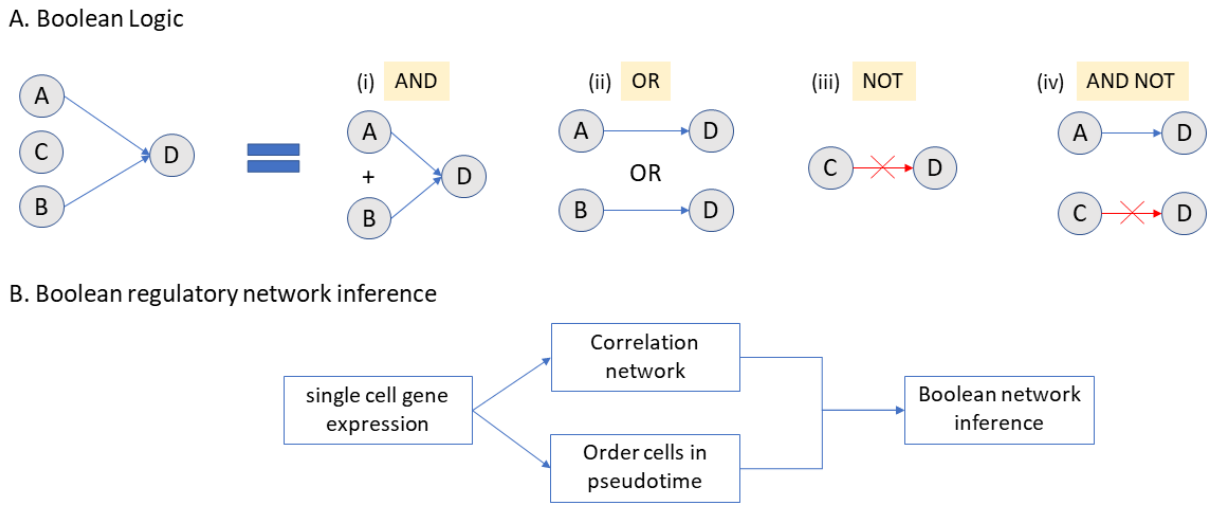


Figure 6.1. Single-cell molecular profiles allow inference of regulatory network models. (A) Boolean logic functions. Four different scenarios may explain a regulatory relationship where A and B activate D, but C does not. (i) A and B act together to activate D, described by the AND function; (ii) either A or B can activate D alone, described by the OR function; (iii) C does not activate D, described by the NOT function; (iv) A is required to activate D but C is not, but this must occur in combination. This is described by the AND NOT function (B) Schematic of the network inference method starting from gene expression profiling using single-cell qRT-PCR data.

When this method was applied to the single-cell qRT-PCR dataset, Boolean network models were reconstructed to describe the HSC to MEP trajectory and HSC to LMPP trajectory. These models were found to have complex structures, with each gene receiving inputs from multiple regulators and often as a composite of the Boolean functions described in Fig. 6.1A (e.g. (*Notch* AND *Tcf7*)

AND NOT *Etv6* activates *Ets1*). A simplified graphical representation of the networks illustrates the highly connected nature of the networks (Fig. 6.2).

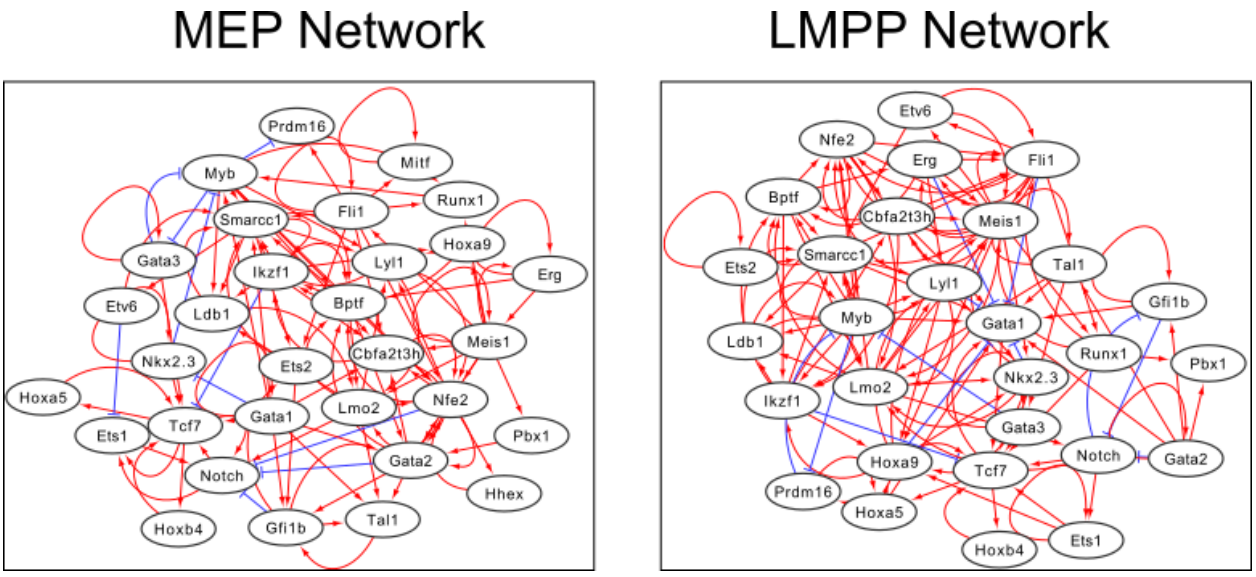


Figure 6.2. Transcriptional regulatory network models for differentiation from HSCs to MEPs or LMPPs. Activation is indicated by a red pointed arrow, and repression by a blue flat-headed arrow. Figure generated by Fiona Hamey for publication (Hamey et al. 2017).

Experimental validation was required to confirm the accuracy of the hybrid network inference method. In this chapter, the work done to validate the regulatory networks from HSCs to MEPs (MEP network) and HSCs to LMPPs (LMPP network) will be presented.

6.1.1. Aims

The aims of this chapter were to:

- Correlate the identified network rules to transcription factor binding patterns
- Experimentally validate regulatory relationships identified from the proposed network model

To address these aims, previously published Chromatin Immunoprecipitation Sequencing (ChIP-seq) data was interrogated in erythroid and myeloid-lymphoid model cell lines (Schütte et al. 2016; Hamey et al. 2017). Comparing the ChIP-seq data with the network rules identified a regulatory relationship which could be validated *in vitro* using luciferase assays.

6.2. Boolean network modelling reveals regulatory relationships within MEP and LMPP differentiation networks

Boolean abstraction was used to model transcriptional regulatory networks along the two differentiation trajectories identified in the qRT-PCR data described in Chapter 5. Modelling the transcriptional networks was performed to gain insight into the regulatory relationships governing the differentiation process. Boolean functions were abstracted for each gene based on pairwise correlations across the data (Hamey et al. 2017). These functions are simplified to activating or repressing relationships in Table 6.1 and Table 6.2, respectively. The genes had rules, or Boolean functions, that were specific to one or both networks, or shared between the two. A previously described and experimentally validated regulatory relationship was present in both regulatory networks, in which *Gata2* activated *Gfi1b* (Moignard et al. 2013). In the LMPP network, several genes inhibited the expression of *Gata1*, which is important in erythroid development (Leonard et al. 1993; Martin et al. 1990; Zon et al. 1993). Conversely, *Nkx2.3* and *Notch* were inhibited in the MEP network; both these genes are involved in the lymphoid lineage, influencing lymphoid tissue development and commitment towards lymphoid fates (Pabst, Zweigerdt, and Arnold 1999; Pabst et al. 2000; Pui et al. 1999).

Table 6.1. Simplified rules: activating relationships in MEP and LMPP networks. Activation between transcription factors in the MEP and LMPP network models, identified through Boolean abstraction. The activation of genes may also be shared between networks. The activating gene is the first column, and its targets are in the MEP network, LMPP network, or Shared Rules columns.

Gene	MEP network	LMPP network	Shared Rules
Bptf	*Gata2*	*Erg*	*Smarcc1* *Nfe2* *Lmo2* *Ikzf1*
Cbfa2t3h	*Gata2*	*Gata1* *Fli1* *Meis1*	*Nfe2* *Ikzf1*
Erg	*Erg*	*Lyl1* *Fli1*	*Meis1* *Bptf*
Ets1	*Notch*	*Hoxa9* *Ets1*	*Tcf7*
Ets2	*Gfi1b*	*Ets2*	*Smarcc1*
Etv6	---	*Fli1* *Meis1*	*Smarcc1*
Fli1	*Mitf* *Prdm16* *Ets2*	*Etv6* *Meis1* *Erg*	*Runx1* *Cbfa2t3h* *Smarcc1*

Gene	MEP network	LMPP network	Shared Rules
	Fli1		*Nfe2*
Gata1	*Myb* *Tal1* *Gfi1b* *Gata2* *Gata1* *Cbfa2t3h* *Smarcc1*	---	*Tcf7*
Gata2	*Cbfa2t3h* *Nfe2* *Bptf*	*Pbx1* *Gata1* *Gata2*	*Tal1* *Gfi1b*
Gata3			*Gata3* *Tcf7*
Gfi1b	*Tal1* *Gata2* *Ets2*	*Gata1* *Gfi1b*	---
Hhex	*Nfe2*	*Hhex*	---
Hoxa5	---	*Hoxa9* *Gata1* *Prdm16*	*Tcf7*
Hoxa9	---	*Meis1* *Nkx2.3*	*Lyl1* *Ikzf1*
Hoxba4	---	---	*Tcf7*
Ikzf1	---	---	*Ldb1* *Hoxa9* *Cbfa2t3h* *Smarcc1* *Bptf*
Ldb1	---	*Myb*	*Smarcc1* *Lmo2*
Lmo2	*Tal1*	*Nkx2.3* *Meis1*	*Ldb1* *Lyl1* *Notch* *Nfe2* *Bptf*
Lyl1	---	*Hoxa9* *Nkx2.3*	*Smarcc1* *Nfe2* *Lmo2*
Meis1	*Pbx1* *Gata2* *Meis1*	*Nkx2.3* *Etv6*	*Hoxa9* *Runx1* *Fli1* *Erg* *Cbfa2t3h* *Nfe2* *Lmo2*
Mitf	---	---	*Mitf*
Myb	*Runx1* *Cbfa2t3h*	*Myb* *Gata1* *Ikzf1*	*Ldb1* *Lyl1* *Smarcc1* *Nfe2* *Bptf*

Gene	MEP network	LMPP network	Shared Rules
Nfe2	*Hhex* *Gata2*	*Fli1* *Meis1*	*Lyl1* *Cbfa2t3h* *Lmo2* *Bptf*
Nkx2.3	*Nkx2.3*	*Hoxa9* *Meis1* *Lmo2*	*Tcf7*
Notch	---	*Lmo2*	*Tcf7* *Ets1*
Pbx1	*Gata2*	---	---
Prdm16	---	*Hoxa5*	---
Runx1	*Myb*	*Pbx1* *Fli1* *Tcf7* *Meis1*	---
Smarcc1	*Etv6* *Ets2*	*Gata1* *Fli1*	*Ldb1* *Lyl1* *Ikzf1* *Bptf*
Tal1	*Gfi1b*	*Gata3* *Tal1* *Gata1* *Lmo2*	---
Tcf7	*Notch* *Ets1*	*Gata1*	*Hoxb4* *Hoxa5*

Table 6.2. Simplified rules: repressive relationships in MEP and LMPP networks. Repression between transcription factors in the MEP and LMPP network models, identified through Boolean abstraction. The repression of genes may also be shared between networks. The repressing gene is the first column, and its targets are in the MEP network, LMPP network, or Shared Rules columns.

Gene	MEP only	LMPP only	Shared Rules
Erg	---	*Gata1*	---
Etv6	*Ets1*	---	---
Fli1	---	*Gata1*	---
Gata1	*Nkx2.3*	---	---
Gata2	---	---	*Notch*
Gata3	---	---	*Myb*
Gfi1b	---	---	*Notch*
Hoxa9	---	*Gata1*	---
Ikzf1	---	---	*Tcf7*
Lyl1	---	*Gata1*	---
Myb	*Gata3*	---	*Prdm16*
Nfe2	*Notch*	---	---
Nkx2.3	*Myb*	*Gata1*	---
Notch	---	*Gfi1b*	---
Prdm16	---	*Myb*	---

6.3. Linking ChIP-seq data and regulatory rules to identify relationships to validate *in vitro*

To understand which rules could be validated *in vitro*, ChIP-seq data for 416B and HoxB8-FL cell lines was analysed for binding patterns at genes of interest. ChIP-seq identifies genome-wide binding profiles for transcription factors and other proteins and can be visualised using the UCSC genome browser (W. J. Kent et al. 2002). The 416B cell line is a murine cell line with megakaryocytic potential, whereas the HoxB8-FL cell line was established to investigate myeloid and lymphoid cell differentiation (Dexter et al. 1979; Redecke et al. 2013). Validating regulatory relationships using primary cells would be challenging due to the very limited availability of material. Moreover, *in vitro* cell lines ensure a more homogenous cell population; therefore, these model cell lines were used instead. The 416B cell line is supposed to be transcriptionally similar to MEPs, while the HoxB8-FL cell line should resemble LMPPs and GMPs. Previously, single-cell qRT-PCR data was collected for 416B and HoxB8-FL cells using the same gene set; the data was projected onto the diffusion map described in Chapter 5 (Fig. 6.3A). This confirmed that 416B cells occupied a territory that forms part of the MEP trajectory, and the expression state of HoxB8-FL cells resembled that of primary bone marrow cells from the LMPP trajectory. Therefore, 416B could be used to represent the MEP trajectory, and that the LMPP trajectory could be represented by HoxB8-FL cells (Fig. 6.3B).

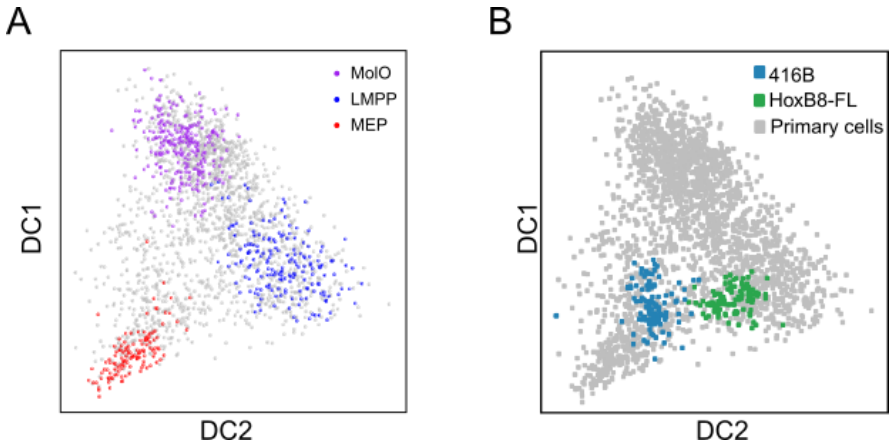


Figure 6.3. Transcriptional profiles of model haematopoietic cell lines occupy territories on MEP and LMPP trajectories on the HSPC qRT-PCR dataset. (A) Diffusion map of qRT-PCR data collected for 2,167 HSPC single cells. The MolO cells are highlighted to represent HSCs (purple); LMPPs (blue) and MEPs (red) are highlighted and represent the end points for the identified differentiation trajectories. (B) Diffusion map of qRT-PCR data collected for 2,167 HSPC single cells with projected cell line data. 416B cells – green; HoxB8-FL cells – blue. The primary HSPC cells are shown in grey.

In addition to qRT-PCR data, ChIP-seq data was previously collected for these cell lines (Schütte et al. 2016; Hamey et al. 2017). However, ChIP-seq data was only available for ERG, FLI1, GATA2, GFI1B, LMO2, LYL1, RUNX1, and SCL. The binding patterns for each of these proteins were analysed for all transcription factor-encoding genes included in the gene set (an example is shown in Fig. 6.4). The presence of peaks indicated that the transcription factor was bound within a specific genomic region. The transcription factor may be differentially bound between the two cell lines or show similar binding patterns in both. For example, *Bptf* is bound by LYL1 in HoxB8-FL cells, but not in 416B cells. *Bptf* is also bound by FLI1 in both cell lines but at distinct genomic regions and is not bound by GATA2 in either cell line. The binding profile may be influenced by the behaviour and expression of the transcription factor as well as the expression of the target gene in the cell line.

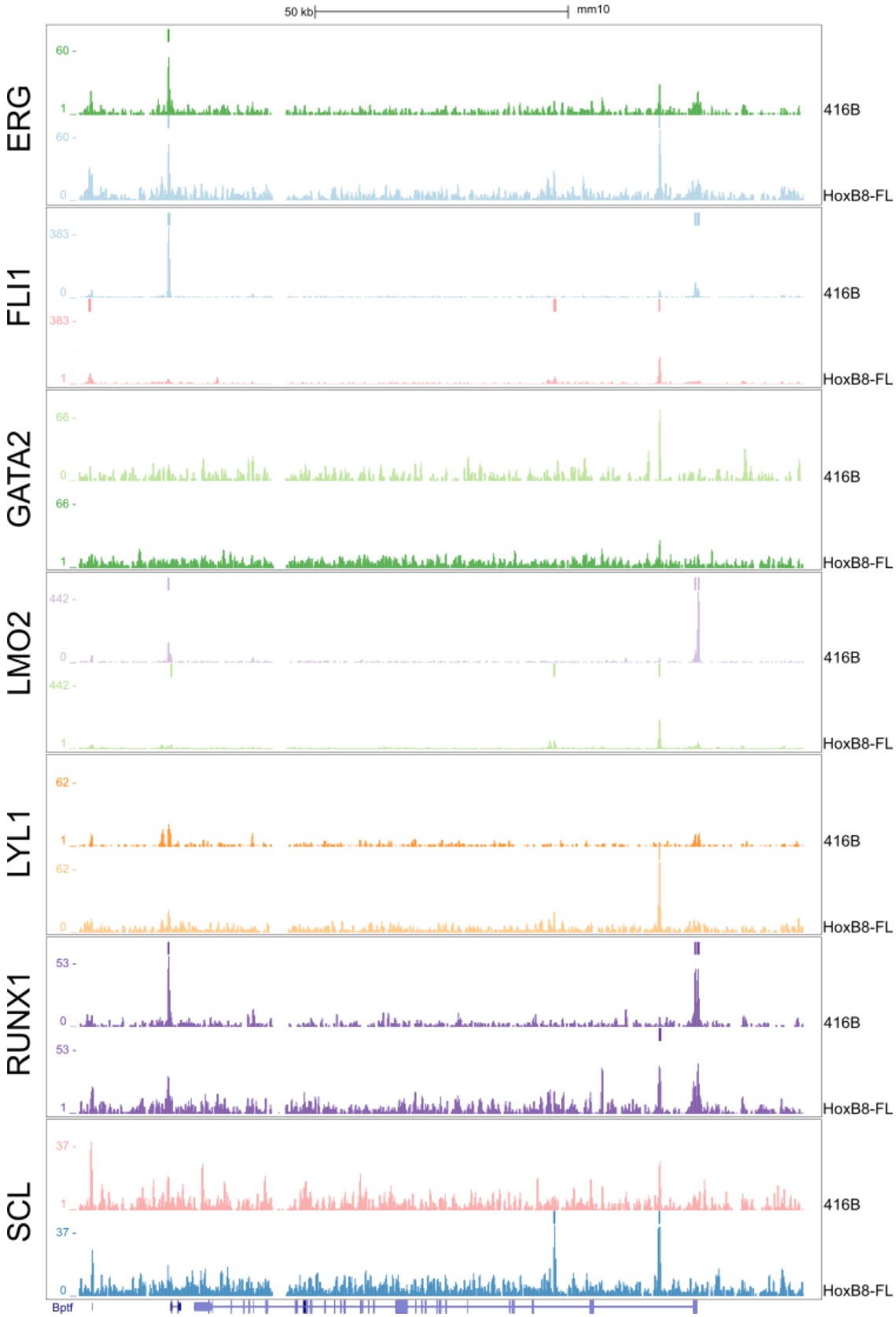


Figure 6.4. ChIP-seq data reveals transcription factor binding patterns in 416B and HoxB8-FL cell lines. ChIP-seq analysis of transcription factor binding in 416B and HoxB8-FL cell lines. *Bptf* was chosen as an example of how the raw data and binding peaks are visualised in the UCSC genome browser. Bars above each track indicate a binding event that was called as a peak.

The network rules for MEP and LMPP trajectories were integrated with the ChIP-seq data for 416B and HoxB8-FL cells to determine which regulatory network rules could be validated in the model cell lines (Fig. 6.5). Transcription factor binding to each gene was recorded in a binary matter (i.e. "yes" or "no") and did not take into account the expression level of the gene in the cell line, which would potentially impact on the binding intensity. Most genes, except for *Prdm16* and *Hoxa5*, were bound by most proteins in either 416B or HoxB8-FL cells lines, or both (Fig.6.5A). However, not all network rules correlated with the binding peaks observed (Fig. 6.5B). By linking these analyses, it was possible to select potential candidates to validate *in vitro*.

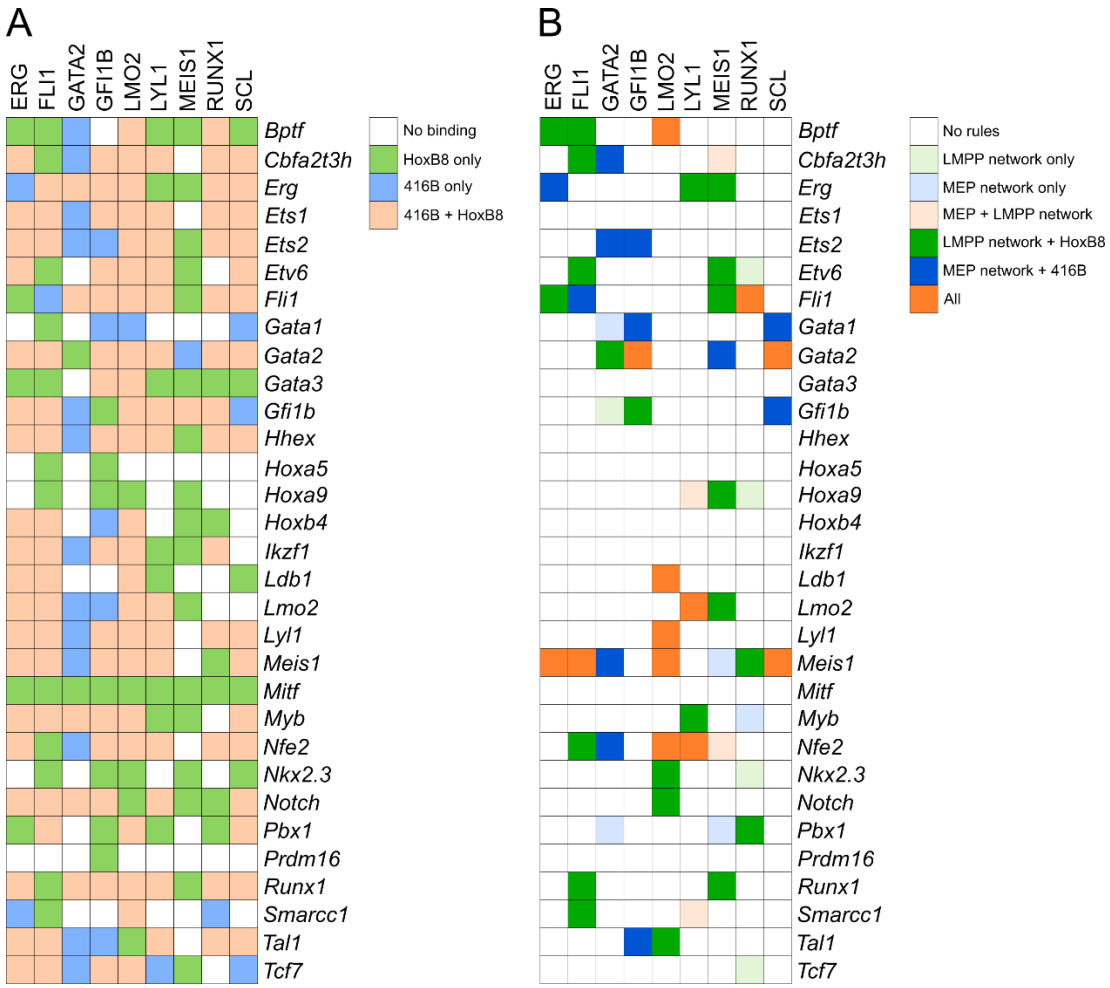


Figure 6.5. Integrating transcription factor binding with network rules to identify regulatory relationships to validate in vitro. (A) Figure showing where transcription factor binding (columns) was observed in transcription factor-encoding genes (rows) in 416B and HoxB8-FL cell lines. Squares are coloured by transcription factor binding: No binding – white; HoxB8 binding only – light green; 416B binding only – light blue; 416B and HoxB8 binding – peach. (B) Figure showing network rules observed in the MEP and LMPP trajectories, and the agreement of the binding patterns with network rules. Squares are coloured by the presence of network rules. No rules – white; LMPP network rule only (no binding in HoxB8-FL cells) – pale green; MEP network rule only (no binding in 416B cells) – pale blue; MEP and LMPP network rule (no binding in either cell line) – pale peach; HoxB8 binding and LMPP rule –green; 416B binding and MEP rule – dark blue; All (shared rules; binding in both cell lines) – orange.

In the MEP network model, GATA2 positively regulated *Cbfa2t3h* and *Nfe2*, which was not seen in the LMPP network model (Fig. 6.6A). Interrogating the ChIP-Seq data for GATA2 showed that GATA2 bound to both *Cbfa2t3h* and *Nfe2* loci in 416B cells. Whilst binding could be seen in the Hoxb8 cells, these binding events were minor compared with regions bound elsewhere in the genome, and therefore not recognised as peaks by the computational algorithm (Fig. 6.6B and C). At the *Cbfa2t3h* locus, two prominent binding peaks were identified at the promoter region in 416B cells. The two peaks represent the previously identified minimal and full promoter; the minimal promoter represents the most conserved region. At the *Nfe2* locus, a prominent peak was identified at the -7kb enhancer region in 416B cells. Single-cell profiling previously performed in our lab showed that *Gata2* is not highly expressed in HoxB8-FL cells, which is consistent with primary bone marrow LMPP cells (Hamey et al. 2017). The ChIP-seq data was in accordance with the single-cell profiling data, as GATA2 binding at both *Cbfa2t3h* and *Nfe2* loci was limited in HoxB8-FL cells. The ChIP-seq data therefore corroborates the *Gata2*-activates-*Cbfa2t3h* and *Gata2*-activates-*Nfe2* rules found only in the MEP network model.

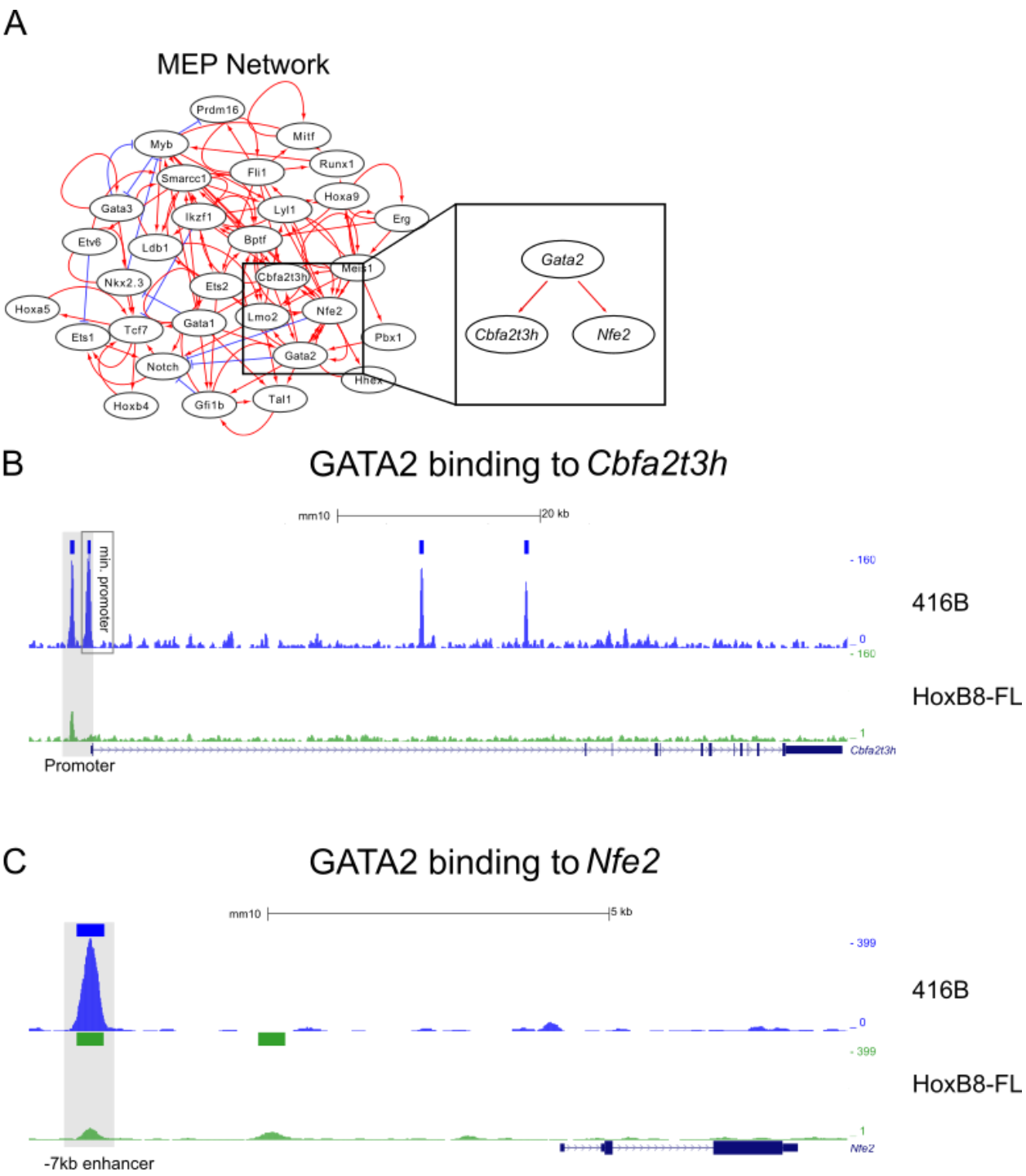


Figure 6.6. Regulatory relationships unique to the MEP network model are supported by transcription factor binding. (A) Diagram of the MEP network model zoomed into the trio of genes with a regulatory pattern identified as unique to the MEP network. The diagram shows the trio within the network model and alone for clarity. (B) ChIP-seq analysis of GATA2 binding in 416B and HoxB8-FL cell lines at the *Cbfa2t3h* locus. The minimal and full promoters are highlighted. The bars above the tracks indicate a binding event that was called as a peak. (C) ChIP-seq analysis of GATA2 binding in 416B and HoxB8-FL cell lines at the *Nfe2* locus. The -7kb enhancer region is highlighted. The bars above the tracks indicate a binding event that was called as a peak.

6.4. *In vitro* validation supports differences in network model connectivity

To validate whether GATA2 binding contributes to the transcriptional activation of *Cbfa2t3h* and *Nfe2* in 416B cells, as predicted by the network model, reporter constructs were generated for the *Cbfa2t3h* minimal and full promoter as well as the *Nfe2* enhancer. The constructs were generated as wild-type versions (WT) or with GATA2 binding site mutations. Luciferase reporter assays were performed to determine the activation of *Cbfa2t3h* and *Nfe2* constructs with and without mutated GATA2 binding sites (n=3 biological replicates). A construct that lacked the promoter/enhancer was included as an empty vector control (pGL2-Basic vector for *Cbfa2t3h* and pGL2-Promoter vector for *Nfe2*). The level of luciferase activity directly corresponded to the activity of the *Cbfa2t3h* promoter or *Nfe2* enhancer (Fig. 6.7). Luciferase assays were performed in 416B cells only as the interrogated rules were specific to the MEP network model.

Luciferase assays showed that the *Cbfa2t3h* promoter and *Nfe2* enhancer regions are active in 416B cells, with the *Cbfa2t3h* promoter being more active than the *Nfe2* enhancer (Fig. 6.7A/B). Mutation of the GATA2 binding sites decreases their activity, as seen by a decrease in luciferase activity normalised against the empty vector control. The GATA2 mutants show a significant fold reduction in luciferase activity compared to the WT control (Fig. 6.7C/D). Specifically, GATA2 mutations caused a 0.48±0.05-fold change in activity at the *Cbfa2t3h* promoter ($p<0.001$), a 0.58±0.05-fold change in activity at the *Cbfa2t3h* minimal promoter ($p<0.001$), and a 0.46±0.03-fold change in activity at the *Nfe2* enhancer ($p<0.05$). These results are consistent with GATA2 activating *Cbfa2t3h* and *Nfe2* during MEP differentiation, as suggested by the network model. Therefore, the luciferase assays validated the regulatory relationships proposed *in silico* between GATA2 and *Cbfa2t3h*, and between GATA2 and *Nfe2*, which was also supported by transcription factor binding.

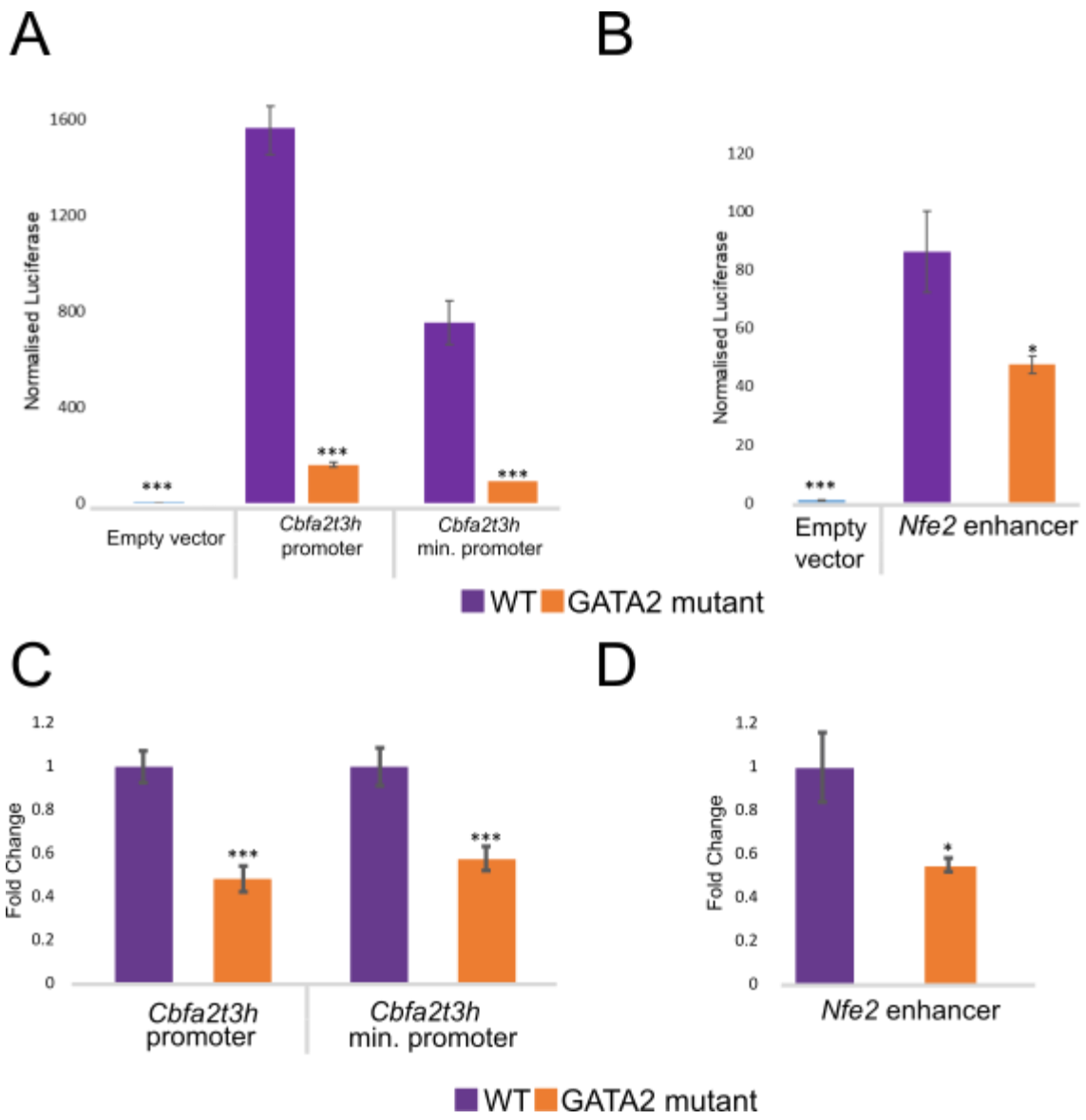


Figure 6.7. In vitro validation of MEP regulatory network rules. Normalised luciferase activity at the (A) *Cbfa2t3h* promoter and (B) *Nfe2* enhancer, comparing the wild-type and GATA2 mutant regulatory regions. Results are normalised to the empty vector (A: pGL2-Basic vector; B: pGL2-Promoter vector). Fold change in luciferase activity at the (C) *Cbfa2t3h* promoter and (D) *Nfe2* enhancer, comparing the wild-type and GATA2 mutant regulatory regions. WT: wild-type; *P <0.05, **P<0.01, ***P<0.001; two-tailed unpaired *t*-Test, $n=3 \pm$ SD

6.5. Conclusions

This chapter focused on validating the hybrid inference method used to identify transcriptional regulatory networks describing HSC differentiation to MEP and LMPP cell fates. The regulatory network inference methods were constructed using qRT-PCR data which profiled 2,167 single cells and included HSCs and early progenitors. Both ChIP-seq and luciferase assays served to validate regulatory relationships predicted by the network model.

Boolean modelling has been used to infer regulatory networks in many studies, including the stem cell field (Dunn et al. 2014; Moignard et al. 2015). However, these studies were either limited by the use of bulk expression data, which obscures heterogeneity in cell fate decisions, or only considered binary gene expression. In doing so, the continuous data is not considered and therefore the accuracy of a model describing how cells transition across differentiation may be affected. To overcome this problem, a hybrid Boolean network inference model, which considers pseudotime ordering of single cells, was developed. Although the model is restricted to binary expression states, it does not exclude cyclical relationships and does not computationally limit the number of genes in the network model, unlike other existing methods (Ocone et al. 2015; Schütte et al. 2016).

Examining the regulatory relationships identified by the network model revealed previously identified relationships, such as the activation of *Gfi1b* by *Gata2*, as well as expected expression patterns in which erythroid genes were inhibited along the lymphoid trajectory, and vice versa (Moignard et al. 2013). The identified MEP network-specific relationships, in which *Gata2* positively regulates *Nfe2* and *Cbfa2t3h*, are consistent with known biological functions of these genes. G*ata2* is a regulator of HSPC function and involved in HSC maintenance and expansion, as well as early haematopoietic cell formation (Rodrigues et al. 2005; K.-C. Lim et al. 2012; Tsai and Orkin 1997). *Gata2* mouse models demonstrate that a homozygous mutation in *Gata2* is embryonic lethal, and knockout models show defects in primary haematopoiesis (Tsai et al. 1994). *Cbfa2t3h* is a component of a transcription factor complex that regulates erythroid and megakaryocytic programs (Goardon et al. 2006; Fujiwara et al. 2010; Hamlett et al. 2008). The gene encodes for ETO2, a corepressor in complex with SCL (encoded by *Tal1*) (Schuh et al. 2005). During differentiation, GATA2 binds and activates *Cbfa2t3h*, causing ETO2 to repress its own promoter, leading to erythroid maturation and a GATA1-driven transcriptional program (Fujiwara et al. 2009). *Nfe2* was originally discovered as an upstream regulator of globin gene expression (Ney et al. 1993). It is expressed in HSCs and erythroid and megakaryocytic lineages, and is required for megakaryocyte maturation and platelet production (Andrews et al. 1993; Shivdasani et al. 1995).

To validate the MEP network specific rule, transcription factor binding patterns were investigated in 416B and HoxB8-FL cell lines, which represent erythroid and myeloid-lymphoid lineages, respectively (Dexter et al. 1979; Redecke et al. 2013). Gene expression profiles for these cell lines were previously obtained by interrogating the same gene set using qRT-PCR. ChIP-seq data showed prominent binding peaks at the *Cbfa2t3h* minimal and full promoter, and the -7kb *Nfe2* enhancer region, in 416B cells; while binding was also observed in HoxB9-FL cells, it was very

limited. Luciferase reporter constructs were generated to experimentally validate GATA2 transcriptional activation and showed that the wild-type constructs had significantly higher activity than constructs without the promoter/enhancer region or with mutated GATA2 binding sites. Therefore, transcription factor binding and luciferase assays validated regulatory relationships proposed *in silico* by the inferred network model. Furthermore, linking regulatory relationships to the MEP regulatory network but not the LMPP network model illustrates how network topology guides interactions between HSPC regulators such as *Gata2* and lineage-restricted genes like *Cbfa2t3h* and *Nfe2*. Identifying and validating simple rules in the network models demonstrates how *in-silico* investigations can drive *in vitro* and *in vivo* studies.

6.5.1. Limitations and future work

Single-cell qRT-PCR is a sensitive method of measuring gene expression but limits the focus of any study to the genes selected by the investigator. In this study, the gene set was myeloid-biased and focused on regulators of stem cell maintenance, limiting discovery of regulatory relationships along the LMPP trajectory. However, a handpicked gene set also results in incomplete network models, as it is not possible to discovery relationships involving novel genes. Our model suggests that *Cbfa2t3h* activates several genes in the network; however, it has been traditionally identified as a corepressor with SCL (Schuh et al. 2005). The described relationships could therefore be explained by the direct function of *Cbfa2t3h* or a double repressive link, but it is not possible to verify this activity without expanding the gene set used. It would be interesting to identify and validate regulatory relationships with a larger, unbiased gene set to uncover more of the complex interactions involved in HSPC differentiation. Furthermore, identifying mechanisms that direct stem cells into specific differentiation trajectories would be a valuable investigation for improving our understanding of the haematopoietic system, as well as our understanding of the genes involved in cell fate and the perturbations that occur in disease.

6.5.2. Summary

In summary, a computational network inference method was applied to single-cell gene expression profiles and identified differences in transcriptional regulatory programs between cells differentiating towards erythroid and lymphoid lineages. Regulatory network rules were validated using transcription factor binding patterns and *in vitro* assays, demonstrating the value of *in silico* network inference for driving further investigation into processes governing differentiation.

Chapter 7: Discussion

7.1. Thesis Overview

The purpose of this thesis was to interrogate the transcriptional landscape of haematopoietic stem and progenitor cell (HSPC) differentiation using single-cell techniques and genome editing technologies. Specifically, single-cell RNA-sequencing (scRNA-seq) and quantitative real-time PCR (qRT-PCR) were used to investigate the structure of the haematopoietic hierarchy. The scRNA-seq dataset formed the basis of an investigation into genes implicated in HSC biology using CRISPR genome editing. The qRT-PCR dataset was used to infer regulatory networks for cell fates and identify regulatory relationships that could be experimentally validated. An overview of this work, as well as its implications for HSPC biology, will be discussed here.

7.1.1. Characterising the haematopoietic transcriptional landscape using single-cell technologies

Two single-cell profiling techniques were used in this thesis to measure gene expression in individual HSPCs: scRNA-seq (Chapter 3) and qRT-PCR (Chapter 5). Interrogation of both datasets revealed they were able to generate structures of the transcriptional landscape that recapitulated known aspects the haematopoietic hierarchy.

A previously published dataset was the basis of the qRT-PCR investigation (N. K. Wilson et al. 2015). This dataset isolated discrete populations of LT-HSCs using four commonly used sorting strategies, as well as discrete progenitor populations: FSR-HSCs, LMPPs, MEPs, CMPs and GMPs. The focus of the study was to resolve heterogeneity within the LT-HSC compartment and therefore the gene set consisted of genes important in HSC biology, skewed towards the myeloid-erythroid lineages. In the work presented in Chapter 5, this dataset was supplemented with intermediate progenitor populations to more comprehensively interrogate the HSPC compartment. The three additional populations were FSR-HSCs, MPPs, and PreMegEs, and were chosen as they are defined populations within the HSPC compartment that should have multi-lineage potential whilst lacking the long-term potential of reconstituting the haematopoietic system in an irradiated mouse (Pronk et al. 2007; Cabezas-Wallscheid et al. 2014).

The four HSC sorting strategies were generally similar in their gene expression, but also demonstrated key differences which may reflect properties of the HSCs themselves or that of

'contaminating' cells present in the impure HSC fractions. HSC4 in particular was the most transcriptionally unique of the four LT-HSC populations, as seen in the dimensionality-reduction visualisations and clustering analysis. It was characterised by less *Cbfa2t3h*-expressing cells and higher *Gata3* expression relative to the other LT-HSC populations, together suggesting that HSC4 cells have less erythroid potential and are instead biased towards the lymphoid lineage (Pandolfi et al. 1995; Goardon et al. 2006). Studies show that *Gata3*, which is essential to T-lymphocyte differentiation, is present in LT-HSCs and involved in regulating their self-renewal and entry into the cell cycle (Ku et al. 2012; Frelin et al. 2013; D. G. Kent et al. 2009). Together with the reduced *Cbfa2t3h* expression, this suggests that the expression pattern seen in HSC4 cells can in fact be attributed to the functional HSCs and that this sorting strategy may enrich for lymphoid-primed HSCs.

Clustering and correlation analysis demonstrated that the transcriptional profiles of these HSPC populations were in line with known biology. Correlation analysis measures the strength of relationships between sets of genes; if two genes are positively correlated, their expression increases simultaneously, whereas a negative correlation describes a situation where the expression of one gene decreases while the expression of another increases. While correlation analysis does not describe cause-and-effect relationships, it can give insights into functional relationships between genes. In the qRT-PCR data described in Chapter 5, a cluster of positive correlations was observed between *Gata2*, *Gfi1b*, and *Tal1*, which is consistent with literature that describes positive *Gata2-Tal1* and *Gata2-Gfi1b* relationships (Moignard et al. 2013; Bloor et al. 2002; J. E. Pimanda et al. 2007). Furthermore, negative correlations were apparent between erythroid and lymphoid genes, such as *Gata1-Nkx2.3* and *Gata1-Notch* (Pabst et al. 2000; Pabst, Zweigerdt, and Arnold 1999). Overall, this demonstrates that the qRT-PCR dataset accurately depicted relationships in haematopoiesis and could be used to further study transcriptional regulation in these populations.

Interestingly, the functional and fate-output properties of CMPs have been increasingly challenged, with studies suggesting that CMPs are not a bipotent population, but rather a heterogeneous population of committed myeloid or erythroid/megakaryocyte progenitors (Paul et al. 2015). The CMP phenotype was further challenged in this study, which showed that CMPs generally clustered with GMPs and had similar correlations between pairs of genes. Clustering of GMPs and CMPs was also described by Wilson *et al.* (2015). In particular, genes such as *Hoxa9* and *Notch* were involved in similar, predominantly negative relationships in CMPs and GMPs, whereas these genes have positive correlations with other transcription factors in MEPs. This suggests that CMPs are in

fact already committed towards a cell fate lineage, and the sorting strategy used in this experiment isolated cells moving towards a myeloid fate. Importantly, bulk analysis may not have captured the similarities between these populations, demonstrating the value of single-cell expression analysis.

Pseudotime analysis identified two differentiation trajectories in the dataset, which described the differentiation of HSCs towards MEPs and LMPPs. The high number of transcription factor genes measured made it possible to construct a regulatory network model for differentiation along these two trajectories. The regulatory rules *Gata2*-activates-*Cbfa2t3h* and *Gata2*-activates-*Nfe2* were specific to the MEP trajectory and confirmed *in vitro*, validating the *in-silico* model. Interestingly, while the progenitor populations were associated with more negative correlations, the trajectories involved mainly activating relationships, and very few inhibiting relationships. A possible explanation is that while repression may be important in defining progenitor cell phenotypes, activating relationships are key in driving differentiation towards these cell fates.

The qRT-PCR dataset was useful for identifying regulatory networks within haematopoiesis but was limited in its gene set and cell sampling method. In the MEP trajectory, there is a large gap between the HSCs and MEPs, suggesting cells that have transcriptional profiles representing an intermediate state between these two phenotypes are still missing, despite the additional sampling of MPPs, PreMegEs, and MEPs. Alternatively, the gene set used may not have been sufficient to differentiate these intermediate populations and order them on the pseudotime trajectory. Furthermore, considerable overlap was seen between GMPs and LMPPs, which separate in the scRNA-seq data based on their transcriptional profiles, highlighting the disadvantages of a limited gene set for discriminating between cell-fates. The scRNA-seq experiment in Chapter 3 describes a transcriptome-wide approach to gene expression profiling, and samples cells using broad, inclusive gates.

Broad gates were used to capture cells outside of conventional, strict sorting gates, which are normally considered ‘contaminating’ cells. This approach was used to see whether the contaminating cells represented intermediates between cell states and could contribute towards a continuous picture of the haematopoietic hierarchy. Index sorting made it possible to retrospectively assign cells to HSPC populations. Dimensionality reduction methods confirmed that the scRNA-seq data recapitulated the haematopoietic hierarchy. The diffusion map method was particularly successful for visualising continuous data and clearly separated three trajectories: erythroid, myeloid, and lymphoid. This separation was also seen using STREAM analysis.

However, SPRING analysis revealed that the scRNA-seq atlas was capturing early differentiation into specific trajectories, for example towards megakaryocytes or B-cells. When the top differentially expressed genes for each branch were plotted on the diffusion map, it showed that the diffusion map method was unable to separate out these early, specific branches. This demonstrates that SPRING was the superior method for capturing the full scope of the differentiation events in the scRNA-seq dataset. Furthermore, the results highlight early lineage choices occurring in HSPCs that were not observed at the population level, in line with other studies that suggest cell-fate restricted cells arise from HSPCs without major transitions through intermediate stages (Laurenti and Göttgens 2018; Notta et al. 2015; Paul et al. 2015; Grün et al. 2016; Velten et al. 2017).

Gene expression analysis using scRNA-seq has more scope for discovering novel population markers and gene regulation trends; however, it is limited by a high dropout rate. Dropout events occur when the expression of a gene is moderate in one cell and undetected in another, usually due to low mRNA in individual cells. The low starting amounts of mRNA in single-cells compared to bulk populations make it more likely that a transcript may not be transcribed and/or amplified during cell processing and therefore goes undetected during sequencing, resulting in a dropout event (Kharchenko, Silberstein, and Scadden 2014; W. V. Li and Li 2018). Furthermore, scRNA-seq analysis only considers highly variable genes, meaning that genes that did not demonstrate changes in expression between different samples or conditions were excluded from further analysis. Overall, scRNA-seq suffers from high technical and biological noise, which make it difficult to accurately infer transcriptional regulatory networks using these gene measurements.

Ultimately, the method of choice for gene expression analysis depends on the question being asked—while scRNA-seq gives a transcriptome-wide overview of the transcriptional landscape, is beneficial for novel marker discovery and offers better discrimination between cell-fates and their differentiation trajectories, it is costly and suffers from dropouts. On the other hand, qRT-PCR is limited to a curated gene set but allows for a deeper analysis of specific sets of genes and confers fewer challenges for studying regulatory networks with currently available computational methods.

7.1.2. Genome editing to interrogate HSC biology

Genome editing using the CRISPR/Cas9 system was used in Chapter 4 to investigate genes implicated in HSC biology. This technology has been applied to investigations of the haematopoietic system to create leukaemic mouse models and CRISPR screening platforms to

identify potential therapeutic targets (overview in Section 1.4.1.1) (Heckl et al. 2014; Tzelepis et al. 2016). In this thesis, a CRISPR screen was designed to determine the effect of perturbing the previously identified MolO (molecularly overlapping) and SuMO (surface marker overlap) genes, using changes in EPCR expression as the primary outcome, and changes in apoptosis and colony output as secondary experiments (N. K. Wilson et al. 2015). The MolO and SuMO genes are associated with a molecularly similar subpopulation of HSCs and repopulating HSCs, respectively; the curated candidate gene list included 16 genes that were most highly expressed in LT-HSCs.

Genotyping analysis performed in this investigation suggests that CRISPR gRNAs were successfully targeting the candidate genes, but the results indicate that the gene perturbations had no significant effect on EPCR expression, apoptosis, nor lineage output. Significant changes in EPCR expression based on the percentage of $EPCR^+$ cells or median EPCR expression after seven days in culture were generally caused by one gRNA out of three for the candidate gene. CRISPR/Cas9 technology has been associated with off-target mutations that may cause genomic instability and disrupt the normal function of genes other than the intended target; therefore, drawing conclusions about the effect of perturbing candidate genes based on one gRNA may incorrectly represent their function (Pattanayak et al. 2013; Cho et al. 2014; Y. Fu et al. 2013; X.-H. Zhang et al. 2015). Pooling the results from individual gRNAs for each perturbation together showed that only *Procr* significantly decreased median EPCR expression in cultured E-SLAM cells.

The loss of function of a single gene may not influence the overall function of cells due to genetic compensation, where related genes are upregulated as a consequence of gene knockout in order to maintain normal function (El-Brolosy and Stainier 2017). Furthermore, possible gene redundancies may mean that perturbing a single gene would have no effect on the biological phenotype as another gene would perform the same function (Nowak et al. 1997). Another possible approach for this CRISPR screen would be to transduce cells with CRISPR gRNAs for multiple genes and include a unique barcode for each gRNA. The transduced cells could then be sequenced to reveal which gRNAs contributed to the phenotype (Dixit et al. 2016). This type of analysis may produce more significant results, as multiple genes would be simultaneously perturbed.

However, the overall lack of significant results in this investigation suggest that there are several weaknesses in the study design. These weaknesses concern the medium conditions used and the

parameters investigated. A further explanation and possible solution for each is offered in the list below:

- *Medium conditions.* The conditions in which the HSCs are cultured are optimal for their expansion. It was hypothesized that the gene perturbations would decrease properties such as HSC maintenance and may therefore increase differentiation and mature cell output. However, any effects of the candidate gene perturbations could be masked by the culture conditions used, because factors such as cell proliferation and differentiation are already being encouraged. It would be interesting to see how the results would be affected if the HSCs were cultured in medium that instead discouraged cell expansion, and whether the gene perturbations could result in increased differentiation in HSC maintenance conditions.
- *Investigating changes in EPCR expression.* EPCR is a well-established marker of HSCs but is also a rarely expressed protein. The hypothesis in this investigation was that perturbing the candidate genes would decrease cells that remain in the primitive HSC state, and would therefore decrease EPCR expression. However, assaying the loss of an already lowly expressed protein is difficult and unreliable to quantify. The study could be strengthened by focusing on surface marker expression that is gained after treatment with the various CRISPR gRNAs. As cells are dividing and differentiating, the more mature cells will express different lineage markers. A lineage cocktail was used in this investigation and therefore obscured any interpretation of the more mature cells. Analysing the individual mature lineages separately may reveal more apparent changes in differentiation towards the diverse lineages. Mature cells will also occur more frequently in culture as they have a higher proliferation rate than HSCs, and therefore could be assayed more reliably.
- *Repopulation potential.* The SuMO genes defined in the Wilson *et al.* study were associated with functional HSCs and it was predicted that their perturbation may cause a change in HSC characteristics (N. K. Wilson et al. 2015). The true measure of a real stem cell is to test its ability to repopulate a mouse. Therefore, transplanting the treated cells into irradiated mice could show whether the gene perturbations prevent reconstitution of the haematopoietic system, as well as their impact on lineage output.

A couple other issues, however, call into question whether the study should be improved, or completely redesigned. HSCs are a rare population in the adult bone marrow, estimated at 1 in 10,000 cells (Szilvassy et al. 1990). As 250 E-SLAM cells were sorted into each well in triplicate

for each gene, the number of genes that could be assayed in each individual screen was severely restricted. Furthermore, the Cas9-transgenic mice also had to be specifically bred for this study, limiting the number of individual screens and repetitions possible based on age and litter size. Instead of using Cas9-transgenic mice, more readily available mice could be used, such as C57BL/6, and transduced with Cas9 vector together with the CRISPR gRNA vector. This would make it much more feasible to obtain repeats for all experiments.

A possible redesign for this study would also assay a more abundant cell type. As the MolO and SuMO genes are associated with HSC maintenance and repopulation capacity, it would be interesting to see whether overexpressing these genes in other populations would restore self-renewal and an HSC state. Specifically, the candidate genes could be overexpressed in multipotent populations with finite or no self-renewal, such as MPPs and FSR-HSCs. These cell types are more abundant in the bone marrow and therefore individual screens could include more candidate genes and could be done more frequently. As this experiment design would be investigating a gain of HSC characteristics, assaying HSC markers such as SLAM ($CD48^-$ $CD150^+$) and EPCR would be appropriate, as their expression should increase and therefore any changes would be easier to detect. This proposed design could address the original aim of the CRISPR screen, which was to identify genes important in HSC biology; furthermore, identifying genes that are able to restore self-renewal and an HSC state could have major implications for stem cell therapies.

7.1.3. Implications of this thesis for HSPC biology

The single-cell gene expression analyses performed in this thesis confirm the known structure of the haematopoietic hierarchy, in which HSCs differentiate into mature cell-fates, a process which is regulated by diverse interactions in complex transcriptional networks. The use of single-cell technologies made several observations possible:

- SPRING analysis of the scRNA-seq dataset revealed early lineage branching that was previously undetected in smaller datasets and bulk expression experiments
- Heterogeneity in HSC populations was demonstrated in the qRT-PCR dataset, revealing possible lymphoid-priming in the HSC4 population
- Clustering and correlation analysis showed that the isolated CMP population was most similar to GMPs and most likely represents cells primed towards a myeloid lineage, rather than a bipotent population

- Single-cell analysis can be used to construct transcriptional networks of differentiation processes based on pseudotime ordering, which were validated *in vivo*.

The scRNA-seq dataset is the first of its kind to describe the whole haematopoietic hierarchy and provides a powerful reference dataset for biologists. The interactive website allows researchers to interrogate gene expression on the HSPC landscape. Several groups have used the dataset to demonstrate the computational algorithms they developed to analyse single-cell gene expression data (Griffiths et al. 2018; H. Chen et al. 2018). Furthermore, researchers can project their single-cell datasets onto our dataset to interrogate changes in gene expression caused by their particular phenotype of interest. In our paper describing the scRNA-seq dataset, we projected young and old HSCs from C57BL/6, DBA/2, and Vwf-EGFP mice to show both types of HSCs cluster with the LT-HSCs from our dataset, but old HSCs form a tighter, more molecularly homogenous population (Kowalczyk et al. 2015; Grover et al. 2016; Nestorowa et al. 2016). This type of analysis could also be performed with disease models, leading to potential discoveries about differences in gene regulation in normal versus leukaemic cells that could inform future therapies.

7.2. Future directions for single-cell biology

This thesis described the use of single cell RNA-sequencing and qRT-PCR to further our knowledge of the transcriptional landscape of HSPC differentiation. Numerous techniques have since been developed that offer a different approach to gene expression analysis or enable it to be paired with perturbation or epigenetic research. This section describes these technologies and their implications for cell-fate research.

7.2.1. Advances in single cell gene expression analysis

Single-cell RNA-sequencing is a powerful approach to collect gene expression measurements for the whole transcriptome of individual cells. However, it is associated with a high cost per cell, especially in plate-based methods that require high volumes of reagents (Picelli et al. 2014).

Microwell technologies such as CytoSeq and Microwell-Seq are plate-based methods designed to decrease costs by reducing the reagents required (Fan, Fu, and Fodor 2015; Han et al. 2018). Single cells are deposited into individual wells together with a library of barcoded beads; after cell lysis, the mRNA is able to hybridize to the beads, making it possible to pool the cells before reverse

transcription, amplification and sequencing (Fan, Fu, and Fodor 2015). Microwell-Seq has recently been used to analyse more than 400,000 single cells from major mouse tissues to construct a mouse cell atlas (Han et al. 2018). The Genesis system (Celsee) and SMARTer ICELL8 Single-Cell System (Takara) are commercially available platforms that use microwell technology for single-cell analysis.

Another recent high-throughput scRNA-seq method is SPLiT-seq (Split Pool Ligation-based Transcriptome sequencing), which can be used to transcriptionally profile thousands of formaldehyde fixed cells (Rosenberg et al. 2018). Individual transcriptomes go through multiple rounds of combinatorial barcoding, which append well-specific barcodes, unique molecular identifiers, and sequencing barcodes. Prior to sequencing, this method does not require any complex instruments or expensive reagents. SPLiT-seq was used to analyse more than 150,000 transcriptomes from mouse brains and spinal cords, identifying over 100 different cell types (Rosenberg et al. 2018).

Droplet-based microfluidic methods have been used widely as they successfully reduce the cost per cell while also increasing throughput (Hamey et al. 2016). These technologies encapsulate single cells in nanolitre droplets with barcoded beads. Within each droplet, the single cell is lysed and the bead dissolves to release barcoded reverse transcription oligonucleotides into solution. Reverse transcription of the polyadenylated mRNA can then occur, barcoding all the cDNA from a single cell with the same barcode. The cells are then pooled and sequenced together. Gene expression profiles can be simultaneously generated for thousands of cells, drastically reducing the cost per cell. The first high-throughput droplet-based methods to be published were Drop-Seq and InDrop (Macosko et al. 2015; Allon M. Klein et al. 2015). The 10x Chromium™ system is a commercially available platform from 10x Genomics that allows researchers to perform droplet-based scRNA-seq without having to generate their own microfluidic devices and reagents (Zheng et al. 2017).

Dahlin *et al.* recently used the 10x Chromium™ system to resolve eight lineage trajectories in mouse bone marrow HSPCs: lymphoid, megakaryocyte, erythroid, neutrophil, monocyte, eosinophil, mast cell, and basophil lineages (Dahlin et al. 2018). Furthermore, they observed reduced *Myc* expression and proliferative defects in a c-Kit mutant mouse model (W^{41}/W^{41}), in which transcriptional profiling revealed the lack of a mast cell lineage entry point. This study profiled 44,802 wild-type HSPCs and 13,815 W^{41}/W^{41} HSPCs, demonstrating how droplet-based

methods can be used to perform informative, large-scale transcriptomic studies at a fraction of the cost of plate-based scRNA-seq (Dahlin et al. 2018).

7.2.2. Combining single-cell gene and protein expression measurements

In this thesis, scRNA-seq was combined with index sorting to record surface marker expression data for all the individual cells sorted. This made it possible to use surface marker protein expression to differentiate between cells and retrospectively assign them to HSPC cell types. Wilson *et al.* also used index sorting to link molecular characterisation of single cells by qRT-PCR with functional studies (N. K. Wilson et al. 2015).

Analysing the proteome can elucidate which proteins are involved in defining cell phenotype and function. However, it is not possible to collect index sorting data using droplet-based approaches as they are incompatible with cytometry. Researchers have therefore developed methods to combine transcriptome profiling with protein expression measurements. CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) and REAP-seq (RNA expression and protein sequencing) use antibodies conjugated to DNA barcodes instead of fluorophores to label cells (Peterson et al. 2017; Stoeckius et al. 2017). In CITE-seq, the antibodies are conjugated to streptavidin, which is bound to the biotinylated DNA barcodes, whereas in REAP-seq, the antibody and DNA barcode are covalently bound (Todorovic 2017). Droplet-based scRNA-seq approaches can then be applied to these DNA-barcode labelled cells to generate both mRNA and protein expression data.

The application of CITE-seq was demonstrated on 8,000 individual cord blood cells, which produced cell profiles consistent with established flow cytometry profiles and improved characterisation of natural killer cells based on protein expression (Stoeckius et al. 2017). REAP-seq was first used to study human naïve CD8+ T-cells using a panel of 80 barcoded antibodies. Peterson *et al.* investigated the activation of these T-cells after treatment with a CD27 agonist and characterised differentially expressed genes and proteins in untreated versus treated cells, demonstrating how this technology can be used to enhance preclinical studies (Peterson et al. 2017). Both these methods are compatible with the 10x Chromium™ platform and therefore can be widely used to simultaneously study gene and protein expression in a high-throughput manner. Both methods are currently limited to measuring surface marker protein expression, but may be extended in the future to measure intracellular proteins as well (Todorovic 2017).

7.2.3. Genome and transcriptome sequencing

New single-cell technologies have been developed to sequence the genomic DNA and mRNA from the same cell. These methods pair established scRNA-seq technologies with whole genome amplification methods. DR-seq (gDNA-mRNA sequencing) first amplifies the nucleic acids prior to physical separation to minimise sample loss (Dey et al. 2015). However, it uses CEL-seq to sequence the transcriptome, which only targets the 3' end of mRNAs (Hashimshony et al. 2012). The cells are also manually selected, preventing high throughput application. An alternative method is G&T-seq (genome and transcriptome sequencing), which captures RNA using biotinylated primers and separates it from DNA using streptavidin-coated magnetic beads; the RNA is processed using the Smart-seq2 protocol and multiple displacement amplification is used to amplify DNA (Macaulay et al. 2015, 2016). Sequencing both the genome and transcriptome from a single cell makes it possible to link genomic and transcriptomic heterogeneity. These technologies have not been applied to the haematopoietic system yet, but could be particularly useful in studying blood disorders where acquired mutations are linked to aberrant function (Hamey et al. 2016).

7.2.4. Combining single-cell transcriptomic and epigenomic measurements

While this thesis focused on transcriptional heterogeneity and regulation of HSPC differentiation, the transcriptome is not the only factor that impacts cell fate decision making. Epigenetic regulation plays an important role during HSC maintenance and differentiation. Chromatin modifications are involved in programming gene expression changes in undifferentiated HSCs as well as differentiating cells (Cui et al. 2009). Several single-cell methods have been designed to interrogate the epigenome (Table 7.1).

Table 7.1. Single-cell methods interrogating the epigenome. Methods are grouped by the target of interest.

Target	Methods
DNA methylation	• Single-cell Reduced Representation Bisulfite Sequencing (scRRBS) (H. Guo et al. 2013) • Single-cell Bisulfite Sequencing (scBS) (Smallwood et al. 2014)
Histone modifications	• Drop-ChIP: combines microfluidics, DNA barcoding and next-generation sequencing to assess the chromatin state of single cells (Rotem et al. 2015)
Chromatin accessibility	• Single-cell Assay for Transposase-Accessible Chromatin (scATAC-seq) • Cusanovich *et al.*: based on cellular indexing (Cusanovich et al. 2015) • Buenrostro *et al.*: based on microfluidics (Buenrostro et al. 2015)
Chromatin arrangement	• Single-cell Hi-C (Nagano et al. 2013)

Combined approaches	• Single-cell genome-wide methylome and transcriptome profiling
	  • scM&T-seq (Angermueller et al. 2016)
	  • scMT-seq (Hu et al. 2016)
	• Single-cell Nucleosome, Methylation and Transcription Sequencing (scNMT-seq) (Clark et al. 2018)
	• Single-cell Trio-seq (Y. Hou et al. 2016)
	• Single-cell analysis of genotype, expression and methylation (sc-GEM) (Cheow et al. 2016)

Buenrostro *et al.* used scATAC-seq to interrogate the chromatin accessibility of 10 HSPC populations and recapitulated the haematopoietic hierarchy (Buenrostro et al. 2018). They also performed scRNA-seq separately and associated the transcriptomic and epigenomic data using computational methods. Recently, new methods have been developed that simultaneously profile the transcriptome and features of the epigenome. Angermueller *et al.* and Hu *et al.* both published methods that combined single-cell genome-wide methylome and transcriptome profiling, called scM&T-seq and scMT-seq, respectively (Angermueller et al. 2016; Hu et al. 2016). Single-cell NMT-seq is a recent method that simultaneously profiles chromatin accessibility, DNA methylation, and the transcriptome (Clark et al. 2018). New technologies are also combining measurements of genomic, transcriptomic, and epigenomic data. Hou *et al*. combined scRNA-seq with scRRBS to create scTrio-seq, which is a single-cell triple-omics approach that simultaneously captures information of genomic copy-number variations, the DNA methylome, and the transcriptome (Y. Hou et al. 2016). An alternative method is sc-GEM, which combines a methylation assay with single-cell qRT-PCR and single-cell genotyping (Cheow et al. 2016).

These methods provide a new single-cell approach for studying heterogeneity during differentiation processes whilst combining the analysis of multiple factors that influence cell fate decision making. Many of these methods require the separation of cellular components of the individual analyses, reducing sensitivity, and may be restricted by the cost of performing the multi-

omics approach. The continued growth of the single-cell field is likely to yield solutions to the current issues facing multi-omics approaches, leading to new insights into HSPC biology.

7.2.5. High-throughput CRISPR screening with gene expression analysis

Genome editing is a powerful approach for studying the roles of particular genes and how perturbing their function affects the system in which they act. This approach can provide insights into gene functions in a normal state but is also important for understanding the effect of gene perturbations in disease states.

In the work described in Chapter 4, a low-throughput approach was used to perturb genes implicated in HSC biology to study their effect on cell phenotype and differentiation. Specifically, genes were targeted using CRISPR gRNAs, which guided the Cas9 nuclease to the gene of interest to cut the double stranded DNA, leading to loss of function mutations. This CRISPR screen was performed in bulk populations and measured changes in surface marker expression and colony output. Researchers have also used CRISPR/Cas9 technology to perform genome-wide knockout screens in individual cells, but these screens required follow-up investigations to dissect gene functions and their effect on the transcriptional network (Shalem et al. 2014; T. Wang et al. 2014).

Advances in single-cell and CRISPR/Cas9 technologies have made it possible to perform high-throughput, genome-wide knockout screens that show how the perturbations influence a cell's molecular profile. CRISP-Seq and Perturb-Seq use a library of barcoded gRNAs to target different genes; these barcodes can then identify the gRNAs present within a cell (Adamson et al. 2016; Dixit et al. 2016; Jaitin et al. 2016). While both approaches pair scRNA-seq with the CRISPR-based perturbations, CRISP-seq was developed on the massively parallel RNA-seq (MARS-seq) platform, whereas Perturb-seq uses a droplet-based approach. Jaitin *et al.* used CRISP-seq to interrogate gene regulation and heterogeneity in the immune system using a pool of gRNAs targeting 22 genes. The responses to the various perturbations were heterogeneous across different cell types, highlighting the need for single-cell analysis (Jaitin et al. 2016). Dixit *et al.* used Perturb-seq to interrogate the consequences of perturbing 24 different transcription factors in the immune system, either alone or in combination (Dixit et al. 2016). In their coordinating study, Adamson *et al.* performed CRISPR interference screens using the Perturb-seq platform to investigate the effect of repressing target genes on the mammalian unfolded protein response (Adamson et al. 2016). Typical gRNAs lack a polyadenylated tail and thus are not detectable by scRNAseq; an alternative approach, CROP-seq, redesigned a CRISPR vector to include the gRNA in a polyadenylated

mRNA transcript (Datlinger et al. 2017). This design makes it possible for the gRNA to be detected by scRNA-seq, circumventing the need for a barcoded gRNA library. Importantly, all these approaches were developed with accompanying computational methods capable of handling the complexity of the resultant data sets. These combined CRISPR perturbation and scRNA-seq methods are powerful tools for gaining functional and molecular insights into biological systems, including haematopoiesis.

7.2.6. Lineage tracing and "real time" cell dynamics

In this thesis, lineage trajectories were inferred from single-cell gene expression data using pseudotime ordering. Single-cell gene expression data represents a snapshot of each cell in a particular gene expression state, and pseudotime trajectory inference methods order these snapshots to reconstruct possible differentiation pathways, describing the gene expression changes occurring along the trajectory (Trapnell et al. 2014; Setty et al. 2016; Haghverdi et al. 2016). However, as these methods rely on snapshots, they lack temporal resolution, missing information such as the length of time a cell resides in a particular molecular state and how many cell divisions occur before developmental processes are observed. Furthermore, snapshot data may miss some of the gene expression dynamics that occur between the captured cell states, potentially leading to misrepresentation of the fate decision making processes (Etzrodt and Schroeder 2017).

Quantitative time-lapse imaging technologies offer an alternative approach to studying transcriptional regulation in single cells. Continuous single-cell imaging that is uninterrupted over several cell divisions provides temporal information about molecular dynamics (Skylaki, Hilsenbeck, and Schroeder 2016). This method was applied to investigate the relationship between GATA1 and PU.1 during differentiation towards megakaryocytic-erythroid and granulocytic-monocytic lineages (Hoppe et al. 2016). Traditionally, it was thought that these transcription factors inhibit each other's expression and can reprogram cells towards their respective lineages. However, this was not detected in live cell imaging; instead, the transcription factors were independently regulated at the start of megakaryocytic-erythroid or granulocytic-monocytic differentiation, and they were reinforcing lineage choices rather than initiating them (Hoppe et al. 2016). These findings show how live-cell imaging can reveal biological relationships that are contrary to snapshot gene expression data. Relating live-cell imaging to pseudotime ordering could add a new dimension to transcriptional regulation research, leading to an improved understanding of how a single HSC can differentiate into multiple distinct mature blood cell types.

Continuous live-cell imaging is a non-invasive technique that preserves information about the past and future of a single-cell. The observed molecular dynamics can therefore be used to identify past cell states and predict future fate decisions (Skylaki, Hilsenbeck, and Schroeder 2016). However, it is a very low-throughput method due to the intensive imaging and computational power required to track a single cell through differentiation. Recently, new methods have been developed that interrogate past and future cell states, which could offer meaningful insights into the structure of the haematopoietic tree.

Single-cell gene expression has been combined with genetic labelling of single cells to reconstruct lineage hierarchies in three related technologies: LINNAEUS (lineage tracing by nuclease-activated editing of ubiquitous sequences), scGESTALT (single cell genome editing of synthetic target arrays for lineage tracing) and ScarTrace (Spanjaard et al. 2017; Raj et al. 2018; Alemany et al. 2018). These technologies use CRISPR/Cas9 technology to randomly cause indels in target genes, called genomic scars, which produce somatic mutations that are heritable through cell divisions (Shapiro 2018). Paired with single-cell transcriptome sequencing, cell type and lineage information are recorded and used to reconstruct lineage trees. LINNAEUS and scGESTALT only use scRNA-seq to quantify the genomic scars, whereas ScarTrace also detects the scars from genomic DNA. These methods have been used in zebrafish to study fate decisions governing embryogenesis, brain development, haematopoiesis, and fin regeneration (Spanjaard et al. 2017; Alemany et al. 2018; Raj et al. 2018). It is challenging to implement these methods in more complex organisms; only very recently has a new technology been suggested for a ‘molecular recorder’ that characterises mammalian fate maps (Chan et al. 2018). Undoubtedly, these lineage tracing methods will continue to be optimised and further developed, as they confer a unique opportunity to characterise the molecular identities and lineage histories of cells. In contrast, a recent publication describes a computational method used to predict a cell’s future from scRNA-seq data. RNA velocity uses information about the ratios of unspliced and spliced mRNA to model the direction in which a cell is moving in the transcriptional space (La Manno et al. 2018). This method can be applied to datasets generated with commonly used single-cell plate- and droplet-based RNA-seq platforms, meaning it could be used to study cell dynamics in already existing datasets. Together, these technologies that look to the past and future of cell fate decisions could be particularly useful in HSPC biology, where many questions are raised about the structure of the haematopoietic hierarchy.

7.3. Concluding remarks

This thesis focused on improving our understanding of the transcriptional regulation underpinning haematopoietic stem and progenitor cell differentiation. Advances in single-cell technologies made in-depth study of transcriptional regulation and heterogeneity in HSPC populations possible, using single-cell gene expression techniques to reconstruct lineage trajectories and regulatory networks. The transcriptional landscape generated in this thesis using scRNA-seq has been made publicly available, providing a powerful resource for the haematopoietic community. The methods described in this thesis should be widely applicable to study haematopoiesis in normal and perturbed cells, furthering our knowledge of haematological diseases with implications for future therapies.

References

Adamson, Britt, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, et al. 2016. "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response." *Cell* 167 (7): 1867-1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

Adolfsson, J, O J Borge, D Bryder, K Theilgaard-Mönch, I Astrand-Grundström, E Sitnicka, Y Sasaki, and S E Jacobsen. 2001. "Upregulation of Flt3 Expression within the Bone Marrow Lin(-)Sca1(+)c-Kit(+) Stem Cell Compartment Is Accompanied by Loss of Self-Renewal Capacity." *Immunity* 15 (4): 659–69.

Adolfsson, J, Robert Månsson, Natalija Buza-Vidas, Anne Hultquist, Karina Liuba, Christina T Jensen, David Bryder, et al. 2005. "Identification of Flt3+ Lympho-Myeloid Stem Cells Lacking Erythro-Megakaryocytic Potential a Revised Road Map for Adult Blood Lineage Commitment." *Cell* 121 (2): 295–306. https://doi.org/10.1016/j.cell.2005.02.013.

Aguilo, F., S. Avagyan, A. Labar, A. Sevilla, D.-F. Lee, P. Kumar, I. R. Lemischka, B. Y. Zhou, and H.-W. Snoeck. 2011. "Prdm16 Is a Physiologic Regulator of Hematopoietic Stem Cells." *Blood* 117 (19): 5057–66. https://doi.org/10.1182/blood-2010-08-300145.

Akashi, Koichi, David Traver, Toshihiro Miyamoto, and Irving L. Weissman. 2000. "A Clonogenic Common Myeloid Progenitor That Gives Rise to All Myeloid Lineages." *Nature* 404 (6774): 193–97. https://doi.org/10.1038/35004599.

Alemany, Anna, Maria Florescu, Chloé S. Baron, Josi Peterson-Maduro, and Alexander van Oudenaarden. 2018. "Whole-Organism Clone Tracing Using Single-Cell Sequencing." *Nature* 556 (7699): 108–12. https://doi.org/10.1038/nature25969.

Alexander, Roger P., Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B. Gerstein. 2010. "Annotating Non-Coding Regions of the Genome." *Nature Reviews Genetics* 11 (8): 559–71. https://doi.org/10.1038/nrg2814.

Ali, Mohamed A. E., Kyoko Fuse, Yuko Tadokoro, Takayuki Hoshii, Masaya Ueno, Masahiko Kobayashi, Naho Nomura, et al. 2017. "Functional Dissection of Hematopoietic Stem Cell Populations with a Stemness-Monitoring System Based on NS-GFP Transgene Expression." *Scientific Reports* 7 (1): 11442. https://doi.org/10.1038/s41598-017-11909-3.

Alon, Uri. 2007. “Network Motifs: Theory and Experimental Approaches.” *Nature Reviews Genetics* 8 (6): 450–61. https://doi.org/10.1038/nrg2102.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69. https://doi.org/10.1093/bioinformatics/btu638.

Andrews, Nancy C., Hediye Erdjument-Bromage, Mark B. Davidson, Paul Tempst, and Stuart H. Orkin. 1993. “Erythroid Transcription Factor NF-E2 Is a Haematopoietic-Specific Basic–Leucine Zipper Protein.” *Nature* 362 (6422): 722–28. https://doi.org/10.1038/362722a0.

Angerer, Philipp, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. 2016. “*Destiny*: Diffusion Maps for Large-Scale Single-Cell Data in R.” *Bioinformatics* 32 (8): 1241–43. https://doi.org/10.1093/bioinformatics/btv715.

Angermueller, Christof, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, et al. 2016. “Parallel Single-Cell Sequencing Links Transcriptional and Epigenetic Heterogeneity.” *Nature Methods* 13 (3): 229–32. https://doi.org/10.1038/nmeth.3728.

Ariki, Reina, Satoru Morikawa, Yo Mabuchi, Sadafumi Suzuki, Mayuka Nakatake, Kentaro Yoshioka, Shinya Hidano, et al. 2014. “Homeodomain Transcription Factor Meis1 Is a Critical Regulator of Adult Bone Marrow Hematopoiesis.” Edited by Kevin D. Bunting. *PLoS ONE* 9 (2): e87646. https://doi.org/10.1371/journal.pone.0087646.

Arinobu, Yojiro, Shin-ichi Mizuno, Yong Chong, Hirokazu Shigematsu, Tadafumi Iino, Hiromi Iwasaki, Thomas Graf, et al. 2007. “Reciprocal Activation of GATA-1 and PU.1 Marks Initial Specification of Hematopoietic Stem Cells into Myeloerythroid and Myelolymphoid Lineages.” *Cell Stem Cell* 1 (4): 416–27. https://doi.org/10.1016/j.stem.2007.07.004.

Azcoitia, Valeria, Miguel Aracil, Carlos Martínez-A, and Miguel Torres. 2005. “The Homeodomain Protein Meis1 Is Essential for Definitive Hematopoiesis and Vascular Patterning in the Mouse Embryo.” *Developmental Biology* 280 (2): 307–20. https://doi.org/10.1016/j.ydbio.2005.01.004.

Balazs, Alejandro B, Attila J Fabian, Charles T Esmon, and Richard C Mulligan. 2006. “Endothelial Protein C Receptor (CD201) Explicitly Identifies Hematopoietic Stem Cells in

Murine Bone Marrow." *Blood* 107 (6): 2317–21. https://doi.org/10.1182/blood-2005-06-2249.

Barrangou, Rodolphe, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. 2007. "CRISPR Provides Acquired Resistance against Viruses in Prokaryotes." *Science* 315 (5819): 1709–12. https://doi.org/10.1126/science.1138140.

Beckmann, Julia, Sebastian Scheitza, Peter Wernet, Johannes C Fischer, and Bernd Giebel. 2007. "Asymmetric Cell Division within the Human Hematopoietic Stem and Progenitor Cell Compartment: Identification of Asymmetrically Segregating Proteins." *Blood* 109 (12): 5494–5501. https://doi.org/10.1182/blood-2006-11-055921.

Bedford, F K, A Ashworth, T Enver, and L M Wiedemann. 1993. "HEX: A Novel Homeobox Gene Expressed during Haematopoiesis and Conserved between Mouse and Human." *Nucleic Acids Research* 21 (5): 1245–49.

Beerman, Isabel, Deepta Bhattacharya, Sasan Zandi, Mikael Sigvardsson, Irving L Weissman, David Bryder, and Derrick J Rossi. 2010. "Functionally Distinct Hematopoietic Stem Cells Modulate Hematopoietic Lineage Potential during Aging by a Mechanism of Clonal Expansion." *Proceedings of the National Academy of Sciences of the United States of America* 107 (12): 5465–70. https://doi.org/10.1073/pnas.1000834107.

Begley, C G, P D Aplan, M P Davey, K Nakahara, K Tchorz, J Kurtzberg, M S Hershfield, B F Haynes, D I Cohen, and T A Waldmann. 1989. "Chromosomal Translocation in a Human Leukemic Stem-Cell Line Disrupts the T-Cell Antigen Receptor Delta-Chain Diversity Region and Results in a Previously Unreported Fusion Transcript." *Proceedings of the National Academy of Sciences of the United States of America* 86 (6): 2031–35.

Behbehani, Gregory K., Sean C. Bendall, Matthew R. Clutter, Wendy J. Fantl, and Garry P. Nolan. 2012. "Single-Cell Mass Cytometry Adapted to Measurements of the Cell Cycle." *Cytometry Part A* 81A (7): 552–66. https://doi.org/10.1002/cyto.a.22075.

Ben-David, Y, E B Giddens, K Letwin, and A Bernstein. 1991. "Erythroleukemia Induction by Friend Murine Leukemia Virus: Insertional Activation of a New Member of the Ets Gene Family, Fli-1, Closely Linked to c-Ets-1." *Genes & Development* 5 (6): 908–18.

Bendall, Sean C., Garry P. Nolan, Mario Roederer, and Pratip K. Chattopadhyay. 2012. “A Deep Profiler’s Guide to Cytometry.” *Trends in Immunology* 33 (7): 323–32. https://doi.org/10.1016/j.it.2012.02.010.

Bendall, Sean C, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe. 2014. “Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development.” *Cell* 157 (3): 714–25. https://doi.org/10.1016/j.cell.2014.04.005.

Bersenev, Alexey, Chao Wu, Joanna Balcerek, and Wei Tong. 2008. “Lnk Controls Mouse Hematopoietic Stem Cell Self-Renewal and Quiescence through Direct Interactions with JAK2.” *Journal of Clinical Investigation* 118 (8): 2832–44. https://doi.org/10.1172/JCI35808.

Bhat, N K, C B Thompson, T Lindsten, C H June, S Fujiwara, S Koizumi, R J Fisher, and T S Papas. 1990. “Reciprocal Expression of Human ETS1 and ETS2 Genes during T-Cell Activation: Regulatory Role for the Protooncogene ETS1.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (10): 3723–27. https://doi.org/10.1073/PNAS.87.10.3723.

Bloor, Adrian J.C., María-José Sánchez, Anthony R. Green, and Berthold Göttgens. 2002. “The Role of the Stem Cell Leukemia (*SCL*) Gene in Hematopoietic and Endothelial Lineage Specification.” *Journal of Hematotherapy & Stem Cell Research* 11 (2): 195–206. https://doi.org/10.1089/152581602753658402.

Bockamp, E O, F McLaughlin, A M Murrell, B Gottgens, L Robb, C G Begley, and A R Green. 1995. “Lineage-Restricted Regulation of the Murine SCL/TAL-1 Promoter.” *Blood* 86 (4): 1502–14.

Boehm, T, L Foroni, Y Kaneko, M F Perutz, and T H Rabbitts. 1991. “The Rhombotin Family of Cysteine-Rich LIM-Domain Oncogenes: Distinct Members Are Involved in T-Cell Translocations to Human Chromosomes 11p15 and 11p13.” *Proceedings of the National Academy of Sciences of the United States of America* 88 (10): 4367–71.

Bonzanni, N, A Garg, A Feenstra, J Schütte, S Kinston, D Miranda-Saavedra, J Heringa, I Xenarios, and B Göttgens. 2013. “Hard-Wired Heterogeneity in Blood Stem Cells Revealed Using a Dynamic Regulatory Network Model.” *Bioinformatics* 29: i80--i88.

Borman, Meredith A., Justin A. MacDonald, and Timothy A.J. Haystead. 2004. “Modulation of Smooth Muscle Contractility by CHASM, a Novel Member of the Smoothelin Family of Proteins.” *FEBS Letters* 573 (1–3): 207–13. https://doi.org/10.1016/j.febslet.2004.08.002.

Brennecke, Philip, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, et al. 2013. “Accounting for Technical Noise in Single-Cell RNA-Seq Experiments.” *Nature Methods* 10 (11): 1093–95. https://doi.org/10.1038/nmeth.2645.

Broad Institute. 2018. “SgRNA Designer: CRISPRko.” 2018. https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design.

Brummelman, J., F. De Paoli, A. Anselmo, P. Novellis, G. Veronesi, M. Alloisio, and E. Lugli. 2017. “P1.07-032 28-Color, 30 Parameter Flow Cytometry to Dissect the Complex Heterogeneity of Tumor Infiltrating T Cells in Lung Cancer.” *Journal of Thoracic Oncology* 12 (11): S2008. https://doi.org/10.1016/j.jtho.2017.09.950.

Bryder, David, Derrick J. Rossi, and Irving L. Weissman. 2006. “Hematopoietic Stem Cells: The Paradigmatic Tissue-Specific Stem Cell.” *The American Journal of Pathology* 169 (2): 338–46. https://doi.org/10.2353/ajpath.2006.060312.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” *Nature* 523 (7561): 486–90. https://doi.org/10.1038/nature14590.

Buenrostro, Jason D, M Ryan Corces, Caleb A Lareau, Beijing Wu, Alicia N Schep, Martin J Aryee, Ravindra Majeti, Howard Y Chang, and William J Greenleaf. 2018. “Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation.” *Cell* 173 (6): 1535-1548.e16. https://doi.org/10.1016/j.cell.2018.03.074.

Bulger, Michael, and Mark Groudine. 2011. “Functional and Mechanistic Diversity of Distal Transcription Enhancers.” *Cell* 144 (3): 327–39. https://doi.org/10.1016/j.cell.2011.01.024.

Burda, P, P Laslo, and T Stopka. 2010. “The Role of PU.1 and GATA-1 Transcription Factors during Normal and Leukemogenic Hematopoiesis.” *Leukemia* 24 (7): 1249–57. https://doi.org/10.1038/leu.2010.104.

Bushey, Ashley M., Elizabeth R. Dorman, and Victor G. Corces. 2008. “Chromatin Insulators: Regulatory Mechanisms and Epigenetic Inheritance.” *Molecular Cell* 32 (1): 1–9. https://doi.org/10.1016/j.molcel.2008.08.017.

Byrne, P V, L J Guilbert, and E R Stanley. 1981. “Distribution of Cells Bearing Receptors for a Colony-Stimulating Factor (CSF-1) in Murine Tissues.” *The Journal of Cell Biology* 91 (3 Pt 1): 848–53.

Cabezas-Wallscheid, Nina, Daniel Klimmeck, Jenny Hansson, Daniel B Lipka, Alejandro Reyes, Qi Wang, Dieter Weichenhan, et al. 2014. “Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis.” *Cell Stem Cell* 15 (4): 507–22. https://doi.org/10.1016/j.stem.2014.07.005.

Capron, Claude, Yann Lécluse, Anna Lila Kaushik, Adlen Foudi, Catherine Lacout, Dalila Sekkai, Isabelle Godin, et al. 2006. “The SCL Relative LYL-1 Is Required for Fetal and Adult Hematopoietic Stem Cell Function and B-Cell Differentiation.” *Blood* 107 (12): 4678–86. https://doi.org/10.1182/blood-2005-08-3145.

Carroll, A J, W M Crist, R T Parmley, M Roper, M D Cooper, and W H Finley. 1984. “Pre-B Cell Leukemia Associated with Chromosome Translocation 1;19.” *Blood* 63 (3): 721–24.

Carroll, Dana. 2017. “Genome Editing: Past, Present, and Future.” *The Yale Journal of Biology and Medicine* 90 (4): 653–59.

Cavazzana-Calvo, Marina, Alain Fischer, Frederic D Bushman, Emmanuel Payen, Salima Hacein-Bey-Abina, and Philippe Leboulch. 2011. “Is Normal Hematopoiesis Maintained Solely by Long-Term Multipotent Stem Cells?” *Blood* 117 (17): 4420–24. https://doi.org/10.1182/blood-2010-09-255679.

Cedar, Howard, and Yehudit Bergman. 2009. “Linking DNA Methylation and Histone Modification: Patterns and Paradigms.” *Nature Reviews Genetics* 10 (5): 295–304. https://doi.org/10.1038/nrg2540.

Challen, Grant A., Nathan C. Boles, Stuart M. Chambers, and Margaret A. Goodell. 2010. “Distinct Hematopoietic Stem Cell Subtypes Are Differentially Regulated by TGF-B1.” *Cell Stem Cell* 6 (3): 265–78. https://doi.org/10.1016/j.stem.2010.02.002.

Challen, Grant A, Deqiang Sun, Mira Jeong, Min Luo, Jaroslav Jelinek, Jonathan S Berg, Christoph

Bock, et al. 2011. “Dnmt3a Is Essential for Hematopoietic Stem Cell Differentiation.” *Nature Genetics* 44 (1): 23–31. https://doi.org/10.1038/ng.1009.

Chan, Michelle, Zachary D Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas Norman, Britt Adamson, Marco Jost, et al. 2018. “Molecular Recording of Mammalian Embryogenesis.” *BioRxiv*, August, 384925. https://doi.org/10.1101/384925.

Chen, Huidong, Luca Albergante, Jonathan Y. Hsu, Caleb A. Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, et al. 2018. “STREAM: Single-Cell Trajectories Reconstruction, Exploration And Mapping of Omics Data.” *BioRxiv*, April. https://doi.org/10.1101/302554.

Chen, James Y., Masanori Miyanishi, Sean K. Wang, Satoshi Yamazaki, Rahul Sinha, Kevin S. Kao, Jun Seita, Debashis Sahoo, Hiromitsu Nakauchi, and Irving L. Weissman. 2016. “Hoxb5 Marks Long-Term Haematopoietic Stem Cells and Reveals a Homogenous Perivascular Niche.” *Nature* 530 (7589): 223–27. https://doi.org/10.1038/nature16943.

Cheow, Lih Feng, Elise T Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S W Tan, et al. 2016. “Single-Cell Multimodal Profiling Reveals Cellular Epigenetic Heterogeneity.” *Nature Methods* 13 (10): 833–36. https://doi.org/10.1038/nmeth.3961.

Chhikara, Nirmal, Mayank Saraswat, Anil Kumar Tomar, Sharmistha Dey, Sarman Singh, and Savita Yadav. 2012. “Human Epididymis Protein-4 (HE-4): A Novel Cross-Class Protease Inhibitor.” Edited by William R. Abrams. *PLoS ONE* 7 (11): e47672. https://doi.org/10.1371/journal.pone.0047672.

Chickarmane, Vijay, Tariq Enver, and Carsten Peterson. 2009. “Computational Modeling of the Hematopoietic Erythroid-Myeloid Switch Reveals Insights into Cooperativity, Priming, and Irreversibility.” Edited by Rustom Antia. *PLoS Computational Biology* 5 (1): e1000268. https://doi.org/10.1371/journal.pcbi.1000268.

Cho, S. W., S. Kim, Y. Kim, J. Kweon, H. S. Kim, S. Bae, and J.-S. Kim. 2014. “Analysis of Off-Target Effects of CRISPR/Cas-Derived RNA-Guided Endonucleases and Nickases.” *Genome Research* 24 (1): 132–41. https://doi.org/10.1101/gr.162339.113.

Choi, AHyun, Anuradha Illendula, John A Pulikkan, Justine E Roderick, Jessica Tesell, Jun Yu, Nicole Hermance, et al. 2017. “RUNX1 Is Required for Oncogenic Myb and Myc Enhancer

Activity in T-Cell Acute Lymphoblastic Leukemia." *Blood* 130 (15): 1722–33. https://doi.org/10.1182/blood-2017-03-775536.

Chua, Siang Ling, Wei Cun See Too, Boon Yin Khoo, and Ling Ling Few. 2011. "UBC and YWHAZ as Suitable Reference Genes for Accurate Normalisation of Gene Expression Using MCF7, HCT116 and HepG2 Cell Lines." *Cytotechnology* 63 (6): 645–54. https://doi.org/10.1007/s10616-011-9383-4.

Clark, Stephen J., Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, et al. 2018. "ScNMT-Seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and Transcription in Single Cells." *Nature Communications* 9 (1): 781. https://doi.org/10.1038/s41467-018-03149-4.

Coifman, R R, S Lafon, a B Lee, M Maggioni, B Nadler, F Warner, and S W Zucker. 2005. "Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps." *Proceedings of the National Academy of Sciences of the United States of America* 102 (21): 7426–31. https://doi.org/10.1073/pnas.0500896102.

Collins, C T, and J L Hess. 2016. "Role of HOXA9 in Leukemia: Dysregulation, Cofactors and Essential Targets." *Oncogene* 35 (9): 1090–98. https://doi.org/10.1038/onc.2015.174.

Collombet, Samuel, Chris van Oevelen, Jose Luis Sardina Ortega, Wassim Abou-Jaoudé, Bruno Di Stefano, Morgane Thomas-Chollier, Thomas Graf, and Denis Thieffry. 2017. "Logical Modeling of Lymphoid and Myeloid Cell Specification and Transdifferentiation." *Proceedings of the National Academy of Sciences* 114 (23): 5792–99. https://doi.org/10.1073/pnas.1610622114.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, et al. 2013. "Multiplex Genome Engineering Using CRISPR/Cas Systems." *Science* 339 (6121): 819–23. https://doi.org/10.1126/science.1231143.

Cooke, Michael P, Susan E Sutton, and Albert E Parker. 2010. "C-Myb Controls Hematopoietic Stem Cell Quiescence and Differentiation Via Regulation of P57Kip2." *Blood* 116 (21).

Corrigan, David J., Larry L. Luchsinger, Mariana Justino de Almeida, Linda J. Williams, Alexandros Strikoudis, and Hans-Willem Snoeck. 2018. "PRDM16 Isoforms Differentially Regulate Normal and Leukemic Hematopoiesis and Inflammatory Gene Signature." *The*

Journal of Clinical Investigation 128 (8): 3250–64. https://doi.org/10.1172/JCI99862.

Crooks, G M, J Fuller, D Petersen, P Izadi, P Malik, P K Pattengale, D B Kohn, and J C Gasson. 1999. "Constitutive HOXA5 Expression Inhibits Erythropoiesis and Increases Myelopoiesis from Human Hematopoietic Progenitors." *Blood* 94 (2): 519–28.

Cui, Kairong, Chongzhi Zang, Tae-Young Roh, Dustin E Schones, Richard W Childs, Weiqun Peng, and Keji Zhao. 2009. "Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation." *Cell Stem Cell* 4 (1): 80–93. https://doi.org/10.1016/j.stem.2008.11.011.

Curtis, David J, Mark A Hall, Leonie J Van Stekelenburg, Lorraine Robb, Stephen M Jane, and C Glenn Begley. 2004. "SCL Is Required for Normal Function of Short-Term Repopulating Hematopoietic Stem Cells." *Blood* 103 (9): 3342–48. https://doi.org/10.1182/blood-2003-09-3202.

Curtis, David J, and Matthew P McCormack. 2010. "The Molecular Basis of Lmo2-Induced T-Cell Acute Lymphoblastic Leukemia." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 16 (23): 5618–23. https://doi.org/10.1158/1078-0432.CCR-10-0440.

Cusanovich, D. A., R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure. 2015. "Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing." *Science* 348 (6237): 910–14. https://doi.org/10.1126/science.aab1601.

Dahlin, Joakim S., and Jenny Hallgren. 2015. "Mast Cell Progenitors: Origin, Development and Migration to Tissues." *Molecular Immunology* 63 (1): 9–17. https://doi.org/10.1016/J.MOLIMM.2014.01.018.

Dahlin, Joakim S., Fiona K. Hamey, Blanca Pijuan-Sala, Mairi Shepherd, Winnie W. Y. Lau, Sonia Nestorowa, Caleb Weinreb, et al. 2018. "A Single-Cell Hematopoietic Landscape Resolves 8 Lineage Trajectories and Defects in Kit Mutant Mice." *Blood* 131 (21): e1–11. https://doi.org/10.1182/blood-2017-12-821413.

Datlinger, Paul, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock.

2017. “Pooled CRISPR Screening with Single-Cell Transcriptome Readout.” *Nature Methods* 14 (3): 297–301. https://doi.org/10.1038/nmeth.4177.

Davidson, Eric H. 2010. “Emerging Properties of Animal Gene Regulatory Networks.” *Nature* 468 (7326): 911–20. https://doi.org/10.1038/nature09645.

Davidson, Eric H. 2009. “Network Design Principles from the Sea Urchin Embryo.” *Current Opinion in Genetics & Development* 19 (6): 535–40. https://doi.org/10.1016/j.gde.2009.10.007.

Deneault, Eric, Sonia Cellot, Amélie Faubert, Jean-Philippe Laverdure, Mélanie Fréchette, Jalila Chagraoui, Nadine Mayotte, Martin Sauvageau, Stephen B. Ting, and Guy Sauvageau. 2009. “A Functional Screen to Identify Novel Effectors of Hematopoietic Stem Cell Activity.” *Cell* 137 (2): 369. https://doi.org/10.1016/J.CELL.2009.03.026.

Dexter, T M, T D Allen, D Scott, and N M Teich. 1979. “Isolation and Characterisation of a Bipotential Haematopoietic Cell Line.” *Nature* 277 (5696): 471–74.

Dey, Siddharth S, Lennart Kester, Bastiaan Spanjaard, Magda Bienko, and Alexander van Oudenaarden. 2015. “Integrated Genome and Transcriptome Sequencing of the Same Cell.” *Nature Biotechnology* 33 (3): 285–89. https://doi.org/10.1038/nbt.3129.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, et al. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” *Cell* 167 (7): 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

Doench, John G, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, et al. 2016. “Optimized SgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9.” *Nature Biotechnology* 34 (2): 184–91. https://doi.org/10.1038/nbt.3437.

Drapkin, Ronny, Hans Henning von Horsten, Yafang Lin, Samuel C. Mok, Christopher P. Crum, William R. Welch, and Jonathan L. Hecht. 2005. “Human Epididymis Protein 4 (HE4) Is a Secreted Glycoprotein That Is Overexpressed by Serous and Endometrioid Ovarian Carcinomas.” *Cancer Research* 65 (6): 2162–69. https://doi.org/10.1158/0008-5472.CAN-04-3924.

Dunn, S-J, G Martello, B Yordanov, S Emmott, and A G Smith. 2014. “Defining an Essential Transcription Factor Program for Naïve Pluripotency.” *Science (New York, N.Y.)* 344 (6188): 1156–60. https://doi.org/10.1126/science.1248882.

Dwyer, Daniel F, Nora A Barrett, K Frank Austen, and Immunological Genome Project Consortium. 2016. “Expression Profiling of Constitutive Mast Cells Reveals a Unique Identity within the Immune System.” *Nature Immunology* 17 (7): 878–87. https://doi.org/10.1038/ni.3445.

Dykstra, Brad, David Kent, Michelle Bowie, Lindsay McCaffrey, Melisa Hamilton, Kristin Lyons, Shang-Jung Lee, Ryan Brinkman, and Connie Eaves. 2007. “Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo.” *Cell Stem Cell* 1 (2): 218–29. https://doi.org/10.1016/j.stem.2007.05.015.

Dzierzak, E., and S. Philipsen. 2013. “Erythropoiesis: Development and Differentiation.” *Cold Spring Harbor Perspectives in Medicine* 3 (4): a011601. https://doi.org/10.1101/cshperspect.a011601.

Eaves, Connie J. 2015. “Hematopoietic Stem Cells: Concepts, Definitions, and the New Reality.” *Blood* 125 (17): 2605–13. https://doi.org/10.1182/blood-2014-12-570200.

Edling, Charlotte E., and Bengt Hallberg. 2007. “C-Kit—A Hematopoietic Cell Essential Receptor Tyrosine Kinase.” *The International Journal of Biochemistry & Cell Biology* 39 (11): 1995–98. https://doi.org/10.1016/J.BIOCEL.2006.12.005.

El-Brolosy, Mohamed A., and Didier Y. R. Stainier. 2017. “Genetic Compensation: A Phenomenon in Search of Mechanisms.” Edited by Cecilia Moens. *PLOS Genetics* 13 (7): e1006780. https://doi.org/10.1371/journal.pgen.1006780.

Emambokus, Nikla R, and Jonathan Frampton. 2003. “The Glycoprotein IIb Molecule Is Expressed on Early Murine Hematopoietic Progenitors and Regulates Their Numbers in Sites of Hematopoiesis.” *Immunity* 19 (1): 33–45.

Etzrodt, Martin, and Timm Schroeder. 2017. “Illuminating Stem Cell Transcription Factor Dynamics: Long-Term Single-Cell Imaging of Fluorescent Protein Fusions.” *Current Opinion in Cell Biology* 49 (December): 77–83. https://doi.org/10.1016/J.CEB.2017.12.006.

Evans, T, and G Felsenfeld. 1989. “The Erythroid-Specific Transcription Factor Eryf1: A New

Finger Protein." *Cell* 58 (5): 877–85.

Fan, H. C., G. K. Fu, and S. P. A. Fodor. 2015. "Combinatorial Labeling of Single Cells for Gene Expression Cytometry." *Science* 347 (6222): 1258367–1258367. https://doi.org/10.1126/science.1258367.

Federzoni, Elena A, Peter J M Valk, Bruce E Torbett, Torsten Haferlach, Bob Löwenberg, Martin F Fey, and Mario P Tschan. 2012. "PU.1 Is Linking the Glycolytic Enzyme HK3 in Neutrophil Differentiation and Survival of APL Cells." *Blood* 119 (21): 4963–70. https://doi.org/10.1182/blood-2011-09-378117.

Felsenfeld, Gary, and Mark Groudine. 2003. "Controlling the Double Helix." *Nature* 421 (6921): 448–53. https://doi.org/10.1038/nature01411.

Fitch, Michael J., Luisa Campagnolo, Frank Kuhnert, and Heidi Stuhlmann. 2004. "Egfl7, a Novel Epidermal Growth Factor-Domain Gene Expressed in Endothelial Cells." *Developmental Dynamics* 230 (2): 316–24. https://doi.org/10.1002/dvdy.20063.

Forsberg, E Camilla, Susan S Prohaska, Sol Katzman, Garrett C Heffner, Josh M Stuart, and Irving L Weissman. 2005. "Differential Expression of Novel Potential Regulators in Hematopoietic Stem Cells." *PLoS Genetics* 1 (3): e28. https://doi.org/10.1371/journal.pgen.0010028.

Frelin, Catherine, Robert Herrington, Salima Janmohamed, Mary Barbara, Gary Tran, Christopher J Paige, Patricia Benveniste, et al. 2013. "GATA-3 Regulates the Self-Renewal of Long-Term Hematopoietic Stem Cells." *Nature Immunology* 14 (10): 1037–44. https://doi.org/10.1038/ni.2692.

Friedman, Alan D. 2002. "Runx1, c-Myb, and C/EBPalpha Couple Differentiation to Proliferation or Growth Arrest during Hematopoiesis." *Journal of Cellular Biochemistry* 86 (4): 624–29. https://doi.org/10.1002/jcb.10271.

Fu, Lin, Huaping Fu, Qingyun Wu, Yifan Pang, Keman Xu, Lei Zhou, Jianlin Qiao, Xiaoyan Ke, Kailin Xu, and Jinlong Shi. 2017. "High Expression of ETS2 Predicts Poor Prognosis in Acute Myeloid Leukemia and May Guide Treatment Decisions." *Journal of Translational Medicine* 15 (1): 159. https://doi.org/10.1186/s12967-017-1260-2.

Fu, Yanfang, Jennifer A Foden, Cyd Khayter, Morgan L Maeder, Deepak Reyon, J Keith Joung, and Jeffry D Sander. 2013. "High-Frequency off-Target Mutagenesis Induced by CRISPR-

Cas Nucleases in Human Cells." *Nature Biotechnology* 31 (9): 822–26. https://doi.org/10.1038/nbt.2623.

Fujiwara, Tohru, Hsiang-Ying Lee, Rajendran Sanalkumar, and Emery H Bresnick. 2010. "Building Multifunctionality into a Complex Containing Master Regulators of Hematopoiesis." *Proceedings of the National Academy of Sciences of the United States of America* 107 (47): 20429–34. https://doi.org/10.1073/pnas.1007804107.

Fujiwara, Tohru, Henriette O'Geen, Sunduz Keles, Kimberly Blahnik, Amelia K Linnemann, Yoon-A Kang, Kyunghee Choi, Peggy J Farnham, and Emery H Bresnick. 2009. "Discovering Hematopoietic Mechanisms through Genome-Wide Analysis of GATA Factor Chromatin Occupancy." *Mol Cell* 36 (4): 667–81. https://doi.org/10.1016/j.molcel.2009.11.001.

Fuller, J F, J McAdara, Y Yaron, M Sakaguchi, J K Fraser, and J C Gasson. 1999. "Characterization of HOX Gene Expression during Myelopoiesis: Role of HOX A5 in Lineage Commitment and Maturation." *Blood* 93 (10): 3391–3400.

Gaj, Thomas, Charles A Gersbach, Carlos F Barbas, and III. 2013. "ZFN, TALEN, and CRISPR/Cas-Based Methods for Genome Engineering." *Trends in Biotechnology* 31 (7): 397–405. https://doi.org/10.1016/j.tibtech.2013.04.004.

Gautier, Emmanuel L, Tal Shay, Jennifer Miller, Melanie Greter, Claudia Jakubzick, Stoyan Ivanov, Julie Helft, et al. 2012. "Gene-Expression Profiles and Transcriptional Regulatory Pathways That Underlie the Identity and Diversity of Mouse Tissue Macrophages." *Nature Immunology* 13 (11): 1118–28. https://doi.org/10.1038/ni.2419.

Gazit, Roi, Pankaj K Mandal, Wataru Ebina, Ayal Ben-Zvi, César Nombela-Arrieta, Leslie E Silberstein, and Derrick J Rossi. 2014. "Fgd5 Identifies Hematopoietic Stem Cells in the Murine Bone Marrow." *The Journal of Experimental Medicine* 211 (7): 1315–31. https://doi.org/10.1084/jem.20130428.

Gekas, C., and T. Graf. 2013. "CD41 Expression Marks Myeloid-Biased Adult Hematopoietic Stem Cells and Increases with Age." *Blood* 121 (22): 4463–72. https://doi.org/10.1182/blood-2012-09-457929.

Georgopoulos, K, M Bigby, J H Wang, A Molnar, P Wu, S Winandy, and A Sharpe. 1994. "The Ikaros Gene Is Required for the Development of All Lymphoid Lineages." *Cell* 79 (1): 143–

56.

Gerrits, Alice, Yang Li, Bruno M. Tesson, Leonid V. Bystrykh, Ellen Weersing, Albertina Ausema, Bert Dontje, et al. 2009. “Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State.” Edited by Greg Gibson. *PLoS Genetics* 5 (10): e1000692. https://doi.org/10.1371/journal.pgen.1000692.

Ginsberg, M H, X Du, T E O’Toole, and J C Loftus. 1995. “Platelet Integrins.” *Thrombosis and Haemostasis* 74 (1): 352–59.

Goardon, Nicolas, Julie A Lambert, Patrick Rodriguez, Philippe Nissaire, Sabine Herblot, Pierre Thibault, Dominique Dumenil, John Strouboulis, Paul-Henri Romeo, and Trang Hoang. 2006. “ETO2 Coordinates Cellular Proliferation and Differentiation during Erythropoiesis.” *The EMBO Journal* 25 (2): 357–66. https://doi.org/10.1038/sj.emboj.7600934.

Goodell, M A, K Brose, G Paradis, A S Conner, and R C Mulligan. 1996. “Isolation and Functional Properties of Murine Hematopoietic Stem Cells That Are Replicating in Vivo.” *The Journal of Experimental Medicine* 183 (4): 1797–1806.

Görgens, André, Stefan Radtke, Michael Möllmann, Michael Cross, Jan Dürig, Peter A. Horn, and Bernd Giebel. 2013. “Revision of the Human Hematopoietic Tree: Granulocyte Subtypes Derive from Distinct Hematopoietic Lineages.” *Cell Reports* 3 (5): 1539–52. https://doi.org/10.1016/J.CELREP.2013.04.025.

Gottgens, B. 2015. “Regulatory Network Control of Blood Stem Cells.” *Blood* 125 (17): 2614–20. https://doi.org/10.1182/blood-2014-08-570226.

Gottgens, B, F McLaughlin, E-O Bockamp, J L Fordham, C G Begley, K Kosmopoulos, A G Elefanty, and A R Green. 1997. “Transcription of the SCL Gene in Erythroid and CD34 Positive Primitive Myeloid Cells Is Controlled by a Complex Network of Lineage-Restricted Chromatin-Dependent and Chromatin-Independent Regulatory Elements.” *Oncogene* 15 (20): 2419–28. https://doi.org/10.1038/sj.onc.1201426.

Goyama, Susumu, Go Yamamoto, Munetake Shimabe, Tomohiko Sato, Motoshi Ichikawa, Seishi Ogawa, Shigeru Chiba, and Mineo Kurokawa. 2008. “Evi-1 Is a Critical Regulator for Hematopoietic Stem Cells and Transformed Leukemic Cells.” *Cell Stem Cell* 3 (2): 207–20. https://doi.org/10.1016/j.stem.2008.06.002.

Griffiths, Jonathan A., Arianne C. Richard, Karsten Bach, Aaron T. L. Lun, and John C. Marioni. 2018. “Detection and Removal of Barcode Swapping in Single-Cell RNA-Seq Data.” *Nature Communications* 9 (1): 2667. https://doi.org/10.1038/s41467-018-05083-x.

Grover, Amit, Alejandra Sanjuan-Pla, Supat Thongjuea, Joana Carrelha, Alice Giustacchini, Adriana Gambardella, Iain Macaulay, et al. 2016. “Single-Cell RNA Sequencing Reveals Molecular and Functional Platelet Bias of Aged Haematopoietic Stem Cells.” *Nature Communications* 7 (January): 11075. https://doi.org/10.1038/ncomms11075.

Grün, Dominic, Mauro J. Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, et al. 2016. “De Novo Prediction of Stem Cell Identity Using Single-Cell Transcriptome Data.” *Cell Stem Cell* 19 (2): 266–77. https://doi.org/10.1016/j.stem.2016.05.010.

Guilbert, L J, and E R Stanley. 1980. “Specific Interaction of Murine Colony-Stimulating Factor with Mononuclear Phagocytic Cells.” *The Journal of Cell Biology* 85 (1): 153–59.

Guo, Guoji, Mikael Huss, Guo Qing Tong, Chaoyang Wang, Li Li Sun, Neil D Clarke, Paul Robson, et al. 2010. “Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst.” *Developmental Cell* 18 (4): 675–85. https://doi.org/10.1016/j.devcel.2010.02.012.

Guo, Guoji, Sidinh Luc, Eugenio Marco, Ta-Wei Lin, Cong Peng, Marc A Kerenyi, Semir Beyaz, et al. 2013. “Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire.” *Cell Stem Cell* 13 (4): 492–505. https://doi.org/10.1016/j.stem.2013.07.017.

Guo, Hongshan, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. 2013. “Single-Cell Methylome Landscapes of Mouse Embryonic Stem Cells and Early Embryos Analyzed Using Reduced Representation Bisulfite Sequencing.” *Genome Research* 23 (12): 2126–35. https://doi.org/10.1101/gr.161679.113.

Haan, Gerald de, Ellen Weersing, Bert Dontje, Ronald van Os, Leonid V Bystrykh, Edo Vellenga, and Geraldine Miller. 2003. “In Vitro Generation of Long-Term Repopulating Hematopoietic Stem Cells by Fibroblast Growth Factor-1.” *Developmental Cell* 4 (2): 241–51.

Haan, Gerald de, Erik Zwart, Bertien Dethmers-Ausema, Ronald van Os, Fiona Hamey, Berthold Göttgens, Leonid Bystrykh, and Seka Lazare. 2017. “Neogenin-1: A New Receptor Critical

for Hematopoietic Stem Cell Function." *Experimental Hematology* 53 (September): S47. https://doi.org/10.1016/J.EXPHEM.2017.06.065.

Haghverdi, Laleh, Florian Buettner, and Fabian J. Theis. 2015. "Diffusion Maps for High-Dimensional Single-Cell Analysis of Differentiation Data." *Bioinformatics* 31 (18): 2989–98. https://doi.org/10.1093/bioinformatics/btv325.

Haghverdi, Laleh, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. 2016. "Diffusion Pseudotime Robustly Reconstructs Lineage Branching." *Nature Methods* 13 (10): 845–48. https://doi.org/10.1038/nmeth.3971.

Hahne, Florian, Nolwenn LeMeur, Ryan R Brinkman, Byron Ellis, Perry Haaland, Deepayan Sarkar, Josef Spidlen, Errol Strain, and Robert Gentleman. 2009. "FlowCore: A Bioconductor Package for High Throughput Flow Cytometry." *BMC Bioinformatics* 10 (1): 106. https://doi.org/10.1186/1471-2105-10-106.

Halfon, S, J Ford, J Foster, L Dowling, L Lucian, M Sterling, Y Xu, et al. 1998. "Leukocystatin, a New Class II Cystatin Expressed Selectively by Hematopoietic Cells." *The Journal of Biological Chemistry* 273 (26): 16400–408.

Hall, M. A., D. J. Curtis, D. Metcalf, A. G. Elefanty, K. Sourris, L. Robb, J. R. Gothert, S. M. Jane, and C. G. Begley. 2003. "The Critical Regulator of Embryonic Hematopoiesis, SCL, Is Vital in the Adult for Megakaryopoiesis, Erythropoiesis, and Lineage Choice in CFU-S12." *Proceedings of the National Academy of Sciences* 100 (3): 992–97. https://doi.org/10.1073/pnas.0237324100.

Hamey, Fiona K, Sonia Nestorowa, Sarah J Kinston, David G Kent, Nicola K Wilson, and Berthold Göttgens. 2017. "Reconstructing Blood Stem Cell Regulatory Network Models from Single-Cell Molecular Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 114 (23): 5822–29. https://doi.org/10.1073/pnas.1610609114.

Hamey, Fiona K, Sonia Nestorowa, Nicola K Wilson, and Berthold Gottgens. 2016. "Advancing Haematopoietic Stem and Progenitor Cell Biology through Single Cell Profiling." *FEBS Letters* 590 (22): 4052–67. https://doi.org/10.1002/1873-3468.12231.

Hamlett, Isla, Julia Draper, John Strouboulis, Francisco Iborra, Catherine Porcher, and Paresh Vyas. 2008. "Characterization of Megakaryocyte GATA1-Interacting Proteins: The

Corepressor ETO2 and GATA1 Interact to Regulate Terminal Megakaryocyte Maturation." *Blood* 112 (7): 2738–49. https://doi.org/10.1182/blood-2008-03-146605.

Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, et al. 2018. "Mapping the Mouse Cell Atlas by Microwell-Seq." *Cell* 172 (5): 1091-1107.e17. https://doi.org/10.1016/j.cell.2018.02.001.

Harrison, D E, and R K Zhong. 1992. "The Same Exhaustible Multilineage Precursor Produces Both Myeloid and Lymphoid Cells as Early as 3-4 Weeks after Marrow Transplantation." *Proceedings of the National Academy of Sciences of the United States of America* 89 (21): 10134–38. https://doi.org/10.1073/PNAS.89.21.10134.

Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai. 2012. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification." *Cell Reports* 2 (3): 666–73. https://doi.org/10.1016/j.celrep.2012.08.003.

Hayashi, Tetsutaro, Norito Shibata, Ryo Okumura, Tomomi Kudome, Osamu Nishimura, Hiroshi Tarui, and Kiyokazu Agata. 2010. "Single-Cell Gene Profiling of Planarian Stem Cells Using Fluorescent Activated Cell Sorting and Its 'Index Sorting' Function for Stem Cell Research." *Development, Growth & Differentiation* 52 (1): 131–44. https://doi.org/10.1111/j.1440-169X.2009.01157.x.

Heckl, Dirk, Monika S Kowalczyk, David Yudovich, Roger Belizaire, Rishi V Puram, Marie E McConkey, Anne Thielke, Jon C Aster, Aviv Regev, and Benjamin L Ebert. 2014. "Generation of Mouse Models of Myeloid Malignancy with Combinatorial Genetic Lesions Using CRISPR-Cas9 Genome Editing." *Nature Biotechnology* 32 (9): 941–46. https://doi.org/10.1038/nbt.2951.

Herblot, Sabine, Ann-Muriel Steff, Patrice Hugo, Peter D. Aplan, and Trang Hoang. 2000. "SCL and LMO1 Alter Thymocyte Differentiation: Inhibition of E2A-HEB Function and Pre-Tα Chain Expression." *Nature Immunology* 1 (2): 138–44. https://doi.org/10.1038/77819.

Hingorani, Ravi, Jun Deng, Jeanne Elia, Catherine Mcintyre, and Dev Mittar. 2011. "Detection of Apoptosis Using the BD Annexin V FITC Assay on the BD FACSVerse™ System."

Hitzler, Johann K, Joseph Cheung, Yue Li, Stephen W Scherer, and Alvin Zipursky. 2003. "GATA1 Mutations in Transient Leukemia and Acute Megakaryoblastic Leukemia of Down

Syndrome." *Blood* 101 (11): 4301–4. https://doi.org/10.1182/blood-2003-01-0013.

Hock, Hanno, Melanie J. Hamblen, Heather M. Rooke, Jeffrey W. Schindler, Shireen Saleque, Yuko Fujiwara, and Stuart H. Orkin. 2004. "Gfi-1 Restricts Proliferation and Preserves Functional Integrity of Haematopoietic Stem Cells." *Nature* 431 (7011): 1002–7. https://doi.org/10.1038/nature02994.

Hock, Hanno, Melanie J Hamblen, Heather M Rooke, David Traver, Roderick T Bronson, Scott Cameron, and Stuart H Orkin. 2003. "Intrinsic Requirement for Zinc Finger Transcription Factor Gfi-1 in Neutrophil Differentiation." *Immunity* 18 (1): 109–20.

Hock, Hanno, Eliza Meade, Sarah Medeiros, Jeffrey W Schindler, Peter J M Valk, Yuko Fujiwara, and Stuart H Orkin. 2004. "Tel/Etv6 Is an Essential and Selective Regulator of Adult Hematopoietic Stem Cell Survival." *Genes & Development* 18 (19): 2336–41. https://doi.org/10.1101/gad.1239604.

Hoppe, Philipp S., Michael Schwarzfischer, Dirk Loeffler, Konstantinos D. Kokkaliaris, Oliver Hilsenbeck, Nadine Moritz, Max Endele, et al. 2016. "Early Myeloid Lineage Choice Is Not Initiated by Random PU.1 to GATA1 Protein Ratios." *Nature* 535 (7611): 299–302. https://doi.org/10.1038/nature18320.

Hosking, Brett M., S.-C. Mary Wang, Meredith Downes, Peter Koopman, and George E. O. Muscat. 2004. "The *VCAM-1* Gene That Encodes the Vascular Cell Adhesion Molecule Is a Target of the Sry-Related High Mobility Group Box Gene, *Sox18*." *Journal of Biological Chemistry* 279 (7): 5314–22. https://doi.org/10.1074/jbc.M308512200.

Hotelling, H. 1933a. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology* 24 (6): 417–41. https://doi.org/10.1037/h0071325.

———. 1933b. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology* 24 (7): 498–520. https://doi.org/10.1037/h0070888.

Hou, Qianqian, Fei Liao, Shouyue Zhang, Duyu Zhang, Yan Zhang, Xueyan Zhou, Xuyang Xia, et al. 2017. "Regulatory Network of GATA3 in Pediatric Acute Lymphoblastic Leukemia." *Oncotarget* 8 (22): 36040–53. https://doi.org/10.18632/oncotarget.16424.

Hou, Yu, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, et al. 2016. "Single-Cell Triple Omics Sequencing Reveals Genetic, Epigenetic and Transcriptomic

Heterogeneity in Hepatocellular Carcinomas." *Cell Research* 26 (3): 304–19. https://doi.org/10.1038/cr.2016.23.

Hsu, Patrick D, Eric S Lander, and Feng Zhang. 2014. "Development and Applications of CRISPR-Cas9 for Genome Engineering." *Cell* 157 (6): 1262–78. https://doi.org/10.1016/j.cell.2014.05.010.

Hsu, Yu-Lin, Shao-Fu Shi, Wan-Lin Wu, Ling-Jun Ho, and Jenn-Haung Lai. 2013. "Protective Roles of Interferon-Induced Protein with Tetratricopeptide Repeats 3 (IFIT3) in Dengue Virus Infection of Human Lung Epithelial Cells." Edited by Amy Lynn Adamson. *PLoS ONE* 8 (11): e79518. https://doi.org/10.1371/journal.pone.0079518.

Hu, Youjin, Kevin Huang, Qin An, Guizhen Du, Ganlu Hu, Jinfeng Xue, Xianmin Zhu, Cun-Yu Wang, Zhigang Xue, and Guoping Fan. 2016. "Simultaneous Profiling of Transcriptome and DNA Methylome from a Single Cell." *Genome Biology* 17 (1): 88. https://doi.org/10.1186/s13059-016-0950-z.

Ichikawa-Shindo, Yuka, Takayuki Sakurai, Akiko Kamiyoshi, Hisaka Kawate, Nobuyoshi Iinuma, Takahiro Yoshizawa, Teruhide Koyama, et al. 2008. "The GPCR Modulator Protein RAMP2 Is Essential for Angiogenesis and Vascular Integrity." *The Journal of Clinical Investigation* 118 (1): 29–39. https://doi.org/10.1172/JCI33022.

Ichikawa, H, K Shimizu, Y Hayashi, and M Ohki. 1994. "An RNA-Binding Protein Gene, TLS/FUS, Is Fused to ERG in Human Myeloid Leukemia with t(16;21) Chromosomal Translocation." *Cancer Research* 54 (11): 2865–68.

Ichikawa, Motoshi, Takashi Asai, Toshiki Saito, Go Yamamoto, Sachiko Seo, Ieharu Yamazaki, Tetsuya Yamagata, et al. 2004. "AML-1 Is Required for Megakaryocytic Maturation and Lymphocytic Differentiation, but Not for Maintenance of Hematopoietic Stem Cells in Adult Hematopoiesis." *Nature Medicine* 10 (3): 299–304. https://doi.org/10.1038/nm997.

Iwasaki, H., Chamorro Somoza, Hirokazu Shigematsu, Estelle A Duprez, Junko Iwasaki-Arai, Shin-Ichi Mizuno, Yojiro Arinobu, et al. 2005. "Distinctive and Indispensable Roles of PU.1 in Maintenance of Hematopoietic Stem Cells and Their Differentiation." *Blood* 106 (5): 1590–1600. https://doi.org/10.1182/blood-2005-03-0860.

Iwasaki, Hiroko, Fumio Arai, Yoshiaki Kubota, Maria Dahl, and Toshio Suda. 2010. "Endothelial

Protein C Receptor-Expressing Hematopoietic Stem Cells Reside in the Perisinusoidal Niche in Fetal Liver." *Blood* 116 (4): 544–53. https://doi.org/10.1182/blood-2009-08-240903.

Izon, D J, S Rozenfeld, S T Fong, L Kömüves, C Largman, and H J Lawrence. 1998. "Loss of Function of the Homeobox Gene Hoxa-9 Perturbs Early T-Cell Development and Induces Apoptosis in Primitive Thymocytes." *Blood* 92 (2): 383–93.

Jaitin, Diego Adhemar, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, et al. 2014. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types." *Science* 343 (6172): 776–79. https://doi.org/10.1126/science.1247651.

Jaitin, Diego Adhemar, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. 2016. "Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq." *Cell* 167 (7): 1883-1896.e15. https://doi.org/10.1016/j.cell.2016.11.039.

Jardin, F., J.-P. Jais, T.-J. Molina, F. Parmentier, J.-M. Picquenot, P. Ruminy, H. Tilly, et al. 2010. "Diffuse Large B-Cell Lymphomas with CDKN2A Deletion Have a Distinct Gene Expression Signature and a Poor Prognosis under R-CHOP Treatment: A GELA Study." *Blood* 116 (7): 1092–1104. https://doi.org/10.1182/blood-2009-10-247122.

Jin, Bilian, Yajun Li, and Keith D Robertson. 2011. "DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?" *Genes & Cancer* 2 (6): 607–17. https://doi.org/10.1177/1947601910393957.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* 337 (6096): 816–21. https://doi.org/10.1126/science.1225829.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods." *Biostatistics* 8 (1): 118–27. https://doi.org/10.1093/biostatistics/kxj037.

Jolliffe, Ian. 2011. "Principal Component Analysis." In *International Encyclopedia of Statistical Science*, 1094–96. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_455.

Joughin, Brian A., Edwin Cheung, R. Krishna Murthy Karuturi, and Julio Saez-Rodriguez. 2010. "Cellular Regulatory Networks." *Systems Biomedicine*, January, 57–108. https://doi.org/10.1016/B978-0-12-372550-9.00004-3.

Kamps, M P, and D Baltimore. 1993. "E2A-Pbx1, the t(1;19) Translocation Protein of Human Pre-B-Cell Acute Lymphocytic Leukemia, Causes Acute Myeloid Leukemia in Mice." *Molecular and Cellular Biology* 13 (1): 351–57.

Karsunky, Holger, Hui Zeng, Thorsten Schmidt, Branko Zevnik, Reinhart Kluge, Kurt Werner Schmid, Ulrich Dührsen, and Tarik Möröy. 2002. "Inflammatory Reactions and Severe Neutropenia in Mice Lacking the Transcriptional Repressor Gfi1." *Nature Genetics* 30 (3): 295–300. https://doi.org/10.1038/ng831.

Kaushansky, K. 1995. "Thrombopoietin: The Primary Regulator of Megakaryocyte and Platelet Production." *Thrombosis and Haemostasis* 74 (1): 521–25.

Kent, David G, Michael R Copley, Claudia Benz, Stefan Wöhrer, Brad J Dykstra, Elaine Ma, John Cheyne, et al. 2009. "Prospective Isolation and Molecular Characterization of Hematopoietic Stem Cells with Durable Self-Renewal Potential." *Blood* 113 (25): 6342–50. https://doi.org/10.1182/blood-2008-12-192054.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12 (6): 996–1006. https://doi.org/10.1101/gr.229102.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden. 2014. "Bayesian Approach to Single-Cell Differential Expression Analysis." *Nature Methods* 11 (7): 740–42. https://doi.org/10.1038/nmeth.2967.

Khoramian Tusi, Betsabeh, Samuel L Wolock, Caleb Weinreb, Yung Hwang, Daniel Hidalgo, Rapolas Zilionis, Ari Waisman, Jun Huh, Allon M Klein, and Merav Socolovsky. 2018. "Emergence of the Erythroid Lineage from Multipotent Hematopoiesis." *BioRxiv*. https://doi.org/10.1101/261941.

Kiel, Mark J, Glenn L Radice, and Sean J Morrison. 2007. "Lack of Evidence That Hematopoietic Stem Cells Depend on N-Cadherin-Mediated Adhesion to Osteoblasts for Their Maintenance." *Cell Stem Cell* 1 (2): 204–17. https://doi.org/10.1016/j.stem.2007.06.001.

Kiel, Mark J, Omer H Yilmaz, Toshihide Iwashita, Osman H Yilmaz, Cox Terhorst, and Sean J Morrison. 2005. “SLAM Family Receptors Distinguish Hematopoietic Stem and Progenitor Cells and Reveal Endothelial Niches for Stem Cells.” *Cell* 121 (7): 1109–21. https://doi.org/10.1016/j.cell.2005.05.026.

Kim, Shine Young, Sang-Hyun Hwang, Eun Joo Song, Ho Jin Shin, Joo Seop Jung, and Eun Yup Lee. 2010. “Level of *HOXA5* Hypermethylation in Acute Myeloid Leukemia Is Associated with Short-Term Outcome.” *The Korean Journal of Laboratory Medicine* 30 (5): 469. https://doi.org/10.3343/kjlm.2010.30.5.469.

Kimura, S, A W Roberts, D Metcalf, and W S Alexander. 1998. “Hematopoietic Stem Cell Deficiencies in Mice Lacking C-Mpl, the Receptor for Thrombopoietin.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (3): 1195–1200. https://doi.org/10.1073/PNAS.95.3.1195.

Kitamura, Yukihiko, Eiichi Morii, Tomoko Jippo, and Akihiko Ito. 2002. “Effect of MITF on Mast Cell Differentiation.” *Molecular Immunology* 38 (16–18): 1173–76. https://doi.org/10.1016/S0161-5890(02)00058-5.

Klein, A. M., and B. D. Simons. 2011. “Universal Patterns of Stem Cell Fate in Cycling Adult Tissues.” *Development* 138 (15): 3103–11. https://doi.org/10.1242/dev.060103.

Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 2015. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells.” *Cell* 161 (5): 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

Klemszs, Michael J, Richard A Makib, Thalia Papayannopoulouv, Jason Moorell, and Robert Hromas. 1993. “Characterization of the Ets Oncogene Family Member, Fli-L*.” *The Journal of Biological Chemistry* 268 (8): 5769–73.

Kocabas, F., J. Zheng, S. Thet, N. G. Copeland, N. A. Jenkins, R. J. DeBerardinis, C. Zhang, and H. A. Sadek. 2012. “Meis1 Regulates the Metabolic Phenotype and Oxidant Defense of Hematopoietic Stem Cells.” *Blood* 120 (25): 4963–72. https://doi.org/10.1182/blood-2012-05-432260.

Koike-Yusa, Hiroko, Yilong Li, E-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke

Yusa. 2014. “Genome-Wide Recessive Genetic Screening in Mammalian Cells with a Lentiviral CRISPR-Guide RNA Library.” *Nature Biotechnology* 32 (3): 267–73. https://doi.org/10.1038/nbt.2800.

Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. 2015. “The Technology and Biology of Single-Cell RNA Sequencing.” *Molecular Cell* 58 (4): 610–20. https://doi.org/10.1016/j.molcel.2015.04.005.

Komor, Martina, Saskia Güller, Claudia D. Baldus, Sven de Vos, Dieter Hoelzer, Oliver G. Ottmann, and Wolf-K. Hofmann. 2005. “Transcriptional Profiling of Human Hematopoiesis During In Vitro Lineage-Specific Differentiation.” *Stem Cells* 23 (8): 1154–69. https://doi.org/10.1634/stemcells.2004-0171.

Komorowska, Karolina, Alexander Doyle, Martin Wahlestedt, Agatheeswaran Subramaniam, Shubhranshu Debnath, Jun Chen, Shamit Soneji, et al. 2017. “Hepatic Leukemia Factor Maintains Quiescence of Hematopoietic Stem Cells and Protects the Stem Cell Pool during Regeneration.” *Cell Reports* 21 (12): 3514–23. https://doi.org/10.1016/j.celrep.2017.11.084.

Kondo, M, I L Weissman, and K Akashi. 1997. “Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow.” *Cell* 91 (5): 661–72.

Koni, P A, R Sacca, P Lawton, J L Browning, N H Ruddle, and R A Flavell. 1997. “Distinct Roles in Lymphoid Organogenesis for Lymphotoxins Alpha and Beta Revealed in Lymphotoxin Beta-Deficient Mice.” *Immunity* 6 (4): 491–500. https://doi.org/10.1016/S1074-7613(00)80292-7.

Koressaar, T., and M. Remm. 2007. “Enhancements and Modifications of Primer Design Program Primer3.” *Bioinformatics* 23 (10): 1289–91. https://doi.org/10.1093/bioinformatics/btm091.

Kornberg, RD, and Y Lorch. 1992. “Chromatin Structure and Transcription.” *Annual Review of Cell Biology* 8 (1): 563–87. https://doi.org/10.1146/annurev.cb.08.110192.003023.

———. 1999. “Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome.” *Cell* 98 (3): 285–94. https://doi.org/10.1016/S0092-8674(00)81958-3.

Kornblau, S. M., Y. H. Qiu, N. Zhang, N. Singh, S. Faderl, A. Ferrajoli, H. York, A. A. Qutub, K. R. Coombes, and D. K. Watson. 2011. “Abnormal Expression of FLI1 Protein Is an Adverse Prognostic Factor in Acute Myeloid Leukemia.” *Blood* 118 (20): 5604–12.

https://doi.org/10.1182/blood-2011-04-348052.

Kouzarides, Tony. 2007. “Chromatin Modifications and Their Function.” *Cell* 128 (4): 693–705. https://doi.org/10.1016/j.cell.2007.02.005.

Kovacic, Boris, Andrea Hoelbl-Kovacic, Katrin M Fischhuber, Nicole R Leitner, Dagmar Gotthardt, Emilio Casanova, Veronika Sexl, and Mathias Müller. 2014. “Lactotransferrin-Cre Reporter Mice Trace Neutrophils, Monocytes/Macrophages and Distinct Subtypes of Dendritic Cells.” *Haematologica* 99 (6): 1006–15. https://doi.org/10.3324/haematol.2013.097154.

Kowalczyk, Monika S, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. 2015. “Single Cell RNA-Seq Reveals Changes in Cell Cycle and Differentiation Programs upon Aging of Hematopoietic Stem Cells.” *Genome Research* 25 (12): 1860–72. https://doi.org/10.1101/gr.192237.115.

Krumsiek, Jan, Carsten Marr, Timm Schroeder, and Fabian J. Theis. 2011. “Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network.” Edited by Maurizio Pesce. *PLoS ONE* 6 (8): e22649. https://doi.org/10.1371/journal.pone.0022649.

Ku, C.-J., T. Hosoya, I. Maillard, and J. D. Engel. 2012. “GATA-3 Regulates Hematopoietic Stem Cell Maintenance and Cell-Cycle Entry.” *Blood* 119 (10): 2242–51. https://doi.org/10.1182/blood-2011-07-366070.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Landry, Joseph W, Subhadra Banerjee, Barbara Taylor, Peter D Aplan, Alfred Singer, and Carl Wu. 2011. “Chromatin Remodeling Complex NURF Regulates Thymocyte Maturation.” *Genes & Development* 25 (3): 275–86. https://doi.org/10.1101/gad.2007311.

Laurenti, Elisa, and Berthold Göttgens. 2018. “From Haematopoietic Stem Cells to Complex Differentiation Landscapes.” *Nature* 553 (7689): 418–26. https://doi.org/10.1038/nature25022.

Laurenti, Elisa, Barbara Varnum-Finney, Anne Wilson, Isabel Ferrero, William E Blanco-Bose, Armin Ehninger, Paul S Knoepfler, et al. 2008. “Hematopoietic Stem Cell Function and Survival Depend on C-Myc and N-Myc Activity.” *Cell Stem Cell* 3 (6): 611–24. https://doi.org/10.1016/j.stem.2008.09.005.

Lawrence, H J, C D Helgason, G Sauvageau, S Fong, D J Izon, R K Humphries, and C Largman. 1997. “Mice Bearing a Targeted Interruption of the Homeobox Gene HOXA9 Have Defects in Myeloid, Erythroid, and Lymphoid Hematopoiesis.” *Blood* 89 (6): 1922–30.

Lawrence, H J, S Rozenfeld, C Cruz, K Matsukuma, A Kwong, L Kömüves, A M Buchberg, and C Largman. 1999. “Frequent Co-Expression of the HOXA9 and MEIS1 Homeobox Genes in Human Myeloid Leukemias.” *Leukemia* 13 (12): 1993–99.

Layer, Justin H., Catherine E. Alford, W. Hayes McDonald, and Utpal P. Davé. 2016. “LMO2 Oncoprotein Stability in T-Cell Leukemia Requires Direct LDB1 Binding.” *Molecular and Cellular Biology* 36 (3): 488–506. https://doi.org/10.1128/MCB.00901-15.

Lee, Tong Ihn, and Richard A. Young. 2000. “Transcription of Eukaryotic Protein-Coding Genes.” *Annual Review of Genetics* 34 (1): 77–137. https://doi.org/10.1146/annurev.genet.34.1.77.

Leonard, M, M Brice, J D Engel, and T Papayannopoulou. 1993. “Dynamics of GATA Transcription Factor Expression during Erythroid Differentiation.” *Blood* 82 (4): 1071–79.

Lerner, Adam, and Paul M Epstein. 2006. “Cyclic Nucleotide Phosphodiesterases as Targets for Treatment of Haematological Malignancies.” *The Biochemical Journal* 393 (Pt 1): 21–41. https://doi.org/10.1042/BJ20051368.

Lever, Jake, Martin Krzywinski, and Naomi Altman. 2017. “Points of Significance: Principal Component Analysis.” *Nature Methods* 14 (7): 641–42. https://doi.org/10.1038/nmeth.4346.

Levine, Jacob H., Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, et al. 2015. “Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells That Correlate with Prognosis.” *Cell* 162 (1): 184–97. https://doi.org/10.1016/j.cell.2015.05.047.

Levine, M., and E. H. Davidson. 2005. “Gene Regulatory Networks for Development.” *Proceedings of the National Academy of Sciences* 102 (14): 4936–42. https://doi.org/10.1073/pnas.0408031102.

Ley, Timothy J., Li Ding, Matthew J. Walter, Michael D. McLellan, Tamara Lamprecht, David E. Larson, Cyriac Kandoth, et al. 2010. "*DNMT3A* Mutations in Acute Myeloid Leukemia." *New England Journal of Medicine* 363 (25): 2424–33. https://doi.org/10.1056/NEJMoa1005143.

Li, L, Jan Y. Lee, Jennifer Gross, Sang-Hyun Song, Ann Dean, and Paul E. Love. 2010. "A Requirement for Lim Domain Binding Protein 1 in Erythropoiesis." *The Journal of Experimental Medicine* 207 (12): 2543–50. https://doi.org/10.1084/jem.20100504.

Li, Qiliang, Kenneth R Peterson, Xiangdong Fang, and George Stamatoyannopoulos. 2002. "Locus Control Regions." *Blood* 100 (9): 3077–86. https://doi.org/10.1182/blood-2002-04-1104.

Li, Wei Vivian, and Jingyi Jessica Li. 2018. "An Accurate and Robust Imputation Method ScImpute for Single-Cell RNA-Seq Data." *Nature Communications* 9 (1): 997. https://doi.org/10.1038/s41467-018-03405-7.

Li, Xuan, and Wolf-Dietrich Heyer. 2008. "Homologous Recombination in DNA Repair and DNA Damage Tolerance." *Cell Research* 18 (1): 99–113. https://doi.org/10.1038/cr.2008.1.

Liang, Jingjing, Jun Lyu, Meng Zhao, Dan Li, Mingzhu Zheng, Yan Fang, Fangzhu Zhao, et al. 2017. "Tespa1 Regulates T Cell Receptor-Induced Calcium Signals by Recruiting Inositol 1,4,5-Trisphosphate Receptors." *Nature Communications* 8 (June): 15732. https://doi.org/10.1038/ncomms15732.

Lieu, Y. K., A. Kumar, A. G. Pajerowski, T. J. Rogers, and E. P. Reddy. 2004. "Requirement of C-Myb in T Cell Development and in Mature T Cell Function." *Proceedings of the National Academy of Sciences* 101 (41): 14853–58. https://doi.org/10.1073/pnas.0405338101.

Lieu, Y. K., and E. P. Reddy. 2009. "Conditional C-Myb Knockout in Adult Hematopoietic Stem Cells Leads to Loss of Self-Renewal Due to Impaired Proliferation and Accelerated Differentiation." *Proceedings of the National Academy of Sciences* 106 (51): 21689–94. https://doi.org/10.1073/pnas.0907623106.

Lim, Kim-Chew, Tomonori Hosoya, William Brandt, Chia-Jui Ku, Sakie Hosoya-Ohmura, Sally A Camper, Masayuki Yamamoto, and James Douglas Engel. 2012. "Conditional Gata2 Inactivation Results in HSC Loss and Lymphatic Mispatterning." *J Clin Invest* 122 (10): 3705–17. https://doi.org/10.1172/JCI61619.

Lim, Kwang-il. 2015. "Recent Advances in Developing Molecular Tools for Targeted Genome

Engineering of Mammalian Cells." *BMB Reports* 48 (1): 6–12. https://doi.org/10.5483/BMBREP.2015.48.1.165.

Lindström, Sara. 2012. "Flow Cytometry and Microscopy as Means of Studying Single Cells: A Short Introductional Overview." *Methods in Molecular Biology (Clifton, N.J.)* 853 (January): 13–15. https://doi.org/10.1007/978-1-61779-567-1_2.

Liu, Na, Jingru Zhang, and Chunyan Ji. 2013. "The Emerging Roles of Notch Signaling in Leukemia and Stem Cells." *Biomarker Research* 1 (1): 23. https://doi.org/10.1186/2050-7771-1-23.

Loughran, Stephen J, Elizabeth A Kruse, Douglas F Hacking, Carolyn A de Graaf, Craig D Hyland, Tracy A Willson, Katya J Henley, et al. 2008. "The Transcription Factor Erg Is Essential for Definitive Hematopoiesis and the Function of Adult Hematopoietic Stem Cells." *Nature Immunology* 9 (7): 810–19. https://doi.org/10.1038/ni.1617.

Luc, S., K. Anderson, S. Kharazi, N. Buza-Vidas, C. Boiers, C. T. Jensen, Z. Ma, L. Wittmann, and S. E. W. Jacobsen. 2008. "Down-Regulation of Mpl Marks the Transition to Lymphoid-Primed Multipotent Progenitors with Gradual Loss of Granulocyte-Monocyte Potential." *Blood* 111 (7): 3424–34. https://doi.org/10.1182/blood-2007-08-108324.

Luesink, Maaike, Iris H I M Hollink, Vincent H J van der Velden, Ruth H J N Knops, Jan B M Boezeman, Valérie de Haas, Jan Trka, et al. 2012. "High GATA2 Expression Is a Poor Prognostic Marker in Pediatric Acute Myeloid Leukemia." *Blood* 120 (10): 2064–75. https://doi.org/10.1182/blood-2011-12-397083.

Lun, Aaron T. L., Karsten Bach, and John C. Marioni. 2016. "Pooling across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts." *Genome Biology* 17 (1): 75. https://doi.org/10.1186/s13059-016-0947-7.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

Macaulay, Iain C, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, et al. 2015. "G&T-Seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes." *Nature Methods* 12 (6): 519–22. https://doi.org/10.1038/nmeth.3370.

Macaulay, Iain C, Mabel J Teng, Wilfried Haerty, Parveen Kumar, Chris P Ponting, and Thierry

Voet. 2016. “Separation and Parallel Sequencing of the Genomes and Transcriptomes of Single Cells Using G&T-Seq.” *Nature Protocols* 11 (11): 2081–2103. https://doi.org/10.1038/nprot.2016.138.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Makarova, Kira S, Nick V Grishin, Svetlana A Shabalina, Yuri I Wolf, and Eugene V Koonin. 2006. “A Putative RNA-Interference-Based Immune System in Prokaryotes: Computational Analysis of the Predicted Enzymatic Machinery, Functional Analogies with Eukaryotic RNAi, and Hypothetical Mechanisms of Action.” *Biology Direct* 1 (1): 7. https://doi.org/10.1186/1745-6150-1-7.

Malaise, Muriel, Daniel Steinbach, and Selim Corbacioglu. 2009. “Clinical Implications of C-Kit Mutations in Acute Myelogenous Leukemia.” *Current Hematologic Malignancy Reports* 4 (2): 77–82. https://doi.org/10.1007/s11899-009-0011-8.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. 2013. “RNA-Guided Human Genome Engineering via Cas9.” *Science* 339 (6121): 823–26. https://doi.org/10.1126/science.1232033.

Malik, Sohail, and Robert G. Roeder. 2010. “The Metazoan Mediator Co-Activator Complex as an Integrative Hub for Transcriptional Regulation.” *Nature Reviews Genetics* 11 (11): 761–72. https://doi.org/10.1038/nrg2901.

Mandal, Pankaj K., Leonardo M.R. Ferreira, Ryan Collins, Torsten B. Meissner, Christian L. Boutwell, Max Friesen, Vladimir Vrbanac, et al. 2014. “Efficient Ablation of Genes in Human Hematopoietic Stem and Effector Cells Using CRISPR/Cas9.” *Cell Stem Cell* 15 (5): 643–52. https://doi.org/10.1016/j.stem.2014.10.004.

Maniatis, T, S Goodbourn, and J A Fischer. 1987. “Regulation of Inducible and Tissue-Specific Gene Expression.” *Science* 236 (4806): 1237–45.

Manno, Gioele La, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. “RNA Velocity of Single Cells.” *Nature* 560

(7719): 494–98. https://doi.org/10.1038/s41586-018-0414-6.

Mao, Qi, Li Wang, Ivor W. Tsang, and Yijun Sun. 2017. “Principal Graph and Structure Learning Based on Reversed Graph Embedding.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (11): 2227–41. https://doi.org/10.1109/TPAMI.2016.2635657.

Marke, René, Frank N van Leeuwen, and Blanca Scheijen. 2018. “The Many Faces of IKZF1 in B-Cell Precursor Acute Lymphoblastic Leukemia.” *Haematologica* 103 (4): 565–74. https://doi.org/10.3324/haematol.2017.185603.

Martin, David I. K., Leonard I. Zon, George Mutter, and Stuart H. Orkin. 1990. “Expression of an Erythroid Transcription Factor in Megakaryocytic and Mast Cell Lineages.” *Nature* 344 (6265): 444–47. https://doi.org/10.1038/344444a0.

Maslah, N, B Cassinat, E Verger, J-J Kiladjian, and L Velazquez. 2017. “The Role of LNK/SH2B3 Genetic Alterations in Myeloproliferative Neoplasms and Other Hematological Disorders.” *Leukemia* 31 (8): 1661–70. https://doi.org/10.1038/leu.2017.139.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. “Transcriptional Regulatory Elements in the Human Genome.” *Annual Review of Genomics and Human Genetics* 7 (1): 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623.

Matsumoto, Akinobu, Shoichiro Takeishi, Tomoharu Kanie, Etsuo Susaki, Ichiro Onoyama, Yuki Tateishi, Keiko Nakayama, and Keiichi I Nakayama. 2011. “P57 Is Required for Quiescence and Maintenance of Adult Hematopoietic Stem Cells.” *Cell Stem Cell* 9 (3): 262–71. https://doi.org/10.1016/j.stem.2011.06.014.

McKercher, S R, B E Torbett, K L Anderson, G W Henkel, D J Vestal, H Baribault, M Klemsz, et al. 1996. “Targeted Disruption of the PU.1 Gene Results in Multiple Hematopoietic Abnormalities.” *The EMBO Journal* 15 (20): 5647–58.

Melet, Fabrice, Benny Motro, Derrick J Rossi, Liqun Zhang, and Alan Bernstein. 1996. “Generation of a Novel Fli-1 Protein by Gene Targeting Leads to a Defect in Thymus Development and a Delay in Friend Virus-Induced Erythroleukemia.” *Molecular and Cellular Biology*. Vol. 16.

Meng, Y-S, H Khoury, J E Dick, and M D Minden. 2005. “Oncogenic Potential of the Transcription Factor LYL1 in Acute Myeloblastic Leukemia.” *Leukemia* 19 (11): 1941–47.

https://doi.org/10.1038/sj.leu.2403836.

Miller, Cindy L, and Becky Lai. 2005. “Human and Mouse Hematopoietic Colony-Forming Cell Assays.” *Methods in Molecular Biology (Clifton, N.J.)* 290: 71–89.

Miller, Michelle Erin, Patty Rosten, Madeleine E Lemieux, Courteney Lai, and R Keith Humphries. 2016. “Meis1 Is Required for Adult Mouse Erythropoiesis, Megakaryopoiesis and Hematopoietic Stem Cell Expansion.” *PloS One* 11 (3): e0151584. https://doi.org/10.1371/journal.pone.0151584.

Miranda-Saavedra, Diego, and Berthold Göttgens. 2008. “Transcriptional Regulatory Networks in Haematopoiesis.” *Current Opinion in Genetics & Development* 18 (6): 530–35. https://doi.org/10.1016/j.gde.2008.09.001.

Miyoshi, H, K Shimizu, T Kozu, N Maseki, Y Kaneko, and M Ohki. 1991. “T(8;21) Breakpoints on Chromosome 21 in Acute Myeloid Leukemia Are Clustered within a Limited Region of a Single Gene, AML1.” *Proceedings of the National Academy of Sciences of the United States of America* 88 (23): 10431–34.

Moignard, Victoria, and Berthold Göttgens. 2014. “Transcriptional Mechanisms of Cell Fate Decisions Revealed by Single Cell Expression Profiling.” *BioEssays* 36 (4): 419–26. https://doi.org/10.1002/bies.201300102.

Moignard, Victoria, Iain C Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J Calero-Nieto, Sarah Kinston, et al. 2013. “Characterization of Transcriptional Networks in Blood Stem and Progenitor Cells Using High-Throughput Single-Cell Gene Expression Analysis.” *Nature Cell Biology* 15 (4): 363–72. https://doi.org/10.1038/ncb2709.

Moignard, Victoria, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, et al. 2015. “Decoding the Regulatory Network of Early Blood Development from Single-Cell Gene Expression Measurements.” *Nature Biotechnology* 33 (3): 269–76. https://doi.org/10.1038/nbt.3154.

Morishita, K, E Parganas, C L William, M H Whittaker, H Drabkin, J Oval, R Taetle, M B Valentine, and J N Ihle. 1992. “Activation of EVI1 Gene Expression in Human Acute Myelogenous Leukemias by Translocations Spanning 300-400 Kilobases on Chromosome Band 3q26.” *Proceedings of the National Academy of Sciences of the United States of America*

89 (9): 3937–41. https://doi.org/10.1073/PNAS.89.9.3937.

Morita, Yohei, Hideo Ema, and Hiromitsu Nakauchi. 2010. “Heterogeneity and Hierarchy within the Most Primitive Hematopoietic Stem Cell Compartment.” *The Journal of Experimental Medicine* 207 (6): 1173–82. https://doi.org/10.1084/jem.20091318.

Morrison, Sean J, Nobuko Uchida, and Irving L. Weissman. 1995. “The Biology of Hematopoietic Stem Cells.” *Annual Review of Cell and Developmental Biology* 11 (1): 35–71. https://doi.org/10.1146/annurev.cb.11.110195.000343.

Morrison, Sean J, A.M. Wandycz, H.D. Hemmati, D.E. Wright, and I.L. Weissman. 1997. “Identification of a Lineage of Multipotent Hematopoietic Progenitors.” *Development* 124 (10).

Morrison, Sean J, and I L Weissman. 1994. “The Long-Term Repopulating Subset of Hematopoietic Stem Cells Is Deterministic and Isolatable by Phenotype.” *Immunity* 1 (8): 661–73.

Mossadegh-Keller, Noushine, Sandrine Sarrazin, Prashanth K. Kandalla, Leon Espinosa, E. Richard Stanley, Stephen L. Nutt, Jordan Moore, and Michael H. Sieweke. 2013. “M-CSF Instructs Myeloid Lineage Fate in Single Haematopoietic Stem Cells.” *Nature* 497 (7448): 239–43. https://doi.org/10.1038/nature12026.

Muller-Sieburg, Christa E, Rebecca H Cho, Lars Karlsson, Jing-F Huang, and Hans B Sieburg. 2004. “Myeloid-Biased Hematopoietic Stem Cells Have Extensive Self-Renewal Capacity but Generate Diminished Lymphoid Progeny with Impaired IL-7 Responsiveness.” *Blood* 103 (11): 4111–18. https://doi.org/10.1182/blood-2003-10-3448.

Muller-Sieburg, Christa E, Rebecca H Cho, Marilyn Thoman, Becky Adkins, and Hans B Sieburg. 2002. “Deterministic Regulation of Hematopoietic Stem Cell Self-Renewal and Differentiation.” *Blood* 100 (4): 1302–9.

Mullighan, Charles G., Xiaoping Su, Jinghui Zhang, Ina Radtke, Letha A.A. Phillips, Christopher B. Miller, Jing Ma, et al. 2009. “Deletion of *IKZF1* and Prognosis in Acute Lymphoblastic Leukemia.” *New England Journal of Medicine* 360 (5): 470–80. https://doi.org/10.1056/NEJMoa0808253.

Murre, Cornelis. 2007. “Defining the Pathways of Early Adult Hematopoiesis.” *Cell Stem Cell* 1

(4): 357–58. https://doi.org/10.1016/j.stem.2007.09.008.

Nagano, Takashi, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. 2013. "Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure." *Nature* 502 (7469): 59–64. https://doi.org/10.1038/nature12593.

Naik, Shalin H., Ton N. Schumacher, and Leïla Perié. 2014. "Cellular Barcoding: A Technical Appraisal." *Experimental Hematology* 42 (8): 598–608. https://doi.org/10.1016/j.exphem.2014.05.003.

Nakajima, Hideaki, and Hiroyoshi Kunimoto. 2014. "TET2 as an Epigenetic Master Regulator for Normal and Malignant Hematopoiesis." *Cancer Science* 105 (9): 1093–99. https://doi.org/10.1111/cas.12484.

Narula, Jatin, Aileen M. Smith, Berthold Gottgens, and Oleg A. Igoshin. 2010. "Modeling Reveals Bistability and Low-Pass Filtering in the Network Module Determining Blood Stem Cell Fate." Edited by Anand R. Asthagiri. *PLoS Computational Biology* 6 (5): e1000771. https://doi.org/10.1371/journal.pcbi.1000771.

Nestorowa, Sonia, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. 2016. "A Single Cell Resolution Map of Mouse Haematopoietic Stem and Progenitor Cell Differentiation." *Blood*. https://doi.org/10.1182/blood-2016-05-716480.

Nettey, Leonard, Amber J. Giles, and Pratip K. Chattopadhyay. 2018. "OMIP-050: A 28-Color/30-Parameter Fluorescence Flow Cytometry Panel to Enumerate and Characterize Cells Expressing a Wide Array of Immune Checkpoint Molecules." *Cytometry Part A* 93 (11): 1094–96. https://doi.org/10.1002/cyto.a.23608.

Ney, P A, N C Andrews, S M Jane, B Safer, M E Purucker, S Weremowicz, C C Morton, S C Goff, S H Orkin, and A W Nienhuis. 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): 5604–12.

North, T. E., Terryl Stacy, Christina J Matheny, Nancy A Speck, and Marella F T R de Bruijn. 2004. "Runx1 Is Expressed in Adult Mouse Hematopoietic Stem Cells and Differentiating

Myeloid and Lymphoid Cells, But Not in Maturing Erythroid Cells." *Stem Cells* 22 (2): 158–68. https://doi.org/10.1634/stemcells.22-2-158.

Notta, Faiyaz, Sasan Zandi, Naoya Takayama, Stephanie Dobson, Olga I Gan, Gavin Wilson, Kerstin B Kaufmann, et al. 2015. "Distinct Routes of Lineage Development Reshape the Human Blood Hierarchy across Ontogeny." *Science (New York, N.Y.)* 351 (6269): aab2116. https://doi.org/10.1126/science.aab2116.

Nowak, Martin A., Maarten C. Boerlijst, Jonathan Cooke, and John Maynard Smith. 1997. "Evolution of Genetic Redundancy." *Nature* 388 (6638): 167–71. https://doi.org/10.1038/40618.

Nutt, Stephen L., Donald Metcalf, Angela D'Amico, Matthew Polli, and Li Wu. 2005. "Dynamic Regulation of PU.1 Expression in Multipotent Hematopoietic Progenitors." *The Journal of Experimental Medicine* 201 (2): 221–31. https://doi.org/10.1084/jem.20041535.

O'Neil, Jennifer, Jennifer Shank, Nicole Cusson, Cornelis Murre, and Michelle Kelliher. 2004. "TAL1/SCL Induces Leukemia by Inhibiting the Transcriptional Activity of E47/HEB." *Cancer Cell* 5 (6): 587–96. https://doi.org/10.1016/j.ccr.2004.05.023.

Ocone, Andrea, Laleh Haghverdi, Nikola S Mueller, and Fabian J Theis. 2015. "Reconstructing Gene Regulatory Dynamics from High-Dimensional Single-Cell Snapshot Data." *Bioinformatics (Oxford, England)* 31 (12): i89-96. https://doi.org/10.1093/bioinformatics/btv257.

Ogata, Kiyoyuki, Chikako Satoh, Mikiko Tachibana, Hideya Hyodo, Hideto Tamura, Kazuo Dan, Takafumi Kimura, Yoshiaki Sonoda, and Takashi Tsuji. 2005. "Identification and Hematopoietic Potential of CD45 $^{-}$ Clonal Cells with Very Immature Phenotype (CD45 $^{-}$ CD34 $^{-}$ CD38 $^{-}$ Lin $^{-}$) in Patients with Myelodysplastic Syndromes." *Stem Cells* 23 (5): 619–30. https://doi.org/10.1634/stemcells.2004-0280.

Ogawa, M. 1993. "Differentiation and Proliferation of Hematopoietic Stem Cells." *Blood* 81 (11): 2844–53.

Ogbourne, S, and T M Antalis. 1998. "Transcriptional Control and the Role of Silencers in Transcriptional Regulation in Eukaryotes." *The Biochemical Journal* 331 (Pt 1) (April): 1–14.

Okada, S, H Nakauchi, K Nagayoshi, S Nishikawa, Y Miura, and T Suda. 1992. “In Vivo and in Vitro Stem Cell Function of C-Kit- and Sca-1-Positive Murine Hematopoietic Cells.” *Blood* 80 (12).

Okuda, T, J van Deursen, S W Hiebert, G Grosveld, and J R Downing. 1996. “AML1, the Target of Multiple Chromosomal Translocations in Human Leukemia, Is Essential for Normal Fetal Liver Hematopoiesis.” *Cell* 84 (2): 321–30.

Olsson, Andre, Meenakshi Venkatasubramanian, Viren K. Chaudhri, Bruce J. Aronow, Nathan Salomonis, Harinder Singh, and H. Leighton Grimes. 2016. “Single-Cell Analysis of Mixed-Lineage States Leading to a Binary Cell Fate Choice.” *Nature* 537 (7622): 698–702. https://doi.org/10.1038/nature19348.

Olszewski, Maureen A., John Gray, and Deborah J. Vestal. 2006. “*In Silico* Genomic Analysis of the Human and Murine Guanylate-Binding Protein (GBP) Gene Clusters.” *Journal of Interferon & Cytokine Research* 26 (5): 328–52. https://doi.org/10.1089/jir.2006.26.328.

Orkin, Stuart H, and Leonard I Zon. 2008. “Hematopoiesis: An Evolving Paradigm for Stem Cell Biology.” *Cell* 132 (4): 631–44. https://doi.org/10.1016/j.cell.2008.01.025.

Osawa, Mitsujiro, Tomoyuki Yamaguchi, Yukio Nakamura, Shin Kaneko, Masafumi Onodera, Ken-Ichi Sawada, Armin Jegalian, Hong Wu, Hiromitsu Nakauchi, and Atsushi Iwama. 2002. “Erythroid Expansion Mediated by the Gfi-1B Zinc Finger Protein: Role in Normal Hematopoiesis.” *Blood* 100 (8): 2769–77. https://doi.org/10.1182/blood-2002-01-0182.

Osborne, Geoffrey W. 2011. “Chapter 21 – Recent Advances in Flow Cytometric Cell Sorting.” In *Methods in Cell Biology*, 102:533–56. https://doi.org/10.1016/B978-0-12-374912-3.00021-3.

Ozer, Ufuk, Karen W Barbour, Sarah A Clinton, and Franklin G Berger. 2015. “Oxidative Stress and Response to Thymidylate Synthase-Targeted Antimetabolites S.” *Molecular Pharmacology* 88: 970–81. https://doi.org/10.1124/mol.115.099614.

Pabst, O, R Förster, M Lipp, H Engel, and H H Arnold. 2000. “NKX2.3 Is Required for MAdCAM-1 Expression and Homing of Lymphocytes in Spleen and Mucosa-Associated Lymphoid Tissue.” *The EMBO Journal* 19 (9): 2015–23. https://doi.org/10.1093/emboj/19.9.2015.

Pabst, O, R Zweigerdt, and H H Arnold. 1999. “Targeted Disruption of the Homeobox Transcription Factor Nkx2-3 in Mice Results in Postnatal Lethality and Abnormal

Development of Small Intestine and Spleen." *Development* 126 (10): 2215–25.

Pandolfi, Pier Paolo, Matthew E. Roth, Alar Karis, Mark W. Leonard, E. Dzierzak, Frank G. Grosveld, James Douglas Engel, and Michael H. Lindenbaum. 1995. "Targeted Disruption of the GATA3 Gene Causes Severe Abnormalities in the Nervous System and in Fetal Liver Haematopoiesis." *Nature Genetics* 11 (1): 40–44. https://doi.org/10.1038/ng0995-40.

Pascall, John C., Louise M. C. Webb, Eeva-Liisa Eskelinen, Silvia Innocentin, Noudjoud Attaf-Bouabdallah, and Geoffrey W. Butcher. 2018. "GIMAP6 Is Required for T Cell Maintenance and Efficient Autophagy in Mice." Edited by Jianhua Zhang. *PLOS ONE* 13 (5): e0196504. https://doi.org/10.1371/journal.pone.0196504.

Pastor, William A., L. Aravind, and Anjana Rao. 2013. "TETonic Shift: Biological Roles of TET Proteins in DNA Demethylation and Transcription." *Nature Reviews Molecular Cell Biology* 14 (6): 341–56. https://doi.org/10.1038/nrm3589.

Pattabiraman, D R, and T J Gonda. 2013. "Role and Potential for Therapeutic Targeting of MYB in Leukemia." *Leukemia* 27 (2): 269–77. https://doi.org/10.1038/leu.2012.225.

Pattanayak, Vikram, Steven Lin, John P Guilinger, Enbo Ma, Jennifer A Doudna, and David R Liu. 2013. "High-Throughput Profiling of off-Target DNA Cleavage Reveals RNA-Programmed Cas9 Nuclease Specificity." *Nature Biotechnology* 31 (9): 839–43. https://doi.org/10.1038/nbt.2673.

Paul, Franziska, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, et al. 2015. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors." *Cell* 163 (7): 1663–77. https://doi.org/10.1016/j.cell.2015.11.013.

Pearson, Karl. 1901. "On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72. https://doi.org/10.1080/14786440109462720.

Perez-Garcia, Arianne, Alberto Ambesi-Impiombato, Michael Hadler, Isaura Rigo, Charles A LeDuc, Kara Kelly, Chaim Jalas, et al. 2013. "Genetic Loss of SH2B3 in Acute Lymphoblastic Leukemia." *Blood* 122 (14): 2425–32. https://doi.org/10.1182/blood-2013-05-500850.

Perié, Leïla, Ken R Duffy, Lianne Kok, Rob J de Boer, and Ton N Schumacher. 2015. "The Branching Point in Erythro-Myeloid Differentiation." *Cell* 163 (7): 1655–62. https://doi.org/10.1016/j.cell.2015.11.059.

Peter, Isabelle, and Eric Davidson. 2015. *Genomic Control Process: Development and Evolution*. Academic Press.

Peterson, Vanessa M, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova, and Joel A Klappenbach. 2017. "Multiplexed Quantification of Proteins and Transcripts in Single Cells." *Nature Biotechnology* 35 (10): 936–39. https://doi.org/10.1038/nbt.3973.

Pettitt, Stephen J, Qi Liang, Xin Y Rairdan, Jennifer L Moran, Haydn M Prosser, David R Beier, Kent C Lloyd, Allan Bradley, and William C Skarnes. 2009. "Agouti C57BL/6N Embryonic Stem Cells for Mouse Genetic Resources." *Nature Methods* 6 (7): 493–95. https://doi.org/10.1038/nmeth.1342.

Peyvandi, Flora, Isabella Garagiola, and Luciano Baronciani. 2011. "Role of von Willebrand Factor in the Haemostasis." *Blood Transfusion = Trasfusione Del Sangue* 9 Suppl 2 (Suppl 2): s3-8. https://doi.org/10.2450/2011.002S.

Pfister, Christina, Marcos S Tatabiga, and Florian Roser. 2011. "Selection of Suitable Reference Genes for Quantitative Real-Time Polymerase Chain Reaction in Human Meningiomas and Arachnoidea." *BMC Research Notes* 4 (1): 275. https://doi.org/10.1186/1756-0500-4-275.

Picelli, Simone, Omid R Faridani, Asa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. "Full-Length RNA-Seq from Single Cells Using Smart-Seq2." *Nature Protocols* 9 (1): 171–81. https://doi.org/10.1038/nprot.2014.006.

Pietras, Eric M, Damien Reynaud, Yoon-A Kang, Daniel Carlin, Fernando J Calero-Nieto, Andrew D Leavitt, Joshua M Stuart, Berthold Göttgens, and Emmanuelle Passegué. 2015. "Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions." *Cell Stem Cell* 17 (1): 35–46. https://doi.org/10.1016/j.stem.2015.05.003.

Pimanda, J. E., K. Ottersbach, K. Knezevic, S. Kinston, W. Y. I. Chan, N. K. Wilson, J.-R. Landry, et al. 2007. "Gata2, Fli1, and Scl Form a Recursively Wired Gene-Regulatory Circuit during

Early Hematopoietic Development." *Proceedings of the National Academy of Sciences* 104 (45): 17692–97. https://doi.org/10.1073/pnas.0707045104.

Pimanda, John E., and Berthold Gottgens. 2010. "Gene Regulatory Networks Governing Haematopoietic Stem Cell Development and Identity." *The International Journal of Developmental Biology* 54 (6–7): 1201–11. https://doi.org/10.1387/ijdb.093038jp.

Pina, Cristina, Cristina Fugazza, Alex J. Tipping, John Brown, Shamit Soneji, Jose Teles, Carsten Peterson, and Tariq Enver. 2012. "Inferring Rules of Lineage Commitment in Haematopoiesis." *Nature Cell Biology* 14 (3): 287–94. https://doi.org/10.1038/ncb2442.

Pina, Cristina, José Teles, Cristina Fugazza, Gillian May, Dapeng Wang, Yanping Guo, Shamit Soneji, et al. 2015. "Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis." *Cell Reports* 11 (10): 1503–10. https://doi.org/10.1016/j.celrep.2015.05.016.

Pineault, Nicolas, Cheryl D Helgason, H Jeffrey Lawrence, and R Keith Humphries. 2002. "Differential Expression of Hox, Meis1, and Pbx1 Genes in Primitive Cells throughout Murine Hematopoietic Ontogeny." *Experimental Hematology* 30 (1): 49–57.

Pirot, Nelly, Virginie Deleuze, Rawan El-Hajj, Christiane Dohet, Fred Sablitzky, Philippe Couttet, Danièle Mathieu, and Valérie Pinet. 2010. "LYL1 Activity Is Required for the Maturation of Newly Formed Blood Vessels in Adulthood." *Blood* 115 (25): 5270–79. https://doi.org/10.1182/blood-2010-03-275651.

Platt, Randall J, Sidi Chen, Yang Zhou, Michael J Yim, Lukasz Swiech, Hannah R Kempton, James E Dahlman, et al. 2014. "CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling." *Cell* 159 (2): 440–55. https://doi.org/10.1016/j.cell.2014.09.014.

Plo, Isabelle, Christine Bellanné-Chantelot, Matthieu Mosca, Stefania Mazzi, Caroline Marty, and William Vainchenker. 2017. "Genetic Alterations of the Thrombopoietin/MPL/JAK2 Axis Impacting Megakaryopoiesis." *Frontiers in Endocrinology* 8: 234. https://doi.org/10.3389/fendo.2017.00234.

Poczobutt, Joanna M., Subhajyoti De, Vinod K. Yadav, Teresa T. Nguyen, Howard Li, Trisha R. Sippel, Mary C. M. Weiser-Evans, and Raphael A. Nemenoff. 2016. "Expression Profiling of Macrophages Reveals Multiple Populations with Distinct Biological Roles in an

Immunocompetent Orthotopic Model of Lung Cancer." *The Journal of Immunology* 196 (6): 2847–59. https://doi.org/10.4049/jimmunol.1502364.

Porcher, C, W Swat, K Rockwell, Y Fujiwara, F W Alt, and S H Orkin. 1996. "The T Cell Leukemia Oncoprotein SCL/Tal-1 Is Essential for Development of All Hematopoietic Lineages." *Cell* 86 (1): 47–57.

Pouget, Claire, Tessa Peterkin, Filipa Costa Simões, Yoonsung Lee, David Traver, and Roger Patient. 2014. "FGF Signalling Restricts Haematopoietic Stem Cell Specification via Modulation of the BMP Pathway." *Nature Communications* 5 (November): 5588. https://doi.org/10.1038/ncomms6588.

Pronk, Cornelis J H, Derrick J Rossi, Robert Månsson, Joanne L Attema, Gudmundur Logi Norddahl, Charles Kwok Fai Chan, Mikael Sigvardsson, Irving L Weissman, and David Bryder. 2007. "Elucidation of the Phenotypic, Functional, and Molecular Topography of a Myeloerythroid Progenitor Cell Hierarchy." *Cell Stem Cell* 1 (4): 428–42. https://doi.org/10.1016/j.stem.2007.07.005.

Pui, J C, D Allman, L Xu, S DeRocco, F G Karnell, S Bakkour, J Y Lee, et al. 1999. "Notch1 Expression in Early Lymphopoiesis Influences B versus T Lineage Determination." *Immunity* 11 (3): 299–308.

Qiu, Peng, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. 2011. "Extracting a Cellular Hierarchy from High-Dimensional Cytometry Data with SPADE." *Nature Biotechnology* 29 (10): 886–91. https://doi.org/10.1038/nbt.1991.

Qiu, Xiaojie, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah Pliner, and Cole Trapnell. 2017. "Reversed Graph Embedding Resolves Complex Single-Cell Developmental Trajectories." *BioRxiv*, February, 110668. https://doi.org/10.1101/110668.

Radonić, Aleksandar, Stefanie Thulke, Ian M Mackay, Olfert Landt, Wolfgang Siegert, and Andreas Nitsche. 2004. "Guideline to Reference Gene Selection for Quantitative Real-Time PCR." *Biochemical and Biophysical Research Communications* 313 (4): 856–62. https://doi.org/10.1016/J.BBRC.2003.11.177.

Radtke, Freddy, Anne Wilson, Stephane J C Mancini, and H Robson MacDonald. 2004. "Notch

Regulation of Lymphocyte Development and Function." *Nature Immunology* 5 (3): 247–53. https://doi.org/10.1038/ni1045.

Rainis, Liat, Tsutomu Toki, John E Pimanda, Ester Rosenthal, Keren Machol, Sabine Strehl, Berthold Göttgens, Etsuro Ito, and Shai Izraeli. 2005. "The Proto-Oncogene ERG in Megakaryoblastic Leukemias." *Cancer Research* 65 (17): 7596–7602. https://doi.org/10.1158/0008-5472.CAN-05-0147.

Raj, Bushra, Daniel E Wagner, Aaron McKenna, Shristi Pandey, Allon M Klein, Jay Shendure, James A Gagnon, and Alexander F Schier. 2018. "Simultaneous Single-Cell Profiling of Lineages and Cell Types in the Vertebrate Brain." *Nature Biotechnology* 36 (5): 442–50. https://doi.org/10.1038/nbt.4103.

Rashid, Ayesha, Irakli Dzneladze, Ruijuan He, John F Woolley, Michael D Jain, and Mark D. Minden. 2016. "CSF1R Is Associated with Poor Overall Survival in AML and Mediates Supportive Interactions Between AML and Stromal Cells in the AML Microenvironment." *Blood* 128 (22).

Rasmussen, Kasper D, Guangshuai Jia, Jens V Johansen, Marianne T Pedersen, Nicolas Rapin, Frederik O Bagger, Bo T Porse, Olivier A Bernard, Jesper Christensen, and Kristian Helin. 2015. "Loss of TET2 in Hematopoietic Cells Leads to DNA Hypermethylation of Active Enhancers and Induction of Leukemogenesis." *Genes & Development* 29 (9): 910–22. https://doi.org/10.1101/gad.260174.115.

Rathinam, Chozhavendan, Robert Geffers, Raif Yücel, Jan Buer, Karl Welte, Tarik Möröy, and Christoph Klein. 2005. "The Transcriptional Repressor Gfi1 Controls STAT3-Dependent Dendritic Cell Development and Function." *Immunity* 22 (6): 717–28. https://doi.org/10.1016/j.immuni.2005.04.007.

Ravasi, Timothy, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B. Bajic, Kai Tan, Altuna Akalin, et al. 2010. "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man." *Cell* 140 (5): 744–52. https://doi.org/10.1016/j.cell.2010.01.044.

Redecke, Vanessa, Ruiqiong Wu, Jingran Zhou, David Finkelstein, Vandana Chaturvedi, Anthony A High, and Hans Hacker. 2013. "Hematopoietic Progenitor Cell Lines with Myeloid and Lymphoid Potential." *Nature Methods* 10 (8): 795–803.

Renders, Simom, Pia Sommerkamp, Jasper Panten, Luisa Ladel, Adriana Przybylla, Petra Zeisberger, Daniel Klimmeck, Katharina Schönberger, Nina Cabezas-Wallscheid, and Andreas Trumpp. 2017. "Neogenin Regulates Hematopoietic Stem Cell Quiescence and Maintenance." *Experimental Hematology* 53 (September): S49. https://doi.org/10.1016/J.EXPHEM.2017.06.072.

Robb, L, N J Elwood, A G Elefanty, F Köntgen, R Li, L D Barnett, and C G Begley. 1996. "The Scl Gene Product Is Required for the Generation of All Hematopoietic Lineages in the Adult Mouse." *The EMBO Journal* 15 (16): 4123–29.

Rodrigues, Neil P, Viktor Janzen, Randolf Forkert, David M Dombkowski, Ashleigh S Boyd, Stuart H Orkin, Tariq Enver, Paresh Vyas, and David T Scadden. 2005. "Haploinsufficiency of GATA-2 Perturbs Adult Hematopoietic Stem-Cell Homeostasis." *Blood* 106 (2): 477–84. https://doi.org/10.1182/blood-2004-08-2989.

Rosenberg, Alexander B., Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas Gray, et al. 2018. "Single-Cell Profiling of the Developing Mouse Brain and Spinal Cord with Split-Pool Barcoding." *Science*, March, eaam8999. https://doi.org/10.1126/SCIENCE.AAM8999.

Rotem, Assaf, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. 2015. "Single-Cell ChIP-Seq Reveals Cell Subpopulations Defined by Chromatin State." *Nature Biotechnology* 33 (11): 1165–72. https://doi.org/10.1038/nbt.3383.

Rothenberg, Ellen V. 2014. "Transcriptional Control of Early T and B Cell Developmental Choices." *Annual Review of Immunology* 32 (1): 283–321. https://doi.org/10.1146/annurev-immunol-032712-100024.

Russell, M, F Thompson, C Spier, and R Taetle. 1993. "Expression of the EVI1 Gene in Chronic Myelogenous Leukemia in Blast Crisis." *Leukemia* 7 (10): 1654–57.

Sakurai, M, H Kunimoto, N Watanabe, Y Fukuchi, S Yuasa, S Yamazaki, T Nishimura, et al. 2014. "Impaired Hematopoietic Differentiation of RUNX1-Mutated Induced Pluripotent Stem Cells Derived from FPD/AML Patients." *Leukemia* 28 (12): 2344–54. https://doi.org/10.1038/leu.2014.136.

Salek-Ardakani, Samira, Gil Smooha, Jasper de Boer, Neil J Sebire, Michelle Morrow, Liat Rainis,

Sandy Lee, Owen Williams, Shai Izraeli, and Hugh J M Brady. 2009. “ERG Is a Megakaryocytic Oncogene.” *Cancer Research* 69 (11): 4665–73. https://doi.org/10.1158/0008-5472.CAN-09-0075.

Saleque, Shireen, Scott Cameron, and Stuart H Orkin. 2002. “The Zinc-Finger Proto-Oncogene Gfi-1b Is Essential for Development of the Erythroid and Megakaryocytic Lineages.” *Genes & Development* 16 (3): 301–6. https://doi.org/10.1101/gad.959102.

Sanchez-Freire, Veronica, Antje D Ebert, Tomer Kalisky, Stephen R Quake, and Joseph C Wu. 2012. “Microfluidic Single-Cell Real-Time PCR for Comparative Analysis of Gene Expression Patterns.” *Nature Protocols* 7 (5): 829–38. https://doi.org/10.1038/nprot.2012.021.

Sanjuan-Pla, Alejandra, Iain C. Macaulay, Christina T. Jensen, Petter S. Woll, Tiago C. Luis, Adam Mead, Susan Moore, et al. 2013. “Platelet-Biased Stem Cells Reside at the Apex of the Haematopoietic Stem-Cell Hierarchy.” *Nature* 502 (7470): 232–36. https://doi.org/10.1038/nature12495.

Sanyal, M., J. W. Tung, H. Karsunky, H. Zeng, L. Selleri, I. L. Weissman, L. A. Herzenberg, and M. L. Cleary. 2007. “B-Cell Development Fails in the Absence of the Pbx1 Proto-Oncogene.” *Blood* 109 (10): 4191–99. https://doi.org/10.1182/blood-2006-10-054213.

Sarrazin, Sandrine, Noushine Mossadegh-Keller, Taro Fukao, Athar Aziz, Frederic Mourcin, Laurent Vanhille, Louise Kelly Modis, et al. 2009. “MafB Restricts M-CSF-Dependent Myeloid Commitment Divisions of Hematopoietic Stem Cells.” *Cell* 138 (2): 300–313. https://doi.org/10.1016/j.cell.2009.04.057.

Sauvageau, G, P M Lansdorp, C J Eaves, D E Hogge, W H Dragowska, D S Reid, C Largman, H J Lawrence, and R K Humphries. 1994. “Differential Expression of Homeobox Genes in Functionally Distinct CD34+ Subpopulations of Human Bone Marrow Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 91 (25): 12223–27.

Sauvageau, G, U Thorsteinsdottir, C J Eaves, H J Lawrence, C Largman, P M Lansdorp, and R K Humphries. 1995. “Overexpression of HOXB4 in Hematopoietic Cells Causes the Selective Expansion of More Primitive Populations in Vitro and in Vivo.” *Genes & Development* 9 (14): 1753–65.

Savoia, Anna, Annalisa Pastore, Daniela De Rocco, Elisa Civaschi, Mariateresa Di Stazio, Roberta Bottega, Federica Melazzini, et al. 2011. “Clinical and Genetic Aspects of Bernard-Soulier Syndrome: Searching for Genotype/Phenotype Correlations.” *Haematologica* 96 (3): 417–23. https://doi.org/10.3324/haematol.2010.032631.

Sawai, Catherine M, Sonja Babovic, Samik Upadhaya, David J H F Knapp, Yonit Lavin, Colleen M Lau, Anton Goloborodko, et al. 2016. “Hematopoietic Stem Cells Are the Major Source of Multilineage Hematopoiesis in Adult Animals.” *Immunity* 45 (3): 597–609. https://doi.org/10.1016/j.immuni.2016.08.007.

Schaniel, Christoph, Yen-Sin Ang, Kajan Ratnakumar, Catherine Cormier, Taneisha James, Emily Bernstein, Ihor R Lemischka, and Patrick J Paddison. 2009. “Smarcc1/Baf155 Couples Self-Renewal Gene Repression with Changes in Chromatin Structure in Mouse Embryonic Stem Cells.” *Stem Cells (Dayton, Ohio)* 27 (12): 2979–91. https://doi.org/10.1002/stem.223.

Scherer, S, and R W Davis. 1979. “Replacement of Chromosome Segments with Altered DNA Sequences Constructed in Vitro.” *Proceedings of the National Academy of Sciences of the United States of America* 76 (10): 4951–55.

Schmidt, M., K. Paes, A. De Maziere, T. Smyczek, S. Yang, A. Gray, D. French, et al. 2007. “EGFL7 Regulates the Collective Migration of Endothelial Cells by Restricting Their Spatial Distribution.” *Development* 134 (16): 2913–23. https://doi.org/10.1242/dev.002576.

Schuh, Anna H, Alex J Tipping, Allison J Clark, Isla Hamlett, Boris Guyot, Francesco J Iborra, Patrick Rodriguez, et al. 2005. “ETO-2 Associates with SCL in Erythroid Cells and Megakaryocytes and Provides Repressor Functions in Erythropoiesis.” *Mol Cell Biol* 25 (23): 10235–50. https://doi.org/10.1128/MCB.25.23.10235-10250.2005.

Schulte, Reiner, Nicola K Wilson, Janine C M Prick, Chiara Cossetti, Michal K Maj, Berthold Gottgens, and David G Kent. 2015. “Index Sorting Resolves Heterogeneous Murine Hematopoietic Stem Cell Populations.” *Experimental Hematology* 43 (9): 803–11. https://doi.org/10.1016/j.exphem.2015.05.006.

Schütte, Judith, Huange Wang, Stella Antoniou, Andrew Jarratt, Nicola K Wilson, Joey Riepsaame, Fernando J Calero-Nieto, et al. 2016. “An Experimentally Validated Network of Nine Haematopoietic Transcription Factors Reveals Mechanisms of Cell State Stability.” *ELife* 5 (February): e11469. https://doi.org/10.7554/eLife.11469.

Scialdone, Antonio, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. 2015. "Computational Assignment of Cell-Cycle Stage from Single-Cell Transcriptome Data." *Methods* 85 (September): 54–61. https://doi.org/10.1016/j.ymeth.2015.06.021.

Scialdone, Antonio, Yosuke Tanaka, Wajid Jawaid, Victoria Moignard, Nicola K Wilson, Iain C Macaulay, John C Marioni, and Berthold Göttgens. 2016. "Resolving Early Mesoderm Diversification through Single-Cell Expression Profiling." *Nature* 535 (7611): 289–93.

Scott, E W, M C Simon, J Anastasi, and H Singh. 1994. "Requirement of Transcription Factor PU.1 in the Development of Multiple Hematopoietic Lineages." *Science (New York, N.Y.)* 265 (5178): 1573–77.

Scott, LM, CI Civin, P Rorth, and AD Friedman. 1992. "A Novel Temporal Expression Pattern of Three C/EBP Family Members in Differentiating Myelomonocytic Cells." *Blood* 80 (7).

Seaman, William E. 2000. "Natural Killer Cells and Natural Killer T Cells." *Arthritis & Rheumatism* 43 (6): 1204–17. https://doi.org/10.1002/1529-0131(200006)43:6<1204::AID-ANR3>3.0.CO;2-I.

Setty, Manu, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. 2016. "Wishbone Identifies Bifurcating Developmental Trajectories from Single-Cell Data." *Nature Biotechnology* 34 (6): 637–45. https://doi.org/10.1038/nbt.3569.

Shalem, Ophir, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei S. Mikkelsen, Dirk Heckl, et al. 2014. "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells." *Science* 343 (6166): 84–87. https://doi.org/10.1126/science.1247005.

Shapiro, Ehud. 2018. "On the Journey from Nematode to Human, Scientists Dive by the Zebrafish Cell Lineage Tree." *Genome Biology* 19 (1): 63. https://doi.org/10.1186/s13059-018-1453-x.

Shi, Junwei, Eric Wang, Joseph P Milazzo, Zihua Wang, Justin B Kinney, and Christopher R Vakoc. 2015. "Discovery of Cancer Drug Targets by CRISPR-Cas9 Screening of Protein Domains." *Nature Biotechnology* 33 (6): 661–67. https://doi.org/10.1038/nbt.3235.

Shields, Benjamin J, Jacob T Jackson, Donald Metcalf, Wei Shi, Qiutong Huang, Alexandra L Garnham, Stefan P Glaser, et al. 2016. "Acute Myeloid Leukemia Requires Hhex to Enable

PRC2-Mediated Epigenetic Repression of Cdkn2a." *Genes & Development* 30 (1): 78–91. https://doi.org/10.1101/gad.268425.115.

Shivdasani, R A, M F Rosenblatt, D Zucker-Franklin, C W Jackson, P Hunt, C J Saris, and S H Orkin. 1995. "Transcription Factor NF-E2 Is Required for Platelet Formation Independent of the Actions of Thrombopoietin/MGDF in Megakaryocyte Development." *Cell* 81 (5): 695–704.

Sieburg, H. B., Rebecca H Cho, Brad Dykstra, Naoyuki Uchida, Connie J Eaves, and Christa E Muller-Sieburg. 2006. "The Hematopoietic Stem Compartment Consists of a Limited Number of Discrete Stem Cell Subsets." *Blood* 107 (6): 2311–16. https://doi.org/10.1182/blood-2005-07-2970.

Silver, Nicholas, Emanuele Cotroneo, Gordon Proctor, Samira Osailan, Katherine L Paterson, and Guy H Carpenter. 2008. "Selection of Housekeeping Genes for Gene Expression Studies in the Adult Rat Submandibular Gland under Normal, Inflamed, Atrophic and Regenerative States." *BMC Molecular Biology* 9 (1): 64. https://doi.org/10.1186/1471-2199-9-64.

Simons, Benjamin D., and Hans Clevers. 2011. "Strategies for Homeostatic Stem Cell Self-Renewal in Adult Tissues." *Cell* 145 (6): 851–62. https://doi.org/10.1016/j.cell.2011.05.033.

Sive, Jonathan I, and Berthold Göttgens. 2014. "Transcriptional Network Control of Normal and Leukaemic Haematopoiesis." *Experimental Cell Research* 329 (2): 255–64. https://doi.org/10.1016/j.yexcr.2014.06.021.

Skylaki, Stavroula, Oliver Hilsenbeck, and Timm Schroeder. 2016. "Challenges in Long-Term Imaging and Quantification of Single-Cell Dynamics." *Nature Biotechnology* 34 (11): 1137–44. https://doi.org/10.1038/nbt.3713.

Slattery, Matthew, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. 2014. "Absence of a Simple Code: How Transcription Factors Read the Genome." *Trends in Biochemical Sciences* 39 (9): 381–99. https://doi.org/10.1016/j.tibs.2014.07.002.

Smallwood, Sébastien A, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. 2014. "Single-Cell Genome-Wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity." *Nature*

Methods 11 (8): 817–20. https://doi.org/10.1038/nmeth.3035.

Smithies, O, R G Gregg, S S Boggs, M A Koralewski, and R S Kucherlapati. 1985. “Insertion of DNA Sequences into the Human Chromosomal Beta-Globin Locus by Homologous Recombination.” *Nature* 317 (6034): 230–34.

Solar, G P, W G Kerr, F C Zeigler, D Hess, C Donahue, F J de Sauvage, and D L Eaton. 1998. “Role of C-Mpl in Early Hematopoiesis.” *Blood* 92 (1): 4–10.

Sood, Raman, Yasuhiko Kamikubo, and Paul Liu. 2017. “Role of RUNX1 in Hematological Malignancies.” *Blood* 129 (15): 2070–82. https://doi.org/10.1182/blood-2016-10-687830.

Sorek, Rotem, C. Martin Lawrence, and Blake Wiedenheft. 2013. “CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea.” *Annual Review of Biochemistry* 82 (1): 237–66. https://doi.org/10.1146/annurev-biochem-072911-172315.

Sorensen, Poul H.B., Stephen L. Lessnick, Dolores Lopez-Terrada, Xian F. Liu, Timothy J. Triche, and Christopher T. Denny. 1994. “A Second Ewing’s Sarcoma Translocation, t(21;22), Fuses the EWS Gene to Another ETS–Family Transcription Factor, ERG.” *Nature Genetics* 6 (2): 146–51. https://doi.org/10.1038/ng0294-146.

Souroullas, George P, Jessica M Salmon, Fred Sablitzky, David J Curtis, and Margaret A Goodell. 2009. “Adult Hematopoietic Stem and Progenitor Cells Require Either Lyl1 or Scl for Survival.” *Cell Stem Cell* 4 (2): 180–86. https://doi.org/10.1016/j.stem.2009.01.001.

Spangrude, G J, S Heimfeld, and I L Weissman. 1988. “Purification and Characterization of Mouse Hematopoietic Stem Cells.” *Science* 241 (4861): 58–62.

Spanjaard, Bastiaan, Bo Hu, Nina Mitic, and Jan Philipp Junker. 2017. “Massively Parallel Single Cell Lineage Tracing Using CRISPR/Cas9 Induced Genetic Scars.” *BioRxiv*, October, 205971. https://doi.org/10.1101/205971.

Spitzer, M. H., P. F. Gherardini, G. K. Fragiadakis, N. Bhattacharya, R. T. Yuan, A. N. Hotson, R. Finck, et al. 2015. “An Interactive Reference Framework for Modeling a Dynamic Immune System.” *Science* 349 (6244): 1259425–1259425. https://doi.org/10.1126/science.1259425.

Stadhouders, Ralph, Alba Cico, Tharshana Stephen, Supat Thongjuea, Petros Kolovos, H Irem Baymaz, Xiao Yu, et al. 2015. “Control of Developmentally Primed Erythroid Genes by

Combinatorial Co-Repressor Actions." *Nature Communications* 6: 8893. https://doi.org/10.1038/ncomms9893.

Ståhlberg, Anders, and Martin Bengtsson. 2010. "Single-Cell Gene Expression Profiling Using Reverse Transcription Quantitative Real-Time PCR." *Methods* 50 (4): 282–88. https://doi.org/10.1016/J.YMETH.2010.01.002.

Stein, Sarah J, and Albert S Baldwin. 2013. "Deletion of the NF-KB Subunit P65/RelA in the Hematopoietic Compartment Leads to Defects in Hematopoietic Stem Cell Function." *Blood* 121 (25): 5015–24. https://doi.org/10.1182/blood-2013-02-486142.

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. "Simultaneous Epitope and Transcriptome Measurement in Single Cells." *Nature Methods* 14 (9): 865–68. https://doi.org/10.1038/nmeth.4380.

Storry, Jill R, Magnus Jöud, Mikael Kronborg Christophersen, Britt Thuresson, Bo Åkerström, Birgitta Nilsson Sojka, Björn Nilsson, and Martin L Olsson. 2013. "Homozygosity for a Null Allele of SMIM1 Defines the Vel-Negative Blood Group Phenotype." *Nature Genetics* 45 (5): 537–41. https://doi.org/10.1038/ng.2600.

Sugano, Yasuyoshi, Masaki Takeuchi, Ayami Hirata, Hirokazu Matsushita, Toshio Kitamura, Minoru Tanaka, and Atsushi Miyajima. 2008. "Junctional Adhesion Molecule-A, JAM-A, Is a Novel Cell-Surface Marker for Long-Term Repopulating Hematopoietic Stem Cells." *Blood* 111 (3): 1167–72. https://doi.org/10.1182/blood-2007-03-081554.

Sulong, Sarina, Anthony V Moorman, Julie A E Irving, Jonathan C Strefford, Zoe J Konn, Marian C Case, Lynne Minto, et al. 2009. "A Comprehensive Analysis of the CDKN2A Gene in Childhood Acute Lymphoblastic Leukemia Reveals Genomic Deletion, Copy Number Neutral Loss of Heterozygosity, and Association with Specific Cytogenetic Subgroups." *Blood* 113 (1): 100–107. https://doi.org/10.1182/blood-2008-07-166801.

Sun, W., and James R Downing. 2004. "Haploinsufficiency of AML1 Results in a Decrease in the Number of LTR-HSCs While Simultaneously Inducing an Increase in More Mature Progenitors." *Blood* 104 (12): 3565–72. https://doi.org/10.1182/blood-2003-12-4349.

Swiers, Gemma, Roger Patient, and Matthew Loose. 2006. "Genetic Regulatory Networks

Programming Hematopoietic Stem Cells and Erythroid Lineage Specification." *Developmental Biology* 294 (2): 525–40. https://doi.org/10.1016/j.ydbio.2006.02.051.

Szilvassy, S J, R K Humphries, P M Lansdorp, A C Eaves, and C J Eaves. 1990. "Quantitative Assay for Totipotent Reconstituting Hematopoietic Stem Cells by a Competitive Repopulation Strategy." *Proceedings of the National Academy of Sciences of the United States of America* 87 (22): 8736–40.

Takaki, S, K Sauer, B M Iritani, S Chien, Y Ebihara, K Tsuji, K Takatsu, and R M Perlmutter. 2000. "Control of B Cell Production by the Adaptor Protein Lnk. Definition Of a Conserved Family of Signal-Modulating Proteins." *Immunity* 13 (5): 599–609.

Tenen, Daniel G. 2003. "Disruption of Differentiation in Human Cancer: AML Shows the Way." *Nature Reviews Cancer* 3 (2): 89–101. https://doi.org/10.1038/nrc989.

Thomas, Matthew D., Christopher S. Kremer, Kodi S. Ravichandran, Klaus Rajewsky, and Timothy P. Bender. 2005. "C-Myb Is Critical for B Cell Development and Maintenance of Follicular B Cells." *Immunity* 23 (3): 275–86. https://doi.org/10.1016/j.immuni.2005.08.005.

Todorovic, Vesna. 2017. "Gene Expression: Single-Cell RNA-Seq—Now with Protein." *Nature Methods* 14 (11): 1028–29.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. 2014. "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells." *Nature Biotechnology* 32 (4): 381–86. https://doi.org/10.1038/nbt.2859.

Tsai, F Y, G Keller, F C Kuo, M Weiss, J Chen, M Rosenblatt, F W Alt, and S H Orkin. 1994. "An Early Haematopoietic Defect in Mice Lacking the Transcription Factor GATA-2." *Nature* 371 (6494): 221–26. https://doi.org/10.1038/371221a0.

Tsai, F Y, and S H Orkin. 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): 3636–43.

Tzelepis, Konstantinos, Hiroko Koike-Yusa, Etienne De Braekeleer, Yilong Li, Emmanouil Metzakopian, Oliver M Dovey, Annalisa Mupo, et al. 2016. "A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia." *Cell*

Reports 17 (4): 1193–1205. https://doi.org/10.1016/j.celrep.2016.09.079.

Umeda, Shigeaki, Kouhei Yamamoto, Toshihiko Murayama, Michihiro Hidaka, Morito Kurata, Takumi Ohshima, Shiho Suzuki, Emiko Sugawara, Fumio Kawano, and Masanobu Kitagawa. 2012. "Prognostic Significance of HOXB4 in *de Novo* Acute Myeloid Leukemia." *Hematology* 17 (3): 125–31. https://doi.org/10.1179/102453312X13376952196250.

Unnisa, Z., J. P. Clark, J. Roychoudhury, E. Thomas, L. Tessarollo, N. G. Copeland, N. A. Jenkins, H. L. Grimes, and A. R. Kumar. 2012. "Meis1 Preserves Hematopoietic Stem Cells in Mice by Limiting Oxidative Stress." *Blood* 120 (25): 4973–81. https://doi.org/10.1182/blood-2012-06-435800.

Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C Faircloth, Maido Remm, and Steven G Rozen. 2012. "Primer3--New Capabilities and Interfaces." *Nucleic Acids Research* 40 (15): e115. https://doi.org/10.1093/nar/gks596.

Valge-Archer, V, A Forster, and T H Rabbitts. 1998. "The LMO1 and LDB1 Proteins Interact in Human T Cell Acute Leukaemia with the Chromosomal Translocation t(11;14)(P15;Q11)." *Oncogene* 17 (24): 3199–3202. https://doi.org/10.1038/sj.onc.1202353.

Vaquerizas, Juan M., Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. 2009. "A Census of Human Transcription Factors: Function, Expression and Evolution." *Nature Reviews Genetics* 10 (4): 252–63. https://doi.org/10.1038/nrg2538.

Vassen, Lothar, Hugues Beauchemin, Wafaa Lemsaddek, Joseph Krongold, Marie Trudel, and Tarik Möröy. 2014. "Growth Factor Independence 1b (Gfi1b) Is Important for the Maturation of Erythroid Cells and the Regulation of Embryonic Globin Expression." Edited by Andrew C. Wilber. *PLoS ONE* 9 (5): e96636. https://doi.org/10.1371/journal.pone.0096636.

Vassen, Lothar, Taro Okayama, and Tarik Möröy. 2007. "Gfi1b:Green Fluorescent Protein Knock-in Mice Reveal a Dynamic Expression Pattern of Gfi1b during Hematopoiesis That Is Largely Complementary to Gfi1." *Blood* 109 (6): 2356–64. https://doi.org/10.1182/blood-2006-06-030031.

Velazquez, Laura, Alec M Cheng, Heather E Fleming, Caren Furlonger, Shirly Vesely, Alan Bernstein, Christopher J Paige, and Tony Pawson. 2002. "Cytokine Signaling and Hematopoietic Homeostasis Are Disrupted in Lnk-Deficient Mice." *The Journal of*

Experimental Medicine 195 (12): 1599–1611.

Velten, Lars, Simon F. Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P. Hennig, Christoph Hirche, et al. 2017. "Human Haematopoietic Stem Cell Lineage Commitment Is a Continuous Process." *Nature Cell Biology* 19 (4): 271–81. https://doi.org/10.1038/ncb3493.

Verbiest, Tom, Simon Bouffler, Stephen L Nutt, and Christophe Badie. 2015. "PU.1 Downregulation in Murine Radiation-Induced Acute Myeloid Leukaemia (AML): From Molecular Mechanism to Human AML." *Carcinogenesis* 36 (4): 413–19. https://doi.org/10.1093/carcin/bgv016.

Vicente, Carmen, Ana Conchillo, María A García-Sánchez, and María D Odero. 2011. "The Role of the GATA2 Transcription Factor in Normal and Malignant Hematopoiesis." *Critical Reviews in Oncology / Hematology* 82: 1–17. https://doi.org/10.1016/j.critrevonc.2011.04.007.

Visvader, J E, X Mao, Y Fujiwara, K Hahm, and S H Orkin. 1997. "The LIM-Domain Binding Protein Ldb1 and Its Partner LMO2 Act as Negative Regulators of Erythroid Differentiation." *Proceedings of the National Academy of Sciences of the United States of America* 94 (25): 13707–12.

Vladimer, Gregory I, Maria W Górna, and Giulio Superti-Furga. 2014. "IFITs: Emerging Roles as Key Anti-Viral Proteins." *Frontiers in Immunology* 5: 94. https://doi.org/10.3389/fimmu.2014.00094.

Vu, Thiet M., Ayako-Nakamura Ishizu, Juat Chin Foo, Xiu Ru Toh, Fangyu Zhang, Ding Ming Whee, Federico Torta, et al. 2017. "Mfsd2b Is Essential for the Sphingosine-1-Phosphate Export in Erythrocytes and Platelets." *Nature* 550 (7677): 524–28. https://doi.org/10.1038/nature24053.

Waddington, CH. 1957. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. Allen & Unwin.

Walter, M J, L Ding, D Shen, J Shao, M Grillot, M McLellan, R Fulton, et al. 2011. "Recurrent DNMT3A Mutations in Patients with Myelodysplastic Syndromes." *Leukemia* 25 (7): 1153–58. https://doi.org/10.1038/leu.2011.44.

Wang, Di, Mingzhu Zheng, Lei Lei, Jian Ji, Yunliang Yao, Yuanjun Qiu, Lie Ma, et al. 2012.

"Tespa1 Is Involved in Late Thymocyte Development through the Regulation of TCR-Mediated Signaling." *Nature Immunology* 13 (6): 560–68. https://doi.org/10.1038/ni.2301.

Wang, J H, A Nichogiannopoulou, L Wu, L Sun, A H Sharpe, M Bigby, and K Georgopoulos. 1996. "Selective Defects in the Development of the Fetal and Adult Lymphoid System in Mice with an Ikaros Null Mutation." *Immunity* 5 (6): 537–49.

Wang, Jianwei, Qian Sun, Yohei Morita, Hong Jiang, Alexander Groß, André Lechel, Kai Hildner, et al. 2012. "A Differentiation Checkpoint Limits Hematopoietic Stem Cell Self-Renewal in Response to DNA Damage." *Cell* 148 (5): 1001–14. https://doi.org/10.1016/J.CELL.2012.01.040.

Wang, Kemeng, Guoqing Wei, and Delong Liu. 2012. "CD19: A Biomarker for B Cell Development, Lymphoma Diagnosis and Therapy." *Experimental Hematology & Oncology* 1 (1): 36. https://doi.org/10.1186/2162-3619-1-36.

Wang, Q, T Stacy, M Binder, M Marin-Padilla, A H Sharpe, and N A Speck. 1996. "Disruption of the Cbfa2 Gene Causes Necrosis and Hemorrhaging in the Central Nervous System and Blocks Definitive Hematopoiesis." *Proceedings of the National Academy of Sciences of the United States of America* 93 (8): 3444–49. https://doi.org/10.1073/pnas.93.8.3444.

Wang, Tim, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. 2014. "Genetic Screens in Human Cells Using the CRISPR-Cas9 System." *Science* 343 (6166): 80–84. https://doi.org/10.1126/science.1246981.

Warren, A J, W H Colledge, M B Carlton, M J Evans, A J Smith, and T H Rabbitts. 1994. "The Oncogenic Cysteine-Rich LIM Domain Protein Rbtn2 Is Essential for Erythroid Development." *Cell* 78 (1): 45–57.

Weber, Brittany Nicole, Anthony Wei-Shine Chi, Alejandro Chavez, Yumi Yashiro-Ohtani, Qi Yang, Olga Shestova, and Avinash Bhandoola. 2011. "A Critical Role for TCF-1 in T-Lineage Specification and Differentiation." *Nature* 476 (7358): 63–68. https://doi.org/10.1038/nature10279.

Weinreb, Caleb, Samuel Wolock, and Allon M Klein. 2018. "SPRING: A Kinetic Interface for Visualizing High Dimensional Single-Cell Expression Data." Edited by Bonnie Berger. *Bioinformatics* 34 (7): 1246–48. https://doi.org/10.1093/bioinformatics/btx792.

Weksberg, David C, Stuart M Chambers, Nathan C Boles, and Margaret A Goodell. 2008. “CD150- Side Population Cells Represent a Functionally Distinct Population of Long-Term Hematopoietic Stem Cells.” *Blood* 111 (4): 2444–51. https://doi.org/10.1182/blood-2007-09-115006.

Wilson, Anne, Elisa Laurenti, Gabriela Oser, Richard C van der Wath, William Blanco-Bose, Maike Jaworski, Sandra Offner, et al. 2008. “Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair.” *Cell* 135 (6): 1118–29. https://doi.org/10.1016/j.cell.2008.10.048.

Wilson, Nicola K., David G. Kent, Florian Buettner, Mona Shehata, Iain C. Macaulay, Fernando J. Calero-Nieto, Manuel Sánchez Castillo, et al. 2015. “Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations.” *Cell Stem Cell* 16 (6): 712–24. https://doi.org/10.1016/j.stem.2015.04.004.

Wilson, Nicola K, Fernando J Calero-Nieto, Rita Ferreira, and Berthold Göttgens. 2011. “Transcriptional Regulation of Haematopoietic Transcription Factors.” *Stem Cell Research & Therapy* 2 (1): 6. https://doi.org/10.1186/scrt47.

Wu, Bing, Yunqi Wang, Chaojun Wang, Gang Greg Wang, Jie Wu, and Yisong Y Wan. 2016. “BPTF Is Essential for T Cell Homeostasis and Function.” *Journal of Immunology (Baltimore, Md. : 1950)* 197 (11): 4325–33. https://doi.org/10.4049/jimmunol.1600642.

Wu, Jia Qian, Montrell Seay, Vincent P Schulz, Manoj Hariharan, David Tuck, Jin Lian, Jiang Du, et al. 2012. “Tcf7 Is an Important Regulator of the Switch of Self-Renewal and Differentiation in a Multipotential Hematopoietic Cell Line.” *PLoS Genetics* 8 (3): e1002565. https://doi.org/10.1371/journal.pgen.1002565.

Wu, Thomas D, and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics (Oxford, England)* 26 (7): 873–81. https://doi.org/10.1093/bioinformatics/btq057.

Xu, Huilei, Yen-Sin Ang, Ana Sevilla, Ihor R Lemischka, and Avi Ma’ayan. 2014. “Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells.” *PLoS Computational Biology* 10 (8): e1003777. https://doi.org/10.1371/journal.pcbi.1003777.

Yamamoto, Ryo, Yohei Morita, Jun Ooehara, Sanae Hamanaka, Masafumi Onodera, Karl Lenhard Rudolph, Hideo Ema, and Hiromitsu Nakauchi. 2013. “Clonal Analysis Unveils Self-Renewing Lineage-Restricted Progenitors Generated Directly from Hematopoietic Stem Cells.” *Cell* 154 (5): 1112–26. https://doi.org/10.1016/J.CELL.2013.08.007.

Yang, Dan, Xiangzhong Zhang, Yong Dong, Xiaofei Liu, Tongjie Wang, Xiaoshan Wang, Yang Geng, et al. 2015. “Enforced Expression of Hoxa5 in Haematopoietic Stem Cells Leads to Aberrant Erythropoiesis in Vivo.” *Cell Cycle* 14 (4): 612–20. https://doi.org/10.4161/15384101.2014.992191.

Yang, Jiyeon, Lixiao Zhang, Caijia Yu, Xiao-Feng Yang, and Hong Wang. 2014. “Monocyte and Macrophage Differentiation: Circulation Inflammatory Monocyte as Biomarker for Inflammatory Diseases.” *Biomarker Research* 2 (1): 1. https://doi.org/10.1186/2050-7771-2-1.

Yarden, Y, W J Kuang, T Yang-Feng, L Coussens, S Munemitsu, T J Dull, E Chen, J Schlessinger, U Francke, and A Ullrich. 1987. “Human Proto-Oncogene c-Kit: A New Cell Surface Receptor Tyrosine Kinase for an Unidentified Ligand.” *The EMBO Journal* 6 (11): 3341–51.

Yeoh, Joyce S.G., Ronald van Os, Ellen Weersing, Albertina Ausema, Bert Dontje, Edo Vellenga, and Gerald de Haan. 2006. “Fibroblast Growth Factor-1 and -2 Preserve Long-Term Repopulating Ability of Hematopoietic Stem Cells in Serum-Free Cultures.” *STEM CELLS* 24 (6): 1564–72. https://doi.org/10.1634/stemcells.2005-0439.

Yoshihara, Hiroki, Fumio Arai, Kentaro Hosokawa, Tetsuya Hagiwara, Keiyo Takubo, Yuka Nakamura, Yumiko Gomei, et al. 2007. “Thrombopoietin/MPL Signaling Regulates Hematopoietic Stem Cell Quiescence and Interaction with the Osteoblastic Niche.” *Cell Stem Cell* 1 (6): 685–97. https://doi.org/10.1016/j.stem.2007.10.020.

Zeng, Hui, Raif Yücel, Christian Kosan, Ludger Klein-Hitpass, and Tarik Möröy. 2004. “Transcription Factor Gfi1 Regulates Self-Renewal and Engraftment of Hematopoietic Stem Cells.” *The EMBO Journal* 23 (20): 4116–25. https://doi.org/10.1038/sj.emboj.7600419.

Zerbino, Daniel R, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46 (D1): D754–61. https://doi.org/10.1093/nar/gkx1098.

Zhang, Su-Jiang, Li-Yuan Ma, Qiu-Hua Huang, Guo Li, Bai-Wei Gu, Xiao-Dong Gao, Jing-Yi Shi, et al. 2008. “Gain-of-Function Mutation of GATA-2 in Acute Myeloid Transformation of Chronic Myeloid Leukemia.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (6): 2076–81. https://doi.org/10.1073/pnas.0711824105.

Zhang, Xiao-Hui, Louis Y Tee, Xiao-Gang Wang, Qun-Shan Huang, and Shi-Hua Yang. 2015. “Off-Target Effects in CRISPR/Cas9-Mediated Genome Engineering.” *Molecular Therapy - Nucleic Acids* 4 (January): e264. https://doi.org/10.1038/MTNA.2015.37.

Zhang, Zhenyue, and Jing Wang. 2006. “MLLE: Modified Locally Linear Embedding Using Multiple Weights.” *Proceedings of the 19th International Conference on Neural Information Processing Systems*. MIT Press.

Zhao, D. M., S. Yu, X. Zhou, J. S. Haring, W. Held, V. P. Badovinac, J. T. Harty, and H.-H. Xue. 2010. “Constitutive Activation of Wnt Signaling Favors Generation of Memory CD8 T Cells.” *The Journal of Immunology* 184 (3): 1191–99. https://doi.org/10.4049/jimmunol.0901199.

Zhao, Jingyao, Xufeng Chen, Guangrong Song, Jiali Zhang, Haifeng Liu, and Xiaolong Liu. 2017. “Uhrf1 Controls the Self-Renewal versus Differentiation of Hematopoietic Stem Cells by Epigenetically Regulating the Cell-Division Modes.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (2): E142–51. https://doi.org/10.1073/pnas.1612967114.

Zhao, Ran, Bu Young Choi, Mee-Hyun Lee, Ann M. Bode, and Zigang Dong. 2016. “Implications of Genetic and Epigenetic Alterations of CDKN2A (P16INK4a) in Cancer.” *EBioMedicine* 8 (June): 30–39. https://doi.org/10.1016/J.EBIOM.2016.04.017.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049. https://doi.org/10.1038/ncomms14049.

Zohren, Fabian, George P Souroullas, Min Luo, Ulrike Gerdemann, Maria R Imperato, Nicola K Wilson, Berthold Göttgens, Georgi L Lukov, and Margaret A Goodell. 2012. “The Transcription Factor Lyl-1 Regulates Lymphoid Specification and the Maintenance of Early T Lineage Progenitors.” *Nature Immunology* 13 (8): 761–69. https://doi.org/10.1038/ni.2365.

Zon, LI, Y Yamaguchi, K Yee, EA Albee, A Kimura, JC Bennett, SH Orkin, and SJ Ackerman. 1993. “Expression of MRNA for the GATA-Binding Proteins in Human Eosinophils and Basophils: Potential Role in Gene Transcription.” *Blood* 81 (12).

Zou, Peng, Hiroki Yoshihara, Kentaro Hosokawa, Ikue Tai, Kaori Shinmyozu, Fujiko Tsukahara, Yoshiro Maru, Keiko Nakayama, Keiichi I Nakayama, and Toshio Suda. 2011. “P57(Kip2) and P27(Kip1) Cooperate to Maintain Hematopoietic Stem Cell Quiescence through Interactions with Hsc70.” *Cell Stem Cell* 9 (3): 247–61. https://doi.org/10.1016/j.stem.2011.07.003.