

Molecular characterisation of *ASXL1*-mutant clonal haematopoiesis



Angus John Groom
St. Catherine's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2024

Acknowledgements

I am grateful for the support and scientific mentorship of my supervisor Professor Paresh Vyas throughout the duration of my DPhil. I would like to thank all the members of the Vyas lab past and present spanning my time as part of the group for their scientific input, advice, support, and feedback. In particular, I would like to thank Sven Turkalj and Dr Asger Jakobsen from whom I learned a lot of molecular biology, experimental design, methodology when I first joined the lab as somewhat of a novice. Sven's eagerness to help me by discussing my scientific ideas, or help with laborious experiments particularly those described in chapter 3, has made my DPhil an incredibly enjoyable experience. Asger in particular I owe a lot of gratitude to- from setting up the MARCH cohort (without which this study would not be possible), to his generosity in sharing his data and bioinformatic code. I am incredibly thankful. Both have been inspirational scientists to work with.

From the Vyas lab I would also like to thank Dr David Cruz Hernandez for his wealth of knowledge of molecular biology, and help in designing my CUT&Tag experiments. Special thanks must also be given to Dr Felix Radtke for his advice and expertise in bioinformatic analysis, and for his willingness to help with experiments. I would also like to acknowledge Simone Riva from the Hughes lab for his help troubleshooting bioinformatic analysis issues. From the WIMM core facilities I would like to acknowledge the Flow Cytometry Facility staff in particular for their help with cell sorting, and the staff of the Single-Cell Facility for support with my experiments.

Additionally I am incredibly grateful to my secondary supervisor, Assoc. Professor Adam Wilkinson for his scientific insight and support throughout the project, and has been a fantastic mentor.

I would like to acknowledge the Lady Tata Memorial Charitable Trust for funding my DPhil.

In particular note I would like to thank my family who have been an enormous source of support throughout my DPhil. Most importantly I would like to thank my wife Eleanor for her love, and patience for my decision to undertake this DPhil. I am immeasurably thankful for her support throughout these testing three years.

Finally, I would like to thank the individual patients of the MARCH cohort, without whom this project would not have been possible.

Abstract

Adult stem cells accumulate somatic mutations throughout life as a result of failure to repair DNA damage. If a somatic mutation confers a fitness advantage upon the stem cell, the resulting clone can expand to be over-represented in that tissue. In the haematopoietic system, this process is known as clonal haematopoiesis (CH). Whilst CH is a common condition amongst older healthy individuals, it has been associated with a greater risk of developing leukaemia. One of the genes that is most frequently found mutated in CH and myeloid malignancies is *ASXL1* - a regulator of histone modification distribution. Mutations in *ASXL1* are associated with poor prognosis of blood cancers, and are likely to be acquired years before the onset of disease. Thus far, the mechanisms by which *ASXL1* confers a clonal expansion upon haematopoietic stem and progenitor cells are poorly understood. Identification of these mechanisms could reveal novel therapeutic strategies for blood cancer.

In this study, the cellular and molecular foundations of *ASXL1*-mutant clonal haematopoiesis in humans are characterised. By applying TARGET-seq+ to 4 *ASXL1*- mutant CH samples and 5 age-matched healthy control samples, this study reveals that the *ASXL1*-mutant clone follows broadly similar differentiation trajectories to healthy cells. However, the *ASXL1*-mutant clone is greatly expanded within the HSC compartment. Additionally, mutant cells show an inefficiency of terminal maturation of the erythroid and lymphoid lineages. Compared to wildtype HSCs within the CH bone marrow, mutant HSCs exhibit a transcriptional shift towards a less-dormant, differentiation-primed state, and an upregulation of polycomb target genes. Interestingly, unlike recent reports linking inflammation to clonal expansion in CH, no significant difference in transcriptional response to inflammation was seen between *ASXL1*-mutant and wildtype HSCs. These results suggest a distinct mechanism of clonal expansion for *ASXL1*-mutant cells.

Here, I also present proof-of-concept results for a novel multiomic protocol-GCUT- capable of simultaneously profiling epigenomic component distribution and genotype of single cells. Through a cell mixing experiment, cell lines were clustered on the basis of histone mark distribution, and identified by high-fidelity genotyping of single nucleotide variants. Future use of this protocol on cells from human *ASXL1*-mutant CH samples may identify the direct genomic targets of mutant *ASXL1*, and reveal novel insights into mechanisms of clonal expansion.

Word Count: 44,600

Contents

List of Figures	ix
List of Tables	xii
List of Abbreviations	xiii
1 Introduction	1
1.1 Clonal haematopoiesis in disease and ageing	1
1.1.1 Impact of somatic mutations as a consequence of ageing . .	1
1.1.2 Overview and history of clonal haematopoiesis in humans . .	3
1.1.3 Common genetic drivers of clonal haematopoiesis of indeter- minate potential (CHIP)	5
1.1.4 Clonal dynamics throughout the human life cycle	8
1.1.5 Clonal haematopoiesis and non-haematological disease . . .	9
1.1.6 Clonal haematopoiesis and haematological malignancy . . .	10
1.2 Epigenetic control of gene expression	12
1.3 Polycomb group and Trithorax group activity and role in haematopoiesis	13
1.4 The structure and function of ASXL proteins	16
1.4.1 ASXL1 as part of the PR-DUB complex	18
1.5 The impact of mutations in <i>ASXL1</i>	19
1.5.1 <i>ASXL</i> gene family mutations in non-haematological disease .	19
1.5.2 <i>ASXL1</i> mutations in myeloid malignancy and clonal haematopoiesis	20
1.5.3 Modelling the impact of <i>ASXL1</i> mutations on haematopoiesis	21
1.6 Potential mechanisms of clonal expansion in <i>ASXL1</i> -mutant CH and open questions	26
1.7 Aims of this thesis	29
2 Materials and Methods	31
2.1 Study participants and ethical approval	32
2.2 Sample collection, processing, targeted sequencing and variant calling	32
2.3 Flow cytometry analysis and cell sorting	33
2.3.1 Primary human bone marrow thawing and staining	33
2.3.2 Primary human bone marrow single cell sorting	34
2.4 TARGET-Seq+ primer design	34
2.5 TARGET-Seq+ genotyping primer testing	36

2.6	TARGET-Seq+ library preparation	39
2.6.1	Lysis buffer preparation	39
2.6.2	Reverse transcription and pre-amplification	39
2.6.3	Whole-transcriptome library preparation and sequencing	40
2.6.4	Targeted single-cell genotyping, library preparation, QC, and sequencing	42
2.7	Single-cell genotyping analysis	43
2.7.1	Data pre-processing	43
2.7.2	Mutation calling	44
2.7.3	Estimation of allelic dropout rate	45
2.7.4	Clonal hierarchy assignment	46
2.8	FACS Index data analysis	47
2.9	Single-cell transcriptome analysis	47
2.9.1	Data pre-processing	47
2.9.2	Quality control, normalization and highly variable gene identification	48
2.9.3	Dimensionality reduction, clustering, and data integration	49
2.9.4	Cluster annotation and reference map projection	50
2.9.5	Differential abundance analysis	51
2.9.6	Differential gene expression analysis and gene set enrichment analysis	51
2.9.7	HSPC subclustering	52
2.9.8	Differential regulon analysis	53
2.10	GCUT library preparation	53
2.10.1	GCUT genotyping primer design	53
2.10.2	Cell line culture	54
2.10.3	Lysis buffer preparation	54
2.10.4	Bulk CUT&Tag of cell lines	55
2.10.5	Single nuclei sorting	55
2.10.6	Preamplification of GCUT libraries	56
2.10.7	Single-cell CUT&Tag library preparation and sequencing	56
2.10.8	Targeted single cell genotyping	57
2.11	GCUT optimisation	57
2.11.1	Primer concentration titration	57
2.11.2	Comparison of different bead purification strategies	57
2.12	GCUT data analysis	58
2.12.1	CUT&Tag data pre-processing	58
2.12.2	Quality control and filtering	58
2.12.3	Normalisation, feature selection, and dimensionality reduction	59
2.12.4	GCUT genotyping sequencing data analysis	59
2.13	Figure production	59

3	Single-cell genotyping and immunophenotyping of <i>ASXL1</i>-mutant clonal haematopoiesis	60
3.1	Brief overview and aims for this chapter	60
3.2	Approach to studying <i>ASXL1</i> -mutant clonal haematopoiesis in humans	61
3.3	Brief introduction to TARGET-Seq+	63
3.4	Introduction to the MARCH cohort and sample selection	65
3.5	Experimental design	68
3.6	Single-cell genotyping analysis of <i>ASXL1</i> -mutant CH samples . . .	70
3.6.1	Data QC and filtering	70
3.6.2	Mutation calling	71
3.6.3	Allelic dropout rate estimation	73
3.7	Clonal and immunophenotypic landscape of haematopoiesis in <i>ASXL1</i> -mutant CH samples	76
3.7.1	NOC008	77
3.7.2	NOC082	79
3.7.3	NOC137	81
3.7.4	NOC180	81
3.8	Chapter conclusions	84
4	Single-cell transcriptome landscape of <i>ASXL1</i>-mutant CH	86
4.1	Main aims for the transcriptomic analyses	86
4.2	Transcriptomic landscape of haematopoiesis in healthy and <i>ASXL1</i> -mutant CH bone marrow at the single-cell level	88
4.2.1	Data QC and filtering	88
4.2.2	Dimensionality reduction and clustering	90
4.2.3	Integration of CH and wildtype control samples	93
4.3	Cell type annotation	96
4.3.1	Cluster labelling	96
4.3.2	Cell type annotation validation using reference map projection	101
4.3.3	Correlation of single-cell transcriptome profiles using different cell type annotations	104
4.4	Analysis of <i>ASXL1</i> -mutant clonal size throughout differentiation . .	105
4.5	Differential abundance analysis of the <i>ASXL1</i> -mutant clone	109
4.6	Chapter conclusions	111

5	Transcriptomic consequences of <i>ASXL1</i> mutation	112
5.1	Analysis strategy, hypotheses for mechanisms of clonal expansion, and aims for this chapter	112
5.2	Non-cell autonomous mechanisms of clonal expansion within the HSC	115
5.2.1	Transcriptomic differences between <i>ASXL1</i> -mutant and non-CH WT HSC/MPP	116
5.2.2	Transcriptomic differences between WT-CH and non-CH WT HSC/MPP	118
5.3	Cell-intrinsic mechanisms of clonal expansion within the HSC/MPP compartment	121
5.4	Identification of subpopulations within the HSC/MPP compartment	125
5.5	Patterns of clonal dynamics in HSPC subpopulations	128
5.6	Cell-intrinsic mechanisms of clonal expansion within HSC subpopulations	130
5.6.1	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH cells in HSC2	130
5.6.2	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH cells in HSC1	134
5.7	Gene regulatory network inference analysis of HSC1	138
5.8	Consequences of <i>ASXL1</i> mutation upon the transcriptome of HSPCs	140
5.8.1	Differential regulon analysis of haematopoietic progenitor compartments	140
5.9	Concluding remarks	144
6	GCUT: a novel method for simultaneous epigenomic profiling and targeted genotyping of single cells	146
6.1	Rationale for the development of GCUT	147
6.2	Review of CUT&Tag chemistry	150
6.3	Adaptation of CUT&Tag to a plate-based single-cell protocol	152
6.4	Development and optimisation of GCUT	156
6.4.1	Aims for benchmarking experiments	156
6.4.2	Modification of CUT&Tag for use in single-cell methods . .	156
6.4.3	Bulk CUT&Tag in human cell lines	157
6.4.4	Development of plate-based single-cell CUT&Tag with genotyping	159
6.5	Benchmarking of GCUT with a cell mixing sequencing experiment .	161
6.5.1	Overview of cell mixing experiment	161
6.5.2	CUT&Tag library preparation of cell lines	162
6.5.3	Genotyping of cell lines	164

6.5.4	Chromatin data pre-processing and QC	168
6.5.5	Dimensionality reduction and clustering	170
6.5.6	CUT&Tag library quality with and without genotyping . . .	175
6.5.7	Integration of genotyping and chromatin data	178
6.5.8	Genomic distribution of histone marks of from GCUT CUT&Tag data	182
6.6	Concluding remarks	184
7	Discussion	185
7.1	The cellular landscape of <i>ASXL1</i> -mutant CH	187
7.2	The impact of <i>ASXL1</i> mutations on the transcriptome of human HSPCs	190
7.3	GCUT: a novel single-cell method to simultaneously assay genotype and histone modification distribution in heterogeneous cell populations	196
7.4	Study limitations and future work	200
Appendices		
A	Appendices	204
A.1	<i>ASXL1</i> -mutant samples of the MARCH cohort.	204
A.2	Antibodies	206
A.3	TARGET-seq+ reagents	207
A.4	GCUT reagents	209
A.5	TARGET-seq+ oligonucleotides	212
A.6	GCUT oligonucleotides	214
A.7	Marker genes of HSC characteristics and subpopulations	216
References		220

List of Figures

1.1	Clonal expansion as a result of somatic mutation	3
1.2	Structure of ASXL1	17
1.3	Putative function of mutant ASXL1	28
2.1	Genotyping primer design	35
2.2	Genotyping primer testing using bulk gDNA or cDNA	37
2.3	Single-cell cDNA library formation	37
2.4	Single-cell genotyping amplicon testing	38
3.1	Overview of the TARGET-seq protocol	64
3.2	Overview of the MARCH cohort	66
3.3	Experimental strategy for single-cell analysis of <i>ASXL1</i> -mutant CH samples.	69
3.4	FACS sorting strategy	70
3.5	Single-cell genotyping quality control	72
3.6	Single-cell genotyping mutation calling	74
3.7	Correlation of single-cell and bulk genotyping	76
3.8	Frequency of <i>ASXL1</i> -mutant cells in HSPC compartments from NOC008	78
3.9	Frequency of <i>ASXL1</i> -mutant cells in HSPC compartments from NOC082	80
3.10	Frequency of <i>ASXL1</i> -mutant cells in HSPC compartments from NOC137	82
3.11	Frequency of <i>ASXL1</i> -mutant cells in HSPC compartments from NOC180	83
4.1	Single-cell transcriptome quality control.	89
4.2	Quality control metrics of filtered single-cell transcriptome libraries.	91
4.3	Dimensionality reduction of the combined dataset.	92
4.4	Harmony integration of <i>ASXL1</i> -mutant CH and non-CH samples.	94
4.5	Cluster composition by immunophenotype of integrated dataset.	95
4.6	Gene set enrichment analysis of clusters for published haematopoietic lineage signatures.	97
4.7	Cluster marker gene expression.	99
4.8	Symphony projection onto a reference map of CH haematopoiesis.	102

4.9	Unsupervised and supervised cell type annotations.	103
4.10	Correlation analysis of unsupervised and supervised cell type annotations.	106
4.11	Distribution of the <i>ASXL1</i> -mutant clone throughout haematopoietic differentiation.	108
4.12	Differential abundance of the <i>ASXL1</i> -mutant clone throughout haematopoietic differentiation.	110
5.1	Overview of mechanisms of clonal advantage and differential gene expression analysis strategy.	114
5.2	Differential gene expression analysis between <i>ASXL1</i> -mutant and non-CH HSC/MPPs.	117
5.2	Differential gene expression analysis between <i>ASXL1</i> -mutant and non-CH HSC/MPPs.	118
5.3	Differential gene expression analysis between WT-CH and non-CH HSC/MPPs.	120
5.3	Differential gene expression analysis between WT-CH and non-CH HSC/MPPs.	121
5.4	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH HSC/MPPs.	122
5.4	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH HSC/MPPs.	123
5.5	Identification of HSPC subpopulations.	126
5.6	Composition of HSPC subclusters.	127
5.7	Clonal composition of HSPC subclusters.	129
5.8	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH HSC2 subpopulations.	131
5.8	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH HSC2 subpopulations.	132
5.9	Heightened inflammatory transcriptomic state of <i>ASXL1</i> -mutant HSC2.	133
5.10	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH HSC1 subpopulation.	136
5.11	Gene set enrichment analysis between <i>ASXL1</i> -mutant and WT-CH HSC1 subpopulation.	137
5.12	Differentially active gene regulatory network analysis between <i>ASXL1</i> -mutant and WT-CH in the HSC1 subpopulation.	139
5.13	Differentially active gene regulatory network analysis between <i>ASXL1</i> -mutant and WT-CH in the LMPP and late GMP populations. . . .	141

5.14	Differentially active gene regulatory network analysis between <i>ASXL1</i> -mutant and WT-CH in the EryP and Pro-B populations.	142
6.1	Overview of the CUT&Tag protocol.	151
6.2	Overview of the GCUT protocol.	155
6.3	Bulk CUT&Tag testing in human cell lines.	158
6.4	GCUT single-cell testing in human cell lines.	160
6.4	GCUT single-cell testing in human cell lines.	161
6.5	GCUT cell mixing experiment overview.	163
6.6	Purified CUT&Tag libraries from GCUT cell mixing experiment. . .	164
6.7	GCUT cell mixing experiment genotyping quality control.	166
6.7	GCUT cell mixing experiment genotyping quality control.	167
6.8	Estimation of allelic dropout rate for GCUT genotyping amplicons. .	168
6.9	Single-cell CUT&Tag library sequencing quality metrics.	169
6.10	Comparison of quality control thresholds for H3K27me3 single-cell CUT&Tag libraries.	172
6.11	Quality metrics of different UNF thresholds for H3K27me3 single-cell CUT&Tag libraries.	173
6.12	Comparison of different UNF thresholds for H2AK119ub single-cell CUT&Tag libraries.	174
6.13	Comparison of different UNF thresholds for H3K4me3 single-cell CUT&Tag libraries.	175
6.14	Single-cell CUT&Tag library quality sequencing metrics with and without genotyping.	177
6.15	Integration of single-cell genotyping and CUT&Tag data for H3K27me3. .	179
6.16	Integration of single-cell genotyping and CUT&Tag data for H2AK119ub. .	180
6.17	Integration of single-cell genotyping and CUT&Tag data for H3K4me3. .	181
6.18	Genomic distribution of histone marks of from GCUT CUT&Tag data. .	183

List of Tables

3.1	Selected <i>ASXL1</i> -mutant samples analysed by TARGET-seq+ . . .	67
4.1	Composition of the entire dataset by sequencing run and sample . .	88
A.1	<i>ASXL1</i> -mutant samples of the MARCH cohort.	205
A.2	Antibodies used for immunophenotypic analysis and FACS sorting of BMMNCs.	206
A.3	Composition of lysis buffer used for TARGET-seq+.	207
A.4	Composition of reverse transcription reaction used for capture of single-cell whole-transcriptome libraries.	207
A.5	Composition of preamplification reaction of whole-transcriptome libraries and genotyping amplicons from single cells.	208
A.6	List of reagents used for the CUT&Tag protocol in GCUT.	209
A.7	Composition of buffers used in the CUT&Tag reactions for GCUT.	210
A.8	Reagents used in lysis buffer for GCUT.	210
A.9	Reagents used in pre-amplification reaction in GCUT.	211
A.10	List of oligonucleotides used throughout TARGET-seq+ experiments.	212
A.11	Genotyping primers used in the RT-PCR and pre-amplification reactions of TARGET-seq+.	212
A.12	Barcoded primers used in the PCR1 genotyping reactions of TARGET- seq+.	213
A.13	Custom i5-indexed primers included in the lysis buffer for GCUT. .	214
A.14	Genotyping primers used in the pre-amplification reactions of GCUT cell mixing experiment.	214
A.15	Barcoded primers used in the PCR1 genotyping reactions of the GCUT cell mixing experiment.	215
A.16	List of genes used to define different HSC characteristics.	216
A.17	HSC subpopulation marker genes.	217
A.18	Differentially expressed genes in <i>ASXL1</i> -mutant HSC1.	218
A.19	Differentially expressed genes in <i>ASXL1</i> -mutant HSC2.	219

List of Abbreviations

AA		Amino acid.
ADO		Allelic dropout.
AML		Acute myeloid leukaemia.
<i>ASXL1</i>		Additional sex combs-like 1.
ATAC-seq		Assay for Transposase Accessible Chromatin and sequencing.
<i>BAP1</i>		BRCA1-associated protein 1.
BMMNC		Bone marrow mononuclear cell.
bp		Base pair.
CCUS		Clonal cytopoenia of undetermined significance.
cDNA		Complementary DNA.
CH		Clonal haematopoiesis.
CHD		Coronary heart disease.
CHIP		Clonal haematopoiesis of indeterminate potential.
CMML		Chronic myelomonocytic leukaemia.
CMP		Common myeloid progenitor.
CNV		Copy number variant.
CUT&Tag		Cleavage under targets and tagmentation.
DGEA		Differential gene expression analysis.
DNA		Deoxyribonucleic acid.
<i>DNMT3A</i>		DNA methyltransferase 3A.
ERCC		External RNA controls consortium.
EryP		Erythroid progenitor.
FACS		Fluorescence activated cell sorting.
FDR		False discovery rate.
FMO		Fluorescence minus-one.
FRiP		Fraction of reads in peaks.
GCUT		Genotyping with Cleavage Under Targets and tagmentation.

gDNA		Genomic DNA.
GMP		Granulocyte-monocyte progenitor.
GoT		Genotyping of transcriptomes.
GoTChA		Genotyping of targets with single-cell chromatin accessibility.
GSEA		Gene set enrichment analysis.
GTAC		Genotyping with Assay for Transposase-Accessible Chromatin.
Hox		Homeobox genes.
HSC		Haematopoietic stem cell.
HSPC		Haematopoietic stem and progenitor cells.
indel		Insertion/deletion (mutation).
JMML		Juvenile myelomonocytic leukaemia.
kDa		Kilo Daltons.
Lin		Lineage.
LSK		Lineage- Sca1+ cKit+ cells.
LT-HSC		Long-term haematopoietic stem cell.
MACS		Magnetic activated cell sorting.
mCA		Mosaic Chromosomal alteration.
MDS		Myelodysplastic syndrome.
MEP		Megakaryocyte erythroid progenitor.
MkP		Megakaryocyte progenitor.
MLP		Multi-lymphoid progenitor.
MPN		Myeloproliferative neoplasm.
MPP		Multipotent progenitor.
PB		Peripheral blood.
PCA		Principal component analysis.
PcG		Polycomb group.
PCR		Polymerase chain reaction.
PR-DUB		Polycomb-repressive deubiquitinase complex.
PRC		Polycomb repressive complex.
Pre-B		B cell precursors.
Pre-Pro-B		Pre-B cell progenitors.

Pro-B		B cell progenitors.
RNA		Ribonucleic acid.
RNA-Seq		RNA sequencing.
RT-PCR		Reverse transcription polymerase chain reaction.
sc		Single cell.
SNP		Single nucleotide polymorphism.
SNV		Single nucleotide variant.
ST-HSC		Short-term haematopoietic stem cell.
<i>TET2</i>		Ten-eleven translocase methylcytosine dioxygenase 2.
TrxG		Trithorax group.
UMAP		Uniform manifold approximation and projection.
UNF		Unique nuclear fragments.
VAF		Variant allele frequency.
WT		Wild type.

Chapter 1

Introduction

Contents

1.1	Clonal haematopoiesis in disease and ageing	1
1.1.1	Impact of somatic mutations as a consequence of ageing	1
1.1.2	Overview and history of clonal haematopoiesis in humans	3
1.1.3	Common genetic drivers of clonal haematopoiesis of indeterminate potential (CHIP)	5
1.1.4	Clonal dynamics throughout the human life cycle	8
1.1.5	Clonal haematopoiesis and non-haematological disease .	9
1.1.6	Clonal haematopoiesis and haematological malignancy .	10
1.2	Epigenetic control of gene expression	12
1.3	Polycomb group and Trithorax group activity and role in haematopoiesis	13
1.4	The structure and function of ASXL proteins	16
1.4.1	ASXL1 as part of the PR-DUB complex	18
1.5	The impact of mutations in <i>ASXL1</i>	19
1.5.1	<i>ASXL</i> gene family mutations in non-haematological disease	19
1.5.2	<i>ASXL1</i> mutations in myeloid malignancy and clonal haematopoiesis	20
1.5.3	Modelling the impact of <i>ASXL1</i> mutations on haematopoiesis	21
1.6	Potential mechanisms of clonal expansion in <i>ASXL1</i>-mutant CH and open questions	26
1.7	Aims of this thesis	29

1.1 Clonal haematopoiesis in disease and ageing

1.1.1 Impact of somatic mutations as a consequence of ageing

One of the hallmarks of cellular ageing is the accumulation of deoxyribonucleic acid (DNA) damage, stemming from exposure to a range of exogenous and endogenous sources throughout the life of an individual [1–3]. Exogenous sources of DNA damage include exposure to ultraviolet light, ionizing radiation, or environmental carcinogens such as those present in inhaled tobacco smoke [4, 5]. Common endogenous sources of DNA damage include the accumulation of reactive oxygen species, spontaneous

deamination of 5-methylcytosine, and mitotic error [6, 7]. Failure of endogenous DNA damage repair mechanisms to restore the original genetic code before DNA replication can result in acquisition of a variety of genetic mutations. These can include single nucleotide variations (SNVs), insertions or deletions (indels), copy number variations (CNVs), or large structural rearrangements [8–10]. Over the lifetime of an individual, this process can result in the development of a rich mosaic of genetic mutations.

Whilst the overwhelming majority of somatic mutations have little to no impact on the ability of a cell to survive, occasionally a mutation can occur that confers a fitness advantage. If this happens in a stem cell, Darwinian principles of selection can cause outgrowth of that stem cell and its progeny. Over time, the resulting clone may continue to grow as a result of selection, or the clone size may fluctuate due to genetic drift. Most of the time these mutations confer negligible functional effect upon the host tissue, but they may also serve as the substrate on which cancer can develop. This process of clonal outgrowth as a result of somatic mutation has been observed in multiple healthy bodily tissues [11].

The principles of clonal selection underlie the classical view of cancer development [12]. Cancerous lesions are thought to result from acquisition of consecutive mutations conferring a competitive and growth advantage. Previously, it has been difficult to accurately assess the genetic evolution of cancer-propagating clones, however recent advances in DNA sequencing have enabled researchers to map the clonal architecture of cancerous lesions [13–16]. Still, the molecular basis of clonal selection, particularly in the pre-cancerous stages remains poorly understood. Investigation into the molecular mechanisms by which cancer-associated mutations cause clonal selection will not only aid in the identification of those pre-cancerous lesions most likely progress to malignancy, but may also reveal novel therapeutic strategies to prevent progression to cancer entirely.

Despite also being observed in numerous bodily tissues [3, 17–20], clonal selection has been best described within the haematopoietic system [21–26]. Haematopoietic stem cells (HSCs) sit at the top of the differentiation hierarchy and have the potential

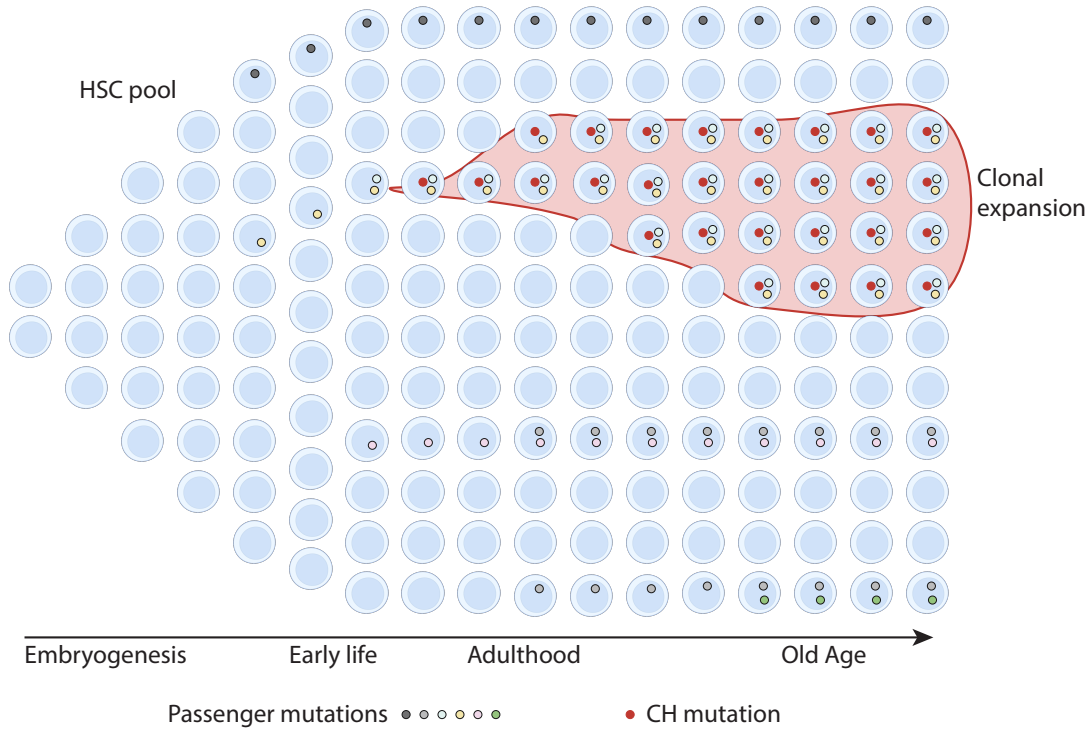


Figure 1.1: Clonal selection of haematopoietic stem cells. Somatic mutations accumulate in stem cells throughout life. Most acquired mutations have a neutral fitness affect upon stem cells. However, if an acquired somatic mutation confers a selective advantage upon the stem cell, the resulting "clone" can grow to be over-represented within the stem cell pool. Accumulation of fitness-conferring mutations is thought to be the underlying mechanism of cancer development. This initial process of clonal expansion in the haematopoietic system is called clonal haematopoiesis.

to reconstitute every cell within the myeloid, lymphoid and megakaryocyte/erythroid lineages [27]. As HSCs divide and differentiate into mature blood cells, the progeny of each single stem cell exponentially increases. Thus, if a somatic mutation conferring selective advantage occurs within an HSC, the resulting mutant clone has the potential to be over-represented as a fraction of the entire haematopoietic system (Figure 1.1). This process has been observed as an age-related phenomenon in humans and has been termed clonal haematopoiesis (CH) [26].

1.1.2 Overview and history of clonal haematopoiesis in humans

Mutations in known driver genes of leukaemia in a haematopoietic stem or progenitor cell (HSPC) can cause preferential expansion of the mutant clone, leading to over-

representation within the bone marrow. Blood cancers were first identified as clonal diseases in 1973, through karyotypic analysis of human chronic myeloid leukaemia samples [21]. However, the first evidence of clonal haematopoiesis in people without haematopoietic malignancies came in the 1990s from studies of non-random X chromosome inactivation in females [22].

Throughout female embryonic development, either the maternal or paternal X chromosome is randomly inactivated in each cell, resulting in a mosaic of two cell populations expressing genes from separate X chromosomes [28]. Through chance, the ratio of maternal:paternal activated X chromosome should be 1:1, however proof of non-random X inactivation estimated the ratio to be over 3:1, suggesting the possibility of clonality [23, 29]. An important observation from these studies was the increased incidence of X-chromosome inactivation in blood cells with age, particularly in the myeloid compartment [29]. However, evidence of the mechanism underlying selection was unclear. Two decades later, *TET2* mutations were discovered in $\approx 5\%$ of elderly females with non-random X-chromosome inactivation [30]. This discovery, combined with the high prevalence of *TET2* mutations in myeloid malignancy [31], provided the first link between acquired somatic mutation and clonal haematopoiesis in healthy individuals.

Further evidence of haematopoietic clonal expansion comes from the identification of pre-leukaemic clones in patients with acute myeloid leukaemia (AML) in disease remission [32, 33]. AML is a clonal haematological malignancy that generally results from the sequential acquisition of somatic mutations within the same founding clone. In an AML patient who was in disease remission, the founding mutation- *AML1/ETO* translocation- was identified in a subset of phenotypically normal HSCs [32]. Unlike *de novo* AML, secondary AML is a consequence of progression from a myelodysplasia such as myelodysplastic syndrome (MDS) or myeloproliferative neoplasm (MPN). In MDS, the founding mutations are usually found within the HSC compartment, whereas the transforming mutations for those cases that progress to AML are only found within the progenitor compartments [34–36]. Therefore, the presence of a founding clone within the HSC compartment in AML patients in

remission is evidence of an initial pre-leukaemic expansion. Further studies also identified HSCs in AML patients with only single pre-leukaemic mutations that were capable of full haematopoietic reconstitution upon transplant into immunodeficient mice [37–39]. This is evidence of initial clonal expansion prior to onset of malignancy.

Three important studies published in 2014 were among the first to identify pre-leukaemic mutations in patients without overt haematological malignancy using whole exome sequencing data from peripheral blood [24–26]. Over 30,000 patients across the three studies were selected specifically for the absence of haematological cancer and were recruited from genome wide association studies for diabetes, solid tumours, and schizophrenia. One important consideration for these studies was the use of peripheral blood. Any clone that causes a small selective advantage, or a neutral effect is likely not to result in a large clone size. Thus small clones would account for only a small proportion of mature blood cells, which constitute most of the peripheral blood. Therefore, only clones bearing mutations with significant selective advantage were likely to be detected.

Across the studies, three common observations were made. Firstly, very few genes bearing mutations were detected, and classical oncogenes were only detected in a very small number of patients. The vast majority of somatic mutations detected ($\approx 80\%$) were in 3 epigenetic regulators: *DNMT3A*, *TET2*, *ASXL1*. Secondly, clonal haematopoiesis was heavily associated with age- around 10-20% of patients over 70 years had CH, whereas it was detected in fewer than 1% of patients under the age of 40. Thirdly, CH was associated with a greater incidence of blood cancer, and age-related diseases such as coronary heart disease.

1.1.3 Common genetic drivers of clonal haematopoiesis of indeterminate potential (CHIP)

One of the most interesting observations from the original 3 studies was the restricted set of genes in which CH mutations occur. Subsequent studies have not only verified these, but have identified new drivers, largely due to the ability of more sensitive techniques than whole exome sequencing to detect smaller clones [40–48]. The vast majority of genetic drivers are genes that encode either epigenetic regulators or RNA

splicing factors. *DNMT3A* and *TET2*, which encode two enzymes with opposing functions for DNA methylation [49, 50], are consistently the most common two genes found mutated in CH, cumulatively accounting for nearly 75% of cases [24–26]. *ASXL1*, a gene that encodes a putative epigenetic regulator [51], is the third most common CH mutation, accounting for between 5-10% of cases. Splicing factors *SF3B1*, *SRSF2*, *PRPF8*, *U2AF1* are all genes frequently mutated in CH. Mutations in oncogenes involved in cellular growth signalling (e.g. *JAK2*, *GNB1*, *GNAS*, *CBL*) and DNA damage response (e.g. *TP53*, *PPM1D*) were also seen in smaller proportion of individuals with CH. Interestingly, of the *JAK2* variants detected in CH, the leukaemia-associated hotspot V617F variant was found almost exclusively.

Notably, transcription factors critical for haematopoietic lineage specification (e.g. *RUNX1*, *GATA2*, *CEBPA*, *NPM1*), which are recurrently found mutated in haematological malignancy (often as transforming mutations), are not found to be mutated in CH. This may suggest that mutations in these transcription factors do not confer fitness advantage to stem cells, but rather progenitor or precursor populations. Additionally, transcription factor mutations are perhaps more likely to bias or block differentiation trajectories, which is a hallmark of haematological cancers.

In addition to somatic variants, several germ line variants have been associated with greater risk of developing CH. Alleles which alter expression of genes involved in the maintenance of genome integrity (e.g. *TERT*, *ATM*, *CHEK2*, *PARP1*), have been associated with a higher risk of developing CH [52, 53]. Germ line variants conferring increased *TERT* expression, the telomerase reverse transcriptase, were associated with the greatest risk of developing CH with a driver mutation [52]. A recent study also associated a germ line polymorphism in the promoter of *TCL1A* with higher growth rate and increased likelihood of CH clone development with mutations in driver genes including *TET2* and *ASXL1*, but not *DNMT3A* [54]. The presence of the SNP was associated with higher expression levels of *TCL1A*. Interestingly, decreased expression of *TCL1A* was associated with *DNMT3A*-mutant CH. Although *TCL1A* is understudied in the context of HSC biology, it is found overexpressed in mature T-cell leukaemias [55, 56], and has been reported to enhance

kinase activity of AKT [57]. Mechanistically, the authors suggested that increased expression of *TCL1A* in CH preserves HSCs in a proliferative state through enhancing AKT-mediated FOXO transcription factor inhibition. Prolonged activation of FOXO target genes in response to cell stressors (e.g. heightened inflammation) have been previously reported to lead to cell cycle arrest or apoptosis [58]. Collectively, these studies suggest that the contextual genetic landscape is a determinant in whether a clone bearing a CH-mutation may grow to become detectable.

Mosaic chromosomal alterations (mCAs) are frequently observed in haematological malignancies and have also been found in patients with CH [59–65]. These have included deletions, duplications, and copy-neutral loss of heterozygosity. The prevalence of mCAs found in CH was found to increase with age, rising from <0.5% at patients under 50 years of age to 2-3% of patients older than 70. One study identified two-thirds of mCA cases had co-occurring mutations in genetic drivers of CH [62]. Certain deletion events identified in this cohort co-localised with genes found mutated in CH (*DNMT3A*, *TET2*, *JAK2*, *ATM*), which could indicate that haploinsufficiency of these genes is sufficient to cause clonal expansion. Notably however, this was not observed for *ASXL1*.

Identification of variants driving clonal selection in CH has been challenging, owing to the modest expansion of CH clones. Studies using targeted deep sequencing with panels of known driver genes of CH have been able to identify CH clones with high sensitivity, but only cover a specific portion of the genome [40–45]. Conversely, unbiased approaches, using whole genome or whole exome sequencing to identify drivers of clonal expansion from peripheral blood samples, have only been able to detect relatively large clones [24–26, 47]. Analyses of patterns of driver mutations have predicted that HSC clonality may occur in a greater proportion of individuals than the observed number of cases of CH with driver mutations [66, 67]. One study reported that only 40% of patients with evidence of clonal expansion had a mutation in known driver genes [47]. Subsequent studies, that utilised somatic variants identified through whole genome sequencing to reconstruct phylogenetic trees, have identified expanded clones in elderly patients without known driver

event [68, 69]. These clones may be attributed to neutral drift, epigenetic alteration, or non-coding variants. Efforts are currently underway to identify the factors governing clonal expansion without driver mutation [70].

1.1.4 Clonal dynamics throughout the human life cycle

The increased incidence of CH with age was a key observation in the three original studies from 2014. This observation inevitably led researchers to ask questions about clonal dynamics throughout life: When do CH clones originate? What factors affect the growth rate of CH clones? Do different clones grow at different rates throughout the lifespan of the individual? Consequently, multiple studies have broadened our understanding of clonal dynamics and architecture throughout haematopoiesis and studies can be categorised into three methodologies. Firstly, rates of clonal expansion from single-time-point data were inferred using the assumed constant rate of passenger mutation acquisition or population genetics theory [54, 66]. Secondly, serial longitudinal measurements of variant allele frequencies from peripheral blood DNA were used to quantify mutation-specific and mutation-independent effects on clone size [68, 71]. Thirdly, phylogenetic trees were constructed from whole genome sequencing of colonies from single HSPCs using somatic variants as cellular barcodes [68, 69].

Taken together, four important conclusions can be drawn from these studies. Firstly, CH mutations are acquired, and mutant clones are established decades before they are detectable in peripheral blood. Phylogenetic analyses have identified that somatic mutations driving CH clonal expansions were acquired in early life [68, 69]. Authors postulated a model of emergence of CH clones by early mutation acquisition, long latency and then rapid selection in later life. However, this was not observed for spliceosome mutations [25, 26]. Instead, it is likely that they were acquired late in life. Secondly, clones bearing a CH mutation expand stably at an exponential rate at old age. Thirdly, by old age the diversity of HSC clonal repertoire is markedly decreased. For most of life haematopoiesis is remarkably polyclonal, however these studies

reveal a significant decrease in clonality at around 75 years of age. Interestingly, many expanded clones did not have mutations in known driver genes of CH.

Finally, mutations in different driver genes have distinct clonal dynamics throughout life. For instance, *DNMT3A*-mutant clones are the most prevalent among younger individuals and become less prevalent as individuals age. Conversely *TET2*-mutant clones increase in prevalence with age, becoming the most common mutant clone in people over 75 [68, 72]. The prevalence of *ASXL1*-mutant clones was relatively constant across ageing. Additionally, the growth rates of different clones were gene-specific: *DNMT3A* and *TP53*-mutant clones expanded at the slowest rate of $\approx 5\%$ per year, whereas *TET2*- and *ASXL1*-mutant clones expanded at $\approx 10\%$ per year [68]. Clones bearing mutations in spliceosome genes such as *SRSF2* and *U2AF1*, were more prevalent in older individuals, and displayed the greatest annual expansion rate of 15-20% per year. These growth rates are estimates based on longitudinal sampling of patients with CH, and are therefore not statistically significant. However, growth rate of clones largely inversely correlated with incidence of CH i.e. those mutations with a higher growth rate were more rare in CH. One reason for this could be that faster growing stem cell clones may exhaust more quickly than slower growing clones, and are therefore detected less often. However, those mutations with a faster growth rate also correlated with a greater risk of developing AML [68]. One reason for this could be that if a clone acquires a transforming mutation that rescues it from exhaustion, a faster growing clone is more likely to out-compete slower growing clones.

1.1.5 Clonal haematopoiesis and non-haematological disease

One of the most intriguing findings from Jaiswal et al. 2014, was the association of CH with an increased risk of overall mortality [26]. This was attributed to an increased incidence of cardiovascular disease including coronary heart disease (CHD) and ischemic stroke. In a subsequent study, the same group found that the presence of a CHIP clone raised the risk of CHD by 1.9 times [73]. Additionally, they found the strongest association between CHIP and CHD for clones with

TET2 or *JAK2* mutations. Murine studies have supported this finding: mice transplanted with bone marrow from *Tet2* knock-out or *Tet2*-haploinsufficient mice showed an acceleration of atherosclerotic plaque formation compared to controls [73, 74]. This suggests that CH is causative in this association with CHD. In these studies, *Tet2*-deficient macrophages were found to over-express pro-inflammatory molecules including IL6, IL-1 β , and this may be indicative of association with other age-related inflammatory diseases.

Other age-related diseases have been associated with CHIP, including chronic liver disease, chronic obstructive pulmonary disease, osteoporosis, chronic kidney disease, gout, type 2 diabetes mellitus, and aplastic anaemia [75–81]. All of these diseases have been linked to a heightened inflammatory state, fuelling a hypothesis that CH increases the likelihood of these diseases through increased expression of inflammatory cytokines. Further evidence for this is a predisposition of CH carriers to acute and chronic infection [82]. Intriguingly, consistent with epidemiological data of a decreased incidence of neurodegenerative disorders in people with myeloid malignancy, CH patients were found to have a decreased incidence of Alzheimer’s disease [83, 84]. However, the exact mechanisms governing this remain to be elucidated.

1.1.6 Clonal haematopoiesis and haematological malignancy

The presence of a CH clone has been associated with ≈ 10 -fold greater risk of developing a subsequent haematological malignancy, although the absolute risk of malignant transformation remains very low ($\approx 0.5 - 1\%$ per year) [25, 26]. In Jaiswal et al. 2014, and Genovese et al. 2014, myeloid malignancies were most common, however some CH carriers also developed lymphoid malignancies. This is not so surprising, given the overlap between the molecular drivers of CH and myeloid malignancy, and the increased incidence of several myeloid malignancies with age [85]. Interestingly, the risk of developing haematological malignancy is the same for patients with a CH clone but without driver mutation [47]. Identifying the drivers behind clonal expansion in these patients is important in understanding the

currently unresolved reasons for this. Under the classic model of cancer development, acquisition of a CH-associated mutation represents the first “hit” of mutational event on the route to malignancy. Caution must be applied when considering CH as a “pre-leukaemic” state however, as the overall risk of transformation is incredibly low, estimated to be $\approx 0.5 - 1\%$ of CH cases per annum [26]. To distinguish clonally restricted haematopoiesis from malignancy or cytopenias, the term “clonal haematopoiesis of indeterminate potential (CHIP)” was introduced [86]. This applied to cases of clonal haematopoiesis with a variant allele frequency of >0.02 of a cancer-associated mutation without frank malignancy.

The risk of transformation to malignancy is influenced by several factors. Increased clone size (measured by VAF) carries the highest association with leukaemic transformation [87, 88]. However, clonal composition, and the presence of a mCAs are also indicators of a higher risk of progression to malignancy. Additionally, mutations in specific genes carry different risks of progression: a single *DNMT3A*-mutant clone has the lowest risk, whereas mutations in *RUNX1*, *IDH1*, *IDH2*, *JAK2*, *TP53*, and splicing factors, have the highest risk [42, 87–89]. Of the three most common mutated genes in CH (*DNMT3A*, *TET2*, and *ASXL1*), *ASXL1* mutations carry the highest risk of malignant transformation [87, 88].

To stratify this risk, the clonal haematopoiesis risk score has been introduced, which classifies patients with CH as high, intermediate or low risk of evolution to malignancy based on several clinical characteristics [89]. It is important to distinguish between CHIP and another CH-associated condition: clonal cytopenia of undetermined significance (CCUS). CCUS applies to patients with CHIP, but also present clinically with unexplained cytopenias without dysplasia- a key feature of myelodysplastic syndrome [90]. Patients with CCUS have a much higher risk of transformation to malignancy than CHIP: the 5-year probability of progression to myeloid malignancy was 82% [91].

The association of the presence of a CH clone with increased incidence of blood cancer and reduction in overall mortality has prompted the question: what are the mechanisms governing clonal expansion? For some CH-associated genes, including

well-described oncogenes such as *TP53* and those encoding signalling molecules, the molecular mechanisms of clonal expansion are relatively obvious. However, for the epigenetic regulators *DNMT3A*, *TET2*, and *ASXL1*, the molecular foundations of clonal expansion are less obvious. In particular, CH-associated mutations in *ASXL1* are perhaps the least well-described, despite stronger association with progression to blood cancer, and frequent occurrence in myeloid malignancies. For these reason, I decided to focus on *ASXL1* mutations in CH in this study.

1.2 Epigenetic control of gene expression

All cells within an organism contain virtually identical DNA, yet vast differences in morphology and function exist between cell types. Quantitative and qualitative differences in gene expression and resultant protein production can provide an explanation for this diversity. Mechanisms that control gene transcription are therefore critical for specifying and maintaining unique gene expression patterns as different tissue types differentiate and proliferate. Epigenetic information is a heritable set of instructions not encoded by the genomic nucleotide sequence. At the molecular level, epigenetic information is encoded by patterns of chemical modifications to chromatin. Although multiple mechanisms of epigenetic transcriptional control have been identified, here we focus on two primary forms: DNA methylation and post-translational histone modification.

The three most common mutated genes observed in CH are *DNMT3A*, *TET2* and *ASXL1*, and all encode epigenetic modifiers. *DNMT3A* and *TET2* are two enzymes that have opposing functions in controlling the pattern of genomic DNA methylation. *DNMT3A* is a de novo methyltransferase that catalyses the reaction converting cytosine to 5-methylcytosine [92]. This reaction usually occurs at CpG dinucleotides, resulting in cell and tissue-specific methylation patterns. DNA methylation rarely occurs at regions of CpG-dense clusters called CpG islands, which are located in an estimated 70% of gene promoters in the human genome [93]. *TET2* however, catalyses the first in a series of chemical reactions in DNA demethylation. *TET2*

is a dioxygenase that converts 5-methylcytosine to 5-hydroxymethylcytosine [94]. DNA methylation is correlated with a repressive transcriptional state.

Distribution of histone post-translational modifications is a determinant in the specification and maintenance of tissue specific transcriptional programmes. Specific amino acid residues, usually located on tails of histone proteins, can undergo chemical modifications including: acetylation, methylation, phosphorylation, and ubiquitination [95]. These are mediated by specific enzymes and can be indicative of a transcriptionally repressive or active chromatin state. ASXL1 is a protein implicated in the regulation distribution of histone modifications, deposited by members of the Polycomb and Trithorax groups[96]. The roles of Polycomb and trithorax group proteins in haematopoiesis are discussed below.

1.3 Polycomb group and Trithorax group activity and role in haematopoiesis

Precise specification and maintenance of transcriptional programmes is crucial for normal tissue development and homeostasis in animals. Two essential systems of epigenetic control of transcription are the actions of the Polycomb group (PcG) and Trithorax group (TrxG) genes. Genes encoded by PcG and TrxG genes form complexes to post-transcriptionally modify amino acid residues of histone proteins at specific genomic loci, resulting in a transcriptionally repressive or active chromatin state respectively [97]. PcG and TrxG proteins were first characterised as regulators of homeotic (*Hox*) genes expression in *Drosophila melanogaster*, which are transcription factors essential for accurate specification of body segmentation [98, 99]. The establishment of inactive *Hox* gene transcriptional states is controlled by PcG proteins. TrxG proteins on the other hand establish activate transcriptional states through the methylation of H3K4 residues [100, 101]. Precise spatiotemporal control of *Hox* gene expression is required for normal development. It is no surprise then, that mammalian homologs of the *Hox* genes also have essential roles in tissue specification, including haematopoiesis [102].

PcG genes encode proteins that can be divided into two functionally distinct protein complexes, which are evolutionarily conserved in humans: PRC1 and PRC2 [103, 104]. PRC1 and PRC2-mediated gene repression is associated with the monoubiquitination of the H2AK119 residue (PRC1) and the trimethylation of the H3K27 residue (PRC2) [105, 106]. Both histone modifications are found enriched at promoter regions and gene bodies of polycomb target genes. Polycomb repressive complexes are thought to be directed to target genes either by complexing with sequence-specific transcription factors, long non-coding RNAs, or by targeting CpG islands [107]. At least six subtypes of the PRC1 complex (PRC1.1-1.6) have been identified in mammals, with an enzymatic core of RING1A/B and one of six PcG ring finger proteins (PCGF1-6) [108]. The RING1A/B subunit is an E3 ubiquitin ligase that catalyses the deposition of monoubiquitin at H2AK119. Two primary PRC2 complexes have also been identified, both encompassing a core of EZH1/2, SUZ12, RBBP4/7, and EED [109]. EZH1/2 are methyltransferases that catalyse the methylation of the H3K27 residue.

Exactly how PcG complexes establish and maintain transcriptional repression is not clear, however there is evidence that PRC1 and PRC2 complexes work in synergy. Canonical PRC1 complexes (PRC1.2, PRC1.4) are defined by the presence of CBX subunits which have the ability to bind H3K27me₃ via the N-terminus chromodomain [110], whereas variant PRC1 complexes (vPRC1; PRC1.1, 1.3, 1.5, 1.6) lack this subunit. Recruitment of vPRC1 complexes is dependent on the intrinsic DNA-binding activity of their subunits [111–113]. This suggests that cPRC1 complexes deposit H2AK119ub at sites of PRC2 activity, whereas vPRC1 complexes act independent PRC2 targets. There is no evidence that PRC1 subunits recognise and bind H2AK119ub for recruitment to target loci. Similarly, H2AK119ub is recognised by the N-terminus ubiquitin-interacting motif of the JARID2 subunit of the PRC2.2 complex [114], directing PRC2 to deposit H3K27me₃ at the same genomic loci. The canonical model of synergistic PcG-mediated gene repression describes deposition of H2K27me₃ at target genes, followed by recruitment of PRC1 and deposition of H2AK119ub [105]. However more recent study of the diversity

of Polycomb complex repertoire has challenged this model [110, 115]. Nonetheless, PRC1 and PRC2 are both important determinants in embryonic and adult stem cell potency and differentiation [116, 117].

PRC1 and PRC2 complexes have been reported to play important roles within mammalian haematopoiesis. PRC1.4 has been linked to the maintenance of both long-term engrafting HSC (LT-HSC) self-renewal and proliferation [118, 119], and also specification of the lymphoid lineage through the repression of critical B cell lymphoid transcription factors [120]. PRC1.1 on the other hand has been linked to HSPC pluripotency escape through the repression of *HoxA* genes in mice [121, 122]. Deficiency of PRC2 components, such as EZH2, SUZ12, and EED also results in loss of HSC self-renewal capacity, reconstitution ability, and lymphoid differentiation implying a key role for PRC2 in maintenance of normal haematopoiesis [123–125]. Additionally, mammalian TrxG proteins have also been heavily implicated in regulating haematopoiesis [126–130], and translocations of the Mixed Lineage Leukaemia (*MLL*) gene are frequently found mutated in AML and ALL [131].

It is interesting that there is significant diversity of PcG and TrxG family members in mammals to regulate gene expression through histone modification. One reason for this could be that there is redundancy in the PcG and TrxG families, and that organisms have evolved multiple complexes to compensate for any loss of function. This creates a robust system for maintaining specific patterns of gene activation or repression. Additionally, different complex members could play a role in targeting the complex to very specific genomic loci, utilising specific DNA binding affinities of their DNA binding domains. This would create precise context-dependent gene expression patterns throughout complex tissues.

Interestingly, novel evidence suggests that in certain contexts, some variant complexes from the same family actually have opposing function at the same genomic loci [132]. In this study, Soto-Feliciano *et al.* showed that in MLL-fusion harbouring leukaemia cell lines, the TrxG-associated MLL1-Menin complex actually represses tumour suppressive transcriptional programmes through the deposition of H3K4me3 at promoter regions, rather than activating them. Previously, MLL1-Menin has

been associated with transcriptional activation of tumour-suppressive programmes [133]. Interestingly, pharmacological inhibition of Menin led to recruitment of MLL3/4-UTX at the same genomic loci where MLL1-Menin was previously bound, and activated gene expression of those tumour-suppressive programmes. It is entirely possible further non-canonical mechanisms of gene regulation exist between PcG and TrxG family members.

Enhancers of trithorax and polycomb (ETP) proteins regulate the activity of both TrxG and PcG proteins. In *Drosophila*, *Asx* (Additional sex combs) was identified as an ETP protein [134], and deletion of *Asx* resulted in deficiency of TrxG and PcG homeotic phenotypes [135]. In humans, the homologs of *Asx* are encoded by *ASXL1-3* [51]. In the following sections the structure and function of ASXL1 and its murine homolog *Asxl1* are discussed.

1.4 The structure and function of ASXL proteins

In humans, the *Additional Sex-combs Like 1* (*ASXL1*) gene is located on the chromosome 20q11.21 and encompasses 13 exons and 12 introns [51]. The *ASXL1* gene encodes a 1,541 amino acid, 165.432 kDa protein, which has been reported to be expressed ubiquitously throughout embryogenesis and adult tissue homeostasis in both mice and human cell lines [136, 137].

The ASXL family of proteins (ASXL1-3) share conserved domains including a plant-homology domain (PHD) at the C-terminus, ASXM1 and ASXM2 domains in the middle of the protein, an ASXH homology (ASXH) domain, and at the N-terminus, an ASXN domain (Figure 1.2). The two terminal domains, ASXN and PHD, have been described as putative genomic binding domains. ASXN (ASX N-terminus domain - also referred to as: HB1 ASXL restriction endonuclease helix-turn-helix/HARE-HTH) has been reported to bind DNA [138], whereas PHD domains have been reported to bind to methylated or unmethylated H3K4 [139]. The prevailing hypothesis for the function of the ASXM1 and ASXM2 domains (523-618 AA; 1009-1123 AA respectively) is mediation of protein-protein interaction with

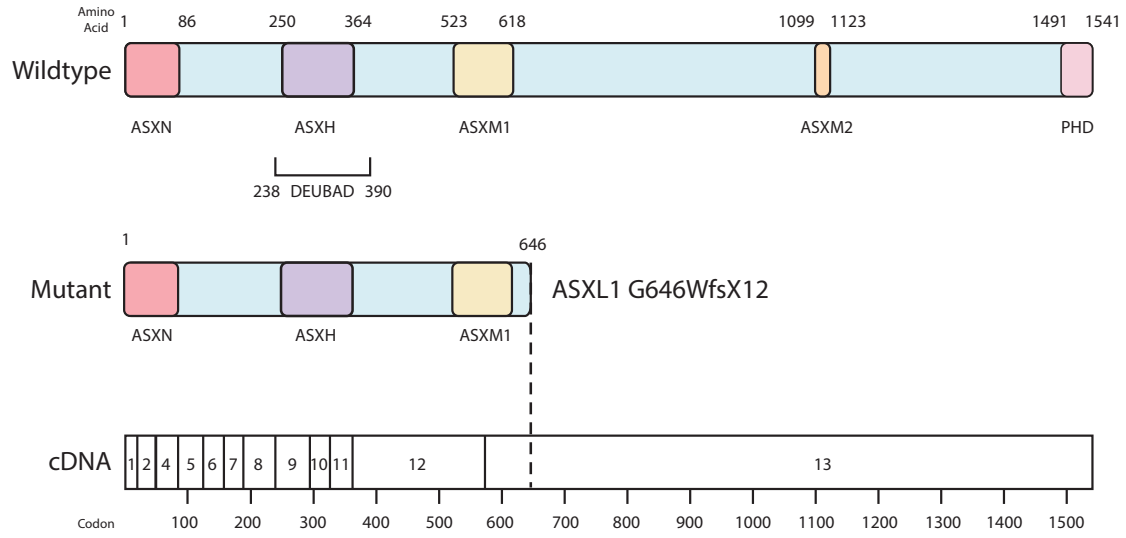


Figure 1.2: Structure of ASXL1 protein and gene. **Top:** Functional domains of ASXL1 protein. ASXN: ASX N-terminus domain; ASXH: Asx homology domain; ASXM: Asx middle domain; DEUBAD: Deubiquitinase adaptor domain; PHD: Plant-homology domain. **Middle:** Structure of mutant ASXL1, resulting from the G646WfsX12 hotspot variant. **Bottom:** Exonic structure of ASXL1.

nuclear hormone receptors including retinoic acid receptors, peroxisome proliferator-activated receptor gamma (PPAR γ), and sex steroid receptors [140, 141]. Finally, the ASXH domain is considered to be the most critical for the chromatin remodelling function of ASXL1, as it contains the deubiquitinase adaptor domain (DEUBAD). Deletion of DEUBAD in ASXL1 leads to the loss of the ability to interact with the deubiquitinating enzyme BRCA1-associated protein (BAP1) [142–144].

Due to the absence of a catalytic domain, ASXL1 protein has not been reported to possess any enzymatic activity. Rather, it is thought that ASXL1 is an integral structural and activity-enhancing component of epigenetic chromatin-modifying protein complexes and transcription factors. Whilst the best described function of ASXL1 is as critical member of the PR-DUB complex [145] (discussed in the following section), several other functions have been described. ASXL1 has previously been reported to bind to members of the PRC2 complex including EZH2 (Enhancer of Zeste Homolog 2), the key histone methyltransferase in the PRC2 complex [146]. In this study, deletion of *ASXL1* in a human AML cell line UKE1 led to a global reduction of H3K27me3 and upregulation of posterior *HOXA* cluster genes.

Subsequent follow-up in constituent *Asxl1*-knockout mice recapitulated the global loss of H3K27me3 at promoter regions of posterior *Hox* genes, suggestive of a role for ASXL1 in regulation of the PRC2-mediated gene repression [147].

Similarly, ASXL1 has also been implicated in trithorax-group mediated transcriptional control by histone methylation. Inoue et al., detected the physical interaction of human ASXL1 with TrxG proteins OGT1 and HCFC1 by co-immunoprecipitation and mass spectrometry in transgenic mouse cell lines [148]. They also reported that these proteins form a complex with ASXL1 that can interact with MLL5. OGT1 and HCFC1 was previously demonstrated to complex with MLL5, promoting trimethylation of H3K4 at promoters of E2F1 and subsequent transcriptional activation of cell cycle-associated genes [149]. Inoue et al., suggest that this complex may explain the observed H3K4me3 loss in HSPCs from *Asxl1*-knockout mouse models [150, 151]. ASXL1 is yet to be linked directly to SET1/COMPASS family member proteins which catalyse the mono-, di- and tri-methylation of H3K4.

1.4.1 ASXL1 as part of the PR-DUB complex

One of the best described functions of ASXL1 is as an integral member of the polycomb-repressive deubiquitinase (PR-DUB) complex. PRC1 complexes catalyse the monoubiquitination of the H2AK119 residue, which is highly correlated with transcriptional repression [105]. Monoubiquitin is removed by BRCA1-associated protein 1 (BAP1) deubiquitinase complexes. The PR-DUB complex was first described in *Drosophila melanogaster*, in which Calypso (the protein with DUB catalytic function) is activated through interaction with Asx, and then deubiquitinates H2AK118ub [145]. In mammals, BAP1- the homolog of Calypso- complexes with ASXL1-3 to form the core of the PR-DUB complex [152]. BAP1 is the key enzymatic component of the PR-DUB complex, which catalyses H2AK119ub removal. Other proteins such as Yin Yang 1 (YY1), forkhead box transcription factors 1/2 (FOXK1/2), O-linked N-acetylglucosamine transferase (OGT), Host cell factor 1 (HCF-1) and methyl-CpG-binding domain family (MBD5/6), have all been associated with BAP1/ASXL1 complexes [153–156]. The precise role for

these proteins in the PR-DUB complex remains to be elucidated, but they may play a role in directing the complex to target loci.

During complex assembly, BAP1 promotes the recruitment of ubiquitin ligases to monoubiquitinate the DEUBAD domain of ASXL1, which provides the substrate on which the C-terminal domain of BAP1 to bind [142, 143]. The ubiquitin C-terminal hydrolase (UCH) domain of BAP1 prevents the elongation of a ubiquitin chain, which prevents degradation by the proteasome. Importantly, ASXL proteins are necessary for the stabilisation of BAP1: in the human HAP1 cell line, BAP1 is not detected upon deletion of ASXL1, despite no change in BAP1 mRNA expression [152]. Upon co-expression in human cell lines, the ASXL1-BAP1 complex can stably catalyse the removal of monoubiquitin from H2AK119 [152]. In a recent publication, Ge and colleagues resolved the crystal structure of the PR-DUB complex to understand specificity for H2AK119ub [157]. They determined that the DEUBAD domain of ASXL1 is essential in forming the ubiquitin binding cleft with the UCH domain of BAP1, and in displacing the C-terminal extension domain of BAP1 to allow interaction with the nucleosome. These results provide insights into the structure and function of ASXL1, furthering understanding of mechanistic impact of mutations.

1.5 The impact of mutations in *ASXL1*

1.5.1 *ASXL* gene family mutations in non-haematological disease

Mutations in *ASXL1* have been associated with a range of non-haematological diseases. Germline *de novo* *ASXL1* mutations have been found in almost 50% of cases in a rare congenital developmental disorder called Bohring-Opitz syndrome [158, 159]. Bohring-Opitz syndrome is characterised clinically by severe developmental delays, intellectual disability, microcephaly, limb anomalies, and facial dysmorphism. Interestingly, *Asxl1* constitutive knockout mouse models displayed similar axial skeletal patterning abnormalities, suggestive of PcG and TrxG activity disruption [147, 160]. Additionally, somatic mutations of *ASXL1* have been reported in a range of non-haematological cancers including colorectal cancer, bladder cancer, cervical cancer, prostate cancer, breast cancer, and melanoma [161, 162].

1.5.2 *ASXL1* mutations in myeloid malignancy and clonal haematopoiesis

Somatic mutations in *ASXL1* are recurrently found in haematological disease, most notably myeloid disorders, accounting for $\approx 43 - 49\%$ of chronic myelomonocytic leukaemia (CMML) patients [163–165], 15–21% of myelodysplastic syndrome (MDS) patients [166–168], 8–10% of myeloproliferative neoplasm (MPN) patients [169–171], 7–8% of juvenile myelomonocytic leukaemia (JMML) patients [172, 173], and 5–17% of acute myeloid leukaemia (AML) patients [174–177]. *ASXL1* mutations are also very common in patients with aplastic anaemia [178, 179]. *ASXL1* mutations in myeloid malignancy have been found to co-occur with mutations in *RUNX1*, *TET2*, *SRSF2*, *NRAS*, *JAK2*, *SETBP1*, *EZH2*, *IDH1*, *TP53*, *STAG2*, *U2AF1* [180–182]. Mutually exclusively occurring mutations include those found in: *DNMT3A*, *FLT3-ITD*, *WT1* and *NPM1* [182, 183].

Importantly, the presence of an *ASXL1* mutation is associated with a poor prognosis, treatment response, or disease progression for all aforementioned myeloid malignancies [168, 184–187]. This has been recognised by the European Leukaemia Network, which has designated *ASXL1* mutations as an unfavourable risk classification for AML [188]. Models in which *Asxl1* expression is altered in mice with leukaemia-associated mutations, have supported this observation of poor prognosis [189, 190]. Haploinsufficiency of *Asxl1* in *JAK2* V617F transgenic mice increased the likelihood of developing MPN-like disease and hastened progression to secondary AML when compared to those mice with both functional *Asxl1* alleles [189].

ASXL1 mutations in myeloid malignancy are exclusively heterozygous. The vast majority of mutations are either frameshift or nonsense mutations, resulting in a premature stop codon [191]. Since most mutations occur in the final exon (13) of the gene, it is predicted that mRNA transcribed from the mutant allele escapes nonsense-mediated decay, which results in the translation of a C-terminally truncated ASXL1 protein. Expression of mutant *ASXL1* in human AML cell lines has confirmed this, detecting a stable C-terminally truncated ASXL1 protein by Western blot and mass spectrometry [192]. The most common mutation in myeloid malignancy, is NM_015338.5:c.1934dup;p.Gly646Trpfs*12 (*ASXL1* c.1934dupG),

accounting for $\approx 37\%$ of mutations in myeloid malignancy [181]. This is a frameshift mutation which represents an addition of a single guanosine nucleotide into an 8-bp guanosine homopolymer. The expression of the C-terminally truncated protein product from this *ASXL1* mutation was detected by western blot in a recent CRISPR-Cas9 model in human CD34+ cord blood cells [144].

Since age-related clonal haematopoiesis is associated with an increased incidence of blood cancer, it is no surprise that somatic *ASXL1* mutations are found in individuals with CH. *ASXL1* is the third most common CH mutation behind those found in *DNMT3A* and *TET2*, accounting for between 5-10% of cases [24–26]. The profile of *ASXL1* mutations in CH is largely similar to that seen in myeloid malignancy; heterozygous mutations are predominantly found within exon 13 and are either frameshift or nonsense. Some cohorts of patients with CH have reported a decreased incidence of the p.G646Trpfs hotspot mutation compared to that seen in myeloid malignancy ([24–26]. This has prompted some to hypothesise that hotspot mutations, like the *DNMT3A* R882H mutation, confer a greater probability of transformation to malignancy compared to more stable, less pathogenic CH mutations [193, 194]. A recent genome-wide analysis of over 200,000 patients with CH reported an increased likelihood of finding an *ASXL1* mutation in males (Female:Male ratio = 0.6, $p = 5.7 \times 10^{-28}$) [53]. Interestingly, smoking has been associated with a greater incidence of *ASXL1* mutation than for other CH mutations [53, 195]. Additionally, compared to *DNMT3A* and *TET2*, mutations in *ASXL1* are also associated with a greater risk of transformation from CH to malignancy [26, 53].

1.5.3 Modelling the impact of *ASXL1* mutations on haematopoiesis

Several mouse models have been generated to investigate the impact of *ASXL1* mutations upon haematopoiesis. Constitutive homozygous knockout of *Asxl1* in mice resulted in the development of an MDS/MPN-like disease [150]. In the same study, this observation was recapitulated in mice with *Asxl1* haploinsufficiency, which more closely resembles the genetics of myelodysplasia. In the bone marrow of both models, reduction of Lineage[−] Sca⁺ c-Kit⁺ (LSK) cells, and myeloid lineage

bias was observed. HSCs from knock-out mice also displayed reduced colony-forming capacity *in vitro* and impaired repopulation capacity *in vivo*.

In a conditional *Asxl1* knockout model, which was restricted to the haematopoietic lineage using a Vav-Cre/Mx1-Cre inducible system, mice also developed myelodysplasia and erythrodysplasia [147]. In this model, the number of HSCs from knockout mice were increased compared to control mice. However knockout HSCs displayed impaired repopulation upon competitive transplantation into recipient mice, which was attributed to a greater rate of apoptosis. Mechanistically, the authors reported that HSCs from knockout mice showed increased expression of posterior *HoxA* genes and the PRC2 target gene and cell cycle regulator: *p16ink4a* [196]. The authors attributed this to a global reduction of the transcriptionally repressive H3K27me3 mark- the distribution of which is partly dependent on H2AK119ub distribution [115]. In *ASXL1*-mutant cells, where H2AK119ub genomic content is reduced due to PR-DUB dysfunction, H3K27me3 distribution by PRC2 (which recognises and binds H2AK119ub) is therefore reduced. These findings suggest that not only is *Asxl1* critical to normal haematopoietic development, but the MDS-like phenotype is likely due to epigenetic dysregulation.

The observation of MDS-like disease in these mouse models led researchers to believe that pathogenic *Asxl1* mutations are loss-of-function. However, very few cases of human myeloid malignancy occur from a single *ASXL1* mutation. Additionally, the observation that *Asxl1*-mutant HSCs have a competitive disadvantage does not recapitulate what is seen in human clonal haematopoiesis. This led to the hypothesis that leukaemia-associated mutations are in fact gain-of-function, and this may explain the discrepancy between observations from knock-out models and human phenotype.

In vitro studies clarified that expression of leukaemia-associated *ASXL1* mutations in human cell lines resulted in the production of a C-terminally truncated ASXL1 protein [197, 198]. Compared to full-length wild-type ASXL1, mutant ASXL1 was shown to have enhanced stability and a higher affinity for complexing with BAP1. Expression of C-terminally truncated mutant ASXL1 in 293T cells

was reported to enhance deubiquitinase activity of BAP1 [197]. This observation was accompanied by a significant global reduction in H2AK119ub, and H3K27me3. Upon expression in transgenic *ASXL1*-mutant murine models, global reduction in H2AK119ub occupancy correlated with reduced posterior *HoxA* gene cluster and *Irf8* gene expression [197, 199, 200]. These observations led researchers to believe that truncating *ASXL1* mutations may have gain-of-function activity rather than loss-of-function.

Accordingly, multiple *ASXL1* mutation knock-in mouse models have been generated. Yang et al., generated an *Asxl1* knock-in mouse model bearing a Y588X *Asxl1* mutation, analogous to the ASXL1 Y591X mutation frequently observed in humans with myeloid malignancies [201]. Consistent with the knockout models, knock-in, but not wildtype mice developed progressive myeloid malignancies including MDS, MPN, and AML-like diseases. Relative to wildtype control mice, the bone marrow of the knock-in mice was composed of a greater proportion of LSK cells, granulocyte-monocyte progenitors (GMPs), and decreased proportion of megakaryocyte-erythrocyte progenitors (MEPs). Mechanistically, these observations were explained by an increase in open chromatin was observed in mutant HSPCs (c-Kit+ cells) around transcription start sites, and increased RNA expression, of genes involved in maintenance of HSC activity and myeloid differentiation.

In contrast, MDS-like disease was not observed in a transgenic overexpression mouse model that utilised the *Rosa26* promoter to drive the expression of a human *ASXL1* E635RfsX15 allele [200]. Mutant *ASXL1* expression reduced the number of LSK cells, immunophenotypic HSC and multipotent progenitors (MPPs) within the bone marrow, but increased MEPs. Interestingly within the MEP compartment there was a significant bias towards immunophenotypic megakaryocyte progenitors compared to erythroid progenitors. Combined with the observation of anaemia and thrombocytosis in peripheral blood, this is indicative of defective erythroid differentiation. In competitive repopulation assays, mutant HSCs displayed impaired repopulation ability with myeloid skewing. These data indicate a competitive disadvantage for mutant HSCs. *ASXL1* mutation was not sufficient to transform

HSPCs, however relative to healthy controls the incidence of myeloid malignancy was higher when co-expressed with a mutant *RUNX1* allele, consistent with what has been observed in malignancy. Consistent with *in vitro* studies, mutant cells were also found to have reduction in both H2AK119ub and H3K4me3 histone marks. Given that overexpression models do not recapitulate a competitive advantage phenotype for HSCs seen in human CH, the accuracy of these models are called into question.

Expression of physiologically relevant levels of mutant *ASXL1* transcripts may produce a faithful recapitulation of CH in mice. Two constitutive *Asxl1* mutation knock-in models bearing the *Asxl1* G463WfsX12 allele have been generated [199, 202]. This murine variant is analogous to the *ASXL1* G646WfsX12 hotspot mutation observed in a range of myeloid malignancies. One model showed an increase in mutant HSPC BM cellularity with an increased HSPC serial replating capacity *in vitro*, but a decreased long term repopulation ability *in vivo* [202]. After 18 months, no haematopoietic diseases were recorded in the knock-in mice, indicating that *Asxl1* mutation alone is not sufficient to induce haematological disease.

However, in the study by Uni et al., an MDS/MPN-like disease in knock-in mice was detected only after an extended period of time [199]. A reduction in the number of immunophenotypic LT-HSCs, common myeloid progenitor (CMP), GMP and MEPs was also seen. Among the mature populations, there was significant myeloid skewing, with a reduction in B and T lymphocytes. *In vitro*, mutant immunophenotypic HSCs also showed increased serial replating and colony-forming capacity compared to wildtype counterparts. To explain these findings, Uni et al., reported a decrease of H2AK119ub occupancy at the PRC1 target gene and cell cycle regulator p16Ink4a, in bone marrow cells from knock-in mice. Even at physiological levels of mutant transcript expression, conflicting results are seen between two independent models.

The key clinical observation of CH and myeloid malignancies is that *ASXL1*-mutant HSPCs have a competitive advantage over wildtype cells. Thus far, murine models have failed to accurately recapitulate this phenotype with conflicting results.

This contrasts with models of loss of function CH-associated mutations in *Dnmt3a* and *Tet2*, whereby mutant HSCs demonstrate an increase of self-renewal [203–206].

One potential explanation for this is that the murine system does not fully simulate the physiological conditions observed in humans. For example, many human pre-leukaemic and leukaemic conditions are age-related. Since the average lifespan of a mouse is roughly 2 years, it is very challenging to simulate the cellular stresses of an aged bone marrow microenvironment. Fujino and colleagues used the transgenic mice model conditionally expressing the *ASXL1* E635RfsX15 allele to assess the impact of mutation with ageing [207]. Whilst they found that the number of immunophenotypic LT-HSCs increased in mutant aged mice relative to controls, the mutant cells once again showed a competitive disadvantage upon secondary transplantation into sublethally irradiated recipient mice. LT-HSCs (CD49f+ in humans; CD150+ in mice) are functionally characterised with multilineage long-term engraftment, compared to the transiently engrafting ST-HSC fraction (CD49f- in humans; CD150- in mice) [208, 209]. Interestingly, when the mutation was induced post-natally the percentage of LSK cells and immunophenotypic ST-HSC, and LT-HSC were increased relative to non-induced control mice. This suggested that expression of mutant *ASXL1* confers a growth advantage on HSPCs in the context of native haematopoiesis.

Another possibility for difference in phenotype between human CH and murine models is the physiological differences between human and murine haematopoiesis. To address this, Köhnke et al., introduced leukaemia and CH-associated *ASXL1* mutations into CD34+ human cord blood cells using CRISPR-Cas9 [144]. *In vitro*, compared to AAVS1 “safe harbour locus” control edited CD34+ cells, mutant HSPCs displayed a reduction in megakaryocyte/erythroid differentiation using colony-forming assays, increased serial replating ability. Upon primary transplantation into recipient mice, mutant edited CD34+ cells displayed increased repopulation ability, coupled with decreased lymphoid output and myeloid-biased differentiation. Myeloid malignancies developed in a small subset of mice transplanted with mutant cells, however since the authors did not genotype the malignant cells, clonal evolution

cannot be ruled out as the cause. Based on these observations, this is the closest model for recapitulating clinical observations of *ASXL1*-mutant pre-leukaemic states in humans.

1.6 Potential mechanisms of clonal expansion in *ASXL1*-mutant CH and open questions

The precise molecular mechanisms by which *ASXL1* mutations confer selective advantage upon HSPCs are elusive. Thus far murine models of *ASXL1* mutations have proved inaccurate at modelling the clinical phenotype seen in humans. One example of this is the inconsistency in results of clonal competition assays involving *ASXL1*-mutant cells. Therefore, proposed mechanisms of clonal expansion from these studies must be treated with caution.

Several groups have used *in vitro* approaches to link *ASXL1* mutations to clonal advantage. One obvious observation from studies performed in human cell lines and primary human cells expressing truncated ASXL1, is the global depletion of H2AK119ub. The proposed mechanism for this is that mutant *ASXL1* increases the stability of the PR-DUB complex, and induces hyperactivity of deubiquitination [144, 197, 198, 210]. Thus, one hypothesis is that mutant *ASXL1* dysregulates expression levels of PRC1 target genes through aberrant PR-DUB complex activity, leading to clonal advantage (Figure 1.3). Identifying genes that are dysregulated by this process would provide insight into the molecular mechanisms dysregulated in *ASXL1* mutant cells.

Köhnke et al demonstrated in *ASXL1*-mutant CRISPR-Cas9 edited CD34+ cord blood cells that increased RNA expression and genomic regions of increased open chromatin and correlated with those where H2AK119ub is lost [144]. Interestingly, they also found highly preserved regions of H2AK119ub occupancy in mutant cells which were enriched for PRC1 target genes. This is perhaps counterintuitive, and contradicts the prevailing hypothesis of mutant ASXL1 function. The authors argued that in mutant cells, PRC1 complexes were still active and specifically deposited H2AK119ub at their target genes. Whereas the H2AK119ub loss in mutant cells

was non-specific. Therefore, relative to the rest of the genome, the H2AK119ub mark was still enriched at PRC1 target genes. From this, one could hypothesise that H2AK119ub-mediated transcriptional repression still occurs at some loci, whereas at others this is lost. This hypothesis could help to explain the observation in *Drosophila melanogaster* that *Asx*-mutants show PcG and TrxG-deficient phenotypes [135]. When performing integrative pathway enrichment analysis on this multimodal dataset, they found an enrichment of HSC and leukaemic stem cell gene signatures in mutant cells, suggesting that *ASXL1* mutation pushes HSPCs into a more *bona fide* HSC-like state. Furthermore, upon pharmacological inhibition of PRC1 and PR-DUB complex members, *ASXL1*-mutant cells were more sensitive than WT cells - indicating an important role for regulation of H2AK119ub distribution in *ASXL1*-mutant cell survival and clonal expansion [144, 210].

Additionally, several molecular pathways have been suggested to play a role in clonal advantage of *ASXL1*-mutant cells. In a murine model in which *ASXL1*-mutant HSPCs did gain a selective advantage *in vivo*, an increase in Akt/mTOR signalling was observed in mutant immunophenotypic LT-HSCs [207]. Enhanced Akt/mTOR signalling through depletion of *Pten* has previously been reported to cause transient expansion and subsequent apoptosis of murine HSCs [211, 212]. The authors suggested that *ASXL1* mutation causes a mild dysregulation of AKT/mTOR signalling in LT-HSCs, resulting in a prolonged expansion of mutant HSCs with delayed apoptosis. The authors also report that mutant-*ASXL1*/BAP1 activated Akt/mTOR signalling in an epigenetics-independent manner. Whilst this may well be true, it is likely that other dysregulated molecular mechanisms caused by aberrant epigenetic control play pivotal roles in *ASXL1*-mutant cells. Notably, Fujino et al., also reported an increase in reactive oxygen species-mediated DNA damage [207]. DNA damage increases the risk of *de novo* mutation so this may provide an explanation for the increased risk of progression to AML in *ASXL1*-mutant CH.

Mounting evidence suggests that changes in environmental factors play a role in selection of mutant stem cells throughout life. It may be that very specific environmental conditions are required for *ASXL1* to confer selective advantage

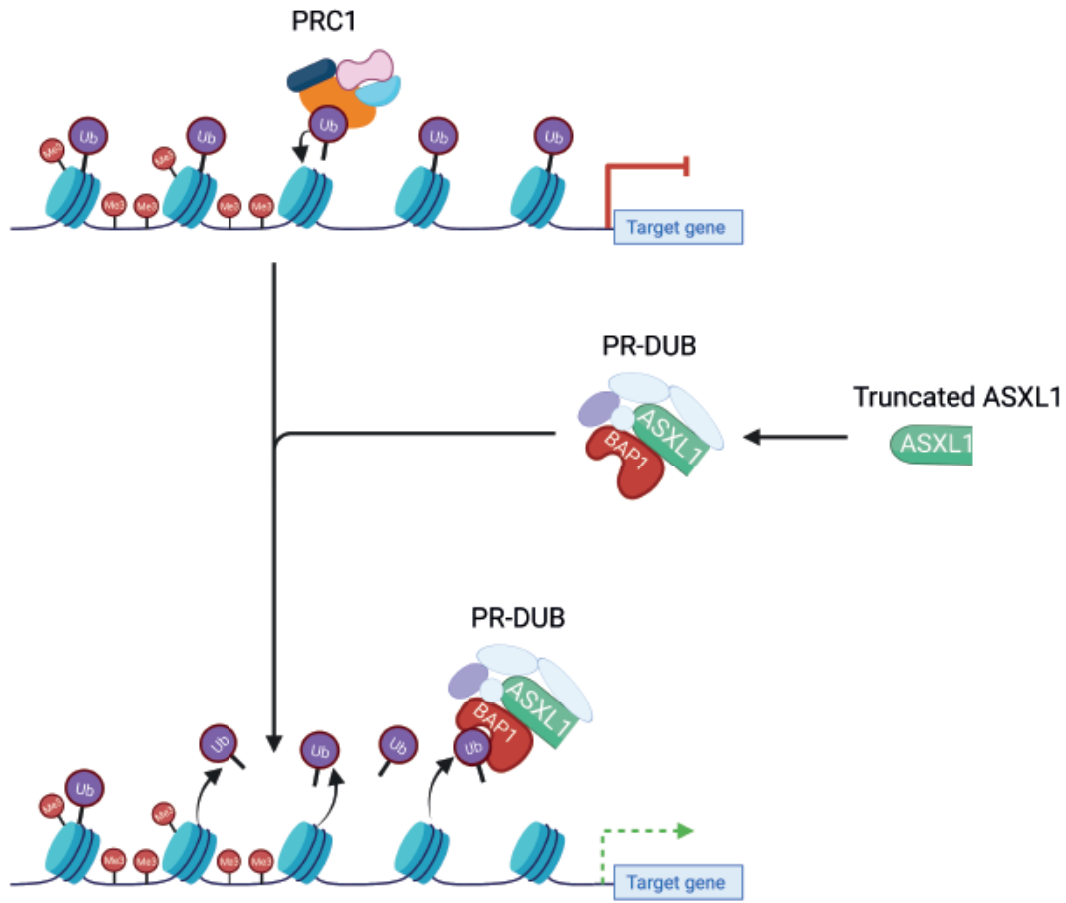


Figure 1.3: Putative function of mutant ASXL1. Polycomb repressive complex 1 monoubiquitinates H2AK119ub at promoter regions and gene bodies of target genes, repressing transcription. C-terminally truncated ASXL1 causes a hyperactivity of the PR-DUB complex, leading to loss of H2AK119ub and an increased likelihood of Polycomb target gene expression.

upon HSCs over time. A heightened inflammatory state is also a hallmark of ageing, and has been proposed as a selective pressure driving clonality in the bone marrow [213, 214]. Evidence supporting this comes from numerous studies in which exposure to inflammatory stimuli causes the functional decline of HSCs [215–218]. If a mutation confers protection against an inflammatory stimulus, mutant HSCs and their progeny are more likely to survive. In mouse models of *DNMT3A*- and *TET2*-mutant CH, mutant HSCs exhibit clonal advantage in response to an inflammatory stimulus [219–223]. In primary human *DNMT3A* and

TET2-mutant CH samples, it was found that mutant HSCs have an attenuated transcriptional response to inflammation [45].

Whether inflammation is the factor driving for *ASXL1*-mutant clonal expansion in CH remains unclear. From animal models, *asxl1*-mutant macrophages and neutrophils from a CRISPR-Cas9 edited zebrafish model displayed raised cytokine level expression compared to controls [224]. Mutant HSPCs from these zebrafish also were reported to be clonally dominant 8 months after fertilisation. Notably, this effect was abrogated when the inflammation suppressor gene, *nr4a1*, was simultaneously knocked out. This could be evidence that *ASXL1* mutation also confers a protective effect upon mutant cells. Increased inflammatory cytokine expression has also been recorded in *ASXL1*-mutant mice and humans [225, 226], however functional experiments are required to assess the role of inflammation in *ASXL1*-mutant clonal expansion. Defining these environmental conditions in humans will be important to develop suitable models to enable these assays.

1.7 Aims of this thesis

Studies have revealed that age-related expansion of HSCs with acquired somatic mutations in known driver genes of leukaemia, is associated with higher probability of developing haematological malignancy. Mutations in *ASXL1*, an epigenetic regulator, have been recurrently identified in patients with clonal haematopoiesis and myeloid malignancy, however the molecular mechanisms by which they confer clonal advantage are unclear. Murine models have failed to accurately recapitulate the phenotype seen in humans with acquired *ASXL1* mutation. To date, only one study has attempted to model *ASXL1* mutations in primary human tissue, albeit *in vitro*. Therefore, there is a need to study *ASXL1*-mutant clonal expansion in humans.

Recent studies have addressed this need in patients with *DNMT3A*- and *TET2*-mutant CH and have provided valuable insight into common and distinct molecular mechanisms of clonal expansion [45, 227]. The association of *ASXL1* mutations with a greater risk of transformation to malignancy, and poor prognosis upon transformation is evidence to suggest the underlying factors governing clonal

expansion may be distinct to that of *DNMT3A* and *TET2*. This study therefore aimed to investigate the cellular and molecular mechanisms by which *ASXL1*-mutant clones gain a selective advantage in primary human tissue from patients with CH.

The aims of this study were:

1. To identify patterns of clonal expansion within immunophenotypically and transcriptomically defined cell populations throughout haematopoiesis in patients with *ASXL1*-mutant CH.
2. To identify the transcriptomic correlates between somatic *ASXL1* mutations and haematopoietic clonal expansion.
3. To develop a novel method capable of assaying differences in epigenomic profile and genotype between single cells from heterogeneous cell populations.

Chapter 2

Materials and Methods

Contents

2.1	Study participants and ethical approval	32
2.2	Sample collection, processing, targeted sequencing and variant calling	32
2.3	Flow cytometry analysis and cell sorting	33
2.3.1	Primary human bone marrow thawing and staining . . .	33
2.3.2	Primary human bone marrow single cell sorting	34
2.4	TARGET-Seq+ primer design	34
2.5	TARGET-Seq+ genotyping primer testing	36
2.6	TARGET-Seq+ library preparation	39
2.6.1	Lysis buffer preparation	39
2.6.2	Reverse transcription and pre-amplification	39
2.6.3	Whole-transcriptome library preparation and sequencing	40
2.6.4	Targeted single-cell genotyping, library preparation, QC, and sequencing	42
2.7	Single-cell genotyping analysis	43
2.7.1	Data pre-processing	43
2.7.2	Mutation calling	44
2.7.3	Estimation of allelic dropout rate	45
2.7.4	Clonal hierarchy assignment	46
2.8	FACS Index data analysis	47
2.9	Single-cell transcriptome analysis	47
2.9.1	Data pre-processing	47
2.9.2	Quality control, normalization and highly variable gene identification	48
2.9.3	Dimensionality reduction, clustering, and data integration	49
2.9.4	Cluster annotation and reference map projection	50
2.9.5	Differential abundance analysis	51
2.9.6	Differential gene expression analysis and gene set enrich- ment analysis	51
2.9.7	HSPC subclustering	52
2.9.8	Differential regulon analysis	53
2.10	GCUT library preparation	53
2.10.1	GCUT genotyping primer design	53
2.10.2	Cell line culture	54
2.10.3	Lysis buffer preparation	54
2.10.4	Bulk CUT&Tag of cell lines	55
2.10.5	Single nuclei sorting	55
2.10.6	Preamplification of GCUT libraries	56

2.10.7 Single-cell CUT&Tag library preparation and sequencing	56
2.10.8 Targeted single cell genotyping	57
2.11 GCUT optimisation	57
2.11.1 Primer concentration titration	57
2.11.2 Comparison of different bead purification strategies . . .	57
2.12 GCUT data analysis	58
2.12.1 CUT&Tag data pre-processing	58
2.12.2 Quality control and filtering	58
2.12.3 Normalisation, feature selection, and dimensionality re- duction	59
2.12.4 GCUT genotyping sequencing data analysis	59
2.13 Figure production	59

2.1 Study participants and ethical approval

Full details of study participants and ethical approval are described in Jakobsen et al., 2024 [45]. In brief, samples were obtained from patients undergoing total hip replacement surgery, enrolled in the Mechanisms of Age-Related Clonal Haematopoiesis (MARCH) study. Prior to surgery, patients provided written informed consent in accordance with the Declaration of Helsinki. Patients were excluded from the study if there was history of haematological cancer, rheumatoid arthritis, septic arthritis, or if the patients were taking oral steroids.

2.2 Sample collection, processing, targeted sequencing and variant calling

Similarly, full details of collection, processing and sequencing are described in Jakobsen et al., 2024 [45]. Briefly, bone marrow aspirates were processed within 24 hours of collection. Mononuclear cells (MNCs) were isolated from bone marrow samples separated by Ficoll density gradient. For select samples, an aliquot of MNCs underwent magnetic activated cell sorting (MACS) enrichment for the CD34+ fraction. Cells were counted, aliquoted, and cryopreserved in 90% Fetal Bovine Serum (FBS) with 10% dimethylsulfoxide (DMSO) in liquid nitrogen until time of use.

Targeted sequencing was performed on bone marrow MNCs. A custom pool of biotinylated capture probes targeting 97 genes frequently found mutated in myeloid

malignancies and clonal haematopoiesis were used to capture specific regions of the genome. PCR was used to amplify captured DNA libraries prior to purification, equimolar pooling, and paired-end read sequencing using an Illumina NextSeq 500. Reads were mapped to hg38, and the resulting BAM files were used as input into either VarDictJava v1.7.0 [228], and MuTect2 [229] for somatic variant calling. Quality control filters were applied post variant calling to exclude possibility of sequencing artifacts. The presence of a driver mutation of VAF > 0.01 was used as a criterium to annotate samples as having CH.

2.3 Flow cytometry analysis and cell sorting

2.3.1 Primary human bone marrow thawing and staining

Cryopreserved samples were thawed at 37°C in a water bath, followed by dropwise addition of 1mL warm FBS. The resulting cell suspension was made up to 10mL with thaw media (IMDM + 20% FBS and 100mg/ml DNase). The cell suspension underwent centrifugation for 10mins at 400g, before the pellet was resuspended in 1mL FACS buffer (IMDM + 20% FBS and 100mg/ml DNase). The suspension was passed through a 35 μ M strainer to filter residual clumps. Cell numbers were estimated using an automated cell counter (NucleoCounter).

Antibody incubations were carried out at 4°C in the dark for 45mins – 1 hour, using flow cytometry staining mixes. Antibody staining mixes were made up as indicated in Table A.2 One antibody staining volume (100 μ L full antibody mix) was used per 5x10⁶ cells in solution. Fluorescence minus-one (FMO) cell suspensions were made up with at least 15,000 cells, and stained with 25 μ L FMO antibody mix per fluorochrome. Proceeding antibody incubations, cells were washed with FACS buffer, centrifuged at 350g for 5 minutes, and resuspended in either FACS buffer (no stain control, or FMO PE-cy5), or live dead staining buffer (FACS buffer with 7-AAD (final concentration 1:100)). Cells were kept on ice for the duration of the sorting process.

2.3.2 Primary human bone marrow single cell sorting

For FACS sorting during the TARGET-seq+ testing or patient sample experiments, either CD34+ (MACS) enriched BMMNCs or unseparated BMMNCs were used. An aliquot of each sample was analysed by flow cytometry prior sorting. A preliminary bulk pre-enrichment sort for 80% CD34+ and 20% Lin+ cells was performed into FACS buffer containing 7-AAD (1:100) for vials with unseparated BMMNCs. Gates were set on a combination of FMO negative population intensities, and identification of clear populations of cells. Dead cells and doublets were excluded.

For single-cell sorting, the sorter was first calibrated for sorting accuracy first using sheath fluid and then cells from the FMO aliquots. Single cells were sorted into plates containing lysis buffer that were kept at 4°C throughout the sorting process.

For CH and healthy control bone marrow samples, populations of Lin-CD34+ cells and Lin+ cells were sorted onto each plate (outlined in Figure 3.4), such that each plate was constituted of 90% Lin-CD34+ and 10% Lin+. For a negative genotyping control, 14 CD34+ cells from the healthy control sample NOC101 were sorted on every on every plate. For a no-template genotyping control, two wells were on every plate were left empty. To minimise RNA degradation, only 192 cells were sorted on each 384 well plate, and a maximum sort time of 15 minutes was set for each plate. Plates were immediately centrifuged, snap frozen on dry ice, and stored at -80°C. The following mean fluorescence intensities were recorded for index sorting analysis: FSC:A, FSC:W, SSC:A, SSC:W, Lineage-Live/Dead (PE-Cy5), CD34 (APC), CD117 (BV785), CD38 (BV421), CD45RA (BB515), CD10 (BV605), CD90 (PE-Cy7), CD49f (PE-DAZZLE), CD123 (PE), CD41 (BV510).

Cells were sorted and analysed using a Sony MA900. For flow cytometry analysis, cell compartments were quantified using FlowJo (v10.10), and index sorting data was analysed using R (v4.3.3).

2.4 TARGET-Seq+ primer design

Targeted genotyping primers were designed to capture the same mutation independently from both gDNA and cDNA during the pre-amplification RT-PCR

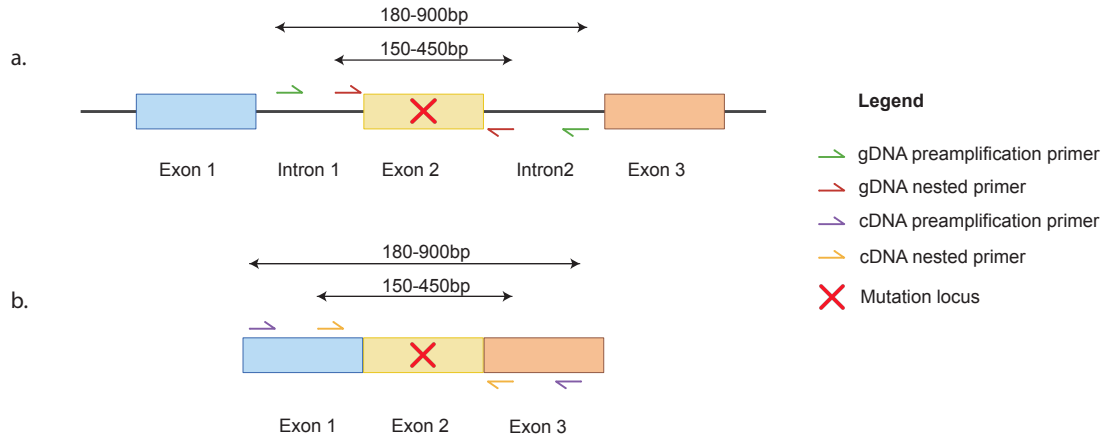


Figure 2.1: Genotyping primer design. Design of primers to capture locus from gDNA (a) and cDNA (b). Pre-amplification primers were designed <900 bp and nested primers were designed inside the pre-amplification primers and <450bp. Primers were designed in the introns (gDNA) or exons (cDNA) flanking the mutation.

step of TARGET-seq+ (Figure 2.1). Primers targeting gDNA were designed to anneal within the intronic regions directly adjacent to the exon carrying the mutation, and cDNA primers were designed to anneal in exonic regions flanking the gDNA primer sites. Therefore, independent amplicons were produced from the gDNA and cDNA for each gene. For genes where this design was not possible, for example those where the exons bearing the mutation were exceptionally long, the same primer pair could be used to produce amplicons from both gDNA and cDNA. However, the nucleic acid of origin would not be able to be determined from these amplicons. Primer pairs were designed using Primer3+ [230] to produce an amplicon length of 180-900bp, and PrimerBLAST [231] was used to check annealing specificity. At least 3 primer pairs were designed for each locus of interest. Pre-amplification primers are detailed in Table A.11.

Primer pairs nested within each preamplification genotyping amplicon generated during RT-PCR pre-amplification were designed to further enrich specifically for the locus of interest during the PCR1 step. As with the RT-PCR primers, PCR1 primers were designed to generate target-specific amplicons from either gDNA or cDNA. Nested primer pairs were designed using PrimerBLAST to a length of 150-450bp. At least 3 nested primer pairs were designed for each locus of interest.

Plate-specific barcode identifier sequences and CS1 and CS2 adapter handles were added to the 5' end of each primer. Nested primers are detailed in Table A.12.

2.5 TARGET-Seq+ genotyping primer testing

For each specific target, combinations of pre-amplification primer pairs were tested using a PCR reaction with either a bulk gDNA or cDNA template extracted from human AML cell line MOLM13. The same conditions used in the pre-amplification PCR reaction including the use of KAPA HiFi HS Polymerase (Roche), were implemented for each reaction. Gel electrophoresis was used to assess primer specificity and amplification efficiency (Figure 2.2). Similarly, combinations of nested PCR1 primers were tested in a PCR reaction with bulk gDNA or cDNA template. The conditions for this reaction replicated those for PCR1, including the use of KAPA2G Robust HS Polymerase (Roche). Again, gel electrophoresis was used for assessing primer specificity and efficiency.

Next, primer combinations were validated using single cells. Primer pair combinations were taken forward for use with patient samples if:

1. a successful cDNA library could be generated from single cells, with no evidence of primer dimer formation or concatenation.

For this, Lin-CD34+ cells from healthy control human bone marrow were sorted into 96-well plates containing 6 μ L of lysis buffer. To test primer compatibility with whole-transcriptome library preparation, RT-PCR reactions were performed for each single cell with combinations of multiplexed cDNA and gDNA preamplification primers corresponding to each patient. For each primer pair combination tested, at least 8 single cells were used. Additionally, 8 single cells underwent RT-PCR as no-genotyping primer positive controls. Purified cDNA libraries from each single cell were assessed by fragment size analysis using a High Sensitivity Fragment Analysis Kit (1 – 6,000 bp) with a Fragment Analyzer (Agilent). Primer pair combinations were excluded if excessive concatemer formation or primer dimers were observed

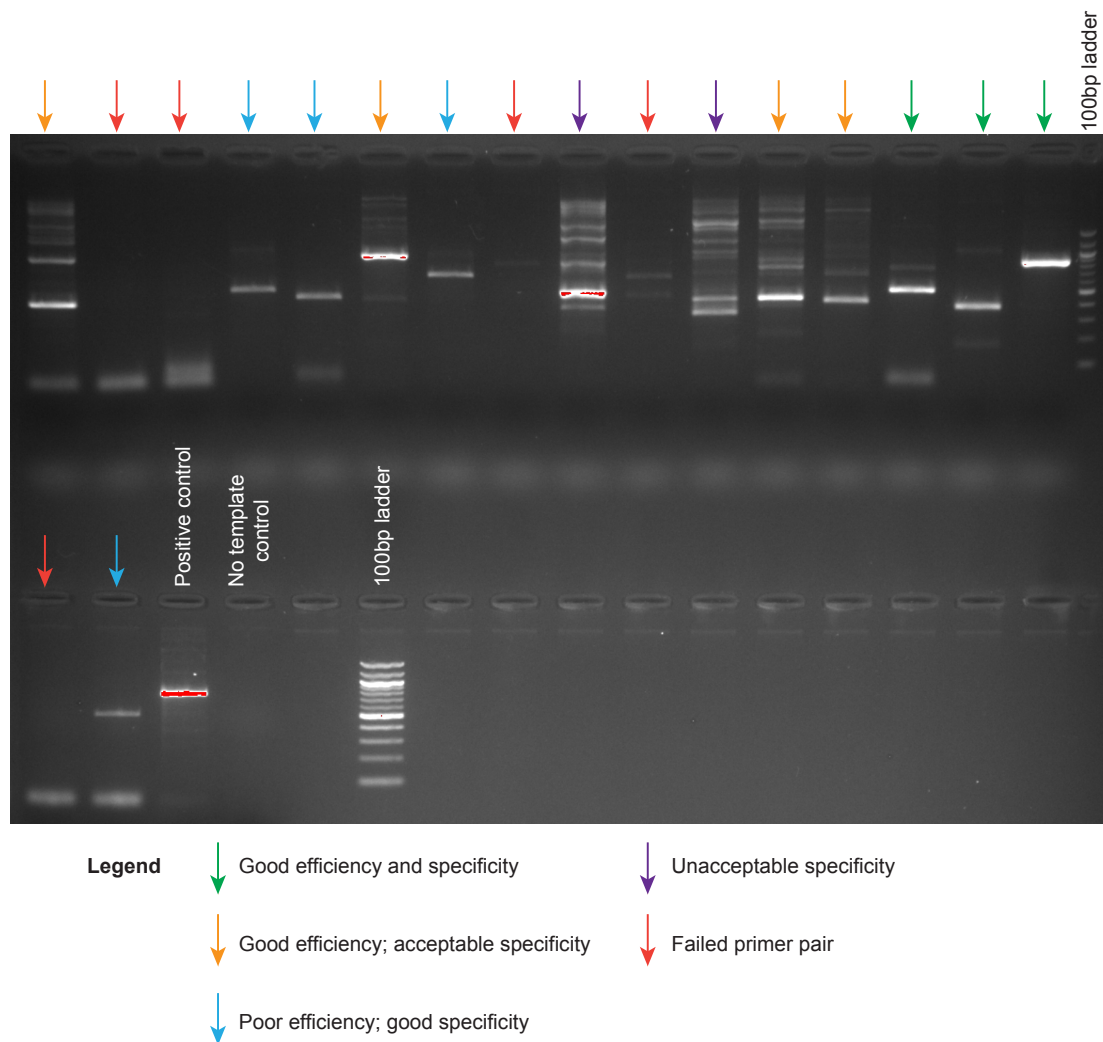


Figure 2.2: Genotyping primer testing using bulk gDNA or cDNA. Representative agarose gel image for testing genotyping primer pairs for specificity and efficiency. Colour of arrow represents efficiency and specificity of primer pair. Pre-amplification and nested primers were tested using the same reaction conditions as TARGET-seq+.

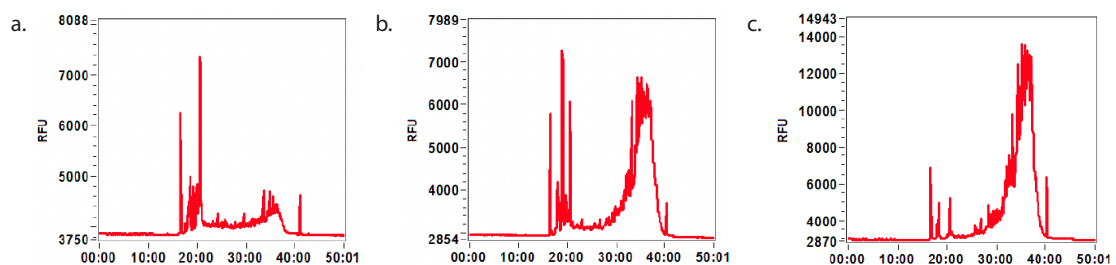


Figure 2.3: Single-cell cDNA library formation. Representative cDNA library traces from single cells. Libraries were generated simultaneously with amplification of genotyping amplicons. a) Poor cDNA yield with excessive primer dimers. b) Good cDNA yield but excessive primer dimers. c) Good cDNA yield but few primer dimers

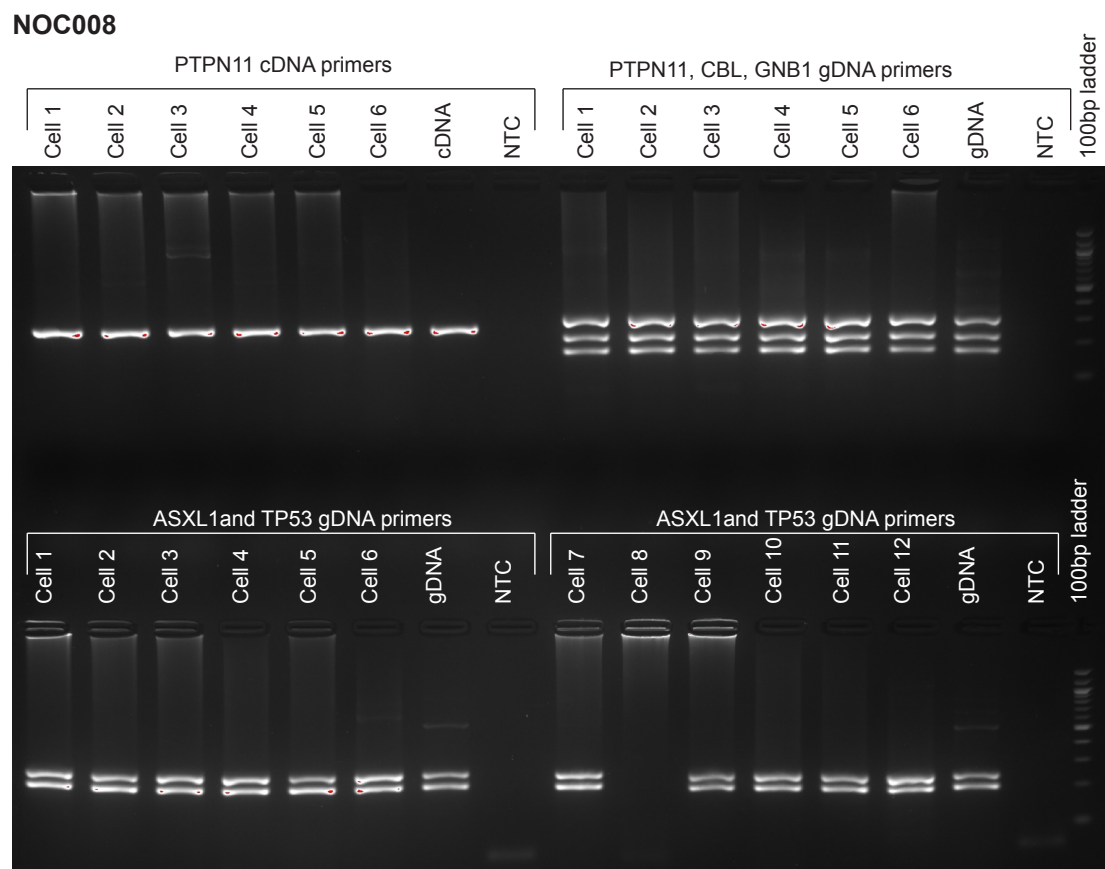


Figure 2.4: Single-cell genotyping amplicon testing. Agarose gel electrophoresis image of single-cell genotyping PCR1 testing. Cell number corresponds to the starting material from a single cell that had undergone RT-PCR. Primer pairs were multiplexed so that multiple genotyping amplicons were produced in each reaction. Successful primer pair mixes produced distinct bands, corresponding to the size of the amplicons.

in the resulting cDNA library. Example fragment analyzer traces of successful and unsuccessful libraries are shown in Figure 2.3.

Multiplexed nested cDNA and gDNA genotyping primers were assessed in a subsequent PCR reaction using a sample of the unpurified cDNA library-genotyping amplicon mix from each single cell. At least 6 single cells were used per primer-pair combination. Gel electrophoresis was used for assessing primer specificity and efficiency (Figure 2.4).

2.6 TARGET-Seq+ library preparation

TARGET-seq+ is an optimised version of TARGET-seq [232]. TARGET-seq+ improves on the original protocol by utilising RT-PCR conditions from SMART-Seq3 [233] to improve sensitivity of transcript detection, resulting in a greater number of cells passing quality control filters. Full details of these modifications applied in the TARGET-Seq+ protocol can be found in Jakobsen et al., 2024 [45]. Oligonucleotides used in TARGET-seq+ can be found in Table A.10.

2.6.1 Lysis buffer preparation

Lysis buffer was prepared in a sterile hood disinfected with RNase inhibitor, inside a sterile clean room. Half-plates (192 wells) of lysis buffer containing well-specific barcoded oligo-dT oligonucleotides were prepared as follows. A lysis buffer stock solution was prepared by mixing poly-ethylene glycol (40%; 5% final concentration), Triton X-100 (10%; 0.1% final), ERCC spike in mix (1:400,000 dilution), dNTPs (10mM; 0.5mM), Protease (1.09 AU/mL; 2.7×10^{-5} AU/mL) and nuclease-free water. To make a stock plate of lysis buffer (enough for 8 sort plates), 25 μ L of this solution was dispensed per well to alternating columns of a fresh plate, using a Mantis robot (Formulatrix). By matching the plate coordinates, 1.79 μ L of individually well-barcoded oligo-dT (HPLC-purified; 10 μ M; 0.5 μ M final concentration) were added to each well of the stock plates using the Integra ViaFlo robot. Sequences of the cell-specific barcoded oligo-dT can be found in Jakobsen et al., 2024 [45]. Using fresh pipette tips, 3 μ L of the lysis buffer-barcoded oligodT mix was transferred to each well of fresh plates. This generated sort plates with unique oligo-dT barcodes in each well. Plates were centrifuged, snap-frozen, and stored at -80°C .

2.6.2 Reverse transcription and pre-amplification

Single-cell reverse transcription of mRNA and subsequent PCR pre-amplification of cDNA libraries (RT-PCR) was performed in a sterile hood disinfected with RNase inhibitor, inside a sterile clean room. Plates with sorted cells were kept on dry ice

until processing. Dilutions of forward and reverse primer mixes ($25\mu\text{M}$ for RT-PCR spike in primers; $100\mu\text{M}$ for pre-amplification spike-in primers) were prepared for each mutant locus. Mastermix for RT-PCR was made up as follows: TrisHCl pH 8.3 (1M; 25mM final concentration), NaCl (1M; 30mM), MgCl_2 (100mM; 2.5mM), GTP (100mM; 1mM), Dithiothreitol (DTT; 100mM; 8mM), $25\mu\text{M}$ spike-in genotyping primers (cDNA primers only; 70nM final concentration) dilutions; nuclease-free water, RNase inhibitor ($40\text{U}/\mu\text{L}$; $0.5\text{U}/\mu\text{L}$), Maxima H-minus reverse transcriptase ($200\text{U}/\mu\text{L}$; $2\text{U}/\mu\text{L}$), HPLC-purified Smart-Seq2 template-switching oligo with locked nucleic acid (SS2-TSO-LNA; $100\mu\text{M}$; $2\mu\text{M}$). This mix was kept on ice and the TSO was added at the last possible moment to avoid degradation. Plates were first incubated at 72°C for 15 minutes in a thermocycler to first lyse the cells. Plates were then transferred to a cold rack (4°C) and $1\mu\text{L}$ of RT-PCR mix was dispensed per well using the Mantis robot (Formulatrix). Plates were immediately sealed, and transferred to a thermocycler to undergo a reverse transcription and template protocol as follows: 42°C for 90 minutes; 10 cycles of 50°C for 2 minutes and 42°C for 2 minutes; 70°C for 5 minutes.

During reverse-transcription, a pre-amplification mix was added as follows: Kapa Hifi HotStart buffer (5X; 1X final concentration), dNTPs (10mM ; 0.3mM), MgCl_2 (100mM ; 0.5mM), Kapa Hifi HotStart polymerase ($1\text{U}/\mu\text{L}$; $0.02\text{U}/\mu\text{L}$), HPLC-purified ISPCR sequence primer ($10\mu\text{M}$; $0.05\mu\text{M}$), spike-in genotyping primer mixes (cDNA and gDNA primers; $100\mu\text{M}$; 400nM), nuclease-free water. After RT, $10\mu\text{L}$ of this mix was added to each well using the Mantis robot. Plates were sealed, centrifuged and transferred to a thermocycler in non-PCR-product-sterile lab, for the following PCR protocol: 98°C for 3 minutes; 21 cycles of: 98°C for 20 seconds, 67°C for 15 seconds, 72°C for 6 minutes; 72°C for 5 minutes. Following PCR, the plates were centrifuged, snap frozen and stored at -80°C .

2.6.3 Whole-transcriptome library preparation and sequencing

Following RT-PCR, each successful well should contain a single-cell cDNA library, and genotyping amplicons spanning the genomic locus targeted by the spiked-in

genotyping primers. An aliquot of each well was diluted with water in a 1:2 ratio in separate plates, for use in targeted single-cell genotyping. At this point, cDNA molecules have well-specific barcodes, but no plate-specific barcodes. Individual plate-specific whole-transcriptome library pools were generated by taking a 1 μ L aliquot from each well per plate, using a Mosquito liquid handler (SPT Labtech). Pools were purified by two separate AMPure bead size exclusions (both 0.6:1 beads:cDNA library ratio), and cDNA libraries were checked using the Agilent Bioanalyzer (High Sensitivity DNA kit), or using a High sensitivity NGS fragment analysis kit (1-6000bp) on a fragment analyser. Libraries were quantified using a Qubit 1x dsDNA HS assay (Life Technologies).

Next, to generate 3'-biased cDNA libraries with cell-specific oligodT barcodes, purified libraries were subjected to tagmentation-based library prep with subsequent PCR amplification using an Illumina Nextera XT DNA library preparation kit. Purified cDNA libraries were first diluted to 0.8ng/ μ L, and 5 μ L of this dilution (4ng) from each 192-cell pool was mixed with 10 μ L tagmentation buffer (TD), and 5 μ L ATM tagmentation enzyme. Each reaction was incubated at 55°C for 10 minutes to tagment the cDNA libraries. Tagmentation reactions were stopped by the addition of 0.2% SDS, to neutralise the Tn5, releasing it from the cDNA. Libraries were amplified in a PCR reaction using NPM polymerase (Illumina), barcoded Illumina Nextera XT i7 forward index primers, custom HPLC-purified i5 index primers. The custom i5 primers are designed to anneal to the ISPCR handle of the barcoded oligodT sequence, resulting in enrichment of the cell-barcode containing 3' cDNA fragments. Sequences of custom i5 primers can be found in Jakobsen et al., 2024 [45]. The combination of the i7 and i5 indexes act as the plate barcodes for the cDNA library. The PCR protocol was as follows: 95°C for 30 seconds; 14 cycles of: 95°C for 10 seconds, 55°C for 30 seconds, 72°C for 30 seconds; 72°C for 5 minutes.

Following amplification, each indexed pool was purified by AMPure bead fragment size exclusion (using a 0.7:1 beads to cDNA ratio), quantified by Qubit 1x dsDNA HS assay (Life Technologies). Tagmented libraries were checked by High sensitivity DNA kit using a Bioanalyzer. Libraries were pooled at an equimolar

ratio, and diluted to 4nM for sequencing. Libraries were sequenced using an Illumina NovaSeq S4 flow cell, and custom primers for Read 1 (P5SEQ) and Index 2 (I5-SEQ). Libraries were sequenced to a target depth of 500,000 reads per cell using the configuration, Read 1: 15bp ; Index 1: 8bp; Index 2: 8bp ; Read 2: 200bp.

2.6.4 Targeted single-cell genotyping, library preparation, QC, and sequencing

Single-cell genotyping was divided into two distinct steps encompassing PCR amplification. The purpose of this is to further enrich for the amplicons of interest, and to add plate- and cell-specific barcodes to the amplicons from each cell, generating uniquely barcoded genotyping libraries from each cell.

First, during PCR1 amplicons spanning the mutations were further enriched for using primers targeting sequences nested within the preamplification amplicons. Plate specific barcodes and CS1 (forward primer)/CS2 (reverse primer) adapter handles were also added to each amplicon by these nested primers. Up to 4 loci were amplified in each PCR1 reaction, and the combination of primers used in each reaction was checked for efficiency as well as heterodimer formation during primer testing. Primer mixes (final concentration 300nM) were added to KAPA 2G Robust HS Ready mix (Roche), and 5 μ L was added to a fresh plate using a Mantis robot. A 1.5 μ L aliquot from each well of genotyping stock plates (containing a 1:2 dilution of the RT-PCR plate product) was added to each of these PCR1 plate using an Integra ViaFlo. These plates underwent a PCR protocol as follows: 95°C for 3 minutes; 24 cycles of: 95°C for 15 seconds, 60°C for 20 seconds, 72°C for 1 minute; 72°C for 5 minutes.

A second PCR reaction (PCR2) was then used to add Illumina sequencing flow cell adapter sequences (P5 & P7) and a 10bp cell-specific barcodes to each amplicon. To perform PCR2, 6.2 μ L reactions are made up of 1 μ L PCR1 product, 4 μ L FastStart High Fidelity polymerase (Sigma-Aldrich), and 1.2 μ L of barcoded primer mix (AccessArray Barcode Library for Illumina Sequencers- 384, Single Direction). This primer mix anneals to the CS1 and CS2 adapter handles on the genotyping amplicons. The PCR protocol for PCR2 was as follows: 95°C

for 10 minutes; 10 cycles of: 95°C for 15 seconds, 60°C for 30 seconds, 72°C for 1 minute; 72°C for 5 minutes.

Following PCR2, 1 μ L from each well was pooled using the Mosquito liquid handler (SPT Labtech). Pools were purified by AMPure bead fragment size exclusion (using a 1.2:1 beads: genotyping library ratio), and quantified by Qubit 1x dsDNA HS assay (Life Technologies). Quality of genotyping libraries was checked with Agilent TapeStation (High Sensitivity D1000 screentape and reagents; Agilent). Libraries were then pooled at an equimolar ratio, accounting for the estimated number of amplicons per pool (number of amplicons per cell X number of cells), and also library size (ng/ μ L). The final pool underwent a final double-sided size exclusion AMPure bead purification (0.6:1, followed by a 1.2:1 beads:library ratio). The purified pool was diluted to 4nM and sequenced using an Illumina NextSeq HighOutput kit (300 cycles) using custom CS1 (Read 1), CS2 (Read 2), and CS2rc (Index read 1) primers, and a configuration of 150bp paired end reads and 8bp index reads.

2.7 Single-cell genotyping analysis

2.7.1 Data pre-processing

A custom TARGET-seq genotyping pipeline was used to pre-process raw single-cell genotyping sequencing reads (<https://github.com/albarmeira/TARGET-seq>). Briefly, raw sequencing reads were first demultiplexed using the cell-specific 384-well barcodes added during PCR2, and the plate-specific barcodes added during PCR1, generating fastq files for each single cell. Reads from fastq files were aligned to hg38 using STAR (v2.7.11a), and using the unique primer and barcode sequences, gDNA and cDNA reads were partitioned into separate bam files. BAM files for each amplicon were manually inspected using integrated genomics viewer to verify the variant had been captured.

Variants were called using mpileup (Samtools v1.17) using the options (`-count-orphans`, `-max-depth 999999`, `-min-BQ 30`, `-ignore-overlaps`), and results were summarised separately by amplicon using a custom pipeline. For amplicons where the variant was an indel, the grep command was used to search for sequences $\pm$

10bp either side of the variant nucleotide within the fastq files for each cell. This was repeated for both the reference and variant sequences and summarised in a .csv file.

For some amplicons- especially indel variants- this custom pipeline was not suitable to call variant reads. Wildtype and variant reads were manually counted using the following unix grep command: `for i in $(ls *_R2*.gz); do fastq=$(basename $i); echo -e "$fastq" | gunzip -dc $i | grep -c [MUTATION SPECIFIC SEQUENCE] "`; `done > ASXL1_MUTATION_WT.txt`. Summary .tsv files of wildtype and variant read counts were compiled and used as input to custom R scripts for downstream analysis. As all reads were assumed to be wildtype or mutant for these loci, it is likely that coverage is modestly underestimated for these loci.

2.7.2 Mutation calling

Summary files from the pre-processing pipeline were first processed by custom R scripts for each individual amplicon. The total coverage for each mutant locus for each cell was first calculated. Low coverage cells (<30 reads) were filtered to avoid unreliable genotyping and were designated “Undetected”. Two wells per plate with no template (no cell) were used as a negative control to check that the amplicon was only present in the wells with cells. Next, the variant allele frequency (VAF) for each cell was calculated by dividing the number of variant reads by the total coverage. To account for inherent error rate in PCR and sequencing, WT control cells (without the variant allele) were sorted in 14 wells on every plate. The VAF for the WT cells (i.e. the error rate) was calculated and used to calculate a minimum mutant VAF threshold for calling mutant and wildtype cells: $((3 \times \text{SD WT VAF} + \text{max WT VAF}) + 0.01)$.

A cell was called mutant when:

1. The number of mutant reads ≥ 10 **and** the VAF $\geq$ the minimum mutant VAF threshold.

A cell was called undetermined when either:

1. The number of mutant reads < 10 and the VAF $\geq$ minimum mutant VAF threshold **OR**
2. The VAF $<$ minimum mutant VAF threshold but VAF $\geq$ the undetermined VAF threshold ($3 \times \text{SD WT VAF} + \text{mean WT VAF}$).

A cell was called WT when:

1. The number of mutant reads < 10 and the VAF $<$ undetermined VAF threshold.

For those mutations where both cDNA and gDNA amplicons were captured, genotyping calls were assigned as follows:

1. A cell was called mutant if the mutation was detected in either the cDNA or the gDNA amplicon.
2. A cell was called WT if the both the cDNA and gDNA calls were WT
3. A cell was called WT if the gDNA amplicon was WT but the cDNA amplicon was undetected or undetermined
4. A cell was called undetermined if the gDNA amplicon was undetected or undetermined, but the cDNA amplicon was WT. This is due to the higher allelic dropout rate of cDNA amplicons.

2.7.3 Estimation of allelic dropout rate

For this dataset, only the WT allelic dropout (ADO) rate was able to be determined. For the mutations that were expected to be heterozygous, the ADO rate was calculated by dividing the number of homozygous mutant cells by the total number of mutant cells. For each cell, thresholds were set to call cells either homozygous reference, homozygous mutant, or heterozygous:

1. Cells were called homozygous mutant when VAF $\geq (1 - \text{minimum mutant VAF threshold})$ or the number of WT reads < 10 .

2. Cells were called homozygous reference when $\text{VAF} < (1 - \text{minimum mutant VAF threshold})$ or the number of mutant reads < 10 .
3. Cells were called heterozygous when the cell was called as mutant but $\text{VAF} < (1 - \text{minimum mutant VAF threshold})$.

2.7.4 Clonal hierarchy assignment

For mutant samples with only an *ASXL1* mutation was present, the clones were designated the same as the genotyping calls. For the two *ASXL1*-mutant CH samples with multiple mutations (NOC008, NOC137), the co-occurrence or independence of mutations was determined by visual inspection by plotting the genotypes in a raster plot. This would determine whether the clonal structure was linear or branching. Using the example mutations A, B, and C, a linear clonal structure would follow: $A \rightarrow AB$, whereas a branching clonal structure would follow: $A \rightarrow AB$ or AC .

For NOC137, a linear clonal structure was unambiguous. The dominant clone harboured the initiating mutation (A; *ASXL1* p.R965X), without the cooperating mutation (B/C; *TET2* p.R1216X), consistent with bulk sequencing VAFs. In the rare case (2 cells) where the cooperating mutation was present without the founding mutation, it was assumed that ADO of the mutant allele had happened for the founding mutation, and these cells were assigned as the daughter clone with the cooperating mutation.

However, for NOC008, the clonal structure was more complicated. Several daughter clones with cooperating mutations (*GNB1* p.K57E, *CBL* p.417R, *TP53* p.Y220C) branched from the dominant clone with the founding mutation (*ASXL1* p.G646fs). A secondary independent clone with an additional founding mutation (*PTPN11* p.D106N) was present with no cooperating mutations. In the rare case where cells were mutant for both founding mutations, this was assumed to be PCR error for the *ASXL1* amplicon. For this *ASXL1* locus, the homopolymer is prone to polymerase error during PCR, so these cells were assigned to the *PTPN11*-mutant clone. Cells with cooperating mutations with undetermined or undetected genotypes were assigned to the founding clone.

2.8 FACS Index data analysis

FCS files were analysed with FlowJo (v10.10). Gates were set according to the MFIs used to set gates on the sorting instrument. Custom R scripts were used to call cell types with the MFI thresholds for each gate. Cell types were defined as follows: **HSC** (Lin-CD34+CD38-CD10-CD45RA-CD90+) **LT-HSC** (Lin-CD34+CD38-CD10-CD45RA-CD90+CD49f+), **ST-HSC** (Lin-CD34+CD38-CD10-CD45RA-CD90+CD49f-), **MPP** (Lin-CD34+CD38-CD10-CD45RA-CD90-), **LMPP** (Lin-CD34+CD38-CD10-CD45RA+), **MLP** (Lin-CD34+CD38-CD10+CD45RA+), **CMP** (Lin-CD34+CD38+CD10-CD45RA-CD123mid), **GMP** (Lin-CD34+CD38+CD10-CD45RA+CD123mid), **MEP** (Lin-CD34+CD38+CD10-CD45RA-CD123lo), **pDC** (Lin-CD34+CD38+CD10-CD45RA+CD123hi), **B-NKT** (Lin-CD34+CD38+CD10+CD45RA+). The cell type assignments were united with single-cell genotyping assignments using cell-identifier as a basis for metadata for single-cell transcriptome analysis.

2.9 Single-cell transcriptome analysis

For the following section, it must be noted that data from two separate sequencing runs were combined for these analyses. Both datasets were analysed identically. Details of both sequencing runs are outlined in Chapter 3.

2.9.1 Data pre-processing

BCL files of raw sequencing intensities were demultiplexed into plate-specific FASTQ files using bcl2fastq (v2.20.0.422) based on unique combinations of custom i5 and i7 barcodes. Using a custom python pipeline, reads were partitioned by cell, mapped, and feature-cell matrix was generated. Cutadapt (v4.9) was used to demultiplex FASTQ files further into cell-specific reads using the unique 14bp cell barcode (Read1). Simultaneously, cutadapt also trimmed poly-A tails and Nextera adapters from Read 2 (cDNA), using the parameters: `-nextseq-trim=0.15`

-A "A20N80;min_overlap=3" -A AGCAACTCTGCGTTGATACCACTGCTT –
minimum-length :25 -e 0.15 –pair-filter=any.

STAR (v2.7.11a) was used to map reads to hg38/GRCh38 reference genome and ERCC92 transcripts, with the GENCODE v38 reference gene annotation. STARsolo (v2.7.11a) was used to generate a gene x cell counts matrix, using the default parameters apart from : –soloType SmartSeq –soloFeatures Gene GeneFull_Ex50pAS. Quality metrics for sequencing and mapping were calculated using FastQC (0.11.9), MultiQC (v1.11), STAR outputs, and Samtools flagstat (v1.17). The data from both sequencing runs was processed by the exact same pipeline but processed at different times.

2.9.2 Quality control, normalization and highly variable gene identification

Single cell transcriptome analysis was performed using two separate packages: the Bioconductor SingleCellExperiment package (v1.24.0) [234] for QC and normalization and the Seurat (v5.0.3) [235] package for all downstream analysis. The feature counts matrix was united with metadata inclusive of genotyping calls and index sorting data, by matching unique cell identifiers, and plate coordinates. Both raw counts matrices from the two different sequencing runs were united at this point.

Cells that met the following criteria were kept in the analysis:

1. The total number of reads per cell $> 25,000$
2. The number of genes detected was between 3,000 and 15,000
3. The percentage of reads mapping to ERCC spike in transcripts < 40
4. The percentage of reads mapping to the mitochondrial genome < 15

Genes were kept in the analysis if they were expressed in greater than or equal to 10 cells.

To account for difference in sequencing depth between, reads were normalised by deconvolution using the scran package (v1.30.2) [236]. This method clusters cells based on library size and normalises pools of cells with calculated size factors.

Highly variable genes were identified using the `BrenneckeGetVariableGenes` function from the `M3Drop` package (v1.28.8) [237, 238]. This function fits a linear model to the relationship between mean gene expression, and the coefficient of variance squared (CV^2) of the ERCC spike in transcripts, which is used to estimate the level of technical noise. Genes were considered highly variable if CV^2 was greater than the estimate of technical noise ($FDR < 0.05$). Ribosomal and mitochondrial genes were excluded from the list of HVGs, leaving 16,163 for use in principal component analysis. At this point, all data was converted to a Seurat object for downstream analysis.

2.9.3 Dimensionality reduction, clustering, and data integration

Prior to principal component analysis, cell cycle scores [239] were added to the metadata using the `CellCycleScoring` function from Seurat. Normalised counts were first scaled using the `ScaleData` function. Principal components ($n=100$) were calculated using the previously identified highly variable gene list using the `RunPCA` Seurat function.

To correct for batch effects, data was integrated by the Harmony package (v1.2.0) [240], using the top 100 PC's, and both sample identifier and investigator as covariates. An elbow plot was used to determine the optimal number of harmony-corrected PC's to account for the most variance between genes. A K-nearest neighbours graph was calculated using the `Seurat::FindNeighbours` function with the top 100 harmony-corrected PC's identified, with $k=5$, the annoy metric = "cosine", and the number of trees = 150. Clusters were then identified using the Louvain clustering implemented by the `Seurat::FindClusters` function, with resolution = 2. Uniform manifold approximation and projection (UMAP) analysis was then performed using `Seurat::RunUMAP` function, again using the top 100 harmony-adjusted PCs, the "cosine" metric, the 5 nearest neighbours, and a minimum distance of 0.05. Integration success was assessed by UMAP visualisation, ensuring heterogeneous representation of both covariates throughout the clusters.

2.9.4 Cluster annotation and reference map projection

Marker genes differentially expressed in each cluster were identified by the RunPrestoAll function from the SeuratWrappers package (v0.3.3), which performs a Wilcoxon Rank Sum test and area under the receiver operator curve (auROC) on normalised, log-transformed gene expression data. Upregulated genes in each cluster were identified using Log2FC threshold of 0.25 and FDR (Benjamini-Hochberg) threshold of 0.05.

To annotate clusters, previously identified differentially expressed marker genes were ranked in descending order by p value, and gene set enrichment analysis was performed using the fgseaMultilevel function from the fgsea package (v1.28.0) [241] on gene signatures obtained from 9 different publications characterising human haematopoiesis [239, 242–249]. Clusters were labelled based on GSEA results, canonical marker gene expression, and immunophenotypic surface marker expression.

To independently validate cluster labels, the dataset was projected onto an in-house reference map of *DNMT3A*-, *TET2*-mutant CH, and healthy control haematopoiesis [45]. Importantly, this reference map was generated using TARGET-seq+ and the same pre-processing pipeline as used in this study. Projection was performed using the Symphony algorithm (v0.1.1) [250], using the same covariates used for Harmony integration, and a sigma parameter of 1. Mapping error was calculated using the calcPerCellMappingMetric from Symphony, and cells were filtered that were ≥ 2.5 median absolute deviations away from the median mapping score. Cell identities were predicted using the knnPredict function, taking the cell type identities from the reference map and $k = 5$ as inputs.

To assess which cluster annotations to use for downstream analysis, immunophenotypic compositions of the different annotations were first inspected. Pseudobulk gene expression profiles were calculated for cell types assigned by manual annotation and those assigned by symphony reference map projection using normalised counts as input into the aggregateAcrossCells function from the scuttle package (v1.12.0). This generated a mean gene expression by cell type matrix of for each annotation method. These matrices were combined into one, and correlation matrix was calculated using

the `cormat` function from the `rstatix` package (v0.7.2), with the parameters `method = "pearson"`, `conf.level = 0.95`. The matrix was hierarchically clustered using the `hclust` function from the `dendextend` package (v1.17.1) [251], with the `method = "complete"` parameter. Results were visualised with the `pheatmap` package (v1.0.12). This process was repeated correlating the different cell type assignments with those from the same dataset as the symphony projection reference map. The final cluster annotation was assigned by a combination of the above methods.

2.9.5 Differential abundance analysis

To compare abundance of mutant and WT cells within CH samples, MELD [252]-an algorithm which calculates the likelihood that a sample or group is composed of a particular cellular state (e.g. genotype)- was utilised. Harmony-corrected PCs were used as input to calculate sample and genotype-associated densities. MELD parameter search was used to determine optimal beta values and Knn. Relative densities for mutant and WT cells were calculated and normalised using an L1 normalization method, so for each cell the sum of the values was 1. The likelihood of finding a mutant cell was then calculated as the ratio of mutant cell density : WT cell density. In both analyses using MELD, these likelihood values were then normalised to the mean likelihood across either the HSC/MPP or the HSC1 compartments. This enabled the estimation of clone expansion or contraction across haematopoiesis. These values were visualised by colouring cells in UMAP space (Figs 4.11d, 5.7d). Finally, these normalised likelihood values were averaged across cells and visualised in Figs 4.11c, 5.7c).

2.9.6 Differential gene expression analysis and gene set enrichment analysis

Differentially expressed genes (DEGs) between mutant and wildtype cells were identified using a linear mixed model implemented by the `dream` pipeline from the `variancePartition` package (v1.32.5) [253]. This is based on `limma-voom` [254], and accounts for covariance. Uncorrected raw counts were used as input and normalised using `scrn` normalisation size factors. The “dream” pipeline fits a

linear mixed model to each gene, and the `variancePartition::eBayes` function was used to perform differential expression testing. The clone was used as the test variable, and sample identifier and patient sex were used as covariates for all analyses. Samples with fewer than 5 cells per group were excluded from the analysis. Benjamini-Hochberg correction was used to adjust p-values for multiple testing, and differentially expressed genes were defined as $\text{FDR} < 0.05$.

Differential gene expression analysis (DGEA) on mini-pseudobulks of cells was used to validate DEGs identified by the dream pathway. Pseudobulk gene expression profiles of cells belonging to either the mutant or wildtype clone per plate were calculated using normalised counts as input into the `aggregateAcrossCells` function from the `scuttle` package [255]. An empirical Bayesian model based on negative binomial dispersions was implemented by the `edgeR` package (v4.0.16) [256] to identify DGEs between pseudobulk replicates. A minimum of 3 cells per replicate was used, and the sample identifier and patient sex were used as covariates in the model's formula. Benjamini-Hochberg correction was used to adjust p-values for multiple testing, and differentially expressed genes were defined as $\text{FDR} < 0.05$.

For gene set enrichment analysis (GSEA), DEGs identified from both the dream mixed model and pseudobulk empirical Bayesian model described above, were first ranked by the z-statistic calculated from both analyses. GSEA was performed using the `fgseaMultilevel` function from the `fgsea` package on gene signatures obtained from both MSigDB v. v2023.2.Hs and published studies. Benjamini-Hochberg correction was used to adjust p-values for multiple testing, and significantly enriched pathways were defined as $\text{FDR} < 0.05$. Results were visualised using `ggplot2` (v3.5.1).

2.9.7 HSPC subclustering

The Self-assembling manifold (SAM) algorithm [257] was used to subcluster the HSC/MPP, EMPP, LMPP, and LMPP cycling cell clusters using harmony-adjusted PCs as input, and sample identifier as the batch. UMAP analysis was performed using SAM-weighted PCs and Louvain clustering was used to call clusters. Sub-clusters were annotated using the original annotations as a guide.

2.9.8 Differential regulon analysis

Transcription factor (TF) regulon activity was inferred using pySCENIC [258]. The arboreto package was first used to infer gene regulatory networks using the single-cell pre-processed raw counts and a list of human transcription factors from Lambert et al., 2018 as input [259]. Pruning of candidate regulons was done using a list of TF motif annotations “motifs-v10nr_clust-nr.hgnc-m0.001-o0.0.tbl”. CisTarget was also used for also used during this step, applying the no drop-out masking parameter, and using the “mc_v10_clust” database of human TF motifs that were annotated at both i) 500bp upstream or 100bp downstream from the transcription start site (TSS) and ii) 10kb around a TSS. AUCell with default parameters was used to quantify regulon activity. To identify differential regulon activity between mutant and wildtype cells within CH samples, WilcoxonAUC was used, implemented by the RunPresto function from the SeuratWrappers package.

2.10 GCUT library preparation

2.10.1 GCUT genotyping primer design

The genotyping strategy for GCUT follows a very similar design to that of TARGET-seq+, including the use of preamplification and nested primers. Primers were designed in a very similar manner to that as described in section 2.4, with the following exceptions. Both preamplification (initial amplification at the same time as the CUT&Tag library) and nested primers were designed for GCUT. As only gDNA is assayed in GCUT, genotyping primers were only designed to target the mutant loci in gDNA. Additionally, since mono-nucleosomal fragments (≈ 147 bp) are preferentially amplified due to the chemistry of CUT&Tag, amplicons were designed to a maximum length of 150bp. This enabled amplification of the mutant loci even if it was to fall within a CUT&Tag fragment. All primers were designed using PrimerBLAST, and prediction of heterodimer formation was done using the OligoAnalyzer tool (Integrated DNA Technologies).

Preamplification primers were first tested with bulk gDNA, using NEB Q5 mastermix, and the following PCR protocol: 72°C for 5 minutes; 98°C for 30

seconds; 20 cycles of: 98°C for 10 seconds, 63°C for 10 seconds; 72°C for 1 minute. Gel electrophoresis was used to assess primer specificity and efficiency of amplicon formation. Nested PCR1 genotyping primers were tested as described in section 2.5. All primers were tested in single cells using the K562 cell line to determine any interference with CUT&Tag library formation, and for primer specificity and efficiency. Sequences of pre-amplification primers and plate-barcoded nested primers can be found in Table A.14 and A.15 respectively.

2.10.2 Cell line culture

K562 and JURKAT cell lines were kindly donated by Dr David Cruz Hernandez, and OCI-AML3 were kindly provided by Vassilena Sharlandijeva and Prof. Tom Milne. K562 and JURKAT were maintained in culture with a media containing 90% RPMI-1640, 10% FBS, and 1% penicillin-streptomycin. OCI-AML3 cell line was maintained in culture with a media containing 80% alpha-MEM, 20% FBS, and 1% penicillin-streptomycin.

2.10.3 Lysis buffer preparation

Plates were prepared in a sterile, PCR-product free clean room under a sterile hood. A stock of 1.1X lysis buffer was first prepared as described in Table A.8, containing Tris HCL (pH 8.0 1M), NaCl (5M), 10% sodium dodecyl sulfate (SDS), nuclease-free water. Using a Mantis robot (Formulatrix), 27 μ L were dispensed into a 384-well plate. Using an Integra ViaFlo robot, 3 μ L of well-specific barcoded i7 primers (100 μ M) were added to each well, making a 10 μ M i7 primer lysis buffer mother plate. Separately, 10 μ M i5 primer-lysis buffer solutions were made, mixing the 1.1X buffer solution with 100 μ M plate-specific barcoded i5 primer oligonucleotides (Table A.13). The Mantis was used to dispense 1 μ L of one solution per well of each plate, so each plates contained an individually barcoded i5 primer. The ViaFlo was then used to transfer 1 μ L of i7-barcoded lysis buffer from the 10 μ M mother plate to each well of the freshly prepared plates, changing tips each time to avoid barcode contamination. This resulted in plates consisting of 2 μ L of lysis buffer with 5 μ M plate-specific i5 primers and 5 μ M well-specific i7 primers. Plates were

snap-frozen on dry ice until sorting. Sequences of well-specific barcoded i7 primers can be found in Turkalj et al., 2023 [260].

2.10.4 Bulk CUT&Tag of cell lines

Buffers were prepared in advance as described in Table A.7. Cell lines were quantified using an automated Nucleocounter and 100,000 cells per condition were aliquoted into a lo-bind Eppendorf tube to minimise cell loss. Cells were first washed with 500 μ L washing buffer and centrifuged in a swinging-bucket centrifuge at 4°C for 5 minutes at 300g. The supernatant was aspirated and the pellet was resuspended in 200 μ L Antibody Buffer to lyse the cell membrane. After 5 minutes incubation on ice, the nuclei were again centrifuged for 5 minutes at 500g. The pellet was resuspended with 100 μ L Antibody buffer with varying dilutions of primary antibody (e.g. 1:50 anti-H3K37me3) and incubated overnight at 4°C in a thermomixer set to 500rpm.

Following primary antibody incubation, nuclei were washed with 200 μ L Dig Wash Buffer, and centrifuged for 5 minutes at 500g at 4°C. Nuclei were resuspended in 200 μ L Dig Wash Buffer with 1:100 secondary antibody and incubated at 25°C for 1 hour in a thermomixer set to 500rpm. Cells were washed again with 200 μ L Dig Wash Buffer, and resuspended 50 μ L 300 Washing Buffer with 1:20 proteinAG-Tn5 fusion protein and incubated for 1 hour at 25°C in a thermomixer at 500rpm. Nuclei were washed once more, before suspension in 200 μ L tagmentation buffer, and incubated for 1 hour at 37°C in a thermomixer at 500rpm. The tagmentation reaction was stopped immediately after incubation by adding 200 μ L STOP buffer. Nuclei were centrifuged once more, resuspended in Washing Buffer with DAPI (1:100), and kept on ice in darkness until sorting.

2.10.5 Single nuclei sorting

For single-nuclei sorting, a sample of nuclei were first analysed to set gates. Due to the use of a single fluorochrome (DAPI), FMOs were not required. The sorter was calibrated for sorting accuracy using sheath fluid. Plates containing lysis buffer were thawed and centrifuged. From each cell line 128 single nuclei were sorted onto each 384-well plate, and two empty wells were left as no template genotyping

controls. Plates were covered with adhesive foil, centrifuged and snap frozen prior to storage at -80°C . Nuclei were sorted using a Sony MA900 sorter and data analysis was performed with FlowJo and R.

2.10.6 Preamplification of GCUT libraries

Pre-amplification was performed in a sterile clean room under a sterile hood. Plates were thawed and incubated in a thermocycler for 15 minutes at 65°C . Next, $1\mu\text{L}$ Tween-20 (10%) was dispensed per well to quench SDS present in the lysis buffer. Plates were covered with an adhesive aluminium foil and centrifuged whilst the PCR mastermix was prepared. A preamplification PCR mix of 2X NEB Q5 mastermix and $10\mu\text{M}$ dilutions of preamplification genotyping primers (prepared with nuclease water), was prepared to a final genotyping primer concentration of 100nM (Table A.9). The Mantis was used to dispense $5\mu\text{L}$ of PCR mix to each well, and the plates were kept on ice, covered with a plastic adhesive covering. Plates were centrifuged and underwent a PCR programme as follows: 72°C for 5 minutes; 98°C for 30 seconds; 98°C for 10 seconds and 63°C for 10 seconds for 27 cycles (cycles determined previously by qPCR); 72°C for 1 minute. Following PCR, 20% of the product was transferred to a genotyping stock plate, containing $4\mu\text{L}$ nuclease-free water. The remainder of the PCR plates were covered and snap frozen until CUT&Tag library preparation.

2.10.7 Single-cell CUT&Tag library preparation and sequencing

Preamplification plates were thawed and contents from each plate were pooled into a 15mL Falcon Tube per plate. Each pool was mixed with 5X volume Qiagen PB with $10\text{--}30\mu\text{L}$ of sodium acetate pH 5.5 and the mixture was passed through a QIAQuick spin column attached to a vacuum pump. The library was eluted in $40\mu\text{L}$ EB (two elutions of $20\mu\text{L}$). AMPure beads were used to purify each library a further 3 times: first, a 0.5:1 beads:sample ratio, followed by two separate 1.2:1 beads:sample ratio. Libraries were quantified by Qubit 1x dsDNA HS Kit, and fragment sizes were assayed using Agilent Bioanalyzer (High Sensitivity DNA Kit). Libraries were pooled at an equimolar ratio and diluted to 4nM prior to sequencing.

CUT&Tag libraries were sequenced using an Illumina NextSeq 500/550 MidOutput kit 150 cycles, using 75bp paired-end reads and 8bp index reads.

2.10.8 Targeted single cell genotyping

Targeted single-cell genotyping was performed as described in section 2.6.4. Purified libraries were pooled at an equimolar ratio and diluted to a concentration of 4nM. Libraries were sequenced using an Illumina MiSeq v3 kit (150 cycles), using custom CS1 (Read 1), CS2 (Read 2), and CS2rc (Index read 1) primers, and a configuration of 75bp paired end reads and 8bp index reads.

2.11 GCUT optimisation

2.11.1 Primer concentration titration

All experiments to determine the optimum concentration of i5 and i7 primers were performed with single cells. Briefly, separate 96-well plates of lysis buffer containing a final concentration of either 500nM, 1 μ M, 2 μ M, or 5 μ M i5 and i7 primers were prepared. Single nuclei from K562 cells that had undergone bulk CUT&Tag using anti-H3K27me3 were sorted into each plate. Plates underwent preamplification using the conditions described in section 2.10.6 and CUT&Tag libraries from 96 nuclei per condition were pooled. Libraries were purified as previously described and assessed using an Agilent Bioanalyzer (High Sensitivity DNA Kit).

2.11.2 Comparison of different bead purification strategies

The experiments comparing CUT&Tag library bead purification strategies were as described as for the primer concentration titration with the following exceptions. Lysis buffer was prepared to contain a final concentration of 5 μ M i5 and i7 primers. Following pooling of single-cell CUT&Tag libraries from each plate, the library from one plate underwent AMPure bead purification using a double-sided purification, using a 0.5:1 followed by a 1.2:1 bead:library ratio. The library from a separate plate underwent the same double-sided purification strategy, followed by an additional

1.2:1 beads:library ratio purification. Purified libraries were assessed using an Agilent Bioanalyzer (High Sensitivity DNA Kit).

2.12 GCUT data analysis

2.12.1 CUT&Tag data pre-processing

Single-cell CUT&Tag sequencing data was pre-processed using a published Snake-make pipeline, originally designed to process plate-based single-cell ATAC data [261, 262]. Raw bcl files were first demultiplexed by bcl2fastq (v2.20.422) into FASTQ files for each cell based on the unique combination of i5 and i7 barcode sequences. Reads were trimmed with cutadapt (v4.9) to remove Nextera sequences from the 3' insert end, before being mapped to hg38 using Hisat2. Reads with low mapping quality (< 30) were excluded by Samtools (v1.17), using the -q 30 parameter. The resulting bam files had duplicate reads were removed with Picard (v3.1.1), were indexed using samtools index, and merged for each single cell using samtools merge (using the ulimit -n 16384 parameter to avoid errors caused by high number of cells). QC metrics from the pre-processing pipeline were calculated using Samtools basicQC. The resulting indexed fragment file was used for downstream analysis. Due to the low number of unique fragments per cell, peaks were not called for this dataset.

2.12.2 Quality control and filtering

Downstream analysis of single-cell CUT&Tag data was performed using custom R scripts and the Signac (v1.13.9) [263] and Seurat (v5.0.3) [235] packages. The fragment file from the pre-processing pipeline and hg38 “EnsDb.Hsapiens.v86” annotations from the Ensembl database were used to make a genome bin x cell matrix using 5kb bins. Genomic bins featuring in fewer than 5 cells were removed, and the resulting bin x cell matrix was used as input, along with the fragment file and genome annotation, to make a chromatin assay object. This was then united with metadata including genotyping calls in Seurat.

QC metrics were calculated using the CountFragments, NucleosomeSignal and TSSEnrichment functions from Signac. Cells with fewer than 100 unique

fragments and a mononucleosomal read percentage of less than 15% were excluded from the analysis.

2.12.3 Normalisation, feature selection, and dimensionality reduction

The following steps were performed on individual Seurat objects that corresponded to different histone marks (e.g. all cells that had been incubated with anti-H3K27me3 primary antibody were included in one object). Data were normalised using term-frequency-inverse document frequency [264] (TF-IDF) to correct for differences in sequencing depth across cells, using a scale factor of 10,000. Variable features were identified using the `Signac::FindTopFeatures` with the “q5” flag, which uses the top 95% of all features. Singular-value decomposition was used to reduce the dimensionality of the data and calculate latent semantic indexing (LSI) components. The LSI components 2:30 were used as input for the K-nearest neighbour (k=5), Louvain clustering and UMAP analyses, implemented by the Seurat package. The first component was excluded as it is highly correlated with sequencing depth. Data were visualised using the `DimPlot` and `CoveragePlot` functions from `Signac`, and `ggplot2` package (v3.5.1).

2.12.4 GCUT genotyping sequencing data analysis

Due to the similarity in genotyping library composition, sequencing data pre-processing and downstream analysis (including the same QC, mutation calling, and ADO thresholds) were performed as previously described in the TARGET-seq+ section.

2.13 Figure production

Figures were produced using a combination of `ggplot2`(v3.5.1), GraphPad Prism (v10.02), Adobe Illustrator (v28.5), and BioRender (<https://www.biorender.com>).

Chapter 3

Single-cell genotyping and immunophenotyping of *ASXL1*-mutant clonal haematopoiesis

Contents

3.1	Brief overview and aims for this chapter	60
3.2	Approach to studying <i>ASXL1</i>-mutant clonal haematopoiesis in humans	61
3.3	Brief introduction to TARGET-Seq+	63
3.4	Introduction to the MARCH cohort and sample selection	65
3.5	Experimental design	68
3.6	Single-cell genotyping analysis of <i>ASXL1</i>-mutant CH samples	70
3.6.1	Data QC and filtering	70
3.6.2	Mutation calling	71
3.6.3	Allelic dropout rate estimation	73
3.7	Clonal and immunophenotypic landscape of haematopoiesis in <i>ASXL1</i>-mutant CH samples	76
3.7.1	NOC008	77
3.7.2	NOC082	79
3.7.3	NOC137	81
3.7.4	NOC180	81
3.8	Chapter conclusions	84

I would like to thank Mr Sven Turkalj and Dr Felix Radtke for their help in the TARGET-seq+ experiments, particularly in processing the genotyping plates. I am also grateful to Dr Asger Jakobsen for processing the NOC137 sample, which I re-analysed in this study.

3.1 Brief overview and aims for this chapter

This chapter sets out the experimental strategy to address the first two aims, set out in the previous chapter: to establish the patterns of *ASXL1*-mutant clonal expansion in throughout human haematopoiesis, and to investigate the transcriptomic correlates between *ASXL1* mutation and clonal expansion. Furthermore, presented

in this chapter are the results of single-cell genotyping and immunophenotyping data from primary human CH samples with an *ASXL1* mutation. These results provide insight into the impact of *ASXL1* mutations on human haematopoietic differentiation.

3.2 Approach to studying *ASXL1*-mutant clonal haematopoiesis in humans

The protein product of *ASXL1* has been functionally characterised primarily as a regulator of genomic H2AK119ub distribution, which is associated with a transcriptionally repressive chromatin state. Therefore, I rationalised that phenotypic consequences of mutation would result from altered transcriptional regulation. Previously, functional models investigating the impact of *ASXL1* mutations upon hematopoiesis have reported genome-wide epigenetic dysregulation. Therefore, developing succinct mechanistic hypotheses linking *ASXL1* mutation to clonal expansion in humans would be challenging. Consequently, to study the phenotypic effects of *ASXL1* mutations upon human haematopoiesis, an unbiased experimental strategy was chosen for this study.

What are the causes of clonal advantage of *ASXL1*-mutant cells in human bone marrow? The ideal strategy to study this question would be to directly compare the transcriptional or epigenetic profiles at sufficiently high resolution to detect subtle differences between phenotypically similar mutant and wildtype cells within the same individual. The ability to compare cell types within the same individual is important as it controls for germline variation, and differences in environmental exposures. To be able to make accurate comparisons between mutant and WT cells, high confidence genotype assignment is critical.

As outlined in chapter one, thus far no method exists that profiles differences in distribution of epigenetic modifications between cells of different genotypes in heterogeneous tissue. Alternatively, studying transcription can aid in understanding the downstream consequences of epigenetic change. Single-cell transcriptomics

has emerged as a powerful tool to distinguish cell populations within heterogeneous populations, and assay differences in gene expression. Two methods currently exist that harness this ability to not only assay transcriptional states, but also determine genotype of single cells: TARGET-seq [232], and Genotyping Of Transcripts (GoT) [265].

TARGET-seq, developed by the Mead lab in Oxford, is a plate-based method that assays three layers of information in single cells: whole transcriptome, genotype, and immunophenotype [232]. During TARGET-seq, single cells are first FACS sorted into plates. A separate reverse transcription reaction for each single cell converts mRNA into a uniquely barcoded cDNA library, which is amplified simultaneously with targeted genotyping amplicons generated from both cDNA and gDNA. TARGET-seq is a highly sensitive method, recording an average of 3,500 genes detected per cell in an experiment assaying primary human haematopoietic tissue. The main advantage of TARGET-seq however, is the accuracy of genotyping. By targeting the mutation in both cDNA and gDNA, the authors reported a 98% genotyping success rate for four mutant loci amplified in each cell.

GoT however, is a droplet-based single-cell method that utilises the 10X genomics platform to combine scRNA-seq with targeted genotyping of mutant loci from cDNA [265]. Heterogeneous populations of cells are mixed with cell-barcoded gel emulsion beads and partitioning oil to generate droplets before RT-PCR with targeted genotyping primers. Although the sensitivity of GoT is not compromised by simultaneous genotyping (the authors reported 2,500 genes detected per cell in human haematopoietic tissue, in line with other droplet-based scRNAseq methods), the accuracy of genotyping is heavily dependent on expression level of the mutant gene. For a highly expressed gene in the cells the authors assayed, like *CALLR*, a genotyping success rate of 60-89% was reported, whereas for the lowly expressed *JAK2*, the genotyping success rate was recorded as only 7%. This is significantly worse than what has been reported for TARGET-seq and could limit the confidence of analyses between cells of different genotypes. One advantage of GoT however is

the ability to assay tens of thousands of cells. Plate-based methods like TARGET-seq are heavily laborious and expensive and which limits the number of cells that can be processed.

No overt haematological clinical phenotype is seen in patients with *ASXL1*-mutant CH. Additionally, the mutant clone is likely to have expanded over many years. Therefore, the transcriptomic differences between mutant and wildtype cells are likely to be subtle. Accurately identifying mutant cells using high-fidelity, accurate genotyping methods is therefore vital to draw high confidence comparisons between clones. Consequently, a modified version of TARGET-seq was chosen to assay bone marrow mononuclear cells from *ASXL1*-mutant CH samples and healthy age-matched control samples.

3.3 Brief introduction to TARGET-Seq+

This section outlines the TARGET-seq protocol, and the modifications made by Jakobsen et al., 2024 [45] to improve sensitivity of whole transcriptome library preparation (TARGET-seq+).

The steps outlining the TARGET-seq+ protocol are shown in Fig 3.1, and detailed in Chapter 2. During TARGET-seq, cells are first sorted into individual wells containing lysis buffer. Whole transcriptome cDNA libraries are generated during the reverse transcription and template switching reactions. Targeted genotyping primers for mutant loci in cDNA are included in these reactions to enrich for amplicons containing a mutation from cDNA. Transcriptome libraries are then amplified by PCR. Spike-in targeted genotyping primers are included in the same reaction, which capture the mutation by generating genotyping amplicons from cDNA and gDNA. An aliquot of the product of this pre-amplification PCR reaction is used for further enrichment of genotyping amplicons. This is achieved with an initial PCR reaction using primers nested within the preamplification genotyping amplicons. A further PCR reaction barcodes the genotyping amplicons fully to correspond to each single cell, and to add Illumina flow cell sequencing adapters. The genotyping libraries are then pooled and sequenced.

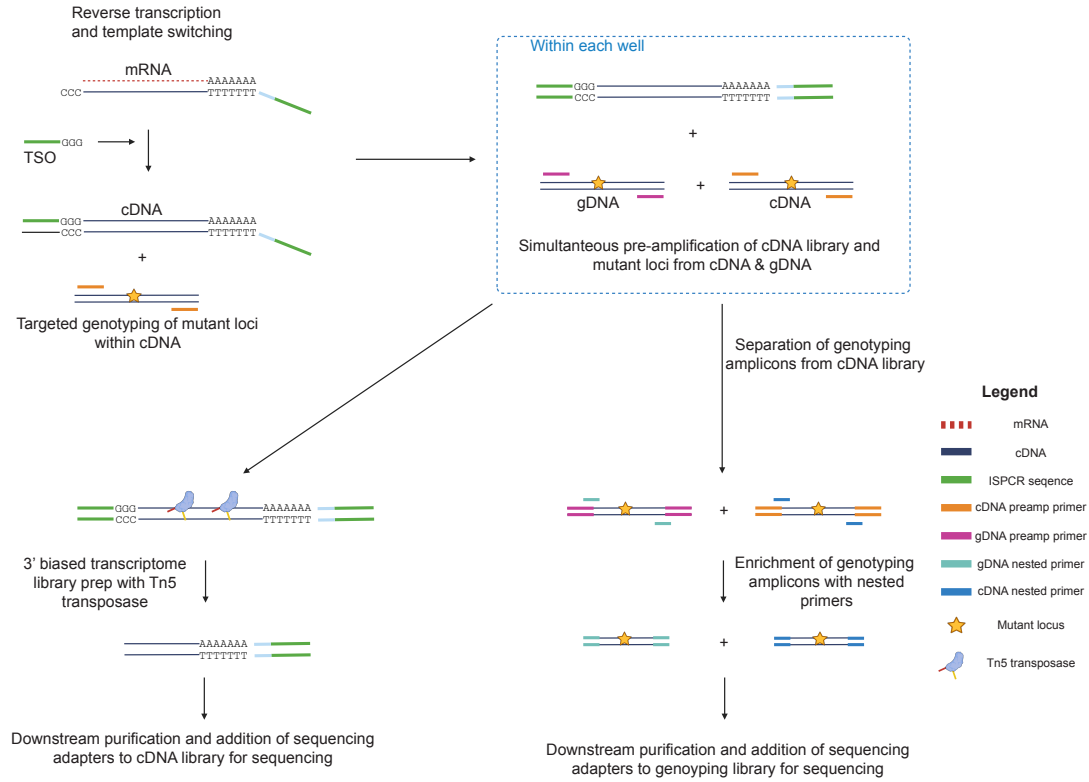


Figure 3.1: Overview of the TARGET-seq protocol. Overview of single-cell whole-transcriptome and genotyping library generation using the TARGET-seq+ protocol, adapted from Rodriguez-Meira et al. 2019 [232]

During transcriptome library preparation, single-cell cDNA libraries that were generated during RT-PCR using cell barcoded oligo-dT oligonucleotides, are pooled by plate. Each pool contains cell-barcoded cDNA libraries from 192 cells on each plate. The libraries are purified and undergo tagmentation with Tn5 enzyme. PCR with custom plate-barcoded primers then enriches specifically for 3' cDNA fragments, generating libraries of cDNA molecules with unique combinations of plate and cell barcodes corresponding to each single cell. Libraries can then be pooled and sequenced.

Jakobsen et al., 2024 modified the TARGET-seq protocol to optimise sensitivity of transcript detection [45]. To achieve this, the authors incorporated elements of Smart-seq3 chemistry [233] into the RT-PCR reaction, such as incorporating the Maxima H-minus reverse transcriptase with buffer including NaCl, 5% poly-ethylene glycol, and GTP. Full details of the modifications can be found in Jakobsen et al.,

2024. The authors reported a 19% increase in the number of genes detected per cell in human Lin-CD34+ cells. When they sequenced transcriptome libraries at a mean depth of 1,000,000 reads per cell, they reported a median of 6,484 genes detected per cell. The modifications improved the detection of lowly expressed genes, which is particularly relevant as it could reveal subtle changes in transcriptional phenotype between cells of different genotypes. The authors used their method to answer very similar questions to those posed in this project, but in the context of human CH with *DNMT3A*- or *TET2*- mutations.

Owing to the increased sensitivity of transcript detection, TARGET-seq+ was chosen to assay *ASXL1*-mutant CH. The use of TARGET-seq+ also enables the direct comparison of dysregulated gene expression between *ASXL1*-mutant CH and *DNMT3A*- or *TET2*-mutant CH.

3.4 Introduction to the MARCH cohort and sample selection

As described in chapter 1, thus far animal models have failed to faithfully recapitulate the clinical observations of CH in humans. This is likely due to failure to account for differences in genetics, lifestyle, and environmental exposures driving heterogeneity in humans. To address why somatic *ASXL1* mutations confer clonal advantage in human CH, the use of primary human tissue will provide the most accurate answers. Additionally, insights gathered from studies in human will aid in the development of more accurate models of CH.

One challenge in studying CH in humans is the lack of availability of human tissue. To study the consequences of *ASXL1* mutations in humans, samples from the Mechanisms of Age-Related Clonal Haematopoiesis (MARCH) study were utilised. This cohort of patients was established by Dr Asger Jakobsen of the Vyas lab, in collaboration with the Nuffield Orthopaedic Centre, Oxford, to further our understanding of clonal haematopoiesis. Full clinical details of the MARCH cohort and ethical practices are found in Jakobsen et al., 2024 [45]. The collection process and variant calling are summarised in Figure 3.2a. In brief, bone marrow was taken

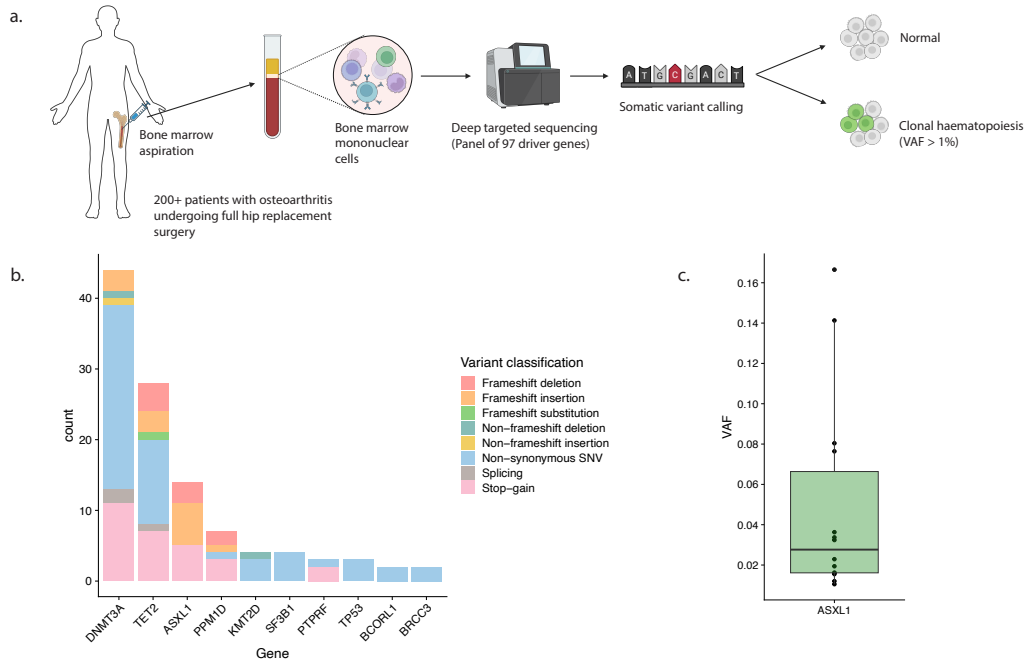


Figure 3.2: Overview of the MARCH cohort. **a)** Overview of sample collection and CH classification in the MARCH study. **b)** Bar plot of the frequency of variants within the top 10 most mutated genes within the MARCH cohort. Bar plot is coloured by the classification of each variant. **c)** Box plot of *ASXL1* variant VAFs, measured by targeted genome sequencing on bulk bone marrow mono-nuclear cells. Each dot represents one patient, since no patients were found to have more than one mutation in *ASXL1*.

from patients with osteoarthritis and no history of haematopoietic malignancy during full hip replacement surgery ($n = 195$; median age = 70). Bone marrow mononuclear cells were isolated by ficoll density gradient separation and subjected to targeted genome sequencing using a panel of 97 known driver genes of CH and myeloid malignancies. Somatic variants were detected using a custom variant calling pipeline, using a threshold of $VAF < 0.01$ to determine the presence of CH for each sample. Using this threshold, 43.5% (85 individuals) of the 195 patients in the cohort were found to have CH with at least one driver mutation.

Of the 85 individuals found to have CH, 13 patients were found to have mutations in *ASXL1* (Table A.1). Consistent with other cohorts of patients with CH [24–26], *ASXL1* was the third most common mutated gene, behind *DNMT3A*, and *TET2* (Figure 3.2b). Nine separate variants of *ASXL1* were found across the 13 patients. The most common mutation was the *ASXL1* p.G646Wfs*12 variant

(n=5), which has previously been reported as a hotspot mutation in myeloid malignancies [181]. However, its incidence in CH has not been reported as frequently [194]. This variant incorporates an additional guanosine residue into an 8-bp guanosine homopolymer, resulting in a frameshift and subsequent premature stop codon. The mean mutant VAF for *ASXL1* of these patients was 0.0526 (range = 0.0105-0.1665). Five patients had single *ASXL1* mutations, and only one patient-NOC064- had two *ASXL1* variants.

Sample	Sex	Age	Gene	Chromosome	Position	Type	Reference	Alternate	VAF	Variant classification	Accession	Exon	cDNA change	Protein change
NOC008	Female	81.6	ASXL1	chr20	32434638	INDEL	A	AG	0.1665	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
			PTPN11	chr2	112450496	SNP	G	A	0.0712	nonsynonymous_SNV	NM_002834	exon3	c.316G>A	p.D106N
			GNB1	chr1	1815790	SNV	T	C	0.0134	nonsynonymous_SNV	NM_002074	exon5	c.169A>G	p.K57E
			CBL	chr11	119278532	SNV	C	G	0.0107	nonsynonymous_SNV	NM_005188	exon9	c.1250C>G	p.P417R
			TP53	chr17	7674872	SNV	T	C	0.0159	nonsynonymous_SNV	NM_000546	exon6	c.659A>G	p.Y220C
NOC082	Female	65.5	ASXL1	chr20	32434638	INDEL	A	AG	0.0804	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
NOC137	Male	73.6	ASXL1	chr20	32435605	SNV	C	T	0.1413	stopgain	NM_015338	exon12	c.2893C>T	p.R965X
			TET2	chr4	105243621	SNV	C	T	0.0405	stopgain	NM_001127208	exon6	c.3646C>T	p.R1216X
NOC180	Female	76.5	ASXL1	chr20	32433562	INDEL	AGACT	A	0.0764	frameshift_deletion	NM_015338	exon11	c.1365_1368del	p.K455fs

Table 3.1: Selected *ASXL1*-mutant samples analysed by TARGET-seq+. Details of the four *ASXL1*-mutant samples that were selected to be analysed by TARGET-seq+. Samples were selected based on size of VAF in bone marrow from bulk targeted DNA sequencing. Other mutations detected in the samples above VAF = 0.01 are also shown.

In order to study the effect of *ASXL1* using primary human samples, sufficient numbers of mutant cells would be needed for meaningful comparisons with wildtype cells. Therefore, a minimum VAF of 0.05 for *ASXL1* (equivalent to $\approx 10\%$ of all cells in the bone marrow) was set as a threshold for sample selection. Of the 13 samples found to have *ASXL1*-mutant CH, 4 samples had mutant VAF > 0.05 (Mean VAF = 0.1162) (Table 3.1). Of these, two samples had the *ASXL1* G646fs hotspot mutation (NOC008 and NOC082). All *ASXL1* mutations resulted in a premature stop codon (three frameshift insertions, and one stop-gain). Additionally, all of these were found within the final two exons (exons 12 and 13), which are predicted to result in expression of a C-terminally truncated protein.

Two cases, NOC008 and NOC137, had additional CH mutations which also need to be considered to account for any confounding effects. NOC008 was also found to have a high-VAF mis-sense mutation in *PTPN11* (VAF =0.0712), and low VAF mis-sense mutations in *CBL* (VAF =0.0107), *GNB1* (VAF =0.0134), and *TP53* (VAF =0.0159). NOC137 had an additional stop-gain mutation in *TET2*

(VAF = 0.0405). The mean age for these patients was 74.3 years. Three patients were female; one was male (NOC137).

Additionally, samples from five healthy patients were selected to act as non-CH controls. No variant was found for VAF > 0.01 in any of the 97 genes in the variant panel. The mean age of these patients was 70.9; four of these were female; one was male.

3.5 Experimental design

To investigate the molecular consequences of *ASXL1* mutations throughout human haematopoiesis, TARGET-seq+ was performed on bone marrow mononuclear cells from the four selected *ASXL1*-mutant CH samples and five age matched control samples. An overview of the experimental strategy for this study is depicted in Figure 3.3.

Clonal haematopoiesis is associated with expansion of a mutant clone, primarily within the haematopoietic stem and progenitor cell (HSPC) compartment. Evidence from murine *in vivo* and *in vitro* models have suggested that *ASXL1*-mutant HSPCs display myeloid-biased differentiation, and reduction in mature lymphoid and erythroid cell types [144, 199, 207]. To investigate how *ASXL1* mutations affect haematopoietic differentiation in humans, a representative cross-section of cell types from the haematopoietic stem cell, and early/late progenitor from the myeloid, lymphoid and erythroid/megakaryocyte lineages were sorted. Since the majority of these cell types are from the immunophenotypic Lin-CD34+ compartment, which composes $\approx 5\%$ of healthy human bone marrow, mononuclear cells from whole bone marrow were first sorted in bulk to enrich for CD34+ cells. Definition of immunophenotypic cell populations for both flow cytometry analysis and sorting is displayed in Figure 3.4a.

Single cells were then sorted into individual wells of plates containing lysis buffer, so that roughly 90% of all cells were immunophenotypic live Lin- CD34+ and 10% of cells were Lin+ (Figure 3.4b). Since the haematopoietic stem and early progenitor cell (HSPC; CD34+CD38-CD10-) compartment take up $\approx 15\%$ of

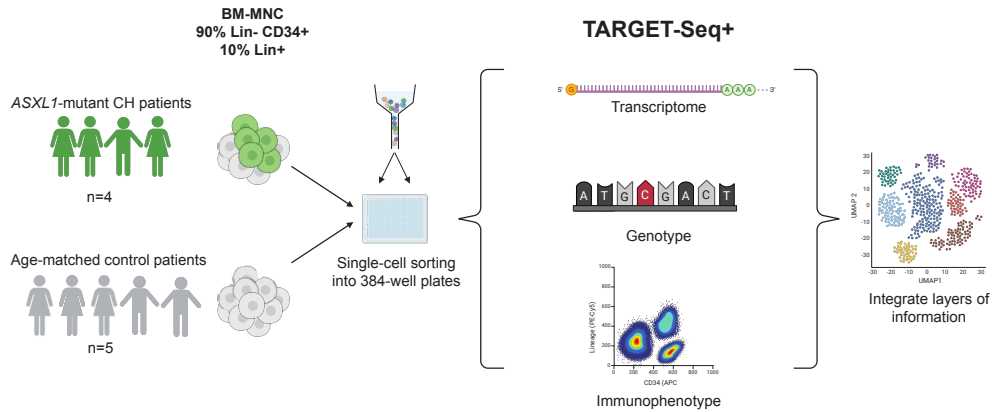


Figure 3.3: Experimental strategy for single-cell analysis of *ASXL1*-mutant CH samples. Schematic of experimental plan

the CD34+ compartment in humans [266], HSPCs were specifically enriched for, comprising 30% of the cells on each plate. Most of the clonal differences are thought to occur in this compartment in CH, so it was important to enrich for them.

On each plate 10% of cells sorted expressed lineage markers, including mature erythroid, myeloid, and lymphoid markers. Full details of the immunophenotypic markers included in the lineage mastermix can be found in Table A.2. The purpose of this was to assay cells from complete differentiation trajectories to identify any differentiation bias in the *ASXL1* mutant clone. Additionally, Lin-CD34+ cells from one healthy control patient (NOC101) was sorted onto every plate. These cells acted as a batch control for both the genotyping (negative control for each locus), and the transcriptome, meaning that any possible batch effect across plates could be identified. Lastly, two wells per plate were left empty, which acted as no template negative controls for single-cell genotyping.

TARGET-seq+ was then performed on each plate, collecting immunophenotype, genotype, and transcriptome data from each cell. By integrating these three data points for each cell, patterns of clonal expansion could be linked to transcriptomic differences across cell types at single-cell resolution.

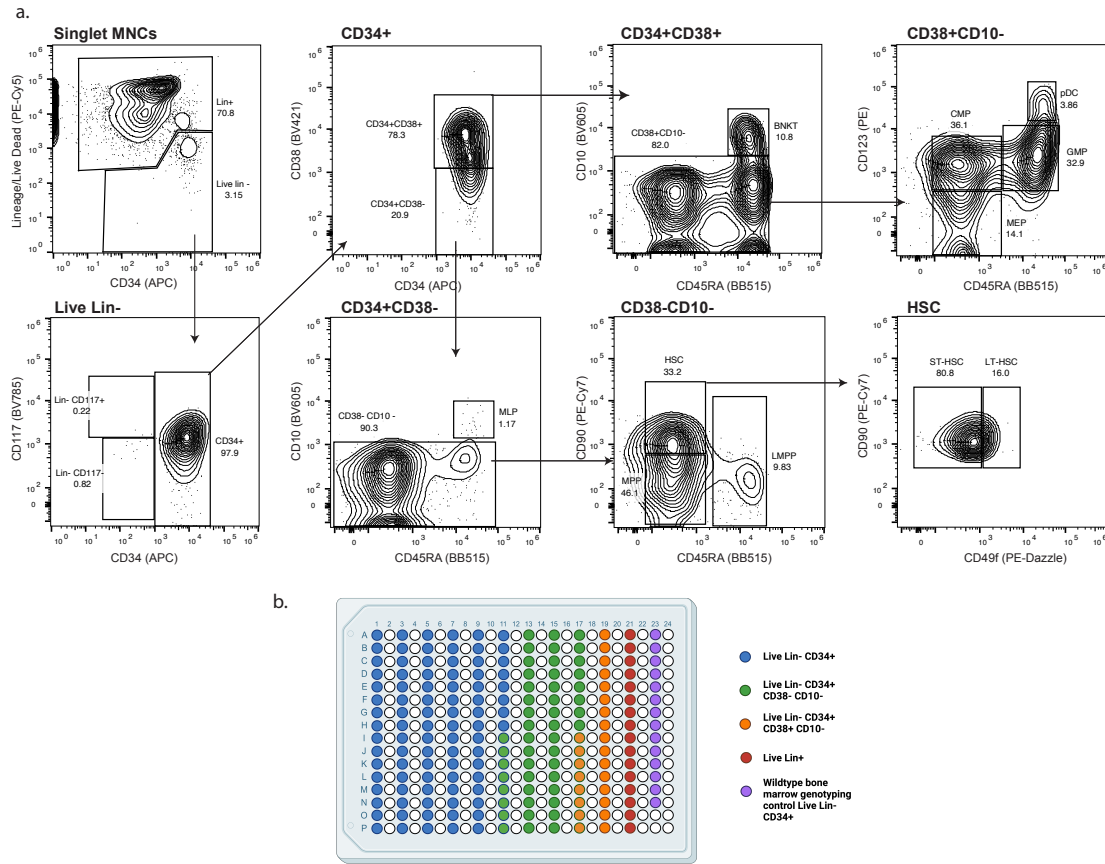


Figure 3.4: FACS sorting strategy for the isolation of single cells . a) Representative flow cytometry contour plots defining immunophenotypic HSPC populations for the NOC101 healthy bone marrow control sample. Each dot represents a single cell. All single mononuclear cells have been passed through two singlet gates. Numbers represent the percentage of cells falling within each gate. **b)** Single-cell FACS plate sorting strategy. Each coloured well corresponds to a different immunophenotypic population, defined as in **a)**. White wells are blank. Cells were sorted into alternating columns to minimise the sorting time, thus minimising mRNA degradation.

3.6 Single-cell genotyping analysis of *ASXL1*-mutant CH samples

3.6.1 Data QC and filtering

A total of 11,400 cells from 4 CH samples and 5 healthy control samples were sorted. The number of cells sorted per sample are detailed in Table 4.1. Cells from healthy WT control samples were not genotyped, unless they had been sorted on the same plate as a CH sample. In total 8,740 cells were genotyped. For the cells

that were genotyped, a mean of 94.33% of cells passed quality control for the gDNA amplicon and 78.47% passed quality control for the cDNA amplicon (Figure 3.5 a,b). In these cells, at least 30 reads covering the variant nucleotide were detected. Importantly, all the blank wells (without a cell) had fewer than 30 reads detected for both the gDNA and cDNA amplicon. Any reads that were detected may be due to cross-well contamination during the genotyping process.

3.6.2 Mutation calling

Next, the sensitivity of single-cell genotyping was investigated. For each cell, a single-cell variant allele frequency (scVAF) was calculated, corresponding to the number of reads containing the variant divided by the total coverage of the locus. Single cell VAFs that varied from 0, 1 (homozygous WT or mutant), or 0.5 (heterozygous) is due to either sequencing error, or allelic dropout. When the single cell VAF distribution for each amplicon was inspected, the pattern for both the cDNA and gDNA amplicons corresponded to cells heterozygous for each mutation only being present in the CH samples (Figure 3.5c,d). Importantly, in the healthy control samples (NOC101 and NOC153) the mean scVAF for both the cDNA and gDNA amplicons was 0.0104, indicating that no heterozygous cells were detected. Any mutant reads that may have been detected in the WT samples are likely due to either cross-well contamination, PCR error, or sequencing error.

The exception to this pattern was seen for the *ASXL1*_G642fs, and *ASXL1*_G646s variants in NOC008 and NOC082 respectively. These two variants are insertions of an extra guanosine nucleotide at either end of an 8bp guanosine homopolymer, and so can be regarded as the same variant. Of note, for these two mutations the maximum scVAF for the CH samples did not reach a maximum of 1. This maximum scVAF is to be expected in some mutant cells (exemplified by the other loci), and indicates amplification of only the mutant allele. Additionally, the mean scVAF for the WT healthy control sample (NOC101) was 0.0143 for these loci- higher than for all other genotyped mutations. This variant has previously been attributed to PCR error [267], due to the phenomenon of polymerase slippage [268]. However,

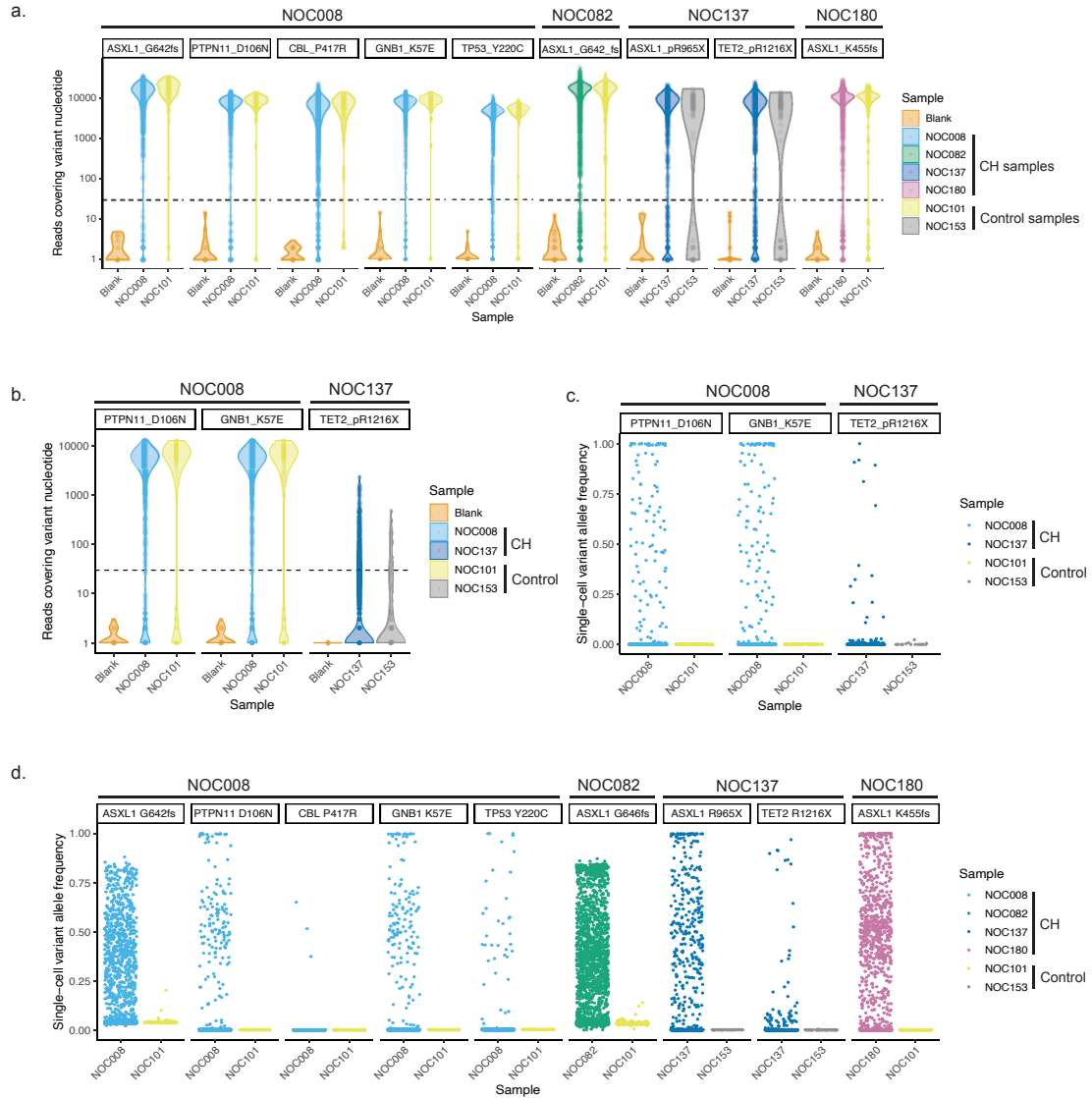


Figure 3.5: Single-cell genotyping quality control. a) Violin plot of reads covering variant nucleotide in gDNA amplicons across all samples. Each dot represents a single cell, coloured by sample. Grouped with each CH sample are the blank wells and WT healthy control samples (either NOC101 or NOC137) that were sorted on the same plates as that CH sample. Horizontal dashed line represents the minimum number of reads across each nucleotide for the amplicon to be considered detected in each cell (n=30). b) As for a) but the reads across the cDNA amplicons for the loci that primers could be designed for. c) Dot plot of single cell VAF distribution (mutant reads/total reads across for each amplicon) for the cDNA amplicons, grouped by sample. Each dot represents a single cell. Grouped with each CH sample are the blank wells and WT healthy control samples (either NOC101 or NOC137) that were sorted on the same plates as that CH sample. d) As for c), but the single-cell VAF distributions for the gDNA amplicons.

more recently a study has shown that this variant is a *bona fide* mutation [269]. The occurrence of mutant reads in the WT sample is likely due to some PCR error or sequencing error (homopolymers can decrease the quality of base calling). This error is therefore likely to be present for the CH samples too. Given that there is a baseline of error from the data from the WT cells, the scVAF from the WT cells can be used to set a threshold for calling mutant cells in the CH samples.

The calculation of thresholds to call genotypes are outlined in chapter 2 and summarised in Figure 3.6a. In brief, cells were called mutant if they had more than 10 mutant reads and the scVAF exceeded the mutant VAF threshold (maximum scVAF from WT cells + 3 x standard deviations). Cells were called WT if they had fewer than 10 mutant reads, and had a scVAF less than the undetermined cell threshold. All other cells were called undetermined, unless they failed QC, in which case they were called as undetected. If gDNA and cDNA amplicons were amplified for each locus, genotypes for each were assigned based on a combination of the two amplicons (criteria outlined in chapter 2). Using these thresholds, across all of the cells that were genotyped, 92.62% cells were successfully assigned a genotype of either mutant or wildtype for each locus (Figure 3.6b; range of mean success rates across loci = 63.3-97%). The number of cells assigned to each genotype per variant locus is outlined in Figure 3.6c.

3.6.3 Allelic dropout rate estimation

Allelic dropout (ADO) is a phenomenon that occurs as a failure of PCR to amplify different alleles at the same rate. ADO occurs when one allele is preferentially amplified over the other. This may occur due to chance, competition for resources during the PCR reaction, or differences in the chemical properties between the mutant and WT alleles. In PCR-based genotyping applications, one serious consequence of ADO is the wrongful assignment of genotypes. For instance, if ADO of the mutant allele occurs in a cell heterozygous for that mutation, only the WT allele will be detected, leading to wrongful assignment of that cell as WT.

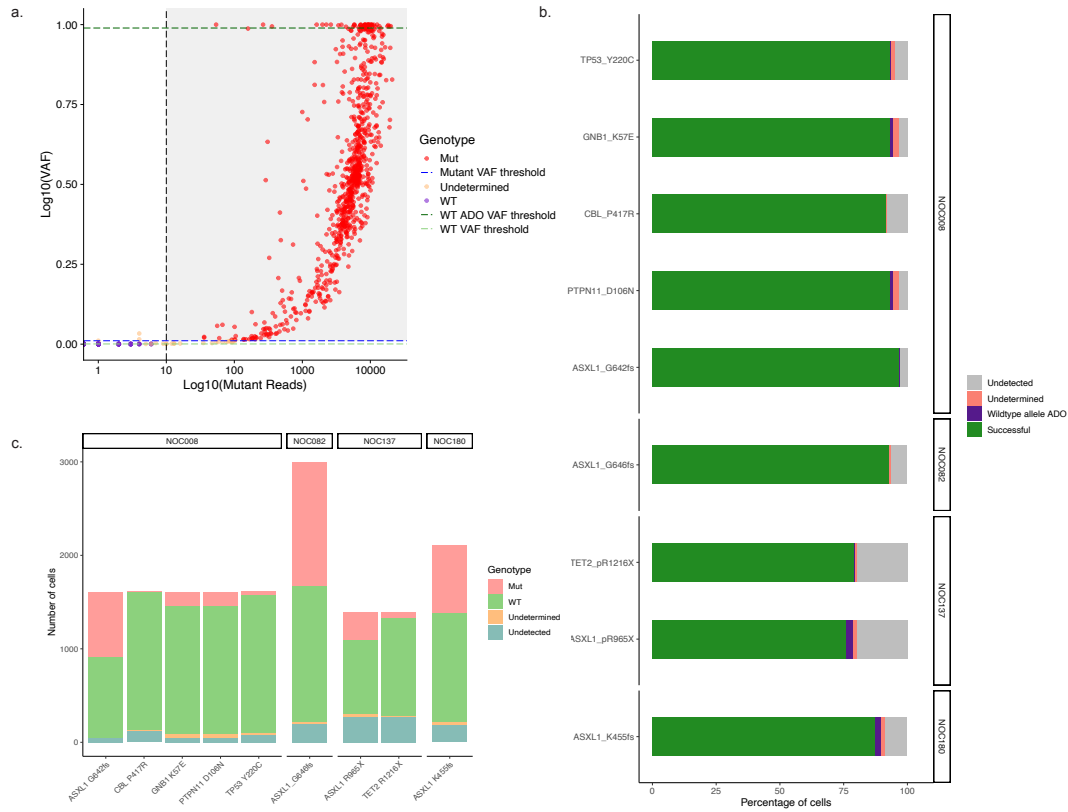


Figure 3.6: Single-cell genotyping mutation calling. **a)** Exemplar scatterplot of process of mutation calling for the ASXL1_K455fs variant locus in NOC180. Each dot represents a cell, and coloured by mutational status. The blue dashed line represents the mutant VAF threshold, and the vertical black dashed line represents the mutant reads threshold ($n=10$). Any cell above these thresholds are called mutant. The light green dashed line represents the WT VAF threshold. Any cell below this threshold is called WT. The dark green dashed line represents the WT ADO threshold. Cells above this threshold are called homozygous mutant, and ADO of the WT allele has occurred for a heterozygous mutation. **b)** Summary of genotyping success in CH samples split by mutant locus. Percentage of cells that were successfully genotyped for each locus are shown in green. Mean percentage of cells that were successfully genotyped for ASXL1: 92.62%. Estimated allelic dropout rate of the WT allele is shown in purple. Mean WT ADO rate across loci: 1.14%. Legend is on the right of the barplot. **c)** Bar plot showing absolute numbers of mutant, wildtype, undetermined and undetected cells per mutant locus in CH samples.

In the case of this study, ADO can have serious implications for the accuracy of downstream analyses.

For heterozygous cells, the ADO rate for the WT allele can be estimated by determining cells that are homozygous mutant. To determine an estimate for WT ADO, a threshold was calculated for each amplicon (WT ADO threshold), as $1 - \text{mutant VAF threshold}$. Cells above this threshold were designated as homozygous mutant, and the percentage of homozygous mutant cells above this was the estimated ADO rate (Figure 3.6a). As wildtype control samples only had wildtype alleles Since homozygous mutant cells could not be identified for the *ASXL1* mutations at the G646fs and G642fs loci in NOC008 and NOC082, this method of calculating ADO was not possible. For all other mutant loci, the mean WT ADO rate was 1.14%, indicating efficient and accurate capture of the mutations in all cells.

The ADO rate for the mutant allele can inversely be determined as the percentage of homozygous WT cells that should be called as mutant. Since *bona fide* WT cells present in these samples will also be homozygous WT for each allele, the percentage of homozygous WT cells is not the WT ADO rate. One method of estimating mutant ADO is detailed in Jakobsen et al., 2024 and Turkalj et al., 2023 [45, 260]. In these publications, authors determined the phasing of heterozygous germline Single Nucleotide Polymorphisms (SNPs) with the mutant and WT alleles within each gDNA genotyping amplicon. If SNPs were in phase with the mutant allele, mutant ADO had occurred in cells that were homozygous reference for both the mutation and the SNP. If SNPs were out of phase with the mutant allele, mutant ADO had occurred in cells where the mutation was homozygous reference, but the SNP was homozygous alternate. By combining the two, the mutant ADO rate could be inferred. Unfortunately, heterozygous germline SNPs documented in the National Institute of Health's SNP database, dbSNP, could not be identified in any of the genotyping amplicons for any loci in the *ASXL1*-mutant CH patients, so mutant ADO rate could not be estimated

Finally, when comparing the percentage of *ASXL1*-mutant cells of the samples that were sorted, with the expected percentage of mutant cells calculated from the

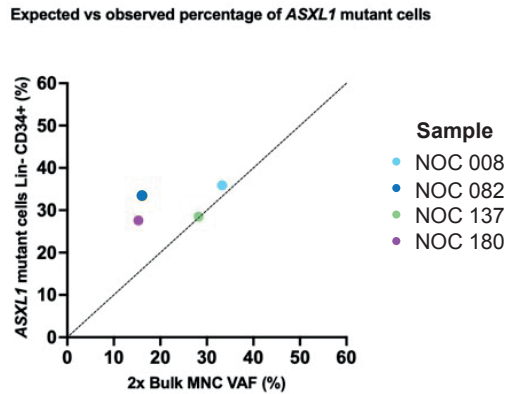


Figure 3.7: Correlation of mutant fractions determined by single-cell genotyping and bulk variant calling. The fraction of *ASXL1*-mutant cells from single cell genotyping versus the cell fraction of the *ASXL1*-mutant clone from bulk bone marrow DNA sequencing. Each dot represents an *ASXL1*-mutant sample. Line of best fit is also plotted.

bulk VAF, a perfect correlation was not observed for all samples (Figure 3.7). This difference was most obvious for NOC082, where over 30% of cells were determined to be mutant by single-cell genotyping. The VAF for *ASXL1* mutation in this sample was 0.0804, recorded when the sample was first collected, which translates to an expected mutant cell percentage of 16.08%. It should be noted however, that the sorted cells are heavily enriched for CD34+ cells, whereas the original variant calling was performed on bone marrow mononuclear cells. If the mutant clone is enriched in the CD34+ compartment, this could account for the observed difference.

3.7 Clonal and immunophenotypic landscape of haematopoiesis in *ASXL1*-mutant CH samples

To gain insight into how human haematopoietic differentiation is altered with *ASXL1* mutation, samples were first analysed for immunophenotype by flow cytometry. Through comparison with healthy control samples, this data can give a wider view of the effect of *ASXL1* mutation upon bone marrow cellularity. One of the main advantages of TARGET-seq+ is the capture of single-cell surface marker expression data during index sorting. By combining index sorting data with single-cell genotyping data, patterns of clonal expansion throughout haematopoiesis can

be established. This analysis provides a more granular view of the specific cellular compartments in which the *ASXL1*-mutant clone has expanded. The following sections explore these analyses on a sample-by-sample basis.

3.7.1 NOC008

As a proportion of all bone marrow mononuclear cells the percentage of live Lin- CD34+ cells in NOC008 was modestly increased compared to the mean of the 5 healthy control samples (Figure 3.8a top). It is worth noting however, that there was some degree of sample variability of the proportion of Live Lin- CD34+ cells between the healthy control patients (range 3.08-8.73%). However, within the Lin- CD34+ compartment in NOC008, immunophenotypic HSCs (which are thought to be the main population in which clonal expansion occurs) were expanded to 12.2% compared to 4.86% in the healthy bone marrow samples (Fig 3.8a bottom). Additionally, B-cell, Natural-killer and T-cell progenitor (B-NKT) cells were expanded in NOC008 composing 25.90% of CD34+ cells.

Next, single-cell genotyping data was combined for the 5 mutant loci in NOC008, so that clonal hierarchy could be inferred. The dominant mutant clone was a single-mutant *ASXL1* clone, comprising 31.76% of the 1,612 cells that were successfully genotyped (Figure 3.8b). Much smaller competing subclones with either a *TP53*, *GNB1*, or *CBL* mutation were also identified, composing 2.11%, 1.05%, 0.19% of the genotyped cells respectively. Interestingly an independent competing CH clone harbouring a *PTPN11* mutation was also identified, comprising 9.49% of cells that were successfully genotyped.

When clonal assignment was combined with surface protein expression data from index sorting, the *ASXL1* clone composed 42.99% of the immunophenotypic HSC compartment (Figure 3.8c). When the HSC compartment was inspected for CD49f expression, a canonical marker expressed in human long-term HSCs [208], the *ASXL1* clone was marginally expanded in the transiently engrafting CD49f-fraction compared the long-term engrafting CD49f+ HSC compartment.

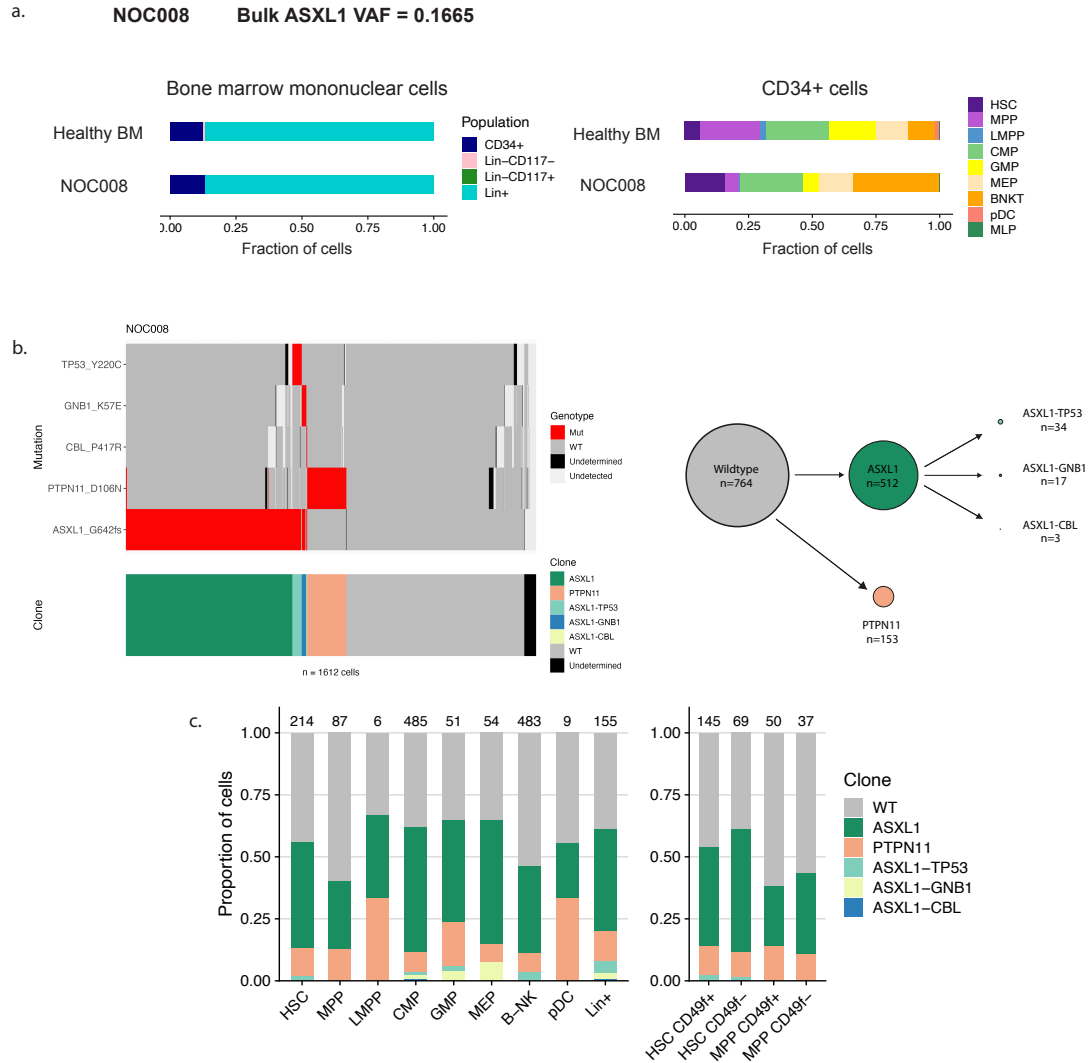


Figure 3.8: Frequency of *ASXL1*-mutant cells in HSPC compartments from NOC008 **a)** Bone marrow cell compartment sizes as determined by flow cytometry. Top: cellular compartments as a fraction of total BMMNCs; Bottom: HSPC compartments as a fraction of Lin-CD34+ cell compartment. As a reference, compartment sizes from healthy age-matched control bone marrow (n=5) is also shown. **b)** Top: raster plot of single-cell genotyping results. Each column represents a single nucleus, and each row represents each targeted genomic locus. Colour represents whether each cell was called as WT (grey), mutant (red), undetermined (black), undetected (white). Middle: Assignment of cell type based on genotyping. Colour legend is on the right. Bottom: Clonal structure inferred from single cell genotyping. Size of each circle is proportional to the size of the wildtype clone. Size of each clone (number of cells) is indicated next to each clone. **c)** Proportion of each clone size within each immunophenotypic HSPC compartment. Absolute number of cells per compartment that were sorted is determined by index sorting, and is displayed above each bar. The right-hand section of the plot is the fraction of each clone size within sub-populations of the HSC compartment, defined by CD49f expression.

Throughout haematopoiesis, the *ASXL1*-mutant clone was expanded relatively equally across the myeloid and erythroid lineages, with no obvious evidence of differentiation bias. The *ASXL1* clone was the largest mutant clone in each immunophenotypic compartment, apart from the lympho-myeloid primed progenitor (LMPP) and plasmacytoid dendritic cell (pDC) compartments where the *PTPN11* clone comprised 33.33% of each cell populations respectively. It is worth noting however, that the cell number for these cell populations are very small, so this observation could be due to sampling error. Importantly however, the *ASXL1* clone composed 41.29% of the immunophenotypic Lin+ compartment (which takes up $\approx 90\%$ of the bone marrow), comparable to the expected mutant fraction from the bulk BMMNC VAF of 0.1665 (33.3%). The same pattern was also seen for *PTPN11* (Lin+ observed % = 12.25%; Bulk VAF= 0.0712 (14.24%)), *ASXL1-TP53* (Lin+ observed % = 4.52%; Bulk VAF= 0.0159 (3.15%)), *ASXL1-GNB1* (Lin+ observed % = 2.58%; Bulk VAF= 0.0134 (2.68%)), and *ASXL1-CBL* (Lin+ observed % = 0.65%; Bulk VAF= 0.0107(2.14%)). This is an indicator of the accuracy of single-cell genotyping, which can measure clone sizes comparable to those measured by variant calling from bulk populations.

3.7.2 NOC082

Since a vial of MACS-enriched CD34+ cells were used to assay this sample, the proportion of CD34+ cells of whole BMMNCs could not be calculated. Within the CD34+ compartment however, the HSC was very modestly contracted relative to the healthy bone marrow (contracted by 0.99%), but the immunophenotypic CMP compartment was greatly expanded (Figure 3.9a). The CMP compartment composed 39.60% of all CD34+ cells, whereas the immunophenotypic GMP compartment was contracted, comprising 2.71% of CD34+ cells. Additionally, the immunophenotypic MEP compartment was expanded by 4.02% to 14.10% when compared with the healthy control samples.

As a single *ASXL1* mutation was present in NOC082, 47.87% of the 2,992 genotyped cells were assigned to a single *ASXL1* clone (Figure 3.9b). When

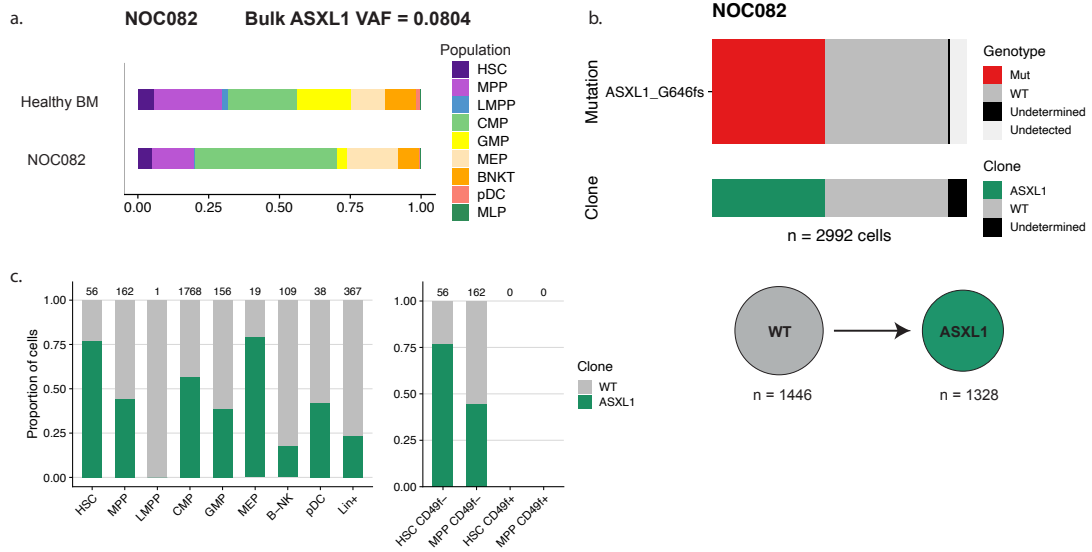


Figure 3.9: Frequency of *ASXL1*-mutant cells in HSPC compartments from NOC082 **a)** Bone marrow cell compartment sizes as determined by flow cytometry. Top: cellular compartments as a fraction of total BMMNCs; Bottom: HSPC compartments as a fraction of Lin-CD34⁺ cell compartment. As a reference, compartment sizes from healthy age-matched control bone marrow (n=5) is also shown. **b)** Top: raster plot of single-cell genotyping results. Each column represents a single nucleus, and each row represents each targeted genomic locus. Colour represents whether each cell was called as WT (grey), mutant (red), undetermined (black), undetected (white). Middle: Assignment of cell type based on genotyping. Colour legend is on the right. Bottom: Clonal structure inferred from single cell genotyping. Size of each circle is proportional to the size of the wildtype clone. Size of each clone (number of cells) is indicated next to each clone. **c)** Proportion of each clone size within each immunophenotypic HSPC compartment. Absolute number of cells per compartment that were sorted is determined by index sorting, and is displayed above each bar. The right-hand section of the plot is the fraction of each clone size within sub-populations of the HSC compartment, defined by CD49f expression.

combined with the index sorting immunophenotyping data, one striking observation was the significant expansion of the *ASXL1* clone within the immunophenotypic HSC compartment (76.79%; Fig 3.9c). Interestingly, no CD49f⁺ cells were detected in this sample. Additionally, 78.95% of the immunophenotypic MEP compartment was composed by the *ASXL1* clone, indicating a possible differentiation bias of the *ASXL1* clone to the erythroid lineage. This is surprising, given that one consistent observation from models of *ASXL1* mutations is a reduction in MEP, and a lineage bias towards megakaryocytic differentiation [144, 199, 207]. Consistent with what was observed in NOC008, the proportion of the *ASXL1* clone in the Lin⁺ population (23.16%), was similar to the predicted proportion of the clone

from the bulk BMMNC VAF (0.0762; 14.124% of cells).

3.7.3 NOC137

For NOC137, the CD34+ cell compartment was found to be slightly expanded as a proportion of BMMNC (14.90%), compared to healthy bone marrow controls (Figure 3.10a). Within the CD34+ compartment, similar patterns to NOC082 and NOC008 can be seen. Whilst the immunophenotypic HSC was modestly expanded, the CMP and MEP compartments were significantly expanded relative to the healthy controls (CMP = 41.60%; MEP = 15.90%).

Single-cell genotyping identified a linear clonal structure, with a single-mutant *ASXL1* clone with a sequentially acquired *TET2* mutation (Fig 3.10b; 21.38% and 5.43% of all genotyped cells respectively). Combined with index sorting data, clonal patterns similar to NOC082 can be observed (Figure 3.10c). The *ASXL1* clone composed 30.71% of the HSC compartment, indicating clonal expansion at the top of the haematopoietic hierarchy. Interestingly the *ASXL1* clone was expanded in the CD49f+ compartment, compared to the CD49f- compartment, which is different to what was observed in the previous two samples. Throughout differentiation, the proportion of the *ASXL1* clone within each immunophenotypic compartment remained relatively consistent, apart from the MEP compartment where it expanded to 37.76% of the 98 sorted cells, indicating possible differentiation bias towards the megakaryocyte/erythroid lineage. Again, the proportion of the mutant clones in the Lin+ compartment was as expected from the bulk BM VAF (*ASXL1*: Lin+ observed= 23.68%; Bulk VAF= 0.1413 (28.26%), *ASXL1-TET2*: Lin+ observed= 7.89%; Bulk VAF= 0.0405 (8.10%)).

3.7.4 NOC180

Surprisingly, NOC180 had a different immunophenotypic profile. The CD34+ compartment was modestly contracted relative to the healthy controls (Figure 3.11a; 6.55%). Within the CD34+ compartment, the size of most immunophenotypic populations were comparable to the control samples. Strikingly, the most expanded compartment was the B-NKT cell compartment, comprising 28.85% of the CD34+

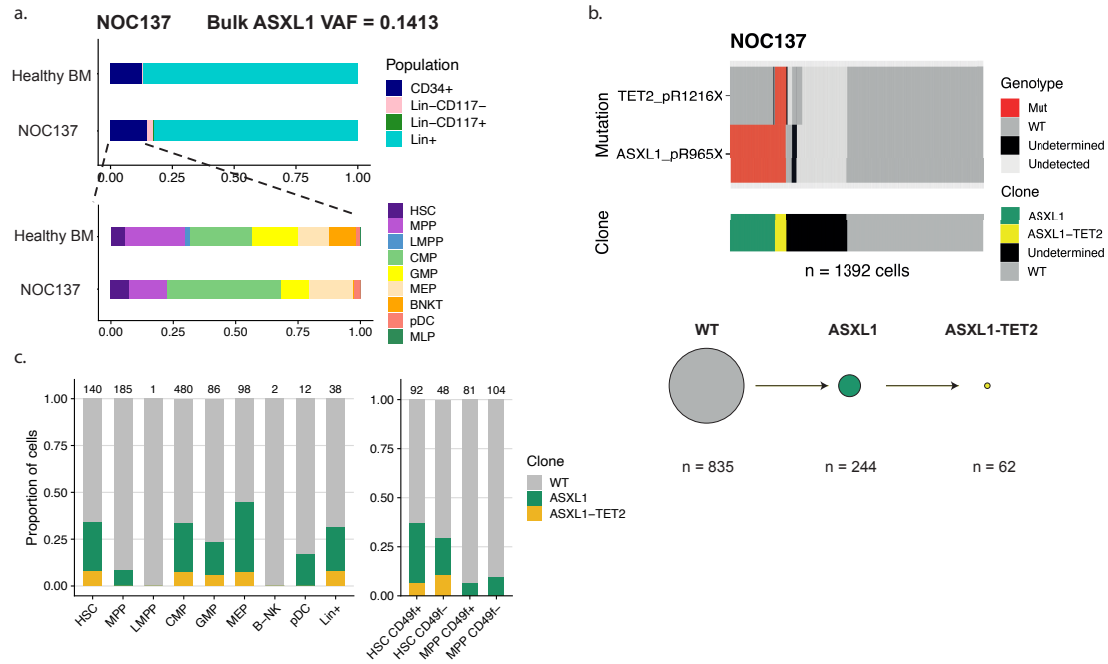


Figure 3.10: Frequency of *ASXL1*-mutant cells in HSPC compartments from NOC137 **a)** Bone marrow cell compartment sizes as determined by flow cytometry. Top: cellular compartments as a fraction of total BMMNCs; Bottom: HSPC compartments as a fraction of Lin-CD34+ cell compartment. As a reference, compartment sizes from healthy age-matched control bone marrow (n=5) is also shown. **b)** Top: raster plot of single-cell genotyping results. Each column represents a single nucleus, and each row represents each targeted genomic locus. Colour represents whether each cell was called as WT (grey), mutant (red), undetermined (black), undetected (white). Middle: Assignment of cell type based on genotyping. Colour legend is on the right. Bottom: Clonal structure inferred from single cell genotyping. Size of each circle is proportional to the size of the wildtype clone. Size of each clone (number of cells) is indicated next to each clone. **c)** Proportion of each clone size within each immunophenotypic HSPC compartment. Absolute number of cells per compartment that were sorted is determined by index sorting, and is displayed above each bar. The right-hand section of the plot is the fraction of each clone size within sub-populations of the HSC compartment, defined by CD49f expression.

cells. This is somewhat different to the other 3 CH samples in which the B-NKT compartment was either contracted or comparable to the control samples.

Another simple linear clonal structure was obvious from the single-cell genotyping data, identifying an *ASXL1* clone that composed 35.68% of all genotyped cells (Figure 3.11b). The *ASXL1* clone was again expanded in the HSC (30.97%), with negligible difference in proportion of CD49f positive or negative fractions. Similarly, the proportions of each immunophenotypic cell compartment were comparable to that of the HSC, with the exceptions of the B-NKT and pDC compartments. The

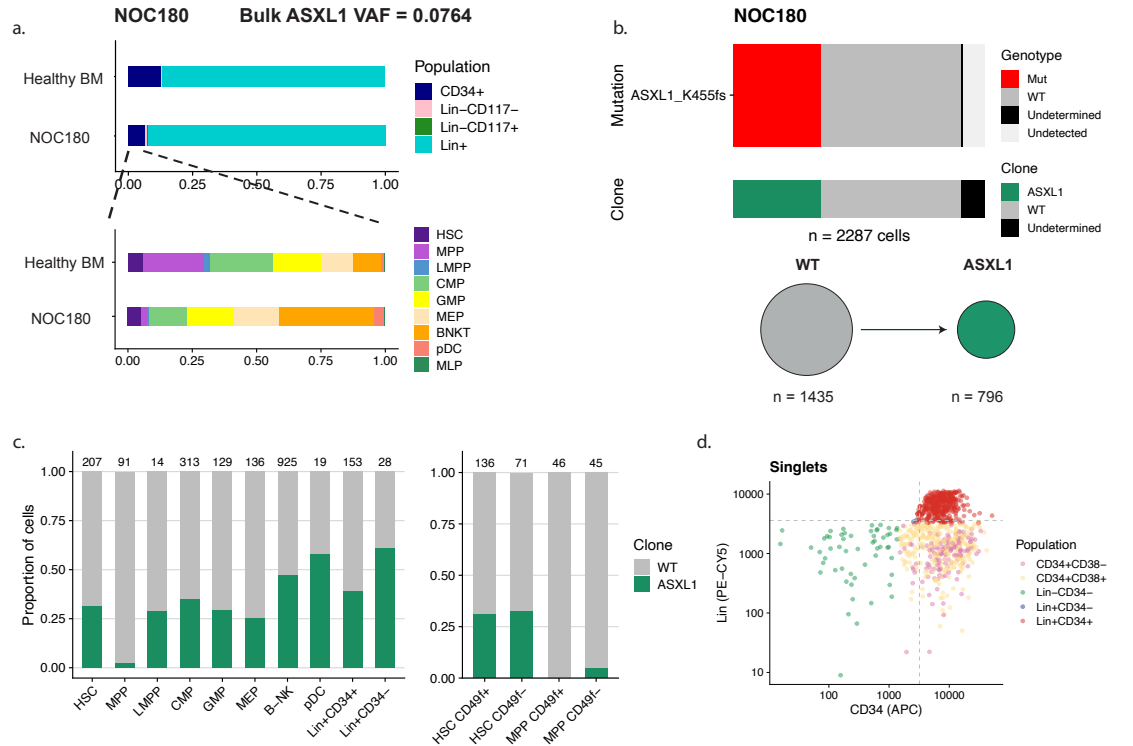


Figure 3.11: Frequency of *ASXL1*-mutant cells in HSPC compartments from NOC180 **a)** Bone marrow cell compartment sizes as determined by flow cytometry. Top: cellular compartments as a fraction of total BMMNCs; Bottom: HSPC compartments as a fraction of Lin-CD34+ cell compartment. As a reference, compartment sizes from healthy age-matched control bone marrow (n=5) is also shown. **b)** Top: raster plot of single-cell genotyping results. Each column represents a single nucleus, and each row represents each targeted genomic locus. Colour represents whether each cell was called as WT (grey), mutant (red), undetermined (black), undetected (white). Middle: Assignment of cell type based on genotyping. Colour legend is on the right. Bottom: Clonal structure inferred from single cell genotyping. Size of each circle is proportional to the size of the wildtype clone. Size of each clone (number of cells) is indicated next to each clone. **c)** Proportion of each clone size within each immunophenotypic HSPC compartment. Absolute number of cells per compartment that were sorted by index sorting, and is displayed above each bar. The right-hand section of the plot is the fraction of each clone size within sub-populations of the HSC compartment, defined by CD49f expression. **d)** Flow cytometry dot plot from index sorting data showing the distribution of sorted Lin+CD34+, Lin+CD34-, and Lin-CD34+ cell populations. Each dot represents a cell, coloured by immunophenotype. Axes represent mean fluorescence intensity for PE-Cy5 (Lineage) and APC (CD34).

immunophenotypic B-NKT compartment was the most abundant in the sample, with 925 cells being sorted. The mutant clone took up 49.75% of this compartment. Again, this is surprising given that models of *ASXL1* mutations have reported a myeloid bias [144, 199, 207]. These results would suggest a lymphoid bias for the mutant clone in this sample. Unexpectedly, the proportion of the *ASXL1* clone in Lin+ cells (69.9%) was over two-fold greater than what was expected from the bulk BMMNC sequencing VAF (VAF = 0.08; expected *ASXL1* clone proportion = 16%).

One explanation for this could be the abundance of Lin+CD34+ cells within the sorted Lin+ cells. In this population, the *ASXL1* clone composed Very few *bona fide* Lin+CD34- cells, which make up the bulk of Lin+ cells, were sorted for this dataset (Fig 3.11d; n=28). Since the abundant B-NKT cells are immunophenotypically CD34+, but are likely to simultaneously express lymphoid markers targeted by the lineage antibody mix, this could explain the imbalance of the sorted Lin+ population. An experiment to verify the accuracy of the genotyping, and the proportions of the mutant clone in each immunophenotypic compartment, would be to perform ddPCR for the mutant locus on immunophenotypic cell compartments sorted from the same sample.

3.8 Chapter conclusions

Throughout this chapter, I have outlined the experimental strategy to study the effects of *ASXL1* mutations on human haematopoiesis in the context of clonal haematopoiesis. By employing a modified version of TARGET-seq, high-fidelity single-cell genotyping was coupled with immunophenotyping and high-sensitivity scRNA-seq to determine the cellular and transcriptomic correlates of clonal expansion with *ASXL1* mutations.

A total of 12,552 cells were sorted from 4 *ASXL1*-mutant CH samples and 5 age-matched healthy control samples. 8,740 cells from the CH samples and 2 CH samples were genotyped and of these, 92.62% were successfully genotyped as either mutant or wildtype for mutant loci. Using this high-fidelity genotyping data, clonal structures within each CH sample were inferred. By analysing clonal composition of

immunophenotypic cell compartments (identified from index sorting data), several patterns of clonal expansion were identified.

Firstly, across the four samples, there was a modest expansion of the CD34+ cell compartment as a proportion of all live bone marrow mononuclear cells. Secondly, the *ASXL1* clone was significantly expanded within the immunophenotypic HSC compartment in all four CH samples. To contextualise this, the human bone marrow contains an estimated 50,000-200,000 immunophenotypic HSCs [270]. The *ASXL1* mutant clone had to originate from a single HSC, so for that clone to take up an average of 43.65% of the HSC compartment across these four samples, this represents a significant expansion of the mutant clone since the initial mutation.

Within the immunophenotypic HSC compartment, in three out of the four CH samples, the mutant clone was expanded to a greater extent in the CD49f- fraction (short-term HSC) compared to the CD49f+ fraction (long-term HSC). Finally, in two patients, there was an expansion immunophenotypic CMP and MEP compartments relative to the 5 healthy bone marrow controls, which correlated with greater *ASXL1* representation within these compartments. Intriguingly, there was significant variability in clonal distribution observed across the 4 CH samples. One speculative hypothesis for this could be plasticity of the *ASXL1*-mutant clone to respond to differing environmental cues between patients. One limitation of this analysis however is that there was a significant amount of heterogeneity between individuals with CH. Due to the limited number of samples analysed, it is therefore difficult to distinguish between the effects of inter-sample heterogeneity and *ASXL1* mutation upon patterns of bone marrow composition.

Chapter 4

Single-cell transcriptome landscape of *ASXL1*-mutant CH

Contents

4.1	Main aims for the transcriptomic analyses	86
4.2	Transcriptomic landscape of haematopoiesis in healthy and <i>ASXL1</i>-mutant CH bone marrow at the single-cell level	88
4.2.1	Data QC and filtering	88
4.2.2	Dimensionality reduction and clustering	90
4.2.3	Integration of CH and wildtype control samples	93
4.3	Cell type annotation	96
4.3.1	Cluster labelling	96
4.3.2	Cell type annotation validation using reference map projection	101
4.3.3	Correlation of single-cell transcriptome profiles using different cell type annotations	104
4.4	Analysis of <i>ASXL1</i>-mutant clonal size throughout differentiation	105
4.5	Differential abundance analysis of the <i>ASXL1</i>-mutant clone	109
4.6	Chapter conclusions	111

I am grateful to Dr Asger Jakobsen for his insight and guidance in single-cell and transcriptome analysis, and for sharing his dataset of WT healthy controls for integration with my dataset.

4.1 Main aims for the transcriptomic analyses

Having established patterns of clonal expansion across immunophenotypically defined cell compartments within *ASXL1*-mutant CH, I next wanted to explore the transcriptomic correlates between *ASXL1* mutation and clonal expansion. Two strategies could be employed to compare different cell states (for example genotype) using single-cell transcriptomic data from multiple samples.

First, analyses could be performed between cells within each sample independently. One advantage of this first approach is that it eliminates any confounding sources of biological variation that may exist between samples. However, vastly different results could be observed when performing the same analyses in different samples. This approach makes it very difficult to draw generalisable conclusions across samples. Given that this project aims to identify common molecular mechanisms of clonal advantage conferred by *ASXL1* mutations, this approach is not suitable. Alternatively, single-cell transcriptomic data from multiple samples could be integrated, before analysing cells from different samples simultaneously. By integrating transcriptomic data from CH samples with that from healthy bone marrow samples, direct comparisons of the transcriptomes can be made between mutant and WT cells from CH and non-CH bone marrow. An integrated approach was therefore chosen for this study, since it offers a higher resolution view of the transcriptional landscape of *ASXL1*-mutant clonal haematopoiesis.

To create a fully integrated dataset, two datasets of single-cell RNA-seq, generated using TARGET-seq+ were combined. First, data from the 12,552 cells from the *ASXL1*-mutant CH dataset, obtained from cells from 4 *ASXL1*-mutant CH samples and 5 WT healthy control samples was used. This was combined with a dataset consisting of transcriptome data from 4,218 cells from four of the five same WT healthy control samples, generated independently by Dr Asger Jakobsen and Dr Bilyana Stoilova for a separate publication, referred to as Jakobsen dataset [45]. The purpose of reanalysing this data was to increase the number of data points for the WT control samples and to increase power of analyses between cells from CH or healthy control samples. The total dataset contained information from 16,756 cells, and Table 4.1 outlines the composition of the complete dataset.

The aims for this chapter were:

1. To create a fully integrated transcriptomic dataset to allow accurate, high-confidence analyses between mutant and wildtype cells at a granular level.
2. To determine cell states based on single-cell transcriptomic profiles.

ASXL1-mutant MARCH samples				ASXL1 sequencing run (Angus)	2021 sequencing run (Asger and Bilyana)	Total cells
Sample	Sex	Age	ASXL1 BM VAF			
NOC008	F	81.6	0.1665	1602		1602
NOC082	F	65.5	0.0804	3026		3026
NOC137	M	73.6	0.1413		1246	1246
NOC180	F	76.5	0.0764	2314		2314
					Total CH cells	8188
Wildtype control MARCH samples						
NOC071	M	80.5	n/a	712	348	1060
NOC072	M	80.5	n/a	356	696	1052
NOC101	F	72	n/a	2114		2114
NOC153	F	80.9	n/a	712	1793	2505
NOC156	F	63.1	n/a	712	1125	1837
					Total WT control cells	8568
					Total cells	16756

Table 4.1: Composition of the entire dataset by sequencing run and sample. Number of cells that were included in the dataset, divided by sample. The *ASXL1*-CH dataset was merged with the Jakobsen dataset, as it contained cells from the same samples, also processed by TARGET-seq+.

3. To validate the patterns of clonal expansion throughout haematopoiesis based on transcriptomically-defined cell compartments.

4.2 Transcriptomic landscape of haematopoiesis in healthy and *ASXL1*-mutant CH bone marrow at the single-cell level

4.2.1 Data QC and filtering

Single-cell transcriptome libraries were sequenced, resulting in an average of 630,743 reads per cell. The two integrated datasets were sequenced using different sequencing flow cells, resulting in different depths of library (mean depth *ASXL1* = 459,135; mean depth Jakobsen = 996,141). Cells were filtered for a minimum number of reads per cell, minimum number of genes detected per cell, and maximum percentages of reads mapping to ERCC-spike in transcripts and mitochondrial genomes (Figure 4.1a-c; thresholds are detailed in Chapter two). Based on these thresholds, of the

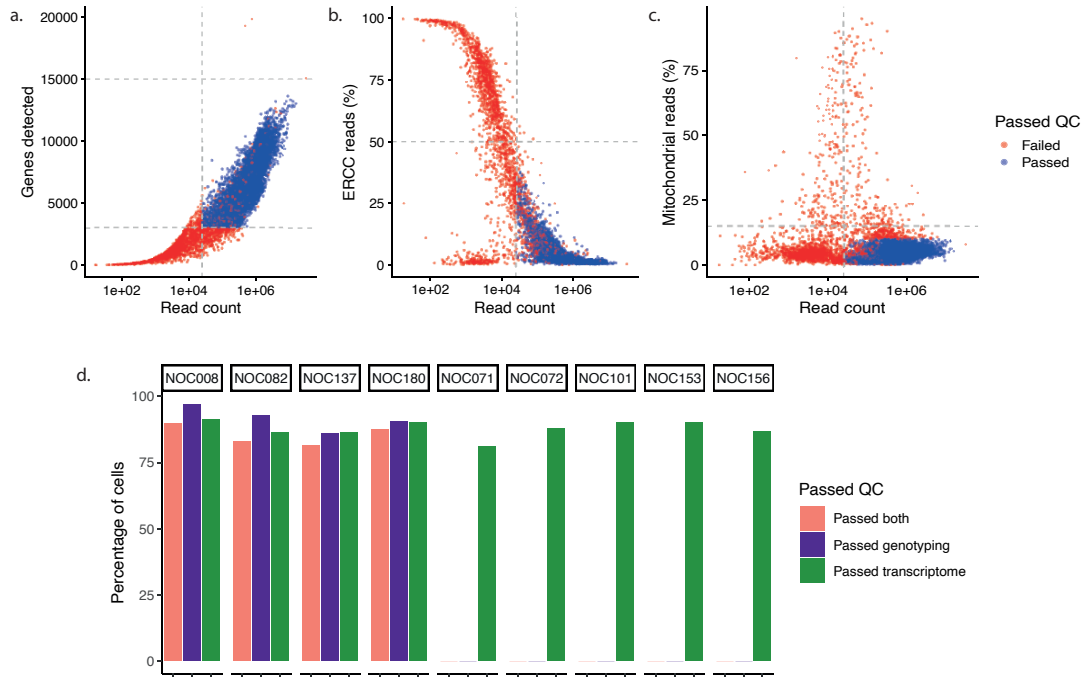


Figure 4.1: Quality control thresholds for single-cell transcriptome libraries. **a)-c)** Scatter plots showing the spread of the number of genes detected per cell, percentages of reads per cell that were an ERCC spike in transcript, and the percentage of reads that mapped to the mitochondrial genome, against the number of reads per cell. Each dot represents a cell. Red dots indicate those that failed QC for both metrics; blue dots indicate those that passed QC for both metrics. Vertical dashed line indicates the threshold for the minimum number of reads per cell (25,000). Horizontal dashed lines indicate the thresholds for the maximum and minimum numbers of detected genes per cell (3,000, and 15,000 respectively), maximum percentage of ERCC reads per cell (50%), and the maximum percentage of mitochondrial reads per cell. **d)** Percentages of cells in each sample passing QC for genotyping (purple), transcriptome (green), and both (pink). N.B. Control healthy bone marrow samples were not genotyped.

total 16,756 cells that were processed, 13,525 cells passed transcriptome QC. For the CH samples, between 77.29-84.80% of cells were also successfully genotyped (Figure 4.1d). Low-quality cells were filtered from the dataset, and of the 6,582 cells from the CH samples within the filtered dataset, 3,616 were WT and 2,832 were had a single *ASXL1* mutation. The remainder were either undetermined or other CH clones.

After filtering the mean number of reads per cell across the combined dataset was 747,302 (range 25,030-15,807,963; Figure 4.2). An average number of 6,710 genes were detected per cell, indicating that this dataset is of high complexity (Figure 4.2b). For context, between 2,000-3,000 genes per cell are detected using

droplet-based scRNAseq methods. As well as a minimum number of genes detected per cell, a maximum threshold was applied to filter any doublets that may have been accidentally sorted into the same well. A mean of 3.21% of reads per cell mapping to ERCC spike-in transcripts and 4.20% of reads per cell mapping to the mitochondrial genome was recorded for this combined dataset, indicating high-quality data (Fig 4.2c,d). When the dataset was partitioned by immunophenotype of single cells, one interesting observation is the increased number of genes detected per cell within the immunophenotypic MEP compartment (mean genes detected per cell = 8,637; Fig 4.2e). This could be correlated with the higher percentage of mitochondrial reads detected per cell, so there may be more mitochondrial genes detected in these cells (Fig 4.2f). Alternatively, there may be inherent differences in mRNA across cell compartments. Additionally, fewer genes were detected in immunophenotypic Lin+ and lymphoid progenitor cell compartments. This may be due to a high rate of mRNA degradation in these cell types, rather than lower rates gene expression [271]. Corroborating these observations are library size differences between cells of different immunophenotypes, despite pooling equal volumes of cDNA library from each single cell.

4.2.2 Dimensionality reduction and clustering

Next, the transcriptome data was normalised to account for differences in library size caused by either sequencing run (detailed above), cell type, or sampling error. Highly variable genes (HVGs) were identified across the dataset ($n = 16,163$ genes), based on a model using ERCC spike-in expression data which accounts for technical variation. Mitochondrial and ribosomal genes were removed from the final HVG list. Cell cycle stage was inferred from normalised gene expression data based on expression of cell cycle marker genes from Kowalczyk et al., 2015 [272]. Principal component analysis (PCA), Louvain clustering, and uniform manifold approximation projection (UMAP) analysis were used to reduce the dimensionality of the data and visualise it.

Based on the first 55 principal components, 25 separate clusters were identified (Figure 4.3a). When cells were coloured by their sample identifier, clusters were

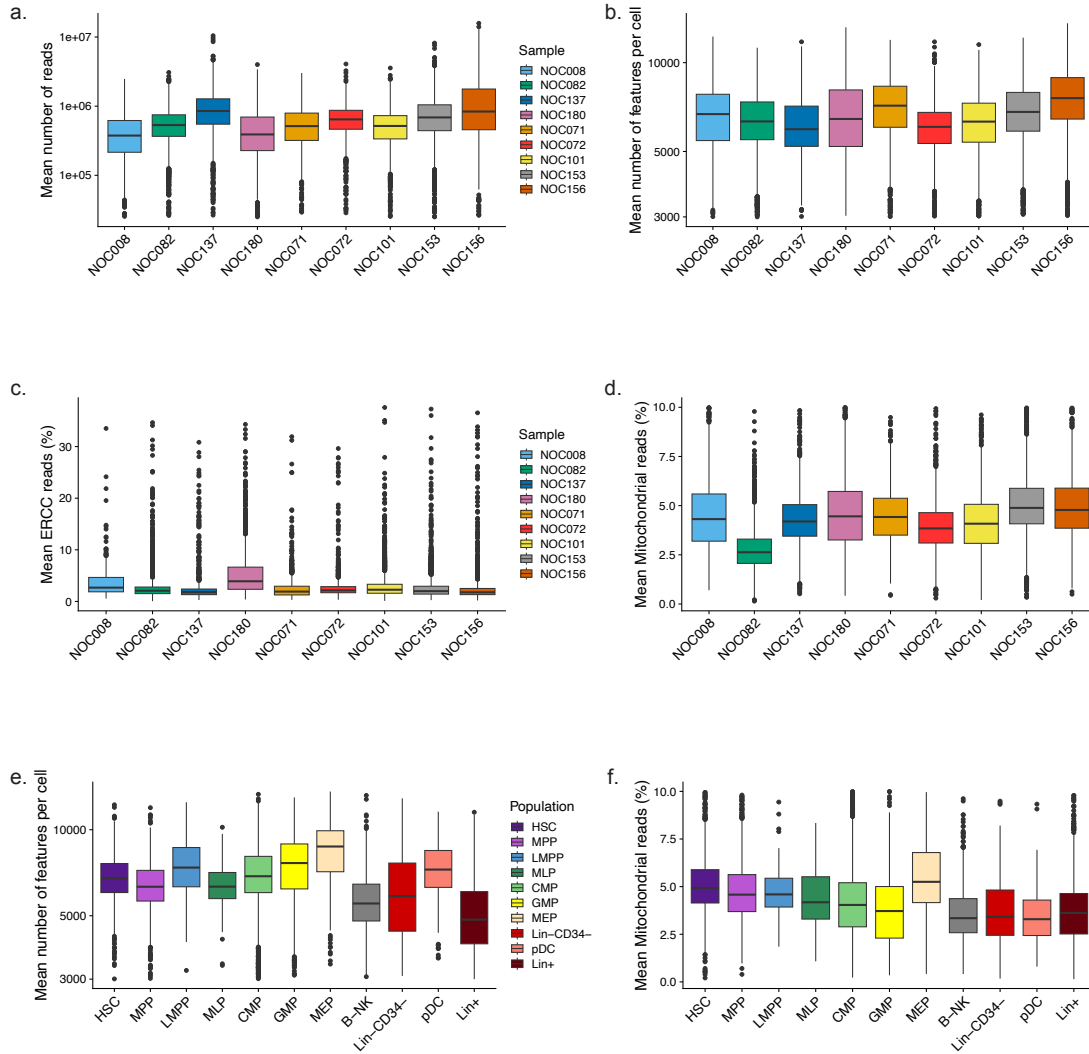


Figure 4.2: Quality control metrics of filtered single-cell transcriptome libraries. Boxplots showing key quality control metrics split either by sample(a-d), or immunophenotype (e-f). Outlier dots represent data from individual cells. **a)** Mean number of reads per cell for each sample. (Mean for dataset = 747,302). **b)** Mean number of genes detected for each sample. (Mean for dataset = 6,710). **c)** Mean percentage of ERCC reads per cell for each sample (Mean for dataset = 3.21%). **d)** Mean percentage of mitochondrial reads per cell for each sample (Mean for dataset = 4.20%). **e)** Mean number of genes detected per cell for each immunophenotypic cell population. **f)** Mean percentage of mitochondrial reads per cell for each immunophenotypic cell population.

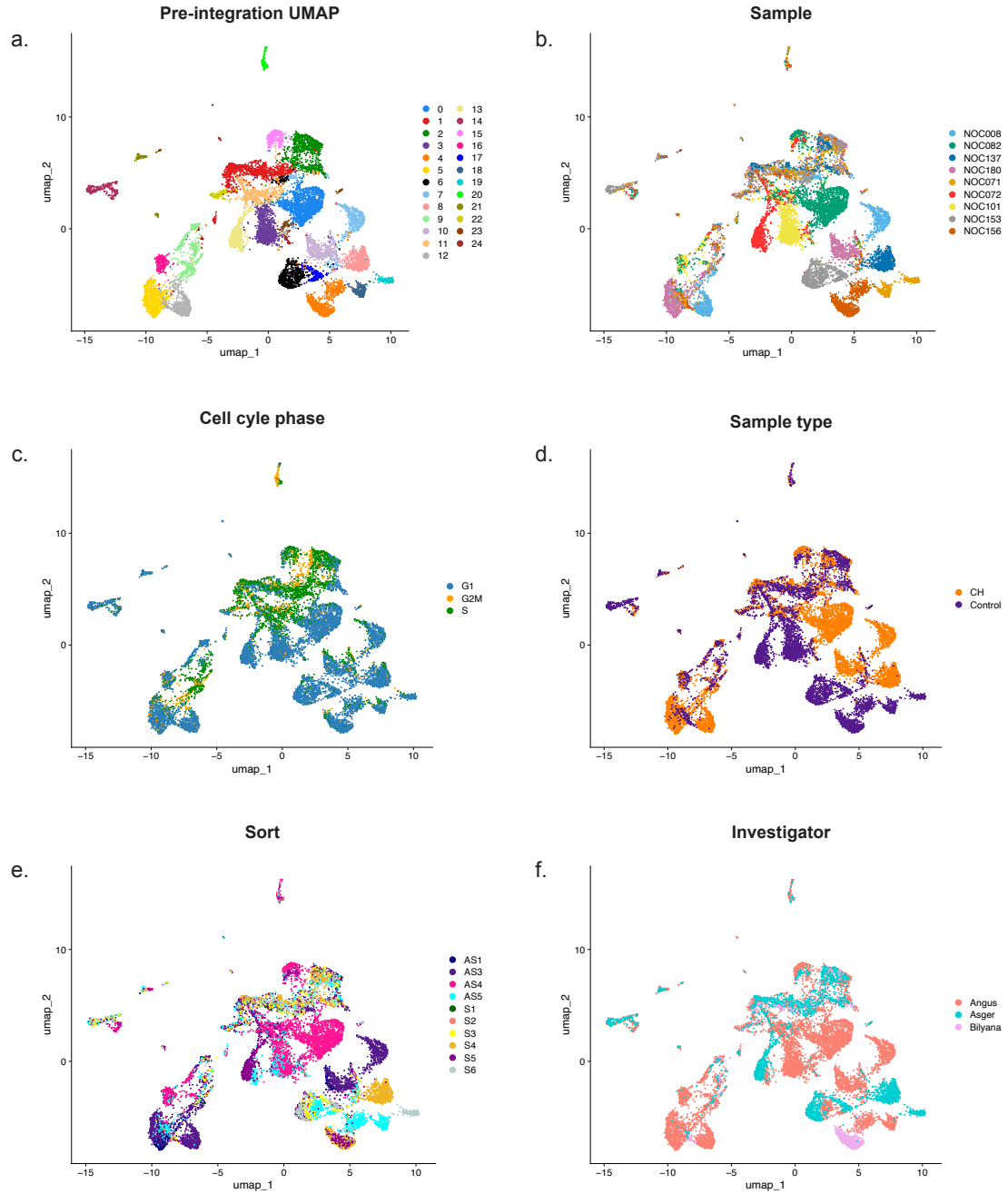


Figure 4.3: UMAP projections of the combined dataset. UMAP projections of single-cell transcriptome data 13,382 single-cells from all 4 ASXL1-mutant CH samples, and 5 age-matched control healthy bone marrow samples. 16,163 highly variable genes and the top 55 PCs were used as input into the UMAP algorithm. Each point represents a cell and are coloured by: Louvain cluster (a), Sample (b), inferred cell-cycle phase (c), Sample type (d), Sort number. (e), and investigator (f).

clearly separated by sample, indicating substantial biological variation between samples (Figure 4.3b). This sample-specific variation appeared more prevalent when the cells were inferred to be in G1 phase rather than G2 or S phase (Figure 4.3c). The same separation was also seen when the cells were highlighted by sample type (CH or control), or sort id (Figure 4.3d,e). These differences are likely due to be driven by the sample-specific biological effects, as cells designated as “CH” (orange dots, Figure 4.3d) were separated by sample, rather than forming separate clusters composed of all CH samples. Finally, variation was identified between cells processed by different investigators (Figure 4.3f). Of note, the cells from the “Asger” and “Bilyana” investigators were sequenced in a separate sequencing run to those from the “Angus” investigator. Also worth noting however is that cells from the same samples were processed by more than one investigator (i.e. “Angus” and “Asger” both processed cells from NOC153, but at separate times). The observed batch effect may be residual variation from differences in sequencing depth that still existed after normalisation.

4.2.3 Integration of CH and wildtype control samples

Confounding biological variation and technical batch effects can skew downstream analyses, leading to inaccurate comparisons. The observed variation stemming from sample and investigator variables therefore needed to be accounted for. To achieve a representative and fully unified landscape of human haematopoiesis, transcriptome data was integrated by the Harmony package [240], using sample identifier and investigator as covariates. Using the Harmony-adjusted principal components, data were again clustered using the Louvain algorithm and visualised by UMAP. When cells were coloured by sample ID or investigator, cells were heterogeneously interspersed throughout all major communities of cells, indicating that the dataset had been fully integrated (Figure 4.4a,b). This was also the case when cells were highlighted by sample type, indicating that the separation seen prior to harmony integration was driven by sample-specific biological differences (Figure 4.4d).

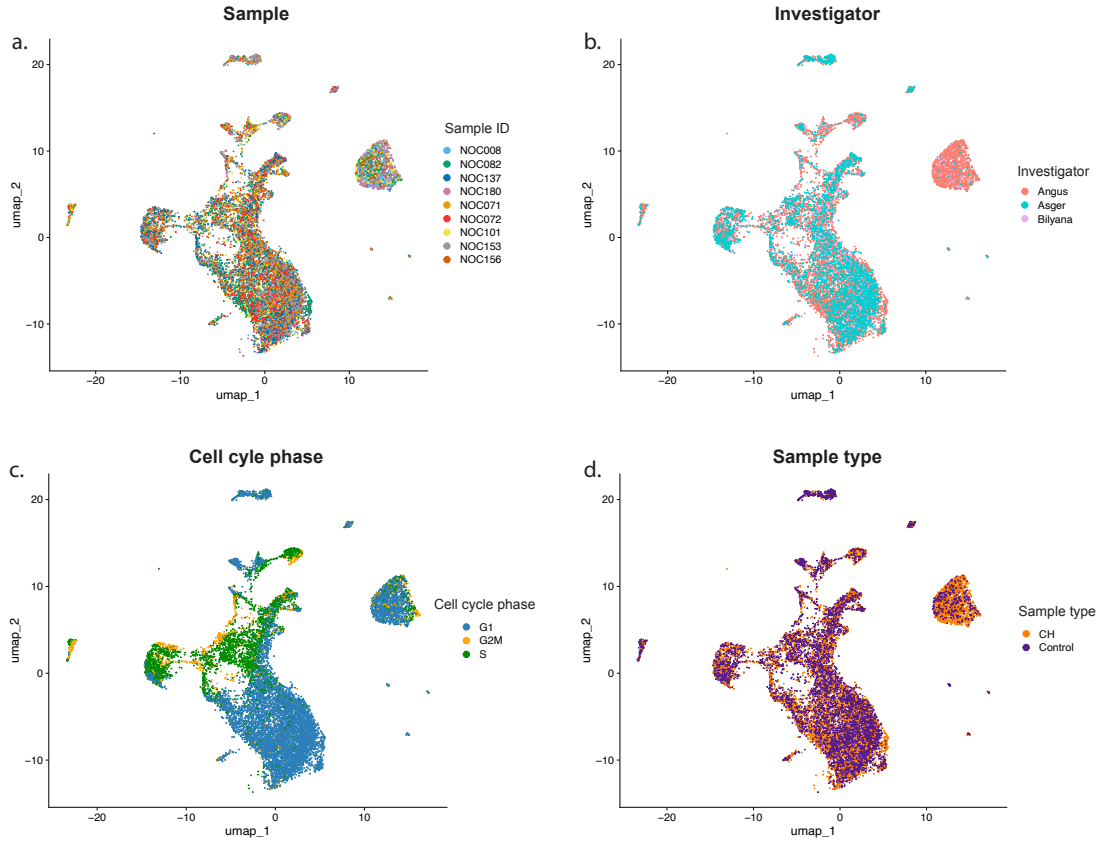


Figure 4.4: UMAP projections of the combined dataset post Harmony integration. Harmony-integrated UMAP projections of single-cell transcriptome data of 13,382 single-cells from all 4 *ASXL1*-mutant CH samples, and 5 age-matched control healthy bone marrow samples. Using sample identifier and investigator as covariates, 16,163 highly variable genes and the top 55 PCs were used as input to calculate harmony-adjusted PCs for Louvain clustering, KNN and UMAP analysis. Each point represents a cell, and are coloured by Sample (a), Investigator (b), inferred cell cycle phase (c), and sample type (d).

After unsupervised clustering with the Louvain algorithm using the harmony-adjusted PCs as input, 39 separate clusters were identified (Figure 4.5a). Clusters ranged in size from 22-1,062 cells (Figure 4.5b). By combining the single-cell immunophenotyping data with the transcriptome data, the resulting UMAP displayed a continuum of cells originating from a clear heterogeneous population of immunophenotypic HSCs and MPPs, and following separate trajectories corresponding to myeloid, erythroid and lymphoid lineages (Figure 4.5c). It is interesting that immunophenotypic HSCs and MPPs are heterogeneously mixed on the basis of gene expression and challenges the notion of a linear differentiation trajectory from

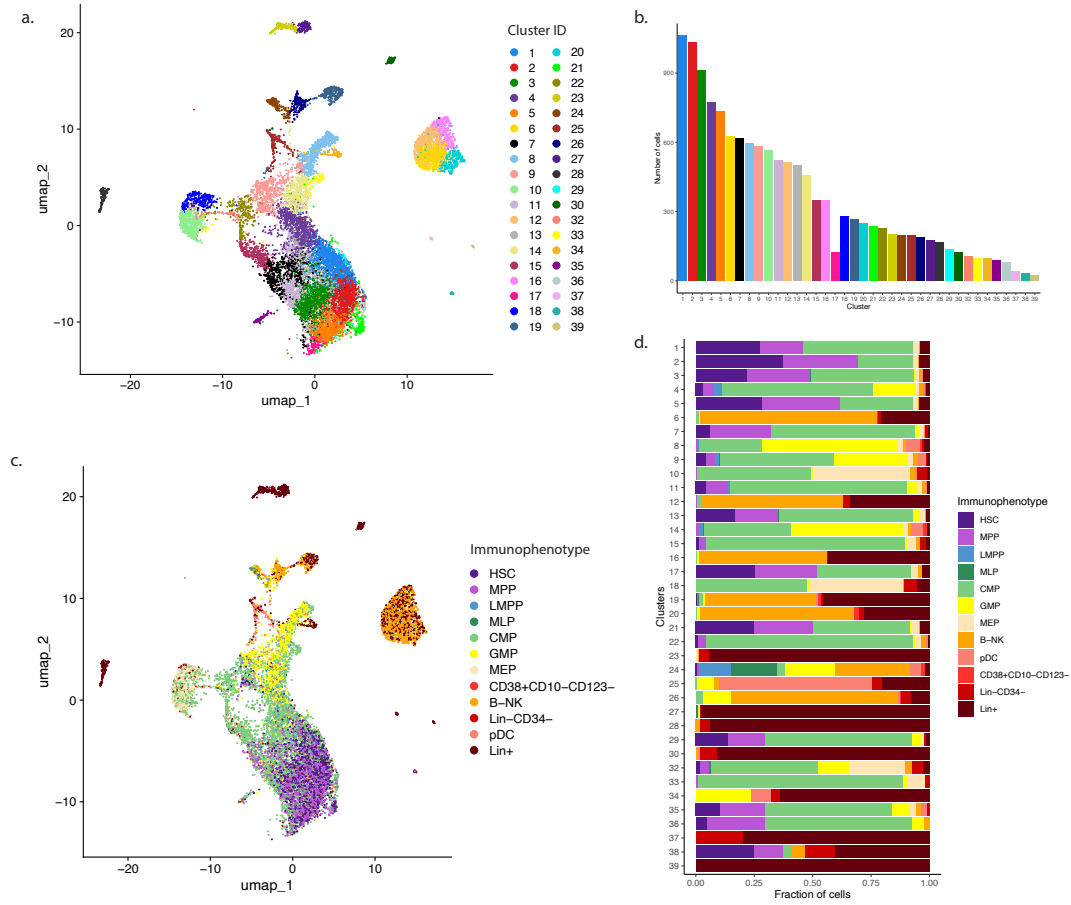


Figure 4.5: UMAP projections of the combined dataset post Harmony integration. **a)** Harmony-adjusted UMAP projection of all cells, coloured by the 39 Louvain clusters. **b)** Bar plot of absolute numbers of cells per identified Louvain cluster. **c)** Harmony-adjusted UMAP projection of all cells, coloured by immunophenotype, from index sorting data. **d)** Composition of the 39 Louvain clusters by immunophenotype, from index sorting.

HSC to MPP. The transcriptomic differences between the two immunophenotypic states could therefore be subtle. Alternatively, if this hypothesis of transcriptional similarity is true, any dropouts as a result of PCR bias or sequencing error may explain this heterogeneous mixing. It should be reiterated at this point that the HSC/MPP fraction is disproportionately large as the CD34+CD38- fraction were enriched for during single-cell sorting. Taking this into account, this UMAP is representative of healthy human haematopoiesis without overt clinical malignant phenotype.

4.3 Cell type annotation

4.3.1 Cluster labelling

Louvain clusters were manually annotated using a combination of immunophenotype composition (Fig 4.5d), gene set enrichment analysis using gene signatures of human haematopoiesis from published studies (Fig 4.6), and canonical marker gene expression (Fig 4.7).

Clusters 1, 2, 3, 5, 17, 21, 29, were some of the largest clusters identified and had the highest composition of immunophenotypic HSCs and MPPs ranging from 29.71-69.51%. These clusters were located at the apex of the main body of cells. Interestingly, the split between the HSC and MPP immunophenotype was relatively even. These clusters were also the most enriched for HSC or MPP-like gene sets.

Another smaller clusters of 523 cells- cluster 11- was also composed of 14.72% CD34+CD38- cells and was located at the apex of the split between the megakaryocyte/erythroid and lympho-myeloid trajectories. Despite some enrichment for CMP/LMPP-like gene signatures, both clusters were strongly enriched for HSC/MPP gene signatures. Despite being immunophenotypically composed of mostly CMPs, cluster 36 was annotated as HSC/MPP due to enrichment in HSC/MPP gene signatures and proximity to the major cell type clusters. Finally, cluster 35 was also labelled as HSC/MPP since it showed immunophenotypic similarity to cluster 36 and showed enrichment in AUCell score for an LT-HSC like gene signature [273]. Since these clusters displayed little transcriptomic distinction and were composed with almost equal ratios of immunophenotypic HSC and MPP, these clusters were merged and annotated as HSC/MPP. This merged cluster showed the highest expression levels of genes identified to be markers of HSCs: *HLF*, *DLK1* and *MEG3* [274–276].

Amongst the early progenitors, the continuum divided into two main trajectories. Immunophenotyping data identified these as the megakaryocyte/erythroid lineage and a lympho-myeloid lineage. At the top of the megakaryocyte/erythroid arm, clusters 7 and 15 were the most immature populations and were mainly composed

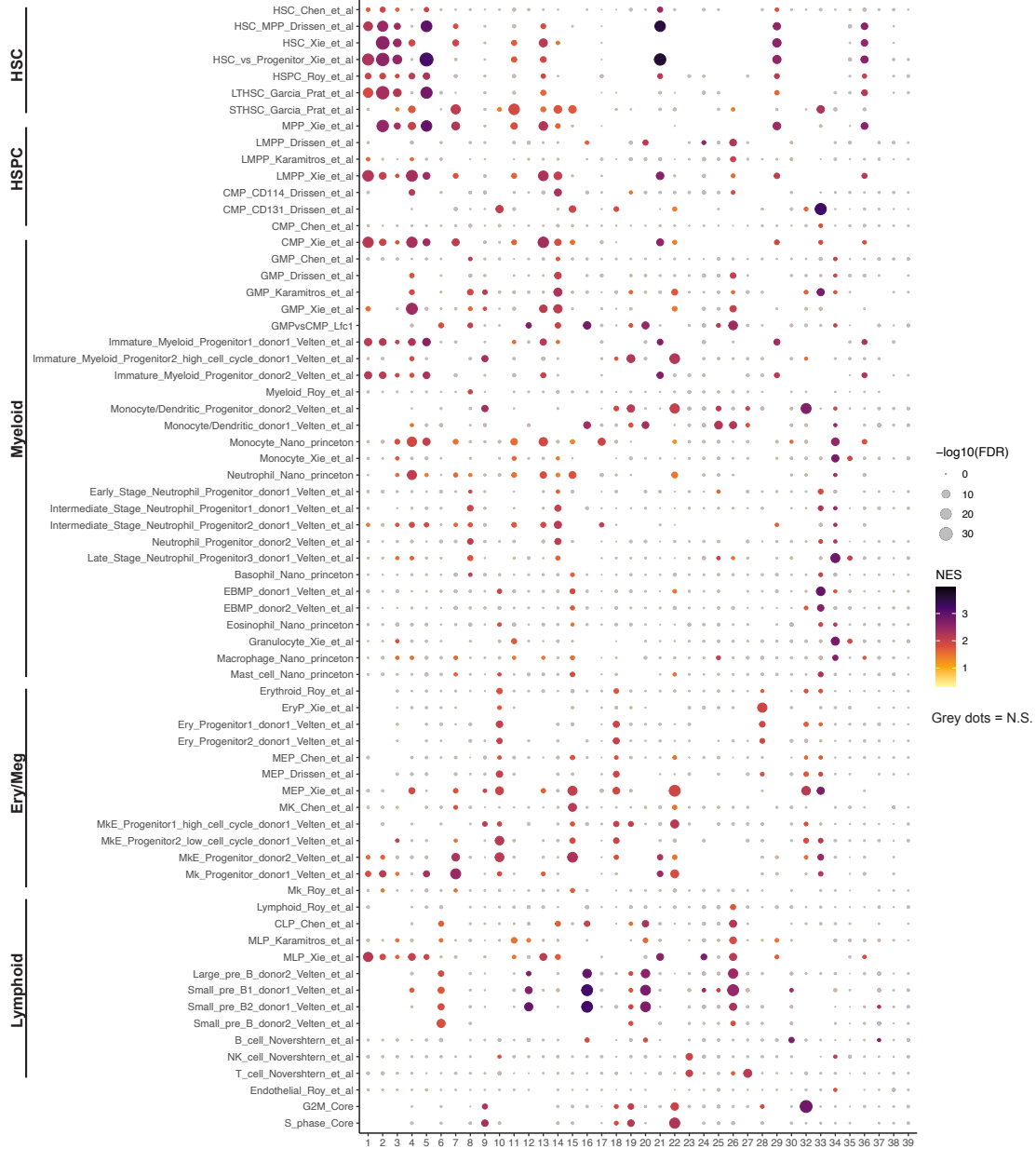


Figure 4.6: Gene set enrichment analysis of clusters for published haematopoietic lineage signatures. Bubble plot of gene set enrichment analysis results for 68 gene signatures defining cell types across lineages of human haematopoiesis, for all 39 Louvain clusters. Size of the dots represents the $-\log_{10}(\text{adjusted P-value})$ for each result (Benjamini-Hochberg adjustment). Colour of the dots represents the normalized enrichment score for each gene set of significant results (FDR < 0.05). Grey dots represent non-significant results (FDR > 0.05).

of immunophenotypic CMP (61.69%; 85.39%) and MPP (26.30%, 2.87%). Consistent with the trajectory of immunophenotype, these clusters were enriched for megakaryocyte/erythroid progenitor and MPP-like gene signatures.

Clusters 7 and 15 also displayed high expression of important megakaryocyte/erythroid transcription factor *GATA1*, and early megakaryocyte markers *FCER1A* and *VWF*. These clusters were therefore merged and annotated as EMPP. Clusters 22, 32, 18, 10, and 28 compose the rest of the megakaryocyte and erythroid trajectory. Like the most immature EMPP population, cluster 22 also expressed *GATA1*, *KLF1*, *CSF2RB*, and *FCER1A*. Since this cluster also expressed megakaryocyte marker genes *PF4*, *GP9*, and *VWF* the most highly out of all the clusters, this was labelled as the megakaryocyte progenitors (MkP).

Clusters 32, 18 and 10, had the highest abundance of immunophenotypic MEPs and were identified as early erythroid progenitors by expression of marker genes *EPOR* and *APOC1*. Expression levels of *CDK1* separated them into cycling and non-cycling progenitors. Finally, cluster 28 appeared the most mature population of the megakaryocyte/erythroid lineage. Immunophenotypically, this cluster was largely composed of Lin+ cells, which includes immunophenotypic markers for late-stage erythroid progenitors. High expression levels of erythroid progenitor marker genes *HBA1* and *GYPA* provided further evidence for cluster 28 to be annotated as such.

At the apex of the lympho-myeloid lineage, clusters 4 and 13 were also immunophenotypically composed mostly of CMPs. Upon inspection of gene set enrichment analysis results, these clusters were strongly enriched for LMPP and CMP gene signatures. Additionally, these clusters strongly expressed LMPP marker genes *HPX*, *PROM1*, *CSF3R*. Expression of both myeloid progenitor marker genes *FLT3* and *MPO* and lymphoid progenitor marker genes *IGHM* and *MZB1* provided evidence for both clusters 4 and 13 to be labelled as LMPP. Similarly, cluster 9 was also enriched for LMPP-like gene signatures and expressed LMPP marker genes. However, cluster 9 was strongly enriched for cell cycle gene signatures and *CDK1* expression, suggesting that these were cycling LMPPs. The expression of genes strongly associated with the cell cycle is likely the reason driving the

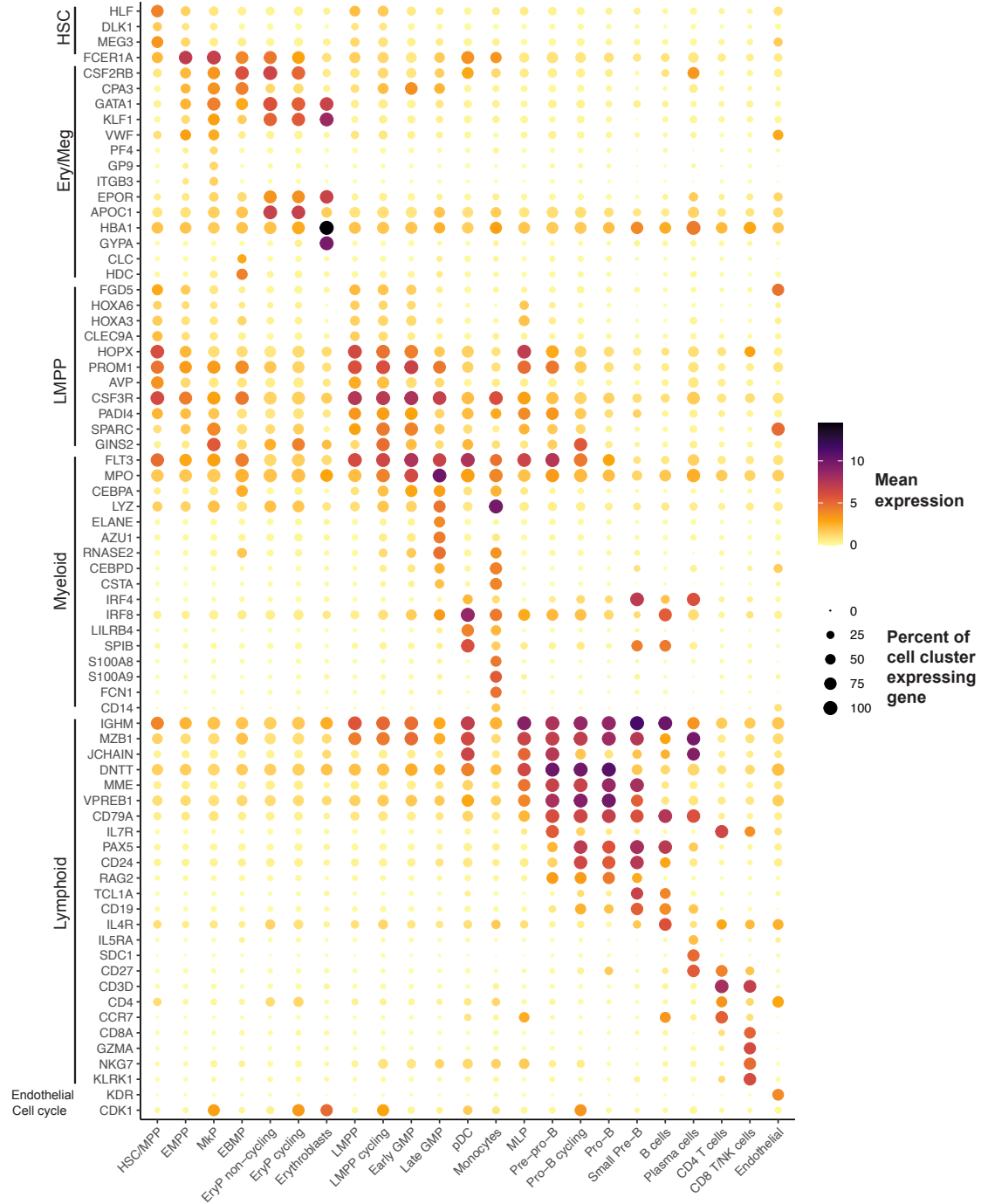


Figure 4.7: Cluster marker gene expression. Bubble plot of canonical marker gene expression across transcriptomically-defined cell clusters, ordered by lineage. The size of each point represents the percentage of each cluster that express the gene (at least one transcript per cell), and the colour of each point represents the mean expression level across constituent cells in each cluster.

union with the MkP cell cluster. Regression of cell cycle genes had deleterious effects on the structure of the whole dataset, and was not included as part of the analysis (data not shown).

Located towards the terminal section of the lympho-myeloid arm, clusters 14 and 8 were composed mainly immunophenotypic GMP (48.14% and 58.63% respectively). These two clusters were enriched for GMP-like and myeloid gene signatures but not lymphoid gene signatures, indicating that these clusters were progenitors committed to the myeloid lineage. Similarly, both clusters showed high expression of key myeloid progenitor marker genes including *MPO*, *CEBPA*, and *FLT3* but had lost expression of lymphoid marker genes, differentiating them from the more immature LMPP. Cluster 8 appeared more immature than cluster 14 and was labelled as early stage GMP. Cluster 14 showed greater expression of later-stage myeloid progenitor marker genes such as *LYZ*, *ELANE*, *AZU* and *RNASE2*, and was labelled Late-stage GMP. Additionally, a small of monocyte/neutrophil progenitors (n= 97 cells) was identified by cluster 34, showing high expression of *LYZ*, *S100A8*, *S100A9*, *FCN1*. Cluster 25, which was 65.31% immunophenotypic plasmacytoid dendritic cell (pDC), and this identity was confirmed through high expression of marker genes *IRF8*, *IRF4*, *SPIB*, and *LILRB4*.

Clusters 24, 26 and 19 composed a separate trajectory projecting from the lympho-myeloid arm. All of the clusters expressed high levels of early lymphoid markers such as *IGHM*, *MZB1*, and *DNTT*. Immunophenotypically, cluster 24 had the greatest composition of CD38-CD10+ MLP (19.70%) of all the clusters in the dataset. Cluster 24 was also enriched for MLP gene signatures and expressed *NKG7* indicating potential to differentiate to NK/T cells, and was therefore labelled MLP.

Clusters 26 and 19 were mainly composed of CD38+CD10+ B-NKT progenitor cells, however cluster 19 also contained a number of Lin+ cells, suggesting this was a more mature lymphoid progenitor population. Cluster 26 mostly retained CD34 expression, whilst gaining expression of B-cell markers and was therefore labelled Pre-Pro B cell. A large population of cells (n = 1,735) expressing B-cell marker genes was composed of clusters 6, 12, 16, 20. Based on expression levels

of B cell marker genes, and level of enrichment for B cell gene signatures, these clusters were virtually indistinguishable and were therefore merged. Like cluster 19, this population started to lose CD34 expression, being composed of both CD34+ immunophenotypic B-NKT cells and Lin+ CD34- cells. Since all of these clusters gain *CD24* expression, this population was labelled as Pro-B (Duperray1990). Cluster 19 expressed a very similar profile of marker genes to the ProB cluster, but was annotated as Pro-B cycling due to the expression of cell cycle marker genes such as *CDK1*. A very small population of cells (cluster 37) gained *TCL1A* and *IL4R* expression and were annotated as Small-pre B cells. Finally, cluster 30, which was immunophenotypically 88.10% Lin+ cells, gained expression of mature B cell marker gene *CCR7* and were annotated as B cells.

Minor populations of mature cells expressing primarily lineage markers were also identified. A population expressing T cell marker gene *CD3D*, almost exclusively composed immunophenotypic Lin+ cells was formed from clusters 23 and 27. Cluster 23 expressed high levels of *CD4*, so were labelled *CD4* T cells. Cluster 27 gained *CD8A*, *GZMA*, *NKG7*, *KLRK1* expression and were labelled CD8 T/ NK cells. Very small populations of Plasma cells and Endothelial cells were also detected (n= 22; n= 32), and identified gain of expression of *SDC1* and *KDR* respectively.

4.3.2 Cell type annotation validation using reference map projection

In previous chapters, the importance of high-confidence genotyping was emphasised to enable accurate comparisons between mutant and wildtype cells when it is expected that differences will be very subtle. Similarly, high-confidence cell type annotation is equally important for high quality analyses. For example, if a cluster of cells that are part of the early HSPC continuum which are not obviously separated, there is the possibility of wrongful cell type assignment. This could potentially lead to an example scenario where two clusters initially annotated as HSCs are compared with each other, where one should have actually been labelled LMPP. This may lead to skewed results. The previously described method that was used to annotate cell clusters is an unsupervised approach.

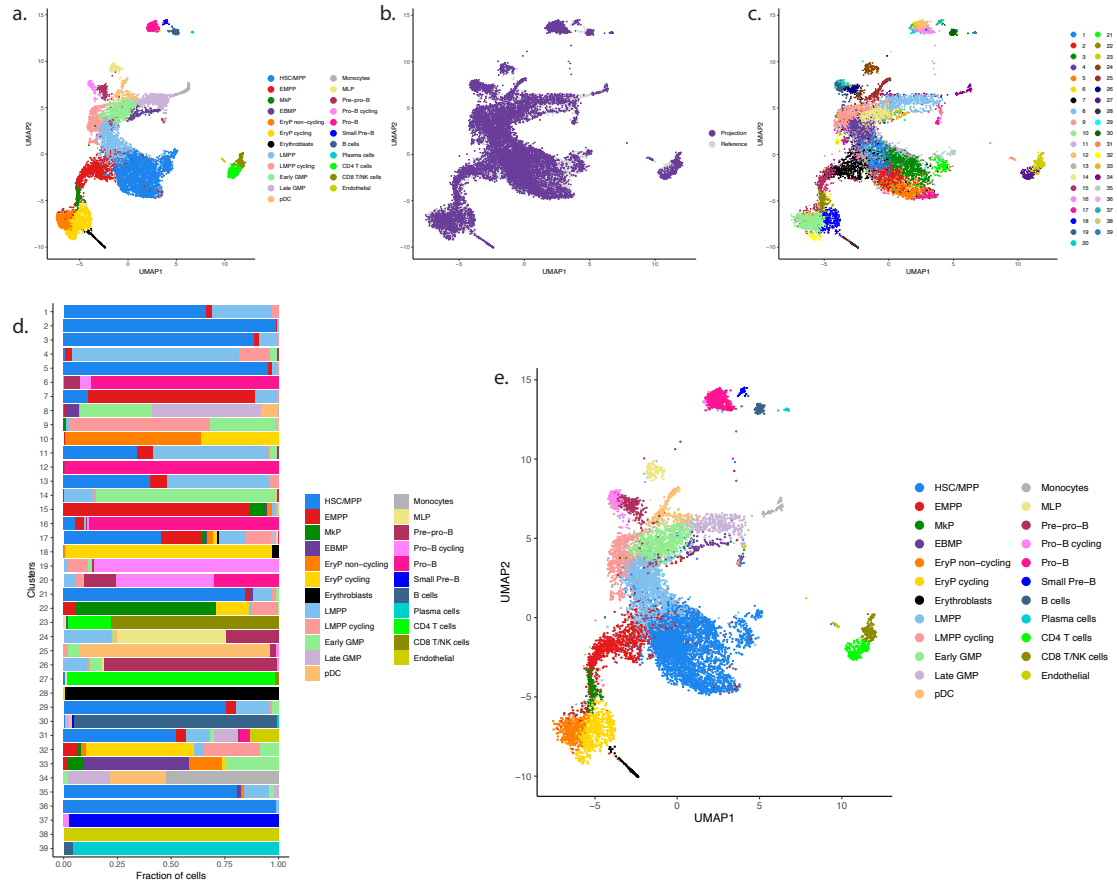


Figure 4.8: Symphony projection onto a reference map of CH haematopoiesis.

a) UMAP projection of the reference dataset from Jakobsen et al., 2024, ($n = 13,247$ cells), coloured by cell type annotation. **b)** UMAP of the symphony projection of the *ASXL1*-CH dataset ($n = 13,382$ cells; coloured purple) onto the reference dataset (grey). **c)** UMAP of the symphony projection of *ASXL1*-CH dataset, coloured by Louvain clusters, identified after Harmony integration. **d)** Fraction of each knn-predicted cell type annotation that compose each Louvain cluster, identified after Harmony integration. **e)** UMAP of the symphony-projected *ASXL1*-CH dataset, coloured by the knn-predicted cell type annotation, inferred by symphony projection.

To give confidence in cell type annotation, and to avoid this possibility, an orthogonal supervised method of annotation was used. Using the Symphony algorithm [250], cells were mapped onto low-dimensional embeddings from a reference map of healthy and CH human bone marrow, produced using TARGET-seq+ (Jakobsen et al., 2024) [45]. Cell types of each mapped cell could then be inferred from the projection, based on the distance to their nearest neighbours from the reference map.

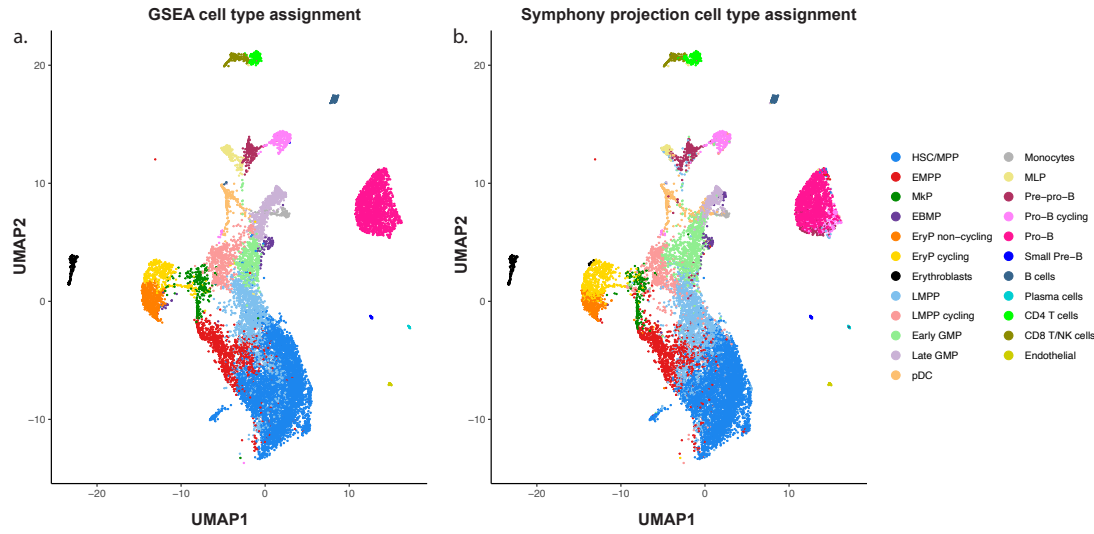


Figure 4.9: Unsupervised and supervised cell type annotations. a) UMAP projection of all cells, coloured by the cell type annotations based on Louvain clustering. b) UMAP projection of all cells, coloured by the knn-predicted cell types inferred from Symphony projection onto a reference dataset.

The reference map had a very similar shape to that seen in the *ASXL1*-CH dataset, with a similarly enriched CD34+CD38- population and three distinct arms corresponding to megakaryocyte/erythroid, myeloid, and lymphoid progenitor differentiation (Figure 4.8a). One difference is that the myeloid and megakaryocyte/erythroid trajectories remain separated, which may be due to a smaller number of cell cycle genes being expressed. When the *ASXL1*-CH dataset was mapped onto the reference map, the cells were heterogeneously dispersed throughout all of the major populations of the reference map (Figure 4.8b). When the Louvain clusters from the *ASXL1*-CH dataset were highlighted on the projection (Figure 4.8c), the location of each cluster reflected what was observed using the unsupervised approach (Figure 4.5a) e.g. the “HSC/MPP” clusters from the *ASXL1*-CH dataset were found in the same low-dimensional space as the HSC/MPP cluster of the reference dataset.

The symphony algorithm predicts cell type for each single cell, based on the proximity to the nearest neighbours from the reference dataset. Next, unsupervised and supervised cell annotation approaches for each Louvain cluster were correlated. For this, the composition of each Louvain cluster with each symphony-predicted cell type was inspected (Fig 4.8d). It has previously been suggested that human

haematopoietic differentiation is a continuum [243]. Therefore, it is not expected that clusters will be 100% composed of the same cell type. This statement may be true at the earliest stem and progenitor level, where cell types are transcriptionally very similar and are not yet committed to a differentiation trajectory. However, as progenitors become increasingly more committed to different lineages, and cell types have to become functionally more specialised, it is likely this continuum becomes a hierarchy.

For the vast majority of clusters, at least 50% of cells were predicted to be the same as the original cell annotation. However, for clusters 8, 11, 13, 17, 20, 32, 33, the major predicted cell type corresponded to the annotated cell type but composed less than 50% of the cluster. This could be because the cells that compose these clusters could be at the point of differentiation commitment to a particular lineage. For example, clusters 11 and 13 are made up mostly of predicted LMPP (53.35 & 47.90% respectively). However, they are also composed of HSC/MPP and EMPP. When inspecting their location on the UMAP, they are located at the point of transition from HSC/MPP to either the megakaryocyte/erythroid or lympho-myeloid trajectories. Therefore, using symphony reference map projection, a cell-specific cell type annotation for this dataset was also generated (Figure 4.8e)

4.3.3 Correlation of single-cell transcriptome profiles using different cell type annotations

As previously explored, using an unbiased clustering approach, cells were annotated based on immunophenotype, marker gene expression, and gene set enrichment analysis. A second cell-specific cell type annotation reference map projection was also obtained using Symphony reference map projection. When cells are highlighted by the different cell type annotations, using the original UMAP embeddings, the two annotations look very similar, indicating high concordance between the two methods (Figure 4.9). However, to be able to compare mutant and wildtype cells accurately, the most suitable cell type annotation must be chosen. Since an analogous study exists that uses the same experimental strategy to explore the mechanisms of clonal expansion in *DNMT3A*- and *TET2*-mutant CH samples

[45], it made sense to benchmark each cell type annotation against this dataset. Any similarities or differences that may exist between *ASXL1*-, *DNMT3A*-, or *TET2*-mutant CH could therefore be reliably assessed.

To assess the accuracy of each annotation, the normalised gene expression profiles from pseudobulks of each cell type annotation was correlated with those of the reference dataset (Figure 4.10a,b). Pearson correlation coefficient scores were calculated for each correlation analysis, and the results were hierarchically clustered. As expected, when each cell type annotation was correlated against itself, the pearson's correlation coefficient was 1, indicating a perfect correlation. For most annotated cell types using both methods, the correlation coefficient with the reference dataset was also close to 1 and were therefore clustered together. The exception to this was for the earliest progenitors (HSC/MPP, LMPP, EMPP), where the gene expression profiles correlated so closely to each other that they clustered as a group, rather than with their reference counterpart. For the unsupervised cell annotation (Figure 4.10a), the mean pearson's correlation coefficient with the respective reference dataset cell type was 0.983. For the symphony cell type annotation the mean correlation coefficient was 0.984. Despite virtually identical correlation with the reference dataset, the symphony cell type annotation was chosen to be used for subsequent analyses.

4.4 Analysis of *ASXL1*-mutant clonal size throughout differentiation

In the previous chapter, patterns of *ASXL1*-mutant clonal dynamics throughout differentiation were investigated in immunophenotypic populations of HSPCs. From this analysis, the *ASXL1*-mutant clone was found to be most expanded in the immunophenotypic HSC and clone size remained relatively constant across progenitor populations from the three lineages. In some samples, there was an expansion of the mutant clone within the immunophenotypic MEP compartment. To validate these findings, the expansion or contraction of the *ASXL1*-mutant clone within each transcriptomically-defined cell state was calculated. As observed

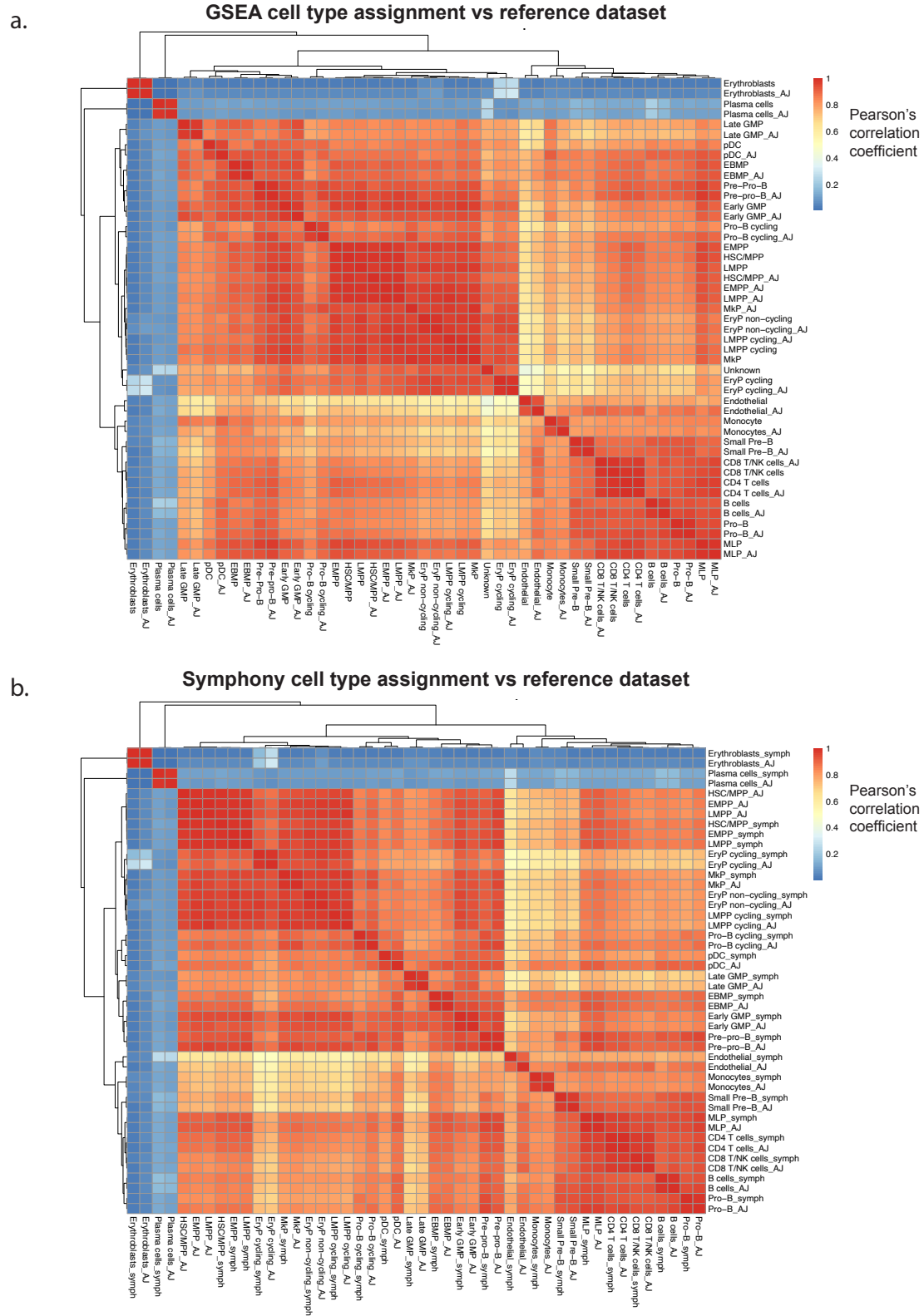


Figure 4.10: Correlation analysis of unsupervised and supervised cell type annotations. Hierarchically-clustered heatmaps correlating pseudobulk normalized gene expression profiles of annotated cell types based on Louvain clustering (a), or inferred from symphony projection (b), with those of annotated cell types from a reference dataset [45]. Annotated cell types from the reference dataset can be identified by the suffix “_AJ”. Each tile is coloured by a Pearson’s correlation coefficient for each individual test.

during cell type annotation, transcriptomic cell type does not perfectly correlate with immunophenotype. However, this analysis can provide granularity for analysis of clonal dynamics throughout haematopoiesis. Additionally, as more cell states were identified including mature Lin+ cell populations, this analysis adds further depth in assessing the clonal dynamics of the *ASXL1*-mutant clone.

For this, the single-cell genotyping data was combined with the transcriptome data. When projecting the *ASXL1*-mutant clone from all 4 CH samples onto the UMAP, the mutant clone was found to be heterogeneously mixed with WT cells and other clones (Figure 4.11a; all in grey). One exception to this was the T-cell populations, where the *ASXL1*-mutant clone was somewhat depleted. When clonal composition for each transcriptomic cell compartment was quantified, within the HSC/MPP compartment the *ASXL1* mutant clone for all 4 samples accounted for $\approx 35\%$ of all cells (Figure 4.11b). This was consistent with observations in the immunophenotypic HSC compartment. The clone was slightly expanded in the progenitor populations from the megakaryocyte/erythroid and myeloid lineages. Strikingly, the *ASXL1*-mutant clone was much smaller in the later stage precursor, or terminally differentiated cell types of all lineages.

Interestingly, within the MLP compartment, the *PTPN11* clone was greatly expanded relative to the clone size within the HSC. This was different to the immunophenotyping data, where the *PTPN11* clone was enriched in the LMPP population. Considering that the *PTPN11* clone is present in only one patient (NOC008), this is likely to be the dominant clone in this compartment. The Pre-pro-B cell, Pro-B and Pro-B cycling also had greater *PTPN11* representation relative to that of the HSC, indicating a relative advantage of the *PTPN11* clone in the B-lymphoid lineage to that of the *ASXL1* clone. Additionally, previously *ASXL1*-mutant cells have been reported to preferentially differentiate towards the myeloid lineage, rather than lymphoid lineage [144, 199, 207]. Since the *PTPN11*-mutant clone is a competing CH clone in this sample, it may be that it has a greater competitive advantage within the lymphoid cell compartments, if there is less competition from the *ASXL1*-mutant clone in these cell types.

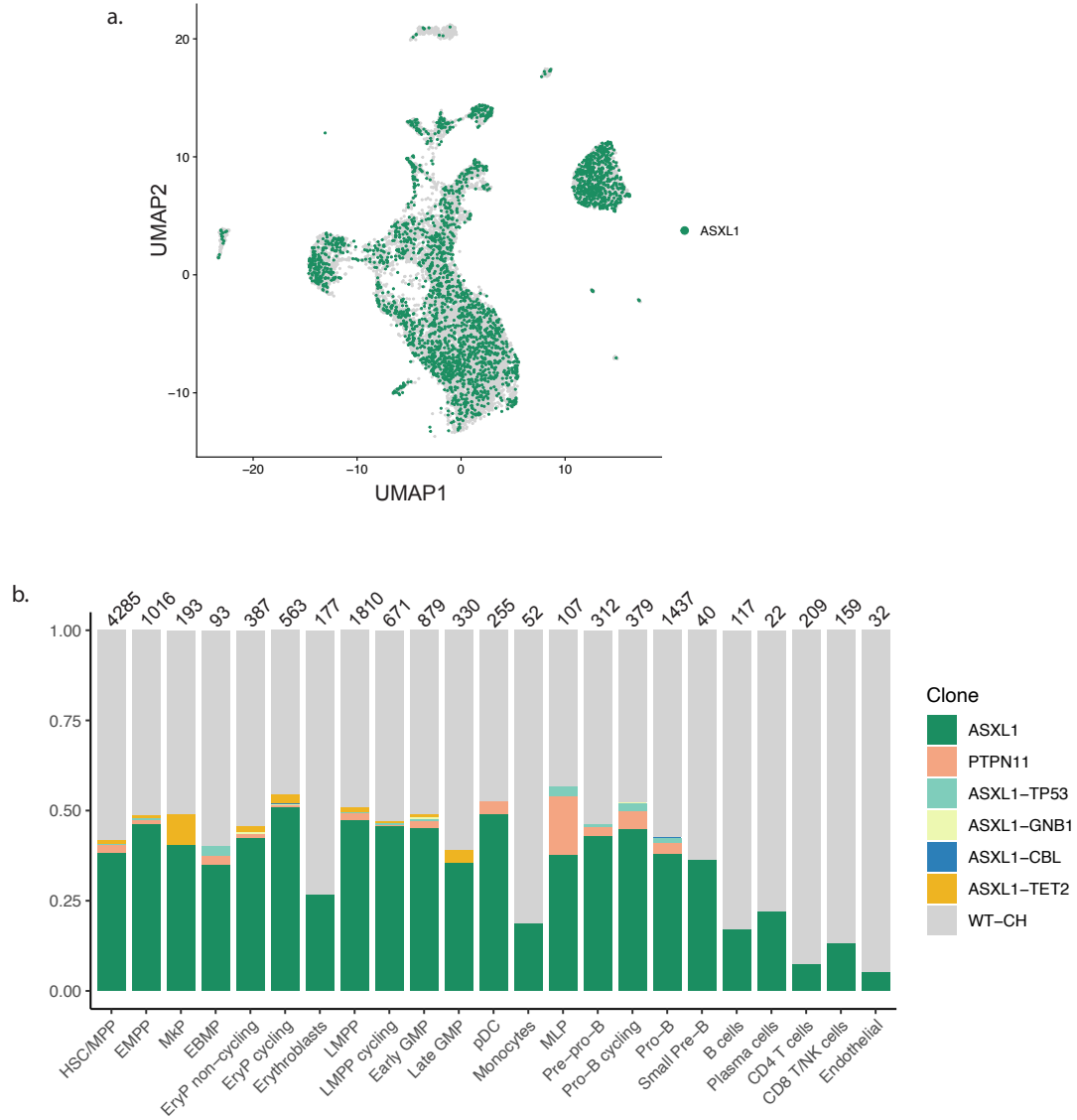


Figure 4.11: Distribution of the *ASXL1*-mutant clone throughout haematopoietic differentiation. **a)** UMAP projection of cells from all samples, with the *ASXL1*-mutant clone highlighted in green. All other cells are coloured grey, regardless of clone. **b)** Clonal composition of each transcriptomically-defined cell cluster (symphony projection). Numbers above bars indicate the number of cells in each cell compartment

4.5 Differential abundance analysis of the *ASXL1*-mutant clone

The evidence presented thus far suggests that the *ASXL1*-mutant clone is most greatly expanded within the HSC/MPP compartment. To further quantify how the *ASXL1*-mutant clone contracts and expands throughout differentiation, the mean densities of mutant and WT clones were compared for each transcriptomic cell compartment within each CH sample. These densities were then normalised to that seen within the HSC/MPP compartment (Figure 4.12a). Thus, the normalised scores can provide insight into any lineage differentiation bias of the mutant clone.

Across the four *ASXL1*-mutant CH samples, both the myeloid and megakaryocyte/erythroid lineages were modestly expanded relative to that of the HSC/MPP. An exception to this was the erythroblast compartment, where the mean likelihood of finding the mutant clone was roughly half that was observed for the HSC/MPP. The compartment with the greatest mean clonal expansion relative to that of the HSC was the monocyte compartment, although this was highly sample dependent. Caution must be applied to this result, as there were only 16 monocytes across the four CH samples. This is interesting however, as truncating *ASXL1* mutations are found in 49% of CMML cases, in which monocytes are greatly expanded [277]. Throughout the lymphoid lineage, the mean *ASXL1* clone size was modestly contracted relative to the HSC/MPP compartment. Most notably, there was a trend to inefficient terminal maturation of the B and T cell lymphoid lineages. One exception to this was the NOC180 sample which consistently did not follow this trend throughout lymphopoiesis. Interestingly, this sample also showed an enrichment for immunophenotypic CD34+CD38+CD10+ B-NKT progenitors.

When the likelihood score of finding the mutant clone relative to the HSC/MPP compartment was mapped on to the UMAP, the differentiation patterns can easily be visualised (Fig 4.12b). The mutant clone is greatly expanded within the HSC/MPP, but not necessarily at the apex of the HSC/MPP. The *ASXL1* mutant clone shows

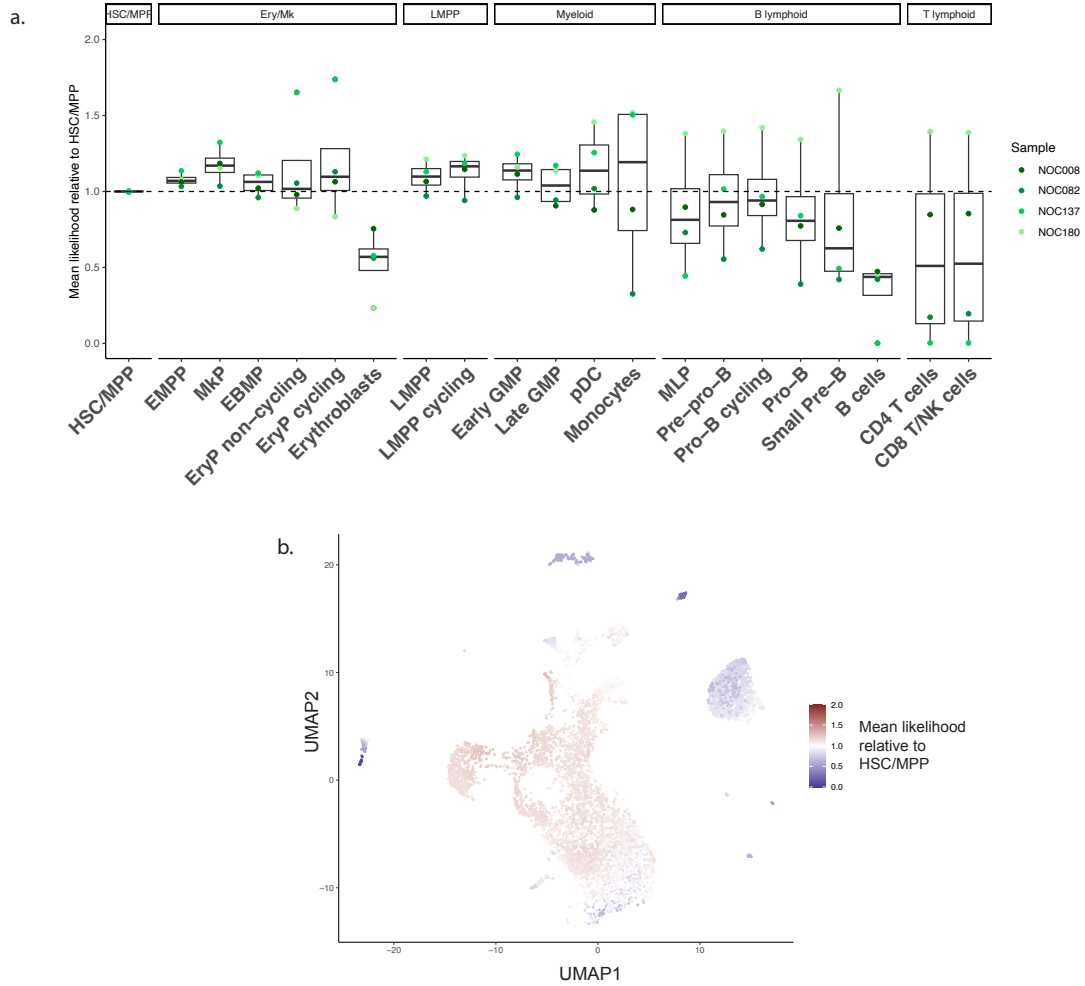


Figure 4.12: Differential abundance of the *ASXL1*-mutant clone throughout haematopoietic differentiation. **a)** Mean likelihood of the finding the *ASXL1*-mutant clone in each transcriptomically-defined cell cluster for each *ASXL1*-CH sample, normalised to that of the HSC/MPP compartment. Each point represents an *ASXL1*-CH sample. The horizontal dashed line at $y=1$ represents the mean likelihood of finding an *ASXL1*-mutant cell in the HSC across all four samples. Points lying above or below that line indicate that relative to the HSC/MPP compartment, the size of the mutant clone is expanded or contracted respectively in that cell compartment. **b)** UMAP embedding of cells from *ASXL1*-mutant samples, highlighted by *ASXL1*-clone likelihood score relative to the HSC/MPP compartment

no overt differentiation bias. However, it appears that the clone does not favour lymphoid differentiation. This does not appear to be a differentiation block as mutant cells still progress along the lymphoid differentiation trajectory, however it is difficult to assess this without looking at the differences in transcriptional programmes of mutant and WT progenitor cells. Additionally, there appears to be an inefficiency of terminal maturation of the erythroid, and B and T-cell lymphoid lineages. The observation that terminal maturation of the myeloid lineage is not affected in the *ASXL1* mutant clone suggests a very modest myeloid differentiation bias. This is concordant with what has been observed in murine *in vivo* and *in vitro* studies, detailed in chapter 1. One hypothesis could be that the mutation confers a subtle lineage bias not through overt expansion, but through the conferral of selective disadvantages to mature megakaryocytes, erythrocytes, and lymphocytes, and a competitive advantage to mature myeloid cells.

4.6 Chapter conclusions

Throughout this chapter, a high-resolution landscape of human haematopoiesis was generated from the transcriptomic profiles from 16,756 single cells. This included both mutant and wildtype cells from 4 *ASXL1*-mutant CH samples and WT cells from 5 healthy control samples. Importantly, although a great deal of transcriptomic variation between samples, upon integration the the resulting dataset was representative of healthy human haematopoiesis without frank malignancy. Furthermore, cells from CH samples were found to be intermingled heterogeneously with cells from the WT healthy control samples, regardless of genotype. The *ASXL1*-mutant clone was found to be expanded most greatly within the HSC/MPP transcriptomic compartment and showed little sign of lineage differentiation bias. There was evidence however of an inefficiency of terminal maturation of the erythroid and lymphoid lineages. These observations provide a clue as to which cell compartments are most affected by *ASXL1* mutations. In the next chapter, downstream analyses between mutant and WT cells are focussed on these cell compartments.

Chapter 5

Transcriptomic consequences of *ASXL1* mutation

Contents

5.1	Analysis strategy, hypotheses for mechanisms of clonal expansion, and aims for this chapter	112
5.2	Non-cell autonomous mechanisms of clonal expansion within the HSC	115
5.2.1	Transcriptomic differences between <i>ASXL1</i> -mutant and non-CH WT HSC/MPP	116
5.2.2	Transcriptomic differences between WT-CH and non-CH WT HSC/MPP	118
5.3	Cell-intrinsic mechanisms of clonal expansion within the HSC/MPP compartment	121
5.4	Identification of subpopulations within the HSC/MPP compartment	125
5.5	Patterns of clonal dynamics in HSPC subpopulations	128
5.6	Cell-intrinsic mechanisms of clonal expansion within HSC subpopulations	130
5.6.1	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH cells in HSC2	130
5.6.2	Differential gene expression analysis between <i>ASXL1</i> -mutant and WT-CH cells in HSC1	134
5.7	Gene regulatory network inference analysis of HSC1 .	138
5.8	Consequences of <i>ASXL1</i> mutation upon the transcriptome of HSPCs	140
5.8.1	Differential regulon analysis of haematopoietic progenitor compartments	140
5.9	Concluding remarks	144

5.1 Analysis strategy, hypotheses for mechanisms of clonal expansion, and aims for this chapter

Thus far, there have been two major observations. First, the major clonal expansion occurs within the transcriptomic HSC/MPP compartment. Secondly, there is a greater difference between the abundance of wildtype and *ASXL1* cells within the mature cell compartments than in the more immature progenitor

compartments. These observations prompt some questions. Why does the mutant clone expand within the HSC/MPP compartment? And what are the molecular mechanisms that could be driving this? Does the mutation have a greater impact upon the transcriptome of more lineage committed cell types? And are the effects of mutation more prominent in particular lineages? In this chapter, transcriptomic profiles from mutant and wildtype cells from both *ASXL1*-mutant CH and WT healthy control samples are used to explore these questions.

Mounting evidence suggests that the acquisition of somatic mutations in *ASXL1* cause clonal expansion within the long-lived haematopoietic stem cell (HSC) compartment. To explore the reasons why this occurs, hypotheses that could explain the observed phenotype must first be made (Figure 5.1a).

First, cell-extrinsic environmental pressures from the CH bone marrow microenvironment might cause the preferential selection of *ASXL1*-mutant HSCs within the HSC pool. Due to the observed heterogeneity in humans, there may be multiple confounding factors that could shape the mechanisms of clonal expansion. These factors must first be identified to contextualise observed differences between mutant and wildtype cells. Secondly, cell-intrinsic differences between mutant and wildtype HSC/MPPs could result in the *ASXL1* mutant cells to outcompeting the wildtype cells for resources. Thirdly it could be a combination of the first two hypotheses, whereby the mutation confers an adaptive advantage to external selective pressures upon the mutant HSCs. To separate these intertwined hypotheses, different groups of cells can be compared.

To address the first hypothesis, transcriptomically defined HSC/MPPs from WT healthy donor cells were compared with mutant or WT HSC/MPPs from *ASXL1*-mutant CH samples. This analysis would identify differences in cellular responses between cells from CH and non-CH bone marrow microenvironments. Putative selective pressures within the CH bone marrow can then be inferred. To determine cell-intrinsic transcriptomic differences resulting from mutation, mutant and WT HSC/MPPs from CH samples were compared. For both of these analyses, differential gene expression analysis (DGEA) with subsequent gene set enrichment

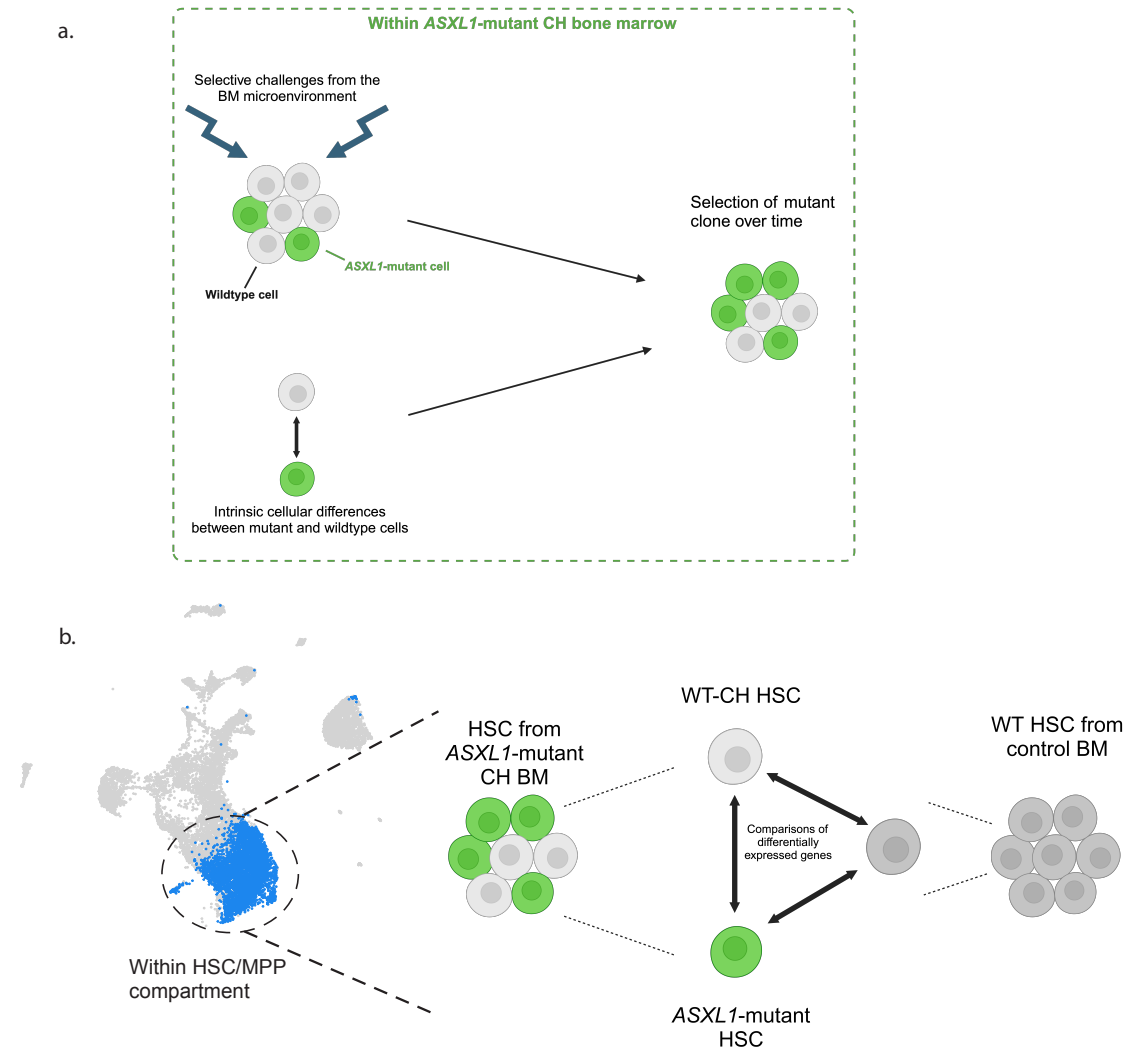


Figure 5.1: Overview of mechanisms of clonal advantage and differential gene expression analysis strategy. **a)** Selection of the *ASXL1*-mutant clone could be due to either: difference in response to selective challenges from the CH bone marrow, or intrinsic cellular advantages of mutant cells leading to clonal expansion, or a mixture of both hypotheses. **b)** Analysis strategy to determine putative mechanisms of clonal expansion within the HSC/MPP compartment.

analysis (GSEA) were used. This enabled identification of putative dysregulated biological processes and cellular components between different groups of cells (Figure 5.1b). Since *ASXL1* plays a reported functional role in epigenetic regulation of gene transcription, this analysis is likely to be the key for identifying probable molecular mechanisms of mutant clonal expansion.

Comparison of transcriptomic profiles between *ASXL1*-mutant and WT cells from CH samples can also be used to identify lineage-specific differences in effect

of *ASXL1* mutation. Since few mature cell types were sorted in this experiment, this comparison was performed between clones within lineage-committed progenitor cell populations.

The aims for this chapter were to:

1. Identify differences in cellular responses between cells from CH and non-CH bone marrow
2. Infer selective challenges present within the CH bone marrow
3. Identify key transcriptional differences between mutant and WT HSC/MPPs from CH bone marrow
4. Identify putative lineage-specific transcriptional responses to *ASXL1* mutation

5.2 Non-cell autonomous mechanisms of clonal expansion within the HSC

External pressures within the CH bone marrow microenvironment with the potential to influence clonal selection were first inferred from transcriptional profiles. For this, differences in cellular responses were identified between cells from CH samples and non-CH samples. To ensure accurate and complete identification of differentially expressed genes, two orthogonal analyses were performed.

First, gene expression profiles from single cells belonging to the two test groups were compared using a linear mixed model implemented by the dream workflow from the limma:voom pipeline [253, 254]. This method of DGEA utilises a linear mixed model approach to account for the confounding sources of variation. For these analyses, the sample identifier, the investigator, and patient sex were used as covariates. Secondly, a mini-pseudobulk approach was used to aggregate normalised gene expression profiles of groups of cells ($n > 3$) from each sort plate. For example, one pseudobulk replicate would be all of the mutant HSC/MPPs from an *ASXL1*-mutant CH sample that was sorted onto one plate. DGEA was performed using DESeq2 on the mini-pseudobulk replicates, defining the same covariates as for the single cell approach.

5.2.1 Transcriptomic differences between *ASXL1*-mutant and non-CH WT HSC/MPP

First, transcriptional profiles between mutant HSC/MPPs from *ASXL1*-mutant CH samples and WT HSC/MPPs from healthy control samples were compared. Using single-cell dream-implemented DGEA, 3,374 genes were found to be significantly ($\text{FDR} < 0.05$) differentially expressed between mutant HSC/MPPs and non-CH WT HSC/MPPs (Figure 5.2b). Of these, 1,448 were found to be upregulated in mutant cells, and 1,926 were upregulated in WT cells.

The most significantly upregulated genes in the mutant HSC/MPPs were *RPS26* (Ribosomal protein S26), a ribosomal 40S subunit, *PDLIM4* (PDZ and LIM domain 4) which is involved in actin cytoskeleton reorganisation, and *MRC2* (Mannose receptor C-Type 2), another cytoskeleton interacting protein implicated in tumorigenesis of multiple malignancies [278–280]. Additional notable upregulated genes were *AXL*, a receptor tyrosine kinase that has been found to confer a self-renewal advantage upon overexpression in CD34+ CML cells [281], and *RUNX3*—a key haematopoietic transcription factor implicated in AML development [282]. Among the most downregulated genes were multiple mitochondrial genome-encoded genes (*MT-CYB*, *MT-ND2,4,5*, *MT-ATP6*, *MT-CO3*), suggesting a decrease in mitochondrial activity in mutant cells compared to non-CH HSCs. Additionally, *RASAL2* (RAS protein activator-like 2) was the most down regulated gene, which is thought to play a role in tumour suppression [283].

When performing gene set enrichment analysis, one striking observation is that *ASXL1*-mutant HSCs were positively enriched for gene signatures of responses to different inflammatory stimuli (Figure 5.2c). These included responses to interferons alpha, beta and gamma stimulation and also nuclear factor kappa-light-chain-enhancer of activated B cells (NF κ B). This may suggest a greater concentration of pro-inflammatory cytokines within the *ASXL1*-CH bone marrow milieu. Also, mutant cells were positively enriched for gene signatures of hypoxic response. Hypoxia is a key physiological feature of the bone marrow niche and is a determinant of the maintenance of HSC quiescence and apoptosis [284]. Alterations of oxygen

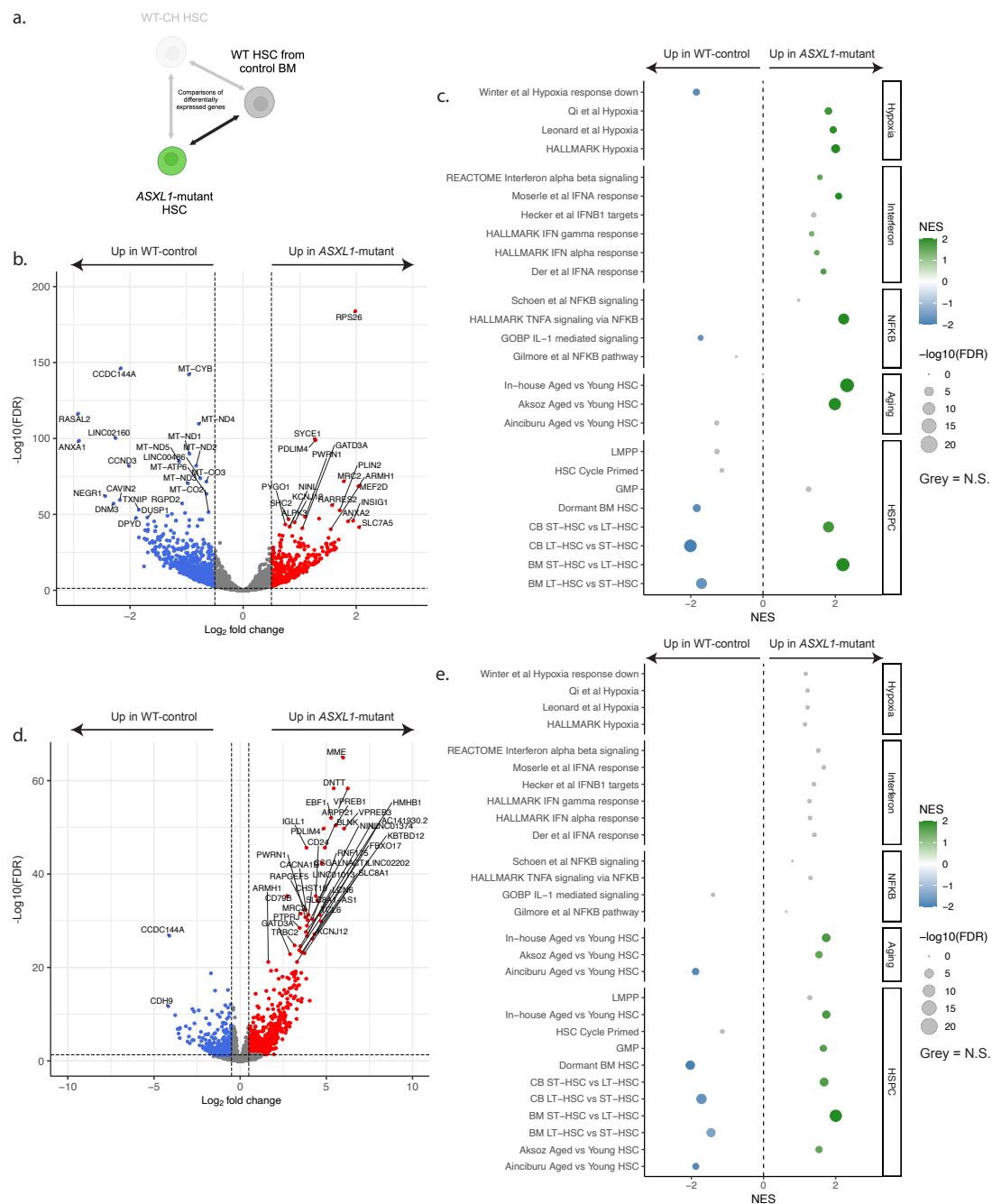


Figure 5.2: Differential gene expression analysis between *ASXL1*-mutant and non-CH HSC/MPPs. **a)** DGEA between mutant HSCs from CH bone marrow, and WT HSCs from non-CH bone marrow. **b)+d)** Volcano plots visualising DEGs between mutant HSCs and WT HSCs from non-CH samples. DEGs were determined either by a linear mixed-model approach comparing gene expression between single cells, whilst compensating for covariates (**b**; Sample identifier and patient sex as covariates), or by pseudobulking gene expression profiles of groups of cells on each plate, such that each clone per plate was a replicate. (*continued on next page*)

Figure 5.2 (previous page): (d). Each point is a gene. Log2FC threshold was set as 0.5, and a significance threshold of FDR < 0.05 (B-H adjustment). Genes were coloured by significant upregulation in CH or WT samples (red and blue respectively). **c) + e)** GSEA results for gene signatures from Hallmark, Reactome, Gene Ontology collections, and published studies, separated by biological category. Size of dots represent significance ($-\text{Log}_{10}(\text{FDR})$). Dots are coloured by NES- green dots are enriched in CH cells, blue dots are enriched in non-CH cells. Grey dots represent non-significant results (FDR < 0.05; B-H adjusted).

levels within the CH bone marrow could act as a selective pressure on mutant and WT cells. Finally, mutant cells were positively enriched for aged HSC and short-term HSC-like gene signatures.

The mini pseudobulk approach identified 1,783 significantly DEGs (FDR < 0.05), of which 1,103 were upregulated and 680 were downregulated in mutant HSC/MPPs (Figure 5.2d). Similar to the single-cell approach *PDLIM4*, *MRC2* were identified amongst the most upregulated genes. Interestingly, multiple genes important in B-cell lineage specification were also found to be upregulated in mutant cells. For example, *EBF1* (early B-cell factor) which is a critical transcription factor in the maintenance of B cell identity [285], and more canonical B cell markers *DNTT*, *VPREB1,3*, *CD79B* [286–288]. Among the downregulated genes, mitochondrially-encoded genes were again identified (*CCDC144A*, *MT-ND1,2,5*), along with *RASAL2*.

When GSEA was performed using this ranked DEG list, mutant cells were again found positively enriched for gene signatures associated with aged HSCs and a more short-term HSC-like phenotype (Table A.16). Whilst mutant cells were also found to be positively enriched for inflammatory response and hypoxia gene sets, these results were statistically not significant after p-value adjustment (All results FDR > 0.05).

5.2.2 Transcriptomic differences between WT-CH and non-CH WT HSC/MPP

Next, the differences between CH and non-CH bone marrow at the transcriptional level were further dissected by performing DGEA between wildtype cells from both sample types (Fig 5.3a). This analysis is important in estimating the differences

between the CH and non-CH bone marrow microenvironment, as it removes the confounding impact of the mutation.

Using the single-cell dream pipeline approach, 4,779 genes were found to be significantly differentially expressed ($\text{FDR} < 0.05$) between CH and non-CH WT HSC/MPPs (Fig 5.3b). Of these, 914 were upregulated in WT-CH HSC/MPPs and 3,865 were downregulated. One notable observation was that B-lymphopoiesis associated genes were again some of the most upregulated genes in the WT-CH HSC/MPPs (*BLNK*, *VPREB1*, *DNTT*, *EBF1*). This could be indicative of an upregulated environmental cue present within the CH bone marrow encouraging B cell lymphoid differentiation, for example IL7 or CXCL12 [289, 290]. Interestingly several genes associated with active $\text{NF}\kappa\text{B}$ signalling including *IRAK3* (Interleukin 1 receptor associated kinase 3), *FKBP5* (FK506-binding protein 5), *NFKBIA* ($\text{NF}\kappa\text{B}$ inhibitor alpha) [291–293] were among the most downregulated genes in WT-CH HSC/MPPs. This may suggest an attenuated response to $\text{NF}\kappa\text{B}$ -mediated inflammatory stimulus within the CH bone marrow. The identified DEGs reaching significance were validated using the mini-pseudobulk approach (Figure 5.3d).

To identify different cellular responses between the WT-CH and non-CH WT cells, GSEA was performed using the ranked DEG lists as input (Fig 5.3c,e). Wildtype HSC/MPPs from CH bone marrow were also found to be positively enriched for gene signatures of response to interferons alpha, beta, and tumour necrosis factor alpha signalling via $\text{NF}\kappa\text{B}$. This is consistent with what has been previously reported linking CH to a heightened inflammatory state within the bone marrow [214, 219–223].

Furthermore, the WT-CH cells were also positively enriched for aged HSC gene signatures, and negatively enriched for long-term, dormant HSC like signatures. Both of these observations have been previously identified in HSC/MPPs from primary human CH bone marrow [45]. Additionally, WT-CH cells were negatively enriched for LMPP and GMP gene signatures, suggestive of less priming towards myeloid differentiation. However, unlike the analysis between mutant and non-CH WT cells, WT-CH HSC/MPPs were not significantly enriched for hypoxia gene

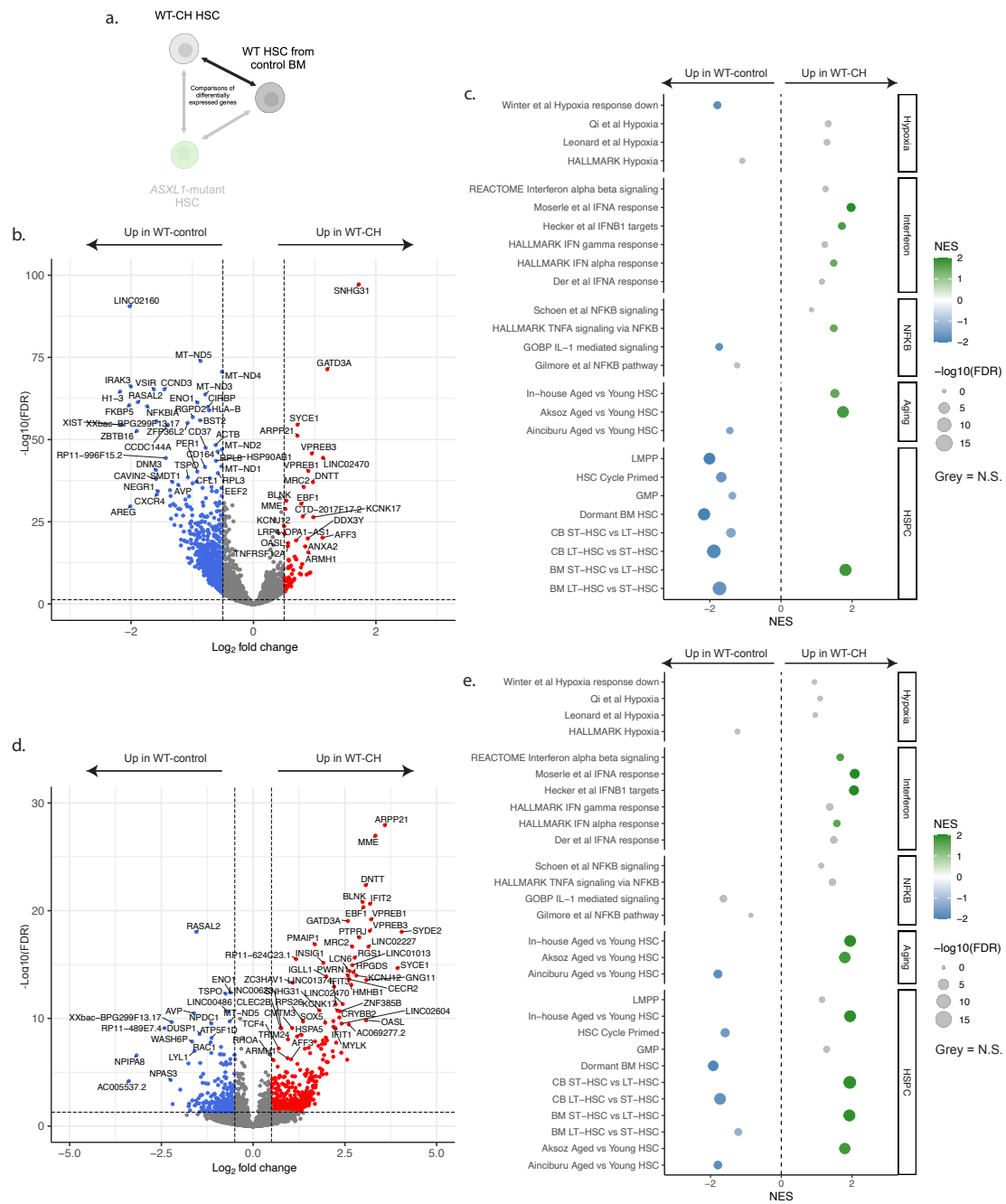


Figure 5.3: Differential gene expression analysis between WT-CH and non-CH HSC/MPPs. **a)** DGEA between WT HSCs from CH samples, and WT HSCs from non-CH samples. **b)+d)** Volcano plots visualising DEGs between WT HSCs from CH samples and WT HSCs from non-CH samples. DEGs were determined either by a linear mixed-model approach comparing gene expression between single cells, whilst compensating for covariates (**b**; Sample identifier and patient sex as covariates), or by pseudobulking gene expression profiles of groups of cells on each plate, such that each clone per plate was a replicate. (*continued on next page*)

Figure 5.3 (previous page): (d). Each point is a gene. Log2FC threshold was set as 0.5, and a significance threshold of FDR < 0.05 (B-H adjustment). Genes were coloured by significant upregulation in CH or WT samples (red and blue respectively). **c)+ e)** GSEA results for gene signatures from Hallmark, Reactome, Gene Ontology collections, and published studies, separated by biological category. Size of dots represent significance (-Log10(FDR)). Dots are coloured by NES- green dots are enriched in CH cells, blue dots are enriched in non-CH cells. Grey dots represent non-significant results (FDR < 0.05; B-H adjusted).

signatures. This difference in response to hypoxia is likely to be a consequence of mutation, rather than a selective pressure within the CH marrow.

In summary, using both single-cell and mini-pseudobulk DGEA approaches, HSC/MPPs from the CH bone marrow independent of mutational status were found to be enriched for inflammatory HSC-like transcriptional phenotypes. Chronic inflammation has been linked to impaired self-renewal potential, increased rate of ageing, myeloid-biased differentiation programmes, and apoptosis in HSCs [215–218], and is therefore likely to be a primary selective pressure within the CH marrow. Further evidence for this hypothesis is that mutant and WT HSC/MPPs are also enriched for aged-HSC gene signatures. Therefore, in the *ASXL1*-mutant patients, the marrow is likely to be more pro-inflammatory than the healthy control non-CH bone marrow. It is tempting to speculate that the mutation may be driving increased expression of pro-inflammatory cytokines, further increasing a hostile environment for aged HSCs. This hypothesis would need to be tested with *in vitro* functional knock-in assays however.

5.3 Cell-intrinsic mechanisms of clonal expansion within the HSC/MPP compartment

Next, I asked how the transcriptional consequences of *ASXL1* mutation confer selective advantage upon HSC/MPPs from the CH marrow. To achieve this, I applied the same single-cell and pseudobulk DGEA approaches to identify DEGs between mutant and WT HSC/MPPs from the CH samples (Figure 5.4a).

Using the single-cell dream DGEA pipeline, 1,253 differentially expressed genes were identified between *ASXL1*-mutant and WT-CH HSC/MPPs that passed the

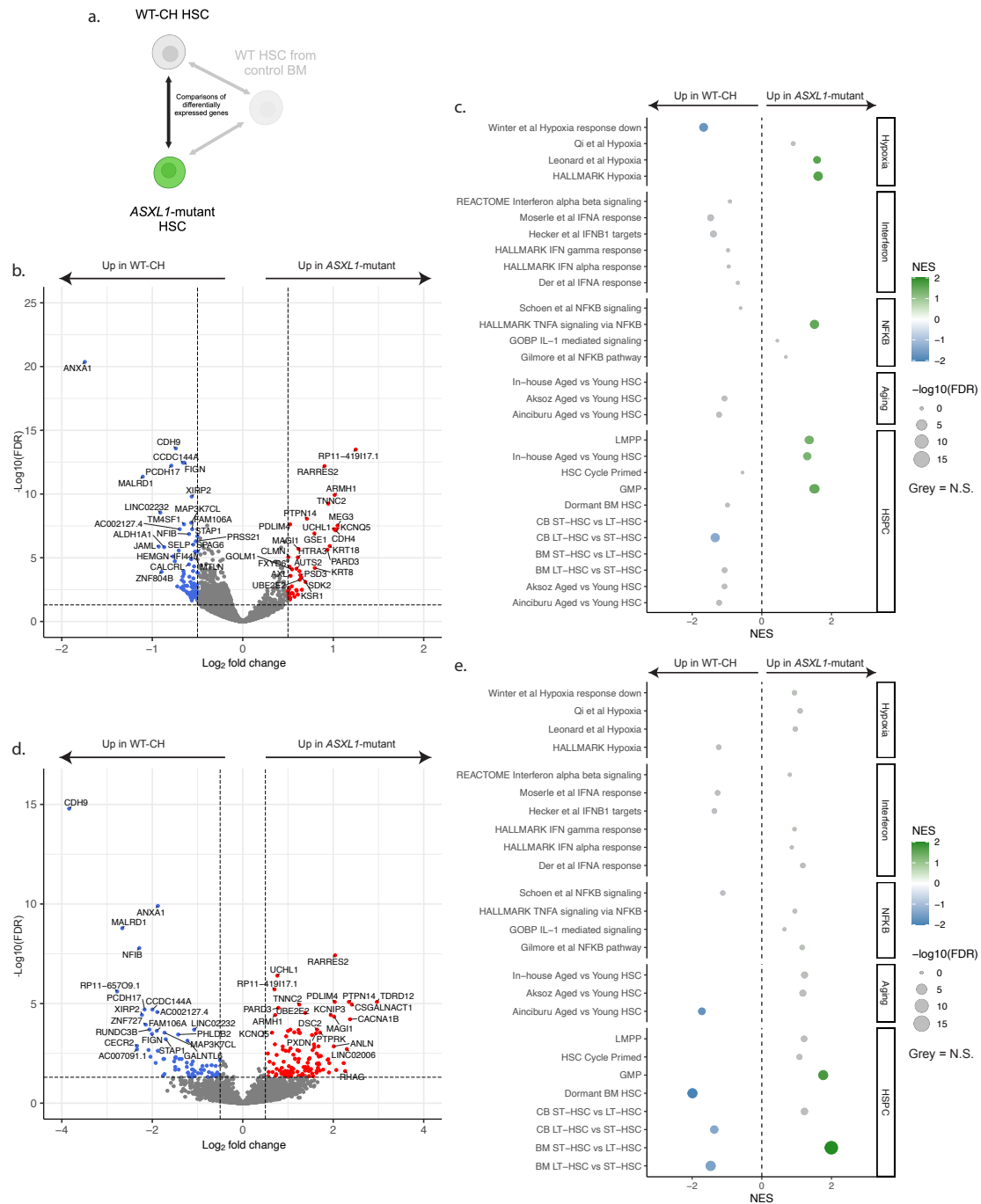


Figure 5.4: Differential gene expression analysis between *ASXL1*-mutant and WT-CH HSC/MPPs. **a)** DGEA between mutant HSCs and WT HSCs, both from CH samples. **b)+d)** Volcano plots visualising DEGs between mutant HSCs and WT HSCs from CH samples. DEGs were determined either by a linear mixed-model approach comparing gene expression between single cells, whilst compensating for covariates (**b**; Sample identifier and patient sex as covariates), or by pseudobulking gene expression profiles of groups of cells on each plate, such that each clone per plate was a replicate. (continued on next page)

Figure 5.4 (previous page): (d). Each point is a gene. Log2FC threshold was set as 0.5, and a significance threshold of $FDR < 0.05$ (B-H adjustment). Genes were coloured by significant upregulation in CH or WT samples (red and blue respectively). **c) + e)** GSEA results for gene signatures from Hallmark, Reactome, Gene Ontology collections, and published studies, separated by biological category. Size of dots represent significance ($-\log_{10}(FDR)$). Dots are coloured by NES- green dots are enriched in mutant cells, blue dots are enriched in WT-CH cells. Grey dots represent non-significant results ($FDR < 0.05$; B-H adjusted).

significance threshold. Interestingly, this is much more than for the *DNMT3A*- and *TET2*- mutant CH, where only 52 and 216 genes in total respectively were found to be differentially expressed between mutant and WT HSCs ($FDR < 0.05$) [45]. This suggests that *ASXL1* mutations may have a greater effect upon global transcription of HSCs than *DNMT3A* and *TET2* mutations.

Amongst the 139 genes significantly upregulated in mutant HSC/MPPs, the most upregulated genes were *RP11-419I17*, a long non-coding RNA sequence, *RARRES2* (retinoic acid receptor responder 2)- a secreted chemotactic protein that exerts its effect through G-protein coupled receptors [294], and *ARMH1* (Armadillo like helical domain containing 1)- a cytosolic protein of unknown function that has found to be upregulated in AML blasts [295]. Other notable genes amongst those most upregulated include *GSE1* (Genetic suppressor element 1)- a subunit of the BRAF35-histone deacetylase complex with oncogenic activity in AML [296], *PARD3* (Partition defective 3)- a polarity protein which regulates asymmetric cell division [297], cytoskeletal proteins (*KRT18*, *MAGI1*, *CLMN*), and potassium gated ion channels proteins (*KCNQ5*, *KCNIP3*).

Those genes most significantly downregulated in mutant cells included *ANXA1* (annexin 1)- an annexin family member with multiple roles including anti-inflammatory properties [298], *CDH9* (Cadherin 9), and *PCDH17* (protocadherin 17)- a tumour suppressor which is downregulated in both ALL and AML [299]. These DEGs were verified using the mini pseudobulk approach described above: of the 207 significant DEGs identified in the pseudobulk analysis, 64.7% were also identified using the single cell approach.

In the previous section, the transcriptional profiles of HSC/MPPs from the *ASXL1*-CH bone marrow were enriched for gene signatures of inflammation and ageing, when compared to non-CH counterparts. To investigate whether *ASXL1* mutations cause difference in transcriptional response to a pro-inflammatory bone marrow microenvironment, GSEA was performed on gene signatures of different inflammatory responses using DEGs identified between *ASXL1*-mutant and WT-CH HSC/MPPs.

Interestingly, no significant differences were observed between mutant and wildtype HSC/MPPs in gene signatures for interferon alpha, beta, and gamma responses, and aged HSCs. However, mutant HSCs were positively enriched for one gene signature of TNF- α signalling. This suggests that both *ASXL1*-mutant and WT-CH HSCs respond very similarly to a heightened inflammatory bone marrow microenvironment postulated to be present in patients with CH. This result contrasts with what has been recorded previously in human *DNMT3A*- and *TET2*-mutant CH [45]. In this study, when compared to WT-CH counterparts, mutant HSC/MPPs were found to have an attenuated transcriptional response to ageing and pro-inflammatory bone marrow milieu. Consequently, in *ASXL1*-mutant CH, clonal advantage is unlikely to be attributed solely to difference in response to inflammation, although it may still play a part. A different hypothesis linking mutation to clonal expansion is therefore required.

As previously identified in this study, *ASXL1*-mutant HSC/MPPs are positively enriched for gene signatures in response to hypoxia. This result is surprising since a hypoxic response has been associated with increased quiescence of murine HSCs [284], yet *ASXL1*-mutant HSC/MPPs are negatively enriched for dormancy gene signatures. Finally, mutant HSCs were also positively enriched for short-term HSC, LMPP and GMP gene signatures, suggesting that mutant HSC/MPPs are less dormant, more active, and more primed towards differentiation than WT-CH HSC/MPPs. From this data, the impact of *ASXL1* mutations within the HSC/MPP compartment are both subtle and unclear. To further clarify the

cell-intrinsic transcriptomic differences between mutant and WT cells, a more granular analysis is required.

5.4 Identification of subpopulations within the HSC/MPP compartment

In the previous section, the transcriptomic differences between mutant and WT HSC/MPPs within CH samples were subtle. One hypothesis to account for this is that *ASXL1* mutations confer clonal advantage upon a specific subpopulation of HSCs. If all HSC/MPPs are analysed together as a heterogeneous population, transcriptomic subtleties between specific subpopulations may be lost. To test this hypothesis transcriptomic HSPCs (HSC/MPP, EMPP, LMPP and LMPP cycling compartments) were subclustered using the self-assembling manifold algorithm (SAM) [257], which prioritises the most biologically distinct features in an unsupervised approach to separate an otherwise seemingly homogenous group of cells. This has previously been used to successfully separate human HSC/MPP populations [45, 273, 300].

Using the SAM algorithm and subsequent Louvain clustering and UMAP projection, 7,782 cells were separated into 10 clusters (Figure 5.5a). When cells were highlighted by the original cell annotations using symphony projection, clusters 0, 2, 3 4, 7, 9 were mainly composed of the HSC/MPP compartment. Clusters 1 and 8 were mainly transcriptomic LMPP; cluster 5 was mainly EMPP, and cluster 6 was mainly LMPP cycling. The resulting UMAP followed a trajectory of differentiation from clusters 0 and 4 down to clusters 5 and 6. This differentiation trajectory was verified by colouring each cell by individual AUCell enrichment score for a gene signature of long-term HSCs [273], and also by highlighting each cell by immunophenotype from index sorting data (Figure 5.5c,d). Combining this information, clusters were annotated, identifying four distinct HSC subpopulations (Figure 5.5e). Marker genes defining each of these four HSC subpopulations are detailed in Table A.17. The HSC1 population (n= 2,079 cells) was labelled by merging clusters 2 and 4. These clusters were merged as they formed a continuum

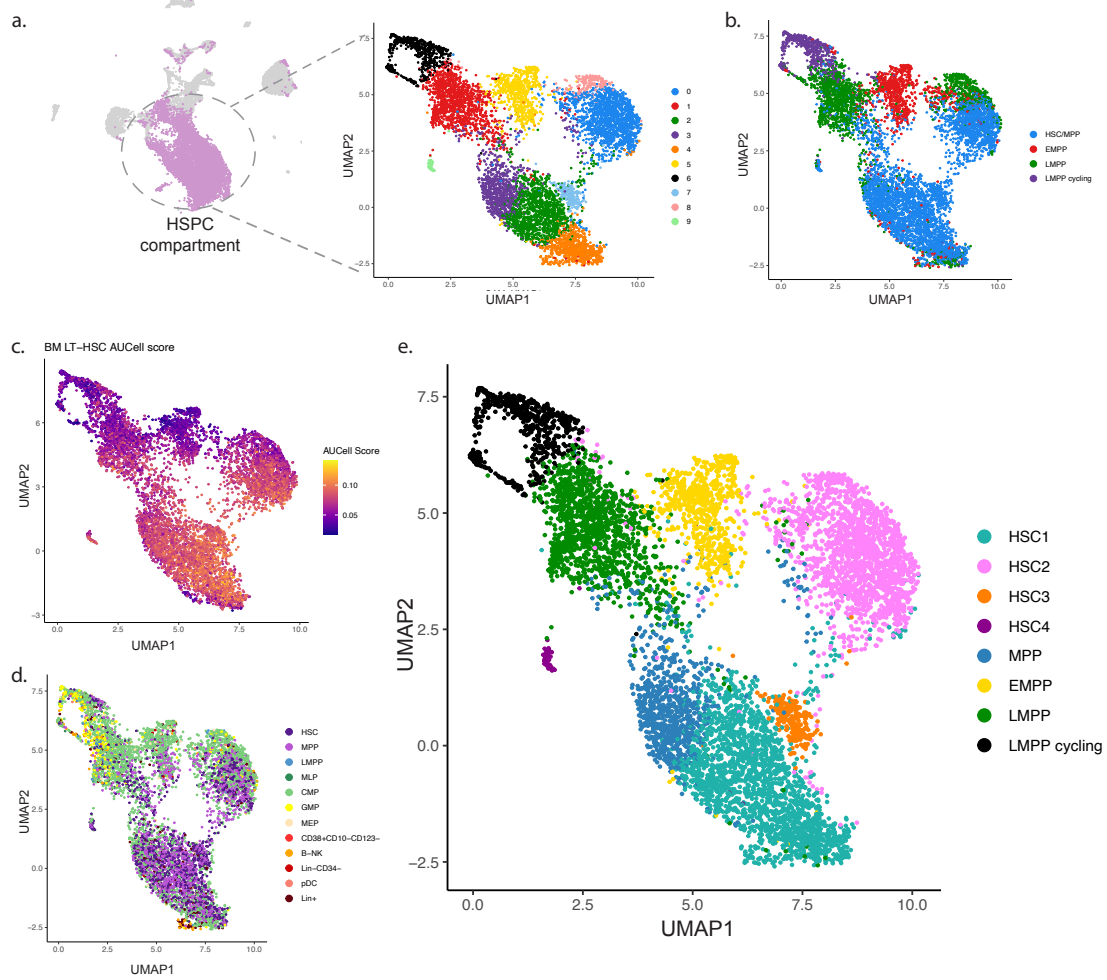


Figure 5.5: Identification of HSPC subpopulations. a) HSC/MPP, EMPP, LMPP, and LMPP cycling compartments (7,782 cells) were subclustered by the self-assembling manifolds (SAM) algorithm using harmony-adjusted PCs as input. The UMAP embedding of the output is coloured by Louvain clusters. b) UMAP embedding of SAM algorithm subclustering, coloured by original cell type annotation. c) UMAP with single cells coloured by AUCell enrichment scores for a gene signature for marker genes expressed by long-term HSCs from human bone marrow. d) UMAP coloured by immunophenotype from index sorting data. e) UMAP with Louvain clusters annotated by subclustered cell type.

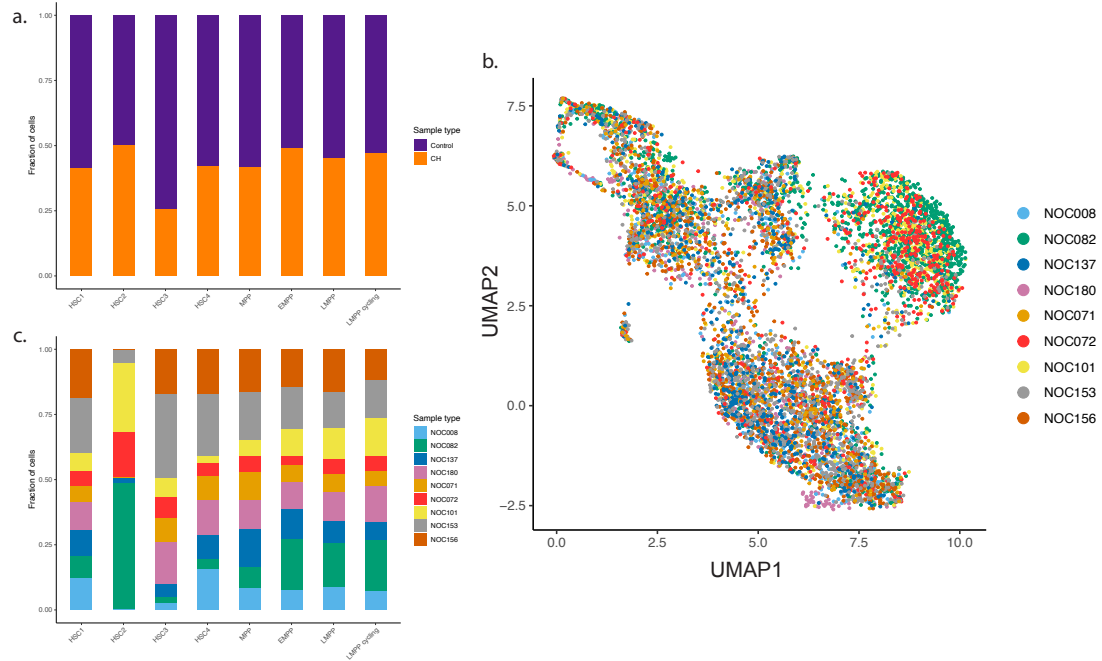


Figure 5.6: Composition of HSPC subclusters. a) Fraction of cells from CH or non-CH samples in each subcluster. b) UMAP coloured by sample identifier. c) Fraction of cells from each individual sample in each subcluster. The four CH samples are at the top of the legend.

with very similar immunophenotypic profiles and similar AUCell enrichment score for the LT-HSC gene signature. Cluster 3 was annotated MPP as it appeared the most mature population of cells in this continuum. HSC2 (n=1,846 cells) was formed from merging clusters 0 and 8, as they formed a distinct major entity of cells. Two further smaller clusters- HSC 3 and 4 formed minor, but distinct entities of transcriptomic and immunophenotypic HSC/MPPs (n = 249 cells; n = 79 cells respectively).

To understand if there was an enrichment of CH samples within specific subpopulations, the composition of each group was inspected. Control samples composed a modestly larger fraction in most subpopulations compared to CH samples (Figure 5.6a). In HSC3 however, cells from CH samples only composed $\approx 25\%$ of cells. Next, sample composition of subpopulations were examined (Figure 5.6b,c). For most populations of cells, samples were distributed heterogeneously, with each sample being represented in each subpopulation. However, notably the HSC2 subpopulation was dominated by the NOC082 CH sample, which accounted

for 47.40% of cells. This population seemed to be composed by specific samples regardless of CH classification. For example, NOC101 and NOC072- both healthy control samples- were overrepresented in this population compared to the other subpopulations (26.00% and 16.79% respectively). Interestingly, one CH sample (NOC180) was not represented at all, and the other CH samples (NOC008, and NOC137) were barely represented in HSC2 (0.32% and 1.84% respectively).

5.5 Patterns of clonal dynamics in HSPC subpopulations

In chapter 4, the major expansion of the *ASXL1*-mutant clone was identified within the HSC/MPP compartment. To understand if selection was occurring within a specific HSC subpopulation, the same strategy to quantify clonal expansion was applied to the SAM-identified subpopulations (Figure 5.7a). The distribution of the *ASXL1*-mutant clone throughout the subpopulations was heterogeneous with the WT cells from the CH and control samples, however there appeared to be a concentration of the mutant clone within the HSC2 compartment. When the clonal composition of the subpopulations were quantified, the HSC2 compartment was composed of the greatest percentage of mutant cells (26.81%; Figure 5.7b). However, as established in the previous section, only one CH sample, NOC082 dominated the HSC2 compartment, making it difficult to conclude that the major expansion of *ASXL1*-mutant clones happens in the HSC2 compartment.

To further quantify patterns of clonal expansion within HSPC subpopulations with greater granularity, the mean densities of the mutant and wildtype clones for each CH sample within each subpopulation were calculated (Figure 5.7c,d). The ratio of mutant/WT clone for each clone was normalised to that of the HSC1 subpopulation, as this was the largest identified HSC population with representation from all 4 CH samples. Relative to the HSC1, the HSC2 population was found to be modestly expanded. However, this is likely to be driven by the over representation of one sample, and it would therefore be right to conclude that the mutant clone is greatly expanded in the HSC2 only in NOC082.

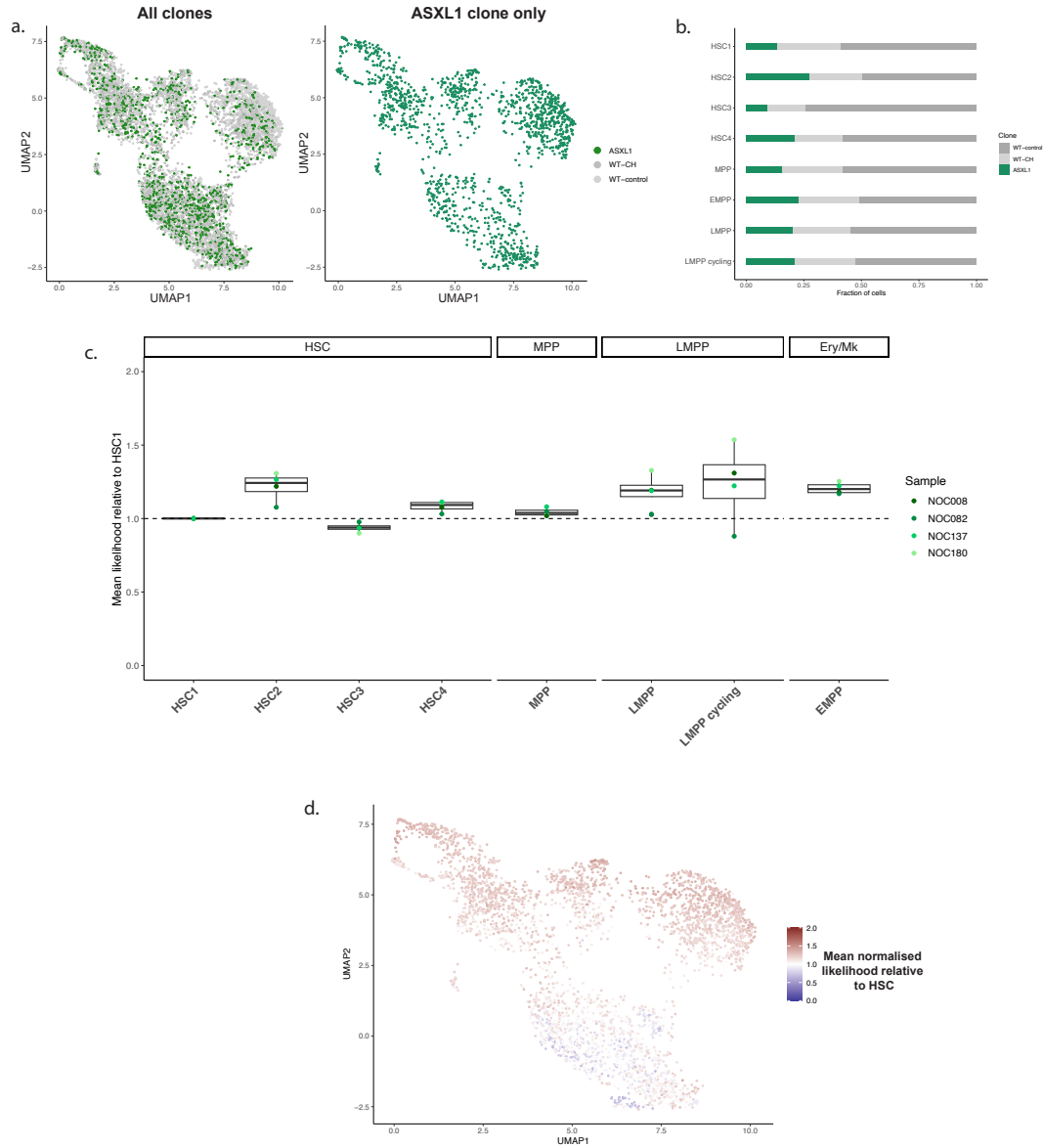


Figure 5.7: Clonal composition of HSPC subclusters. **a)** UMAP embedding with single cells coloured by clone, assigned by single-cell genotyping. Left: coloured by all clones. Right: Projection of only the *ASXL1*-mutant clone. **b)** Fraction of each clone per SAM-identified subcluster. **c)** Mean likelihood of the finding the *ASXL1*-mutant clone in each SAM-identified cell subcluster for each *ASXL1*-CH sample, normalised to that of the HSC1 compartment. Each point represents an *ASXL1*-CH sample. The horizontal dashed line at $y=1$ represents the mean likelihood of finding an *ASXL1*-mutant cell in the HSC1 across all four samples. Points lying above or below that line indicate that relative to the HSC1 compartment, the size of the mutant clone is expanded or contracted respectively in that cell compartment. **d)** UMAP embedding with single cells coloured by likelihood score of each cell being *ASXL1*-mutant, normalized to the HSC1 compartment.

A clear observation from this analysis was that there was little discernible difference in clonal density between the three other HSC subpopulations, and that the greatest number of HSCs compose the HSC1 population. Therefore, a more generalisable conclusion is that within the transcriptomic HSC/MPP population, the mutant clone expands as part of the HSC1 subpopulation. Additionally, the downstream LMPP, LMPP cycling, and EMPP populations were found to be modestly expanded relative to the HSC1, consistent with what was observed in Fig 4.11c. This therefore suggests that the mutant clone has a preference to differentiate towards the progenitor populations relative to their WT counterparts.

5.6 Cell-intrinsic mechanisms of clonal expansion within HSC subpopulations

Next, having identified two major sub-populations of HSC where clonal expansion occurs, I asked: how do *ASXL1* mutations confer clonal advantage to both populations at the transcriptional level? To achieve this, differential gene expression analysis was performed between mutant and WT cells from CH samples for HSC1 and HSC2 (Figures 5.8 and 5.10). HSC3 and HSC4 were not investigated due to insufficient cell number.

5.6.1 Differential gene expression analysis between *ASXL1*-mutant and WT-CH cells in HSC2

First, differential gene expression analysis was used to identify transcriptomic differences between mutant and WT cells in the HSC2 compartment (Fig5.8). Using the single-cell dream-implemented DGEA workflow, 250 DEGs were identified of which only 42 were upregulated in mutant HSC2 and 207 were downregulated (Table A.19). One obvious observation was that genes involved in cell-cell adhesion and cytoskeletal reorganisation including *CDH4*, *SDK2*, *PDLIM4*, *MAGI4*, were upregulated in mutant HSC2. One of the most intriguing, upregulated genes was *LITAF* (lipopolysaccharide-induced TNF-alpha factor)- a protein that directly induces *TNF-α* expression by binding to the promoter of *TNF-α* [301]. As *LITAF* elicits an immune response, *LITAF* is normally expressed in monocytes,

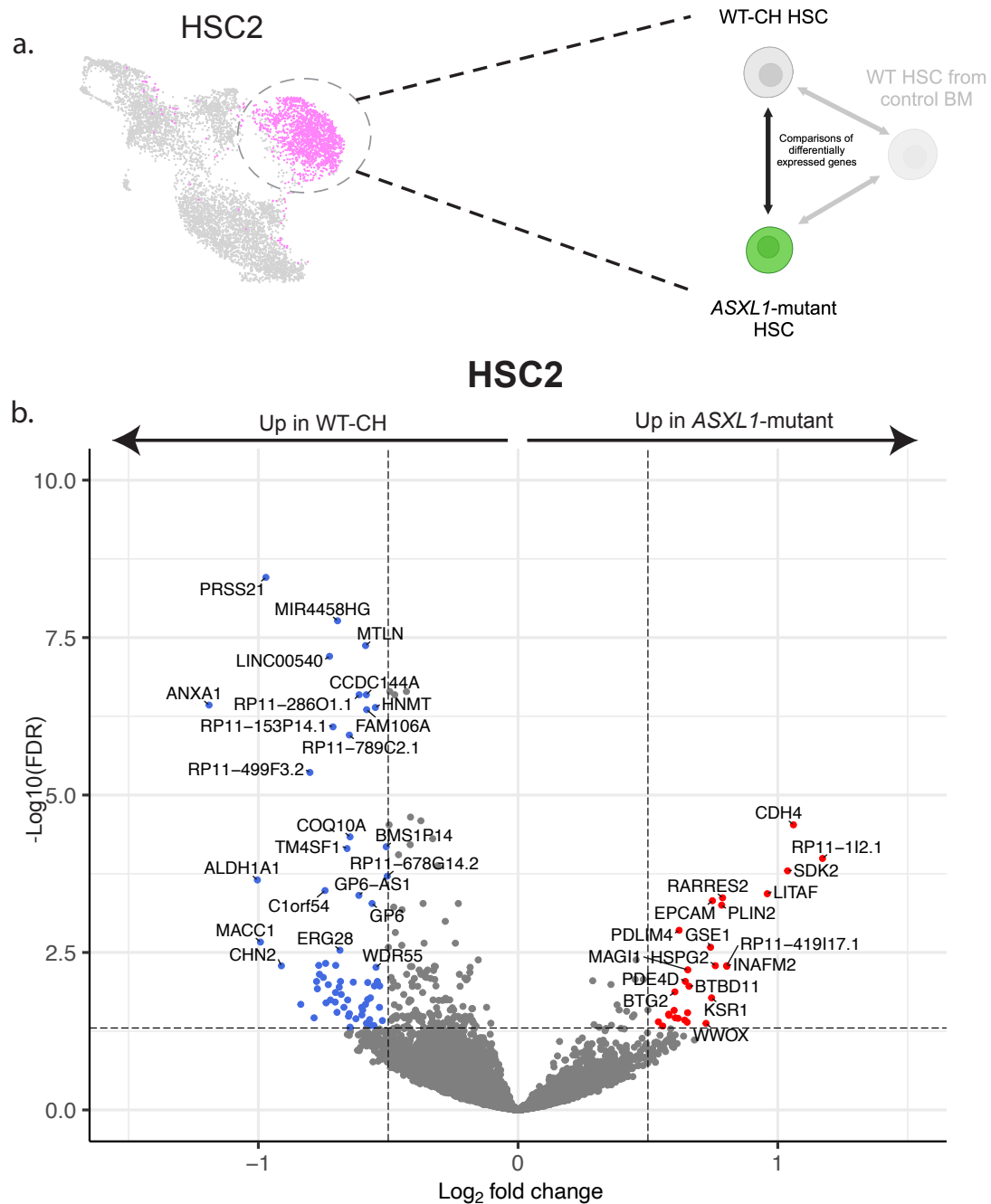


Figure 5.8: Differential gene expression analysis between *ASXL1*-mutant and WT-CH HSC2 subpopulation. **a)** Schematic showing DGEA strategy between mutant and WT HSCs from *ASXL1*-mutant CH samples in the HSC2 (n = 915) subcluster. DGEA was performed accounting for sample and sex batch effects. **b)** Volcano plots visualising differentially expressed genes between mutant HSC2 and WT HSC2 from CH samples. DEGs were determined by a linear mixed-model approach comparing gene expression between single cells, whilst accounting for sample identifier and patient sex batch effects. *(continued on next page)*

Figure 5.8 (previous page): Each point is a gene. Log2FC threshold was set as 0.5, and a significance threshold of FDR < 0.05 (B-H adjustment). Genes were coloured by significant upregulation in mutant or WT cells (red and blue respectively).

macrophages, dendritic cells, and lymphocytes but not normally in HSCs [302]. *LITAF* expression is indicative of inflammatory TNF- α secretion by mutant HSC2, supporting the hypothesis that chronic inflammation is a selective pressure within the CH bone marrow, and is exacerbated by *ASXL1* mutation. Additionally, multiple genes associated with suppression of inflammation were amongst the top downregulated DEGs in mutant HSC2, including *NLRP2* (suppression of NF κ B1 activity by TNF- α), *TRIM21* (modulator of interferon regulatory factor 3 protein stability), and *IRF8* (a transcription factor with both repressive and activator function of interferon-inducible genes) [303–305].

Next, gene set enrichment analysis was used to identify cellular processes and molecular functions commonly dysregulated between mutant and WT HSC2. In the HSC2 subpopulation gene signatures for a more HSC-like transcriptomic phenotype, including dormancy, were upregulated in mutant cells (Figure 5.9a). This is contrary to what was observed when the entire HSC/MPP compartment was analysed together. Signatures for a more progenitor-like state were not significantly enriched for (data not shown), suggesting that mutant HSC2 are more HSC-like than their WT counterparts.

One of the most obvious GSEA results for the HSC2 was the significant enrichment in inflammatory response gene signatures in mutant cells, including increased TGF- β , and TNF- α signalling pathways. This is indicative of a pro-inflammatory population of HSCs within the bone marrow with a heightened response to stimulation. NF κ B activation can lead to cytokine release, and increased expression of cell adhesion molecules [307], correlating with what was observed in the DEG list. These enrichments of inflammation-associated gene signatures were confirmed by AUCell analysis. Highlighting each cell by AUCell enrichment score for these gene sets identifies the HSC2 cluster as having the greatest enrichment (Figure 5.9b,c).

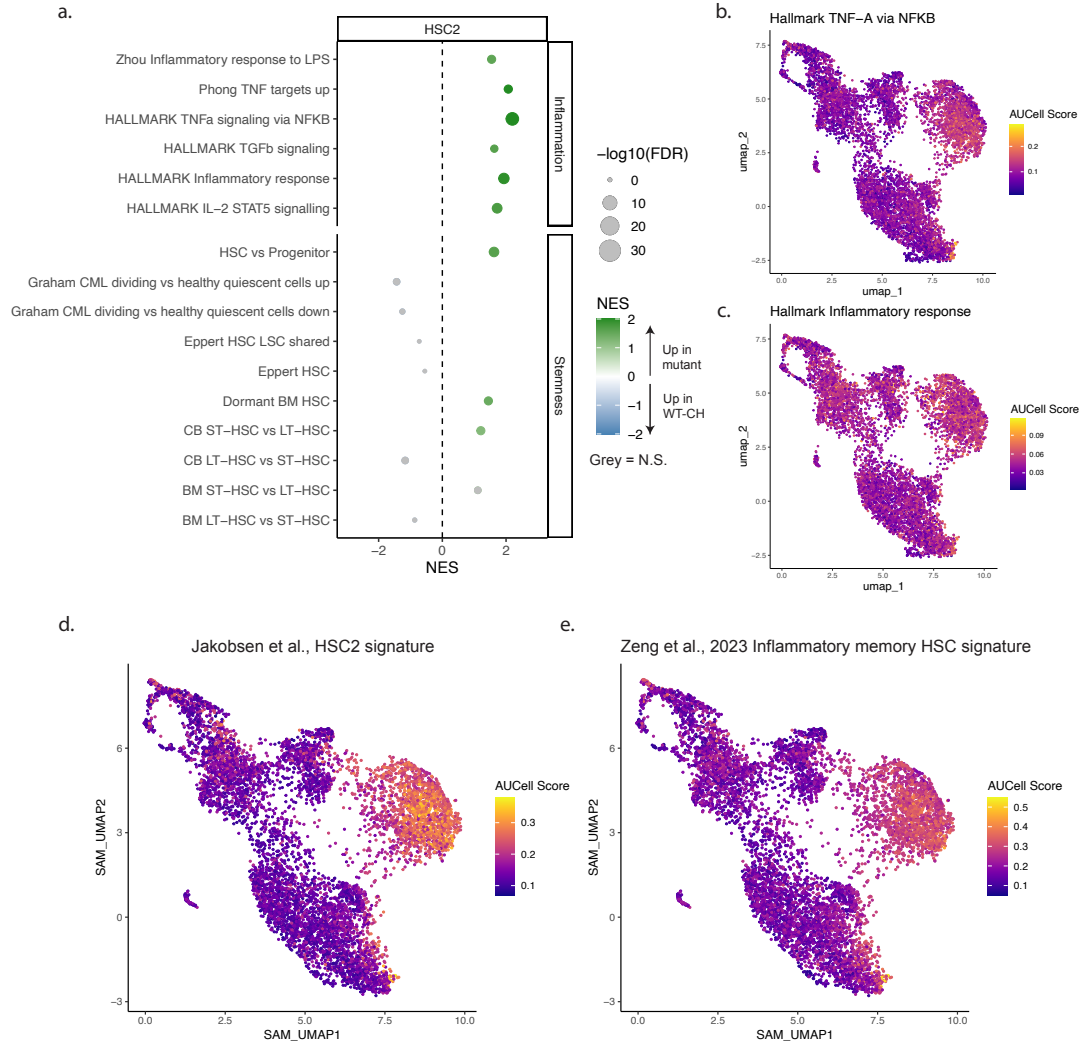


Figure 5.9: Heightened inflammatory transcriptomic state of *ASXL1*-mutant HSC2. a) GSEA results for Hallmark and published study inflammatory and haematopoietic stemness signatures. Ranked lists of DEGs from the dream linear mixed model approaches were used as input. Size of dots represent significance ($-\log_{10}(\text{FDR})$). Dots are coloured by NES- green dots are enriched in mutant cells, blue dots are enriched in WT cells. Grey dots represent non-significant results ($\text{FDR} < 0.05$; B-H adjusted) b) -e) UMAP superimposed with single-cell AUCell signature enrichment scores for Hallmark TNF-a via NFkB (b), Hallmark inflammatory response (c), inflammatory HSC2 population (Jakobsen et al., 2024) [45], and inflammatory memory HSCs (Zeng et al., 2023)[306].

Previously, two independent publications identified similar populations of HSC with a heightened inflammatory response. Jakobsen et al., identified an analogous group of HSCs within the bone marrow of *DNMT3A*- and *TET2*- mutant CH patients [45]. Secondly, Zeng et al., identified a population of “inflammatory memory HSCs” in mice after prior exposure to an inflammatory stimulus (treatment with LPS) [306]. The authors also identified a similar population of HSCs sharing the same transcriptional programme in human patients that had recovered from COVID-19 infection. Highlighting the *ASXL1*-CH data with AUCell enrichment score for signatures from both publications clearly identifies the HSC2 population (Figure 5.9d,e). Since cells from both CH and non-CH samples composed this HSC2 population, this heightened transcriptional inflammatory response is not CH-specific. Differences in response to inflammation between mutant and wildtype cells within the CH sample, however, may explain differences in clonal abundance within this population.

These results may suggest that the patients in which HSC2 are enriched may have been exposed to a prior significant inflammatory insult, regardless whether they had CH or not. When exposed to a second inflammatory stimulus such as an aged bone marrow milieu, this HSC2 population may have a very specific transcriptional response. For the patients who did have a CH clone, over time, the *ASXL1*-mutant HSC2 may outcompete WT HSC2 by entering a more quiescent, HSC-like state, possibly resulting in less stem cell exhaustion.

5.6.2 Differential gene expression analysis between *ASXL1*-mutant and WT-CH cells in HSC1

For HSC1 a total of 234 genes were significantly differentially expressed between mutant and wildtype cells ($\text{FDR} < 0.05$) (Figure 5.10b). Of these, 93 were upregulated in mutant HSC1 and 141 were downregulated (Table A.18). Similar to HSC2, among the top regulated genes were several involved in cytoskeletal reorganisation and cell-cell contact (*MAGI1*, *MARCKS*, *PXDN* and *CDH4*). Increased expression of these genes could be indicative of a more migratory, active HSC phenotype [308]. Further evidence for this comes from several top upregulated genes being involved in

the activation of Rho/RAC signalling pathway (*PLEKHG4B*, *AUTS2*, *FGD1*). Rac GTPases have previously been shown to play a role in a diverse range of functions in mammalian haematopoietic stem and progenitor cells including regulation of stem cell retention, mobilisation, survival, and proliferation [309–311]. Also, *CD9*, a putative cell surface marker of AML leukaemic stem cells [312], was one of the top upregulated genes in mutant HSC1. Interestingly, canonical HSC marker genes *MECOM* (MDS1 and EVI1 complex locus) and *HLF* (Hepatic leukaemia factor) are downregulated in mutant HSC1, suggesting a loss of *bona fide* HSC transcriptional phenotype [313, 314]. Additionally, marker genes of aged HSCs such as *ALDH1A1* and *SELP* were upregulated in WT HSC1 [315], suggesting that mutant HSC1 are transcriptionally more youthful than their WT counterparts.

Throughout this study, multiple analyses suggest that *ASXL1* mutation promotes the transcriptomic shift of HSCs to a less dormant state. To test this in HSC1, gene set enrichment analysis using HSC and progenitor gene signatures was employed (Table A.16). Strikingly, mutant HSC1 were significantly negatively enriched for multiple HSC and dormancy HSC gene signatures, and positively enriched for short-term HSC gene sets (Figure 5.11a). Simultaneously, mutant HSC1 were positively enriched for later stage progenitor gene signatures including those for pre-B cells, GMP, and erythroid progenitors. These results suggest that mutant HSC1 are shifted towards a more progenitor, differentiation-primed HSC-like transcriptomic state rather than a quiescent, less active HSC like state. This hypothesis was correlated with positive enrichment for cell migration and transmembrane ion channel signatures in mutant HSC1, indicative of a more active HSC transcriptional phenotype (Figure 5.11b).

Additionally, mutant HSC1 were significantly negatively enriched for Interferon alpha, beta and gamma signalling. Although acute interferon signalling is required for stem cell cycling, cell cycle exit and quiescence, chronic exposure to high concentrations of interferons alpha and gamma can have deleterious effects upon HSC self-renewal, repopulation capacity, as well as causing DNA damage, and apoptosis [216, 316, 317]. Dampened interferon signalling may also be driving clonal

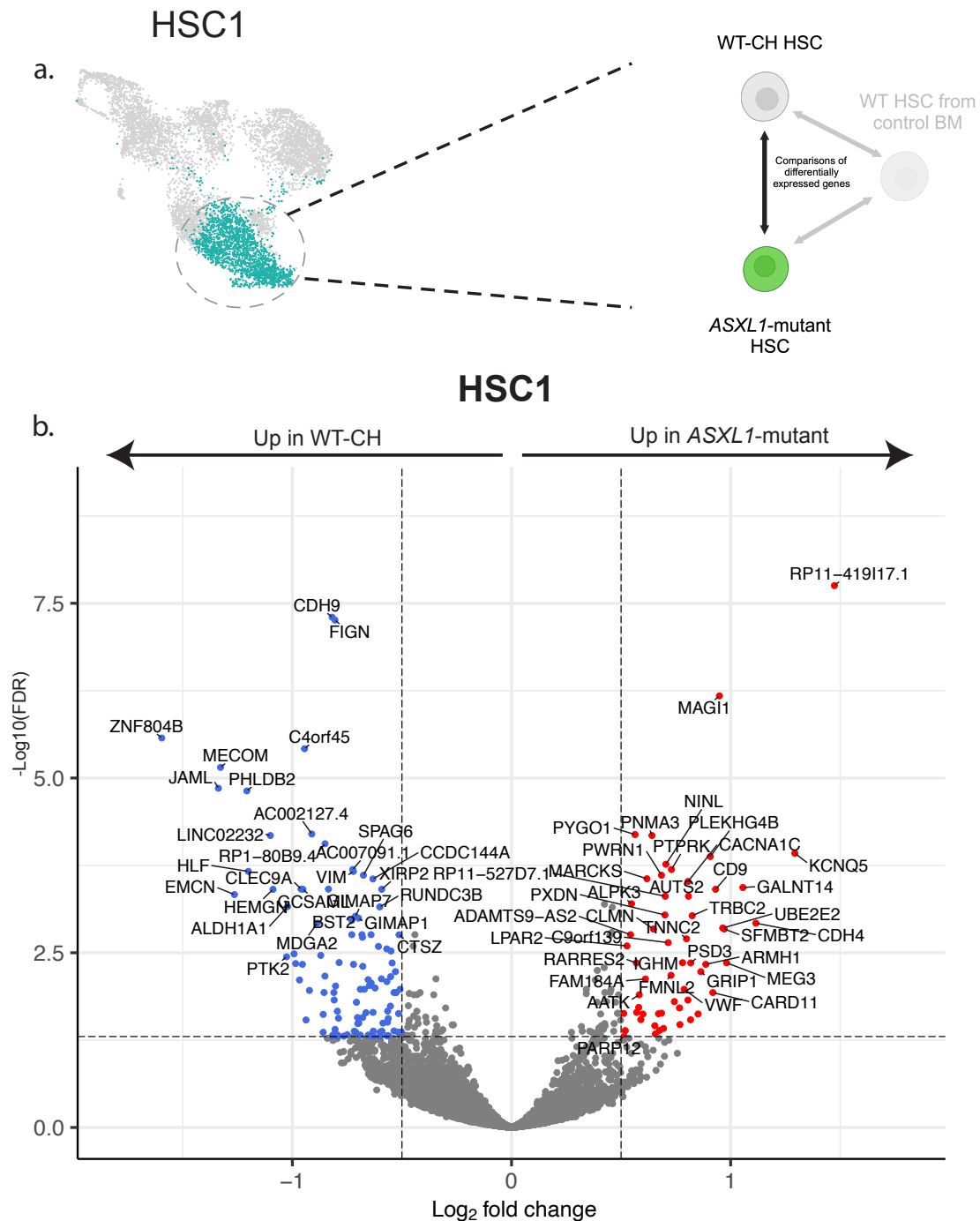


Figure 5.10: Differential gene expression analysis between *ASXL1*-mutant and WT-CH HSC1 subpopulation. **a)** Schematic showing differential gene expression analysis strategy between mutant and WT HSCs from *ASXL1*-mutant CH samples in both the HSC1 ($n = 833$). DGEA was performed accounting for sample and sex batch effects. **b)** Volcano plots visualising differentially expressed genes between mutant HSC1 and WT HSC1 from CH samples. DEGs were determined by a linear mixed-model approach comparing gene expression between single cells, whilst accounting for sample identifier and patient sex batch effects. Each point is a gene. $\log_2\text{FC}$ threshold was set as 0.5, and a significance threshold of $\text{FDR} < 0.05$ (B-H adjustment). Genes were coloured by significant upregulation in mutant or WT cells (red and blue respectively).

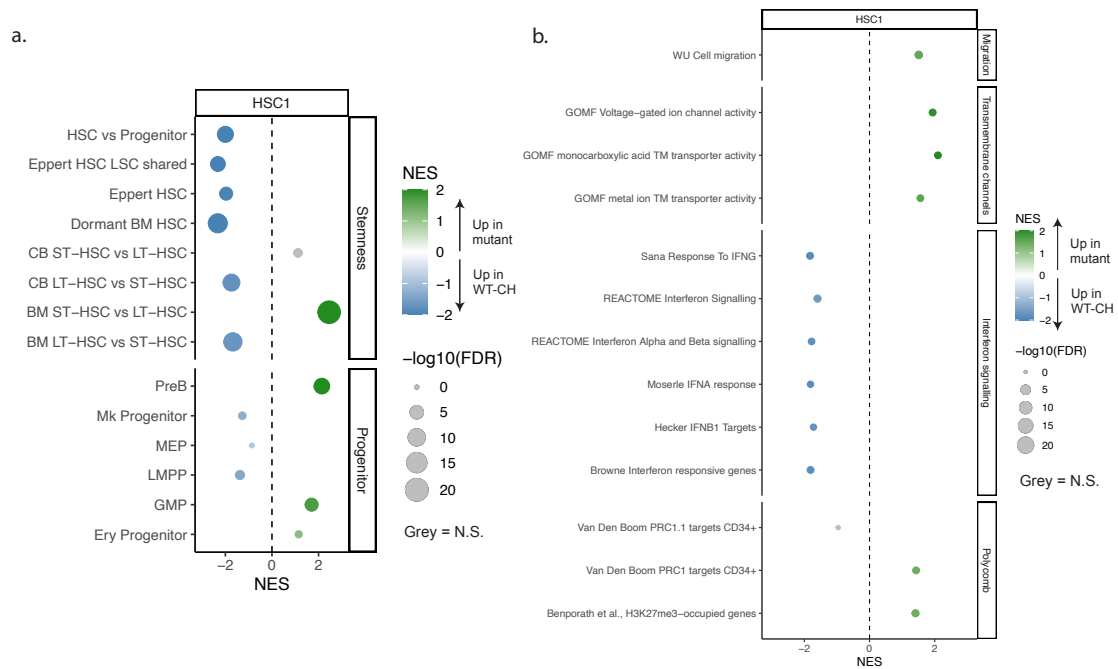


Figure 5.11: Gene set enrichment analysis between *ASXL1*-mutant and WT-CH HSC1 subpopulation. GSEA results for gene ontology and published study haematopoietic stemness (left), transmembrane ion channel activity and polycomb target gene signatures (right). Ranked lists of DEGs from the dream linear mixed model approaches were used as input. Size of dots represent significance ($-\log_{10}(\text{FDR})$). Dots are coloured by NES- green dots are enriched in mutant cells, blue dots are enriched in WT cells. Grey dots represent non-significant results ($\text{FDR} < 0.05$; B-H adjusted).

expansion of mutant HSC1. Differential expression of interferon response genes has occurred in almost all analyses performed thus far, so establishing if these differences are functionally relevant is perhaps challenging. In this context, these differences do appear to be functionally relevant as *ASXL1*-mutant HSC1 display an HSC response akin to acute rather than chronic interferon exposure. This is supported by the loss of quiescence and more active transcriptional phenotype in these cells. For WT HSC1 however, the opposite appears to be true.

Truncated *ASXL1* has previously been shown to enhance polycomb repressive de-ubiquitinase complex activity and increase expression of polycomb genes [144]. One hypothesis may be that mutant *ASXL1* confers clonal advantage through the increased expression of polycomb repressive complex 1 target genes. To test this hypothesis, GSEA was performed using gene sets of PRC1 target genes identified in human CD34+ cord blood cells [318] (Figure 5.11b). Additionally, since it is

thought that polycomb repressive complexes 1 and 2 catalyse the addition of histone post-translational modifications symbiotically [115], gene sets for PRC2 target genes were also used as surrogate [319]. Two of these gene signatures were positively enriched for in mutant HSC1, indicating that dysregulation of PRC1 target genes could be a potential molecular mechanism by which *ASXL1* confers clonal advantage. Analysis of histone mark distribution in *ASXL1*-mutant HSCs would be required to corroborate this hypothesis. Interestingly, *AUTS2*, a member of PRC1.5 complex [320], was one of the most upregulated genes in HSC1 (Figure 5.10b).

5.7 Gene regulatory network inference analysis of HSC1

Dysregulation of transcription factor programmes has the potential to influence HSC stemness and differentiation potential. Differentially active gene regulatory networks (regulons) were inferred from single-cell expression data using pySCENIC (single-cell regulatory network influence and clustering) [258]. In this analysis, one regulon's activity is the expression levels of the inferred targets of a transcription factor (TF). Candidate TF targets are defined through co-expression patterns with the TFs themselves. The list of transcription factor targets are then refined by eliminating those genes without a previously annotated motif for that specific TF at a defined distance from the transcription start site. When SCENIC regulon inference was applied to HSC1, 32 regulons were found to have significantly different activity between mutant and WT cells (Figure 5.12; 28 upregulated in mutant HSC1, 4 downregulated).

Amongst the upregulated regulons in the mutant cells were several involved in the control of differentiation into the lymphoid, myeloid, and megakaryocyte/erythroid lineages. For example, *SOX4*, the top differentially active regulon, has previously been identified as a key regulator of B-cell lymphoid specification [321]. Other upregulated regulons implicated in haematopoietic lineage specification include *RUNX1*, *IRF8*, *SPI1*, *E2F3* [322–325]. This is suggestive of transcriptional priming for differentiation into the haematopoietic lineages in mutant HSC1.

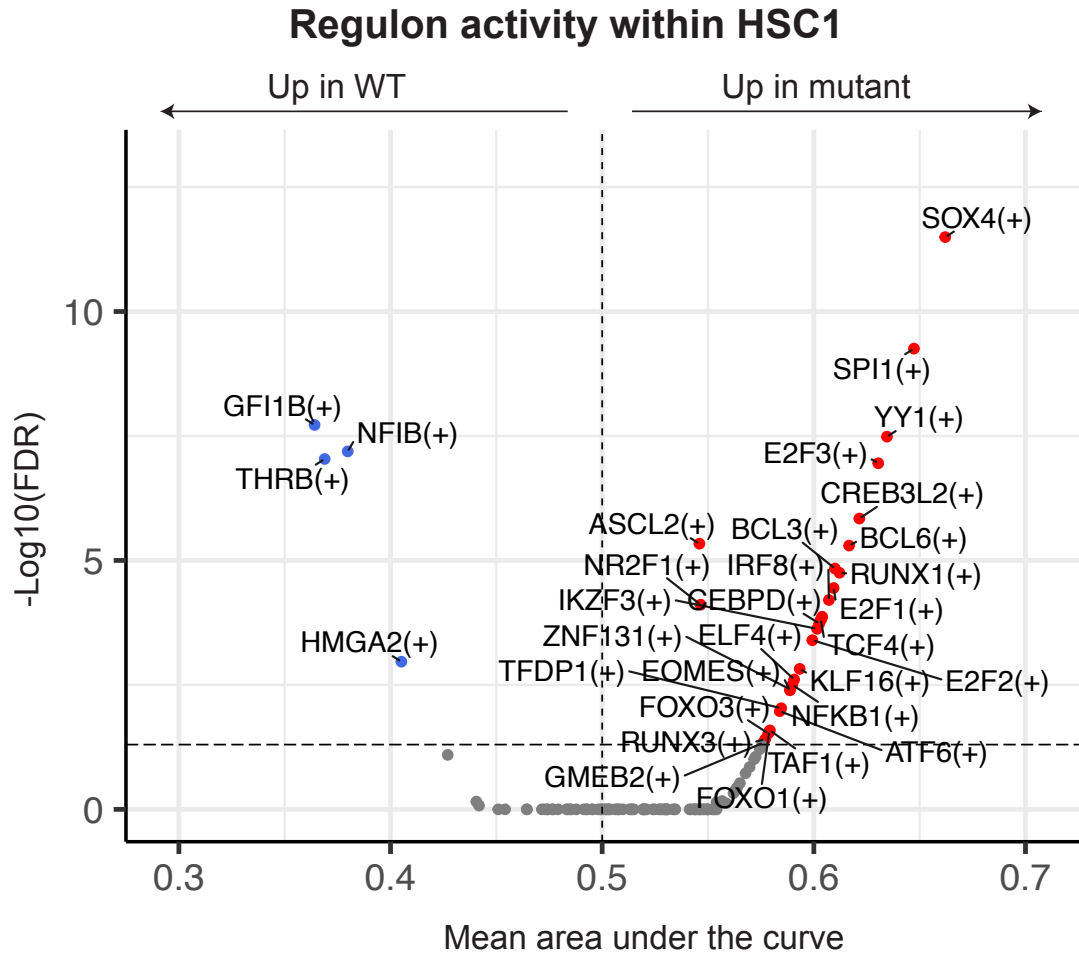


Figure 5.12: Differentially active gene regulatory network analysis between *ASXL1*-mutant and WT-CH in the HSC1 subpopulation. Gene regulatory network (regulon) activity inference between *ASXL1*-mutant and WT-CH in the HSC1 subpopulation. Regulon activity was inferred from single-cell transcriptome data using PySCENIC. Each dot represents a transcription factor regulon. X axis represents the mean area under the curve (AUC) enrichment score for all cells in each group. A significance threshold of FDR < 0.05 (B-H adjustment) was applied. Regulon were coloured by significant upregulation in mutant or WT cells (red and blue respectively).

Concordant with the GSEA results, three of the top downregulated regulons, *GFI1B*, *NFIB*, *HMGA2*, have all been implicated in the maintenance of a quiescent HSC-like state [208, 326, 327]. This is further evidence for a model of a more active, less quiescent HSC state. Finally, the *NFKB* regulon was upregulated in mutant HSC1- evidence of response to a pro-inflammatory microenvironment. Increased *NFκB* activity has also previously been associated with a loss of HSC quiescence [328].

5.8 Consequences of *ASXL1* mutation upon the transcriptome of HSPCs

Evidence from differential abundance analysis of the *ASXL1*-mutant clone throughout haematopoiesis suggests an inefficiency in terminal differentiation in the megakaryocyte/erythroid, and lymphoid lineages. Intriguingly, this did not appear to be the case for the myeloid lineage. A hypothesis may be that the *ASXL1* mutations have differing effects upon cells committed to differentiation between lineages. Additionally, it appears that *ASXL1* has very subtle transcriptional effects upon the HSC compartment. One hypothesis may be that *ASXL1* mutations have a greater effect upon the transcriptome of cells more committed to differentiation than stem cells. This is supported by the fact that *ASXL1* mutations are very common in malignancies and diseases of more mature haematopoietic populations, such as CMML and aplastic anaemia. To test these hypotheses and explore this further, differentially active transcription factor regulons were identified between mutant and WT cells from populations of progenitor cells committed to myeloid, lymphoid, and erythroid differentiation.

5.8.1 Differential regulon analysis of haematopoietic progenitor compartments

Measurement of activity of transcription factors and their putative downstream targets (regulons) can provide insight into differentiation trajectories in specific cell populations. To identify differences between *ASXL1*-mutant and wildtype cells in transcription factor activity throughout differentiation, differential regulon analysis was performed on the transcriptomically-defined LMPP, Late GMP, Erythroid progenitor (EryP), and B cell progenitor (Pro-B) cell populations. These populations were chosen to provide a cross-section of differentiation throughout all three major haematopoietic lineages. Differential gene regulatory network analysis was performed as described in chapter 2, and in section 5.7. It is worth noting that none of the genes encoding the transcription factors mentioned in this section were found to

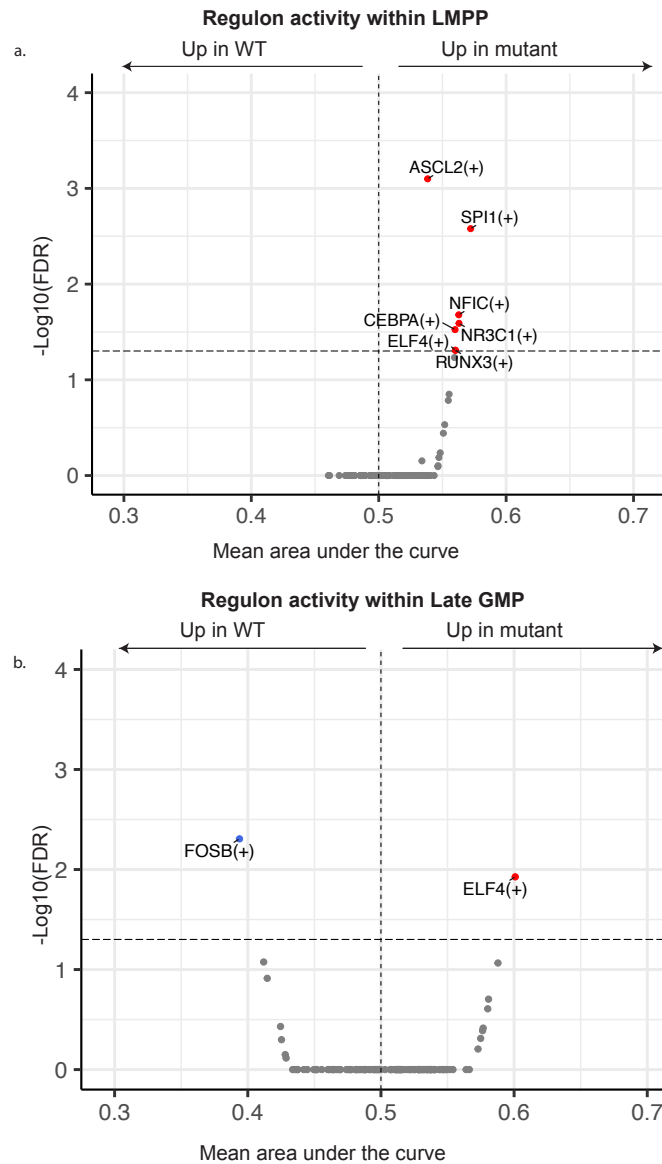


Figure 5.13: Differentially active gene regulatory network analysis between *ASXL1*-mutant and WT-CH in the LMPP and late GMP populations. Gene regulatory network (regulon) activity inference between *ASXL1*-mutant and WT-CH in the LMPP and late GMP populations. Regulon activity was inferred from single-cell transcriptome data using PySCENIC. Each dot represents a transcription factor regulon. X-axis represents the mean area under the curve (AUC) score for all cells in each group. A significance threshold of $\text{FDR} < 0.05$ (B-H adjustment) was applied. Regulon were coloured by significant upregulation in mutant or WT cells (red and blue respectively).

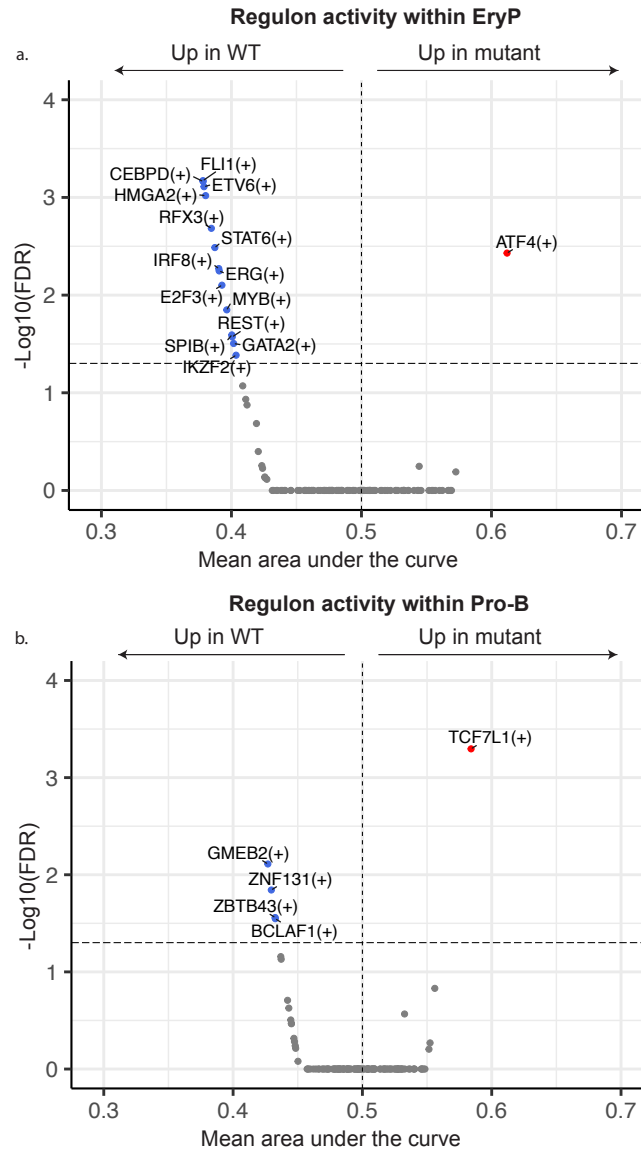


Figure 5.14: Differentially active gene regulatory network analysis between *ASXL1*-mutant and WT-CH in the EryP and Pro-B populations. Gene regulatory network (regulon) activity inference between *ASXL1*-mutant and WT-CH in the EryP and Pro-B populations. Regulon activity was inferred from single-cell transcriptome data using PySCENIC. Each dot represents a transcription factor regulon. X-axis represents the mean area under the curve (AUC) score for all cells in each group. A significance threshold of $FDR < 0.05$ (B-H adjustment) was applied. Regulon were coloured by significant upregulation in mutant or WT cells (red and blue respectively).

be significantly differentially expressed between mutant and WT cells for each cell type (data not shown). This was determined using the dream implementation of the limma-voom workflow, as described earlier, with an adjusted p-value < 0.05 .

Within the LMPP compartment, *ASXL1*-mutant cells showed an enrichment

of regulons associated with myeloid differentiation including *CEBPA*, *SPI1*, *ELF4*, *NFIC* [329–332](Figure 5.13a). This may suggest preferential priming towards myeloid differentiation in mutant LMPPs. Preferential differentiation of *ASXL1*-mutant HSPCs towards the myeloid lineage has been observed in murine *in vivo* and human *in vitro* models, supporting this observation [144, 197]. Of particular interest is the *NFIC* regulon- the *NFIC* gene has been found to be upregulated in blasts isolated from primary human AML samples [332]. In the same study, overexpression of *NFIC* *in vitro* has also been shown to promote monocytic differentiation of CD34+ cells. Further towards terminal myeloid differentiation, in the late GMP compartment very little difference was seen in transcription factor activity (Figure 5.13b). Only two regulons were significantly (FDR < 0.05) differentially active: *FOSB* (enriched in WT cells), and *ELF4* (enriched in mutant cells). Both transcription factors are associated with terminal myeloid differentiation [331, 333]. The paucity of differential regulon activity in late-stage myeloid progenitors suggests that *ASXL1* mutations have little disruptive effect upon myeloid differentiation.

When inspecting differentially active regulons within the erythroid progenitors (EryP), 14 regulons were found to be enriched in WT cells, whereas only one was enriched in mutant cells (Figure 5.14a). Intriguingly, transcription factor regulons associated with erythropoiesis were enriched within WT cells, including *FLI1*, *GATA2*, *HMGA2*, and *MYB* [327, 334–336]. *ATF4* regulon activity, also associated with erythropoiesis [337, 338], was up in mutant EryP, suggesting that erythropoiesis is still regulated within the mutant cell fraction. However, it must also be noted that *ATF4* is also a stress response TF [339–341], suggesting that the upregulation of this regulon may not be specific to erythropoietic differentiation. Given the observation that mutant cells are less likely to be found within the erythroblast population, one might hypothesise that *ASXL1* mutation modestly impairs, but does not arrest terminal erythroid differentiation.

Interestingly, a similar pattern was also observed when regulon activities between mutant and WT Pro-B cells were compared (Figure 5.14b). Five regulons were found to be differentially active between mutant and WT cells. Of the four regulons

found to be more active within the WT Pro-B cell fraction, *ZNF131*, *ZBTB43*, and *BCLAF1* have all been associated with regulation of B-cell lymphopoiesis [342, 343]. *TCF7L1*- a transcription factor which is normally found expressed in human HSCs and progenitors, but not in Pro-B cells- was found to have upregulated activity in mutant Pro-B cells [344]. *TCF7L1* was expressed in both mutant and WT Pro-B cells, however it was not differentially expressed between the two conditions. Taken together, this might suggest that mutant Pro-B cells are less likely to differentiate to mature B cells. A hypothesis may be that *ASXL1* mutations cause an imbalance in transcription factor activity, leading to subtly impaired differentiation into mature cell types.

5.9 Concluding remarks

In this chapter, differential gene expression analysis, gene set enrichment analysis and gene regulatory network inference were used to explore the transcriptomic correlates of *ASXL1* mutations with clonal advantage. At the transcriptomic level, HSC/MPPs from the CH bone marrow responded to inflammatory stimuli regardless of mutational status. Within the CH marrow, differences in transcriptomic profile between mutant and wildtype HSC/MPPs were found to be subtle. To dissect these differences, transcriptomic HSC/MPPs from both CH and healthy samples were subclustered and two major HSC populations were identified. First, a population of HSCs characterised by heightened response to inflammatory stimuli. Mutant HSCs from only one *ASXL1*-CH sample were significantly expanded in this population. A second major population of HSCs accounted for the majority of mutant HSCs from all four CH samples. Mutant HSCs from this population were transcriptionally characterised by an increase in cytoskeletal rearrangement, cell migration, a transition towards an active, less quiescent HSC state, and priming of differentiation programmes to progenitor cell states. This also correlated with an increased expression of polycomb target genes.

Lastly, differential regulon activity analysis was used to explore differences in transcription factor activity throughout differentiation. This analysis revealed

myeloid priming of mutant LMPPs, but no clear genotype-specific differences in myeloid differentiation-associated transcription factor activity of late-stage myeloid progenitors. Additionally, genotype-specific imbalances of erythroid and lymphoid differentiation-associated gene regulatory network activity in EryP and Pro-B populations correlates with observed inefficiencies in terminal maturation of mutant cells. Future work would identify putative molecular mechanisms that underpin this observed difference in transcription factor activity.

Chapter 6

GCUT: a novel method for simultaneous epigenomic profiling and targeted genotyping of single cells

Contents

6.1	Rationale for the development of GCUT	147
6.2	Review of CUT&Tag chemistry	150
6.3	Adaptation of CUT&Tag to a plate-based single-cell protocol	152
6.4	Development and optimisation of GCUT	156
6.4.1	Aims for benchmarking experiments	156
6.4.2	Modification of CUT&Tag for use in single-cell methods	156
6.4.3	Bulk CUT&Tag in human cell lines	157
6.4.4	Development of plate-based single-cell CUT&Tag with genotyping	159
6.5	Benchmarking of GCUT with a cell mixing sequencing experiment	161
6.5.1	Overview of cell mixing experiment	161
6.5.2	CUT&Tag library preparation of cell lines	162
6.5.3	Genotyping of cell lines	164
6.5.4	Chromatin data pre-processing and QC	168
6.5.5	Dimensionality reduction and clustering	170
6.5.6	CUT&Tag library quality with and without genotyping	175
6.5.7	Integration of genotyping and chromatin data	178
6.5.8	Genomic distribution of histone marks of from GCUT CUT&Tag data	182
6.6	Concluding remarks	184

I would like to thank David Cruz Hernandez for his expertise in CUT&Tag and his help in optimisation of the GCUT protocol. I am also grateful to Sven Turkalj for his help in processing single-cell genotyping plates.

6.1 Rationale for the development of GCUT

So far in previous chapters, the transcriptomic differences between *ASXL1*-mutant and wildtype cells from different cell populations have been explored. These results can provide valuable insights into the molecular mechanisms linking somatic mutation to clonal expansion. However, from this analysis it is impossible to determine which of these transcriptomic differences are likely to be direct consequences of the mutation. Dissecting the direct and indirect consequences will clarify which molecular mechanisms are most likely driving clonal expansion, which can then be functionally validated. This may also identify putative therapeutic targets to prevent the growth of clones with malignant potential.

One of the most well-described functions of *ASXL1* has been as an epigenetic regulator. It has been reported that the function of *ASXL1* is to regulate the genomic distribution of post-translational histone modification H2AK119ub, which is associated with transcriptional expression as part of the polycomb-repressive deubiquitinase complex (PR-DUB complex). The PR-DUB complex removes the transcriptionally repressive H2AK119ub from polycomb target gene promoters and gene bodies, increasing the likelihood of polycomb target gene expression. Mutations in *ASXL1* have resulted in the expression of a C-terminally truncated isoform of the protein [198]. Expression of truncated ASXL1 in both human cell lines and primary human CD34+ cord blood cells *in vitro* has resulted in enhanced PR-DUB deubiquitinase activity, subsequent reduction of H2AK119ub, and a growth advantage in serial replating assays [144, 197, 198]. Identifying the genomic loci at which H2AK119ub is lost in *ASXL1*-mutant cells and correlating these with increased gene expression may provide the mechanistic link between *ASXL1* mutation and clonal expansion. Therefore, a logical first step is to understand the differences in the distribution of the H2AK119ub between mutant and WT cells.

Which genomic loci are targeted by *ASXL1*-mutant PR-DUB complex? Researchers have previously attempted to answer this question using bulk assays to

profile the epigenome of genetically modified human and murine cells. However, these approaches fail to account for the vast heterogeneity observed in primary human tissue. For example, upregulated genes in *ASXL1*-mutant cells when compared to wildtype cells are not consistent across CD34+ HSPC compartments. An approach which accounts for this heterogeneity is required to accurately determine differences in H2AK119ub distribution.

An important comparison to determine consequences of *ASXL1* mutation would be between mutant and wildtype cells within the same sample. Therefore, any method to perform this comparison would require separation of mutant and wildtype cells. A challenge of using bulk epigenomic assays on primary human CH bone marrow, is the absence of a biomarker for mutant cells apart from the mutation itself. Therefore, pure populations of mutant cells cannot currently be sorted. To overcome these challenges, an ideal method would simultaneously profile histone mark distribution and genotype of single cells.

A method that enables the use of low cell numbers would be the most applicable for use in assaying *ASXL1*-mutant CH for two reasons. First, the mutation confers a clonal expansion upon the CD34+ HSPC compartment which comprises only $\approx 5\%$ of all bone marrow mononuclear cells. Secondly, the VAF *ASXL1*-mutations in CH are generally low (<0.1), so mutant cells within the CD34+ population are rare. Traditional methods assaying genomic histone mark distribution such as ChIP-seq (Chromatin Immunoprecipitation sequencing) and associated variations require large numbers of cells, and are prone to high background noise, low signal, and epitope masking [345, 346]. In recent years, updated methods such as CUT&RUN (Cleavage Under Targets and Release Under Nuclease) and the associated CUT&Tag (Cleavage Under Targets and Tagmentation) have addressed these problems, generating high-quality base-pair resolution genomic histone mark profiles from low input numbers of cells [346–348].

An important consideration is that highly specific antibodies are required for use in CUT&Tag to obtain a high-quality library from low starting cell numbers. Primary antibodies that work well in ChIP-seq may not work for use in CUT&Tag for this

reason, so candidate antibodies would have to be screened rigorously. Importantly however, CUT&Tag has also been adapted for use in single-cell protocols using both 10X droplet-partition and Takara ICell8 nanowell chip platforms [242, 346, 349].

Several methods already exist uniting single-cell epigenomic profiling with genotyping, which can be informative in designing a new method. Of particular interest are those that incorporate the Assay for Transposase Accessible Chromatin (ATAC-seq), because the chemistry of fragment amplification is almost identical to CUT&Tag, due to the use of Tn5 transposase. GTAC, which was based upon TARGET-Seq and developed in the Vyas lab, is a plate-based method which profiles chromatin accessibility in single cells whilst simultaneously genotyping genomic loci on gDNA [260]. Briefly, ATAC-seq is performed on a bulk population of cells (that can be pre-enriched for specific cellular populations using FACS), before single nuclear sorting into plates. After pre-amplification of ATAC-seq fragments and enrichment of gDNA fragments using targeted primers, the library is split to generate separate genotyping and ATAC-seq libraries corresponding to the same single cells. Crucially, when genotyping 5 separate loci, it can achieve successful genotyping rates of between 82-98% with allelic dropout rates of up to 11.9%.

Genotyping of targets with single-cell chromatin accessibility (GoTChA), developed in the Landau lab, is a droplet-based method also combining single-cell ATAC-seq with targeted capture of mutations from gDNA [350]. Bulk cell populations are transposed with Tn5 tagmentase, prior to partitioning of single nuclei in oil droplets with barcoded beads and custom primers targeting genomic loci of interest. These custom primers also adapt the gDNA genotyping amplicons to bind to the barcoded gel beads. Similar to GTAC, after initial amplification, the PCR product is divided to produce separate genotyping and ATAC libraries. Droplet-based methods are much higher throughput than plate-based techniques, however the genotyping success rate of GoTChA is significantly worse than GTAC. In a cell-mixing experiment testing heterozygous genotyping of the *JAK2*V617 hotspot mutation, genotyping reads covering the locus could only be recovered for between 30.7-78.8%. In the

SET2 cell line, which is biallelic for *JAK2*V617, only 34% of cells were correctly identified as heterozygous, indicating allelic dropout had occurred in 66% of cells.

Requirements for designing a method to assay histone mark distribution in single cells with simultaneous targeted genotyping would include:

1. Low starting cell numbers to assay rare cellular populations.
2. High cell-throughput to maximise use of rare populations and give power to statistical analyses between mutant and wildtype cells.
3. High-sensitivity, base-pair resolution genomic histone modification profiles of single cells.
4. High-accuracy capture of genomic loci with efficient, high-confidence targeted genotyping.

To be able to satisfy three of these requirements, I designed GCUT: Genotyping with Cleavage Under Targets and tagmentation. This is a novel plate-based method that combines histone profiling with high-fidelity genotyping of single cells.

6.2 Review of CUT&Tag chemistry

The principle of Cleavage Under Targets and Tagmentation (CUT&Tag), originally developed by Steven Henikoff's group, is to employ an enzyme-tethering strategy to efficiently cleave gDNA at loci of an epitope of interest in bulk cell populations [346]. CUT&Tag is based on CUT&RUN, which uses antibody-directed micrococcal nuclease (MNase) delivery to cleave gDNA at targeted epitope loci [348]. In CUT&Tag, protein A-conjugated Tn5 transposase is used to cleave gDNA at target loci instead of MNase. CUT&Tag makes use of adapter sequences inserted by Tn5 to simplify downstream library preparation, improving suitability for adaptation to single-cell applications. Successful CUT&Tag libraries can be prepared from as few as 1,000 cells.

The steps of CUT&Tag are outlined in Figure 6.1. Cells are first washed and lysed, leaving the nuclei intact. Nuclei are gently permeabilised so that the integrity

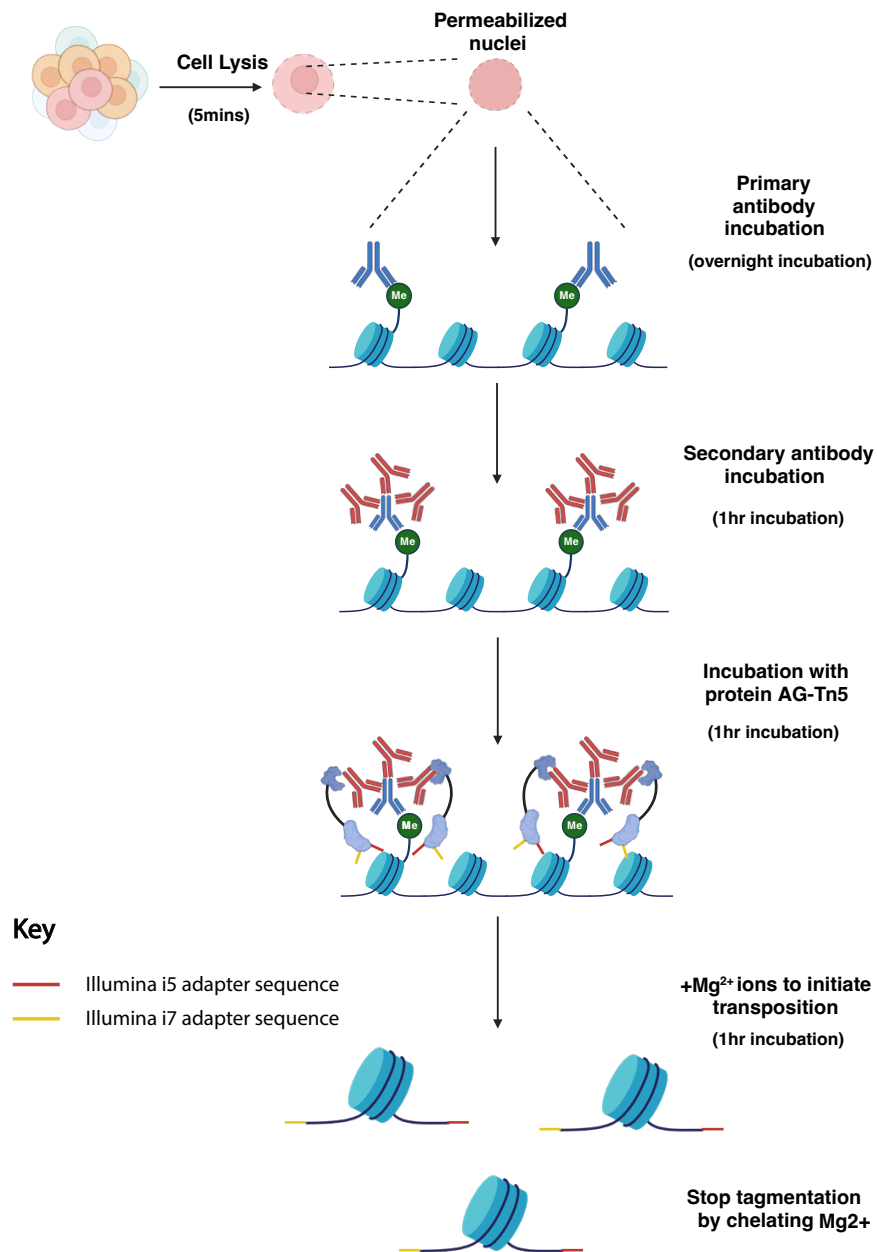


Figure 6.1: Overview of the CUT&Tag protocol. CUT&Tag protocol as first described in Kaya-Okur et al., 2029 [346].

of the nuclear membrane is not fully compromised. Next, nuclei are incubated with primary antibodies against the epitope of interest. Histone modifications are preferred as epitopes of interest as they are covalently bonded to histones, however transcription factors may also be targeted (despite the transient nature of interaction). A subsequent antibody incubation with anti-IgG secondary antibodies increases binding sites for directed enzyme delivery to the loci of interest. Protein AG-tethered Tn5 transposase fusion proteins (pAG-Tn5) were designed so that pAG binds to IgG polypeptides of the secondary antibodies. Thus, Tn5 transposase is directed to the epitope of interest. To maximise the number of enzymes bound at the target loci, nuclei are initially incubated with pAG-Tn5 without activating Mg^{2+} ions. Incubation with Mg^{2+} ions at 37°C activates the Tn5, allowing the simultaneous cleavage of gDNA and transposition of i5 and i7 sequences (loaded into the Tn5). The transposition reaction is terminated through chelation of Mg^{2+} ions with EDTA.

Following termination of the transposition reaction, Tn5 is released from the genome and DNA is purified either by AMPure bead cleanup, or precipitation column. PCR with i5 and i7 primers targeting the transposed i5 and i7 sequences amplifies the fragments of gDNA spanning the epitope of interest. The library can then be purified and sequenced.

6.3 Adaptation of CUT&Tag to a plate-based single-cell protocol

I chose to adapt the bulk CUT&Tag protocol into a single-cell application that can be integrated with a genotyping strategy. This was due to the simplicity of CUT&Tag library preparation, and the efficiency and accuracy of epigenomic profiling. Since plate-based targeted single-cell genotyping methods such as TARGET-seq and GTAC produced much higher quality genotyping data than alternative droplet-based approaches, I chose to pursue a plate-based strategy for GCUT. The ultimate motivation of developing GCUT is to determine differences in histone mark distribution between mutant and WT cells from primary human tissue. Evidence from both my TARGET-seq+ data and previous publications suggests that these

differences will be subtle, so I decided to prioritise high-confidence genotyping for GCUT over developing a high-throughput method.

All the steps in the following paragraphs are summarised in Figure 6.2.

To adapt CUT&Tag into a plate-based single-cell method, there were two options. First, individual cells could be FACS-sorted into plates, and the CUT&Tag protocol could be performed on single cells simultaneously in each well. This strategy presents severe challenges in terms of time required to process a meaningful number of cells, and also cost- the amount of reagents required for this protocol would be unfeasible. Alternatively, CUT&Tag could be performed on a bulk population of cells, and individual nuclei containing the CUT&Tag fragments could be sorted into plates after the transposition reaction. For this, permeabilised nuclei could be stained with DAPI, and DAPI -positive nuclei would be isolated into each well of a PCR plate containing lysis buffer. This bulk CUT&Tag strategy was chosen in other single-cell CUT&Tag methods [242, 349]. Downstream amplification of the CUT&Tag libraries from single-cells could be done in individual wells, using cell-specific combinations of barcoded primers.

For targeted single-cell genotyping, I employed the same strategy used in TARGET-seq and GTAC. Wells containing single nuclei that had undergone CUT&Tag still theoretically contained the whole genome for each cell, so genomic loci of interest could be enriched for using targeted primers. Initially, spiked-in genotyping primers can enrich for loci of interest in the same pre-amplification PCR reaction that amplifies the CUT&Tag library. There is a possibility that genomic loci of interest are within the CUT&Tag fragments. Given that epitopes bound to nucleosomes are targeted, these fragments are likely to be the length of a single nucleosome (147bp). To account for this, genotyping amplicons were designed no longer than 150bp.

Following the preamplification reaction, the CUT&Tag and genotyping libraries are split. As per the TARGET-seq and GTAC genotyping protocols, an aliquot of 20% the PCR product for each cell was taken and used as a template in subsequent genotyping PCR reactions. Primers nested within the original genotyping amplicons

were designed to further enrich for the loci of interest and add plate-identifier barcodes during the PCR1 reaction. Cell-identifier barcodes and sequencing flow cell adapter sequences are then added to each amplicon in PCR2. Both genotyping and CUT&Tag libraries can then be purified and sequenced separately. Full details of the protocol are described in Chapter 2.

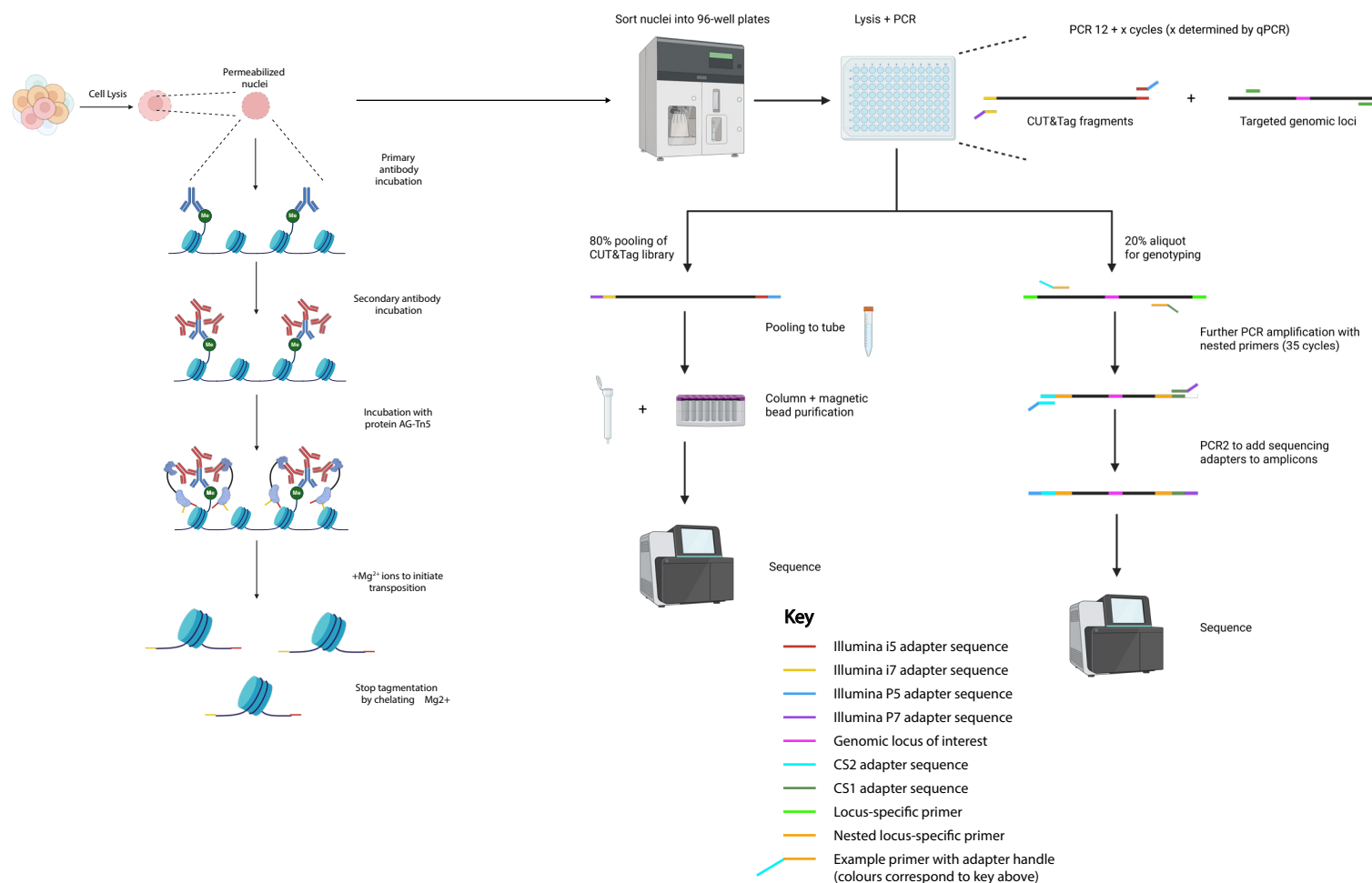


Figure 6.2: Schematic of GCUT. Overview of the GCUT protocol. Full details of experimental protocol are listed in Chapter 2. Primer and sequence legend is in the middle of the figure.

6.4 Development and optimisation of GCUT

6.4.1 Aims for benchmarking experiments

During the development of GCUT, I designed a series of experiments to benchmark each step of the protocol. The aims of these experiments were:

1. To produce a CUT&Tag library from a human cell line using pre-selected primary antibodies.
2. To produce CUT&Tag libraries from single cells of a human cell line.
3. To optimise single-cell CUT&Tag library preparation to produce high-quality libraries.
4. To generate multiple genotyping amplicons from single cells that have undergone CUT&Tag.

6.4.2 Modification of CUT&Tag for use in single-cell methods

One of the biggest foreseen challenges developing a plate-based method compared to a droplet-based method is the ability to assay sufficient cell numbers to produce meaningful results. FACS sorting, which is used to partition cells into individual wells of PCR plates, is inherently inefficient. Depending on the parameters used, based on experience roughly 20-30% of cells processed by a FACS sorter are recovered during single-cell sorting. Therefore, it is critically important to maximise the number of cells prior to sorting. To account for this, I modified the CUT&Tag protocol to minimise the cell loss during this process.

First, centrifugation steps are a primary source of great cell loss. In the original protocol, there is a total of 10 separate centrifugation steps- mostly prior to and following washing steps, which are included to ensure specificity of reagent binding. The cell recovery from a pilot experiment using this protocol was 10% (data not shown). By reducing the number of centrifugation steps to 7, keeping

the wash steps that are critical to ensuring specificity of antibody or enzyme binding, cell recovery was increased to 20% (data not shown). This still represents a significant loss of cells, and further optimisations will be required to improve cell recovery. Secondly, an incubation step with concanavalin A beads in the original CUT&Tag protocol was removed, as this can interfere with the single nuclei sorting step during GCUT. Finally, buffers were modified to include 30% BSA in every buffer (to prevent nuclei from clumping), and to improve nuclear permeabilization (addition of 10% NP-40 to the antibody and Tn5 incubation buffers). Buffer compositions are listed in Table A.7.

6.4.3 Bulk CUT&Tag in human cell lines

To test whether CUT&Tag libraries can be generated using a modified protocol for use in single-cell applications, I designed a bulk CUT&Tag experiment (Fig 6.3a). For these experiments, I used the human cell line K562 as it is triploid [283]. A 1.5-fold greater gDNA content than a diploid cell line would theoretically provide more binding sites for the primary antibody, lowering the sensitivity of the assay in the benchmarking experiments. Briefly, 100,000 K562 cells per condition underwent the modified CUT&Tag protocol, followed by PCR enrichment of the transposed fragments. The optimal number of PCR cycles per condition was assessed by qPCR. Purified libraries underwent fragment size analysis and were quantified. For this experiment, two test conditions of anti-H2AK119ub and anti-H3K27me3 primary antibodies were used. The anti-H3K27me3 primary antibodies also acted as a positive control, as they had been previously verified in a published study [346]. Anti-IgG primary antibodies were used as a negative control condition as there should be no binding of any epitope on the genome.

For both test conditions, a nucleosomal fragment distribution pattern was observed with alternate peaks roughly every 147bp (Figure 6.3b,c). This is consistent with the notion that Tn5 transposase is directed to epitopes bound to nucleosomes and cleaves either side of the nucleosomal complex. In contrast, no library was seen in the negative control condition (Figure 6.3d). Whilst H2AK119ub yielded a

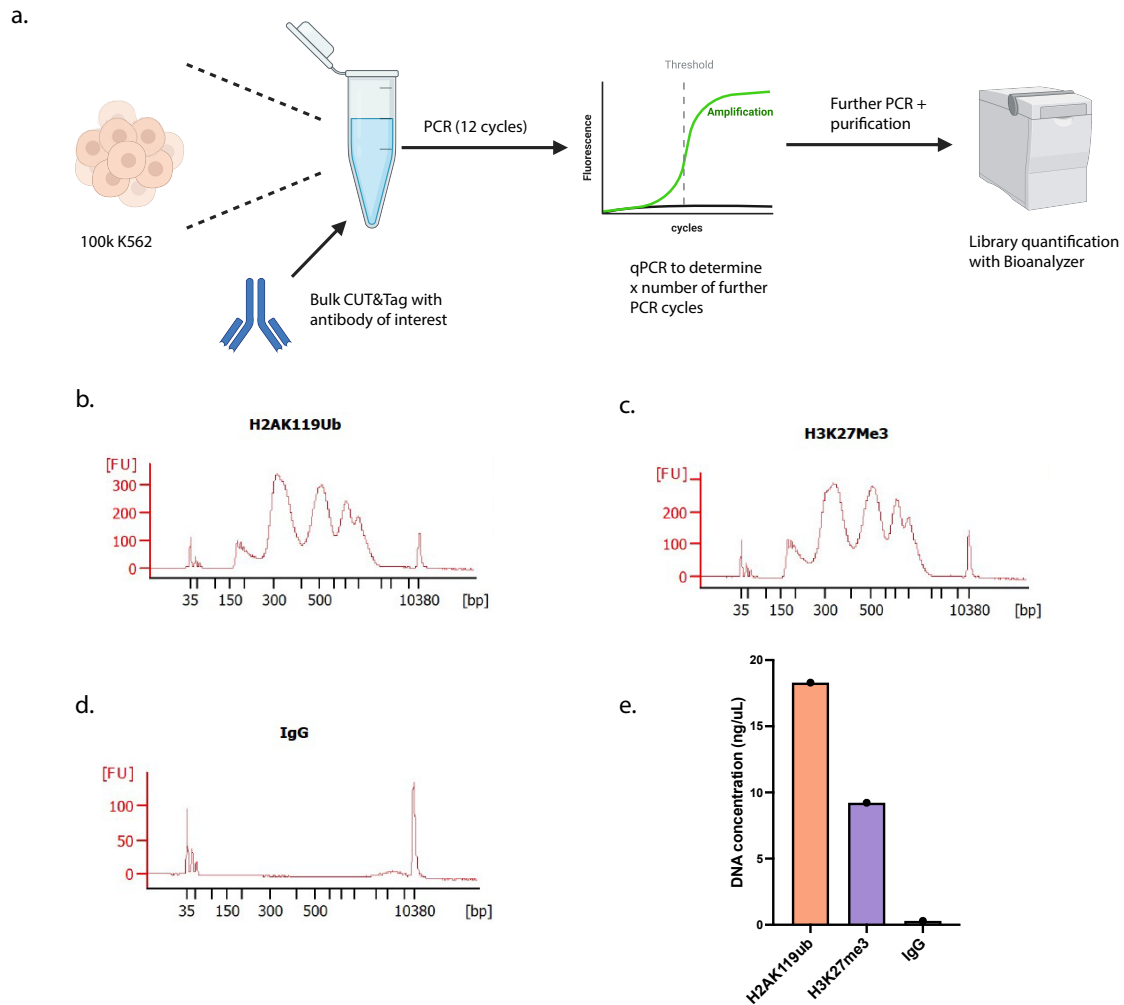


Figure 6.3: Bulk CUT&Tag testing in human cell lines. a) Experimental overview for testing primary antibodies using bulk CUT&Tag. b)-d) Bioanalyzer fragment distribution traces from 100k K562 cells per condition that have undergone bulk CUT&Tag with primary antibodies targeting H2AK119ub (b), H3K27me3 (positive control, c), or IgG (negative control, d). e) DNA concentrations of purified CUT&Tag libraries (ng/μL).

higher concentration of DNA per microliter than the H3K27me3 condition (Figure 6.3e), 5 extra cycles were required to amplify the DNA in this library (16 cycles for H2AK119ub, 11 cycles for H3K27me3), indicative of possible differences in binding efficiency of primary antibodies to their respective epitopes. Alternatively, different amounts of cell loss could contribute to the amount of DNA produced.

6.4.4 Development of plate-based single-cell CUT&Tag with genotyping

Next, I decided to test whether this bulk protocol could be used to generate CUT&Tag libraries from single cells. I designed an experiment for this that simultaneously tested the single-cell genotyping strategy for GCUT, outlined in Figure 6.4a. Briefly, four separate aliquots of 100,000 K562 cells underwent bulk CUT&Tag with either anti- H3K27me3, H2AK119ub, H4K4me3, or IgG primary antibodies. Single nuclei were sorted into 96-well plates and underwent GCUT preamplification with spike-in targeted genotyping primers for mutant loci in K562 (*ASXL1*, *HOXA9*, *CBL*, and *NOTCH1*). An extra plate of H3K27me3 nuclei were sorted and underwent preamplification without genotyping primers. An additional genotyping PCR reaction was performed on aliquots of preamplification product from single cells, using nested genotyping primers for each amplicon and bulk gDNA or no template as control conditions. Pools of the remaining preamplification product from 96 cells were purified, quantified and analysed.

A CUT&Tag library displaying the same nucleosomal fragment as that seen in the bulk experiment was seen in the H3K27 me3 scCUT&Tag condition (without genotyping primers), indicating that CUT&Tag libraries can be generated from single cells (Figure 6.4c). The same result was also observed when genotyping amplicons are simultaneously amplified during preamplification (Figure 6.4d), yielding very similar concentrations of CUT&Tag library (Figure 6.4b). For H3K4me3 and H2AK119ub, nucleosomal fragment distributions were also observed, however the concentration of library yielded was significantly less (Figure 6.4e,f). This indicates either less genomic occupation of each respective histone mark than H3K27me3, or an inefficiency in the CUT&Tag protocol. As seen in the bulk experiment, the anti-IgG negative control condition did not generate a library (Figure 6.4g).

For the genotyping testing, products of the second genotyping reactions were separated by agarose gel electrophoresis. Four separate genotyping amplicons were amplified from gDNA of each single cell that had undergone CUT&Tag (Figure 6.4h). Taken together, these results confirm that CUT&Tag libraries

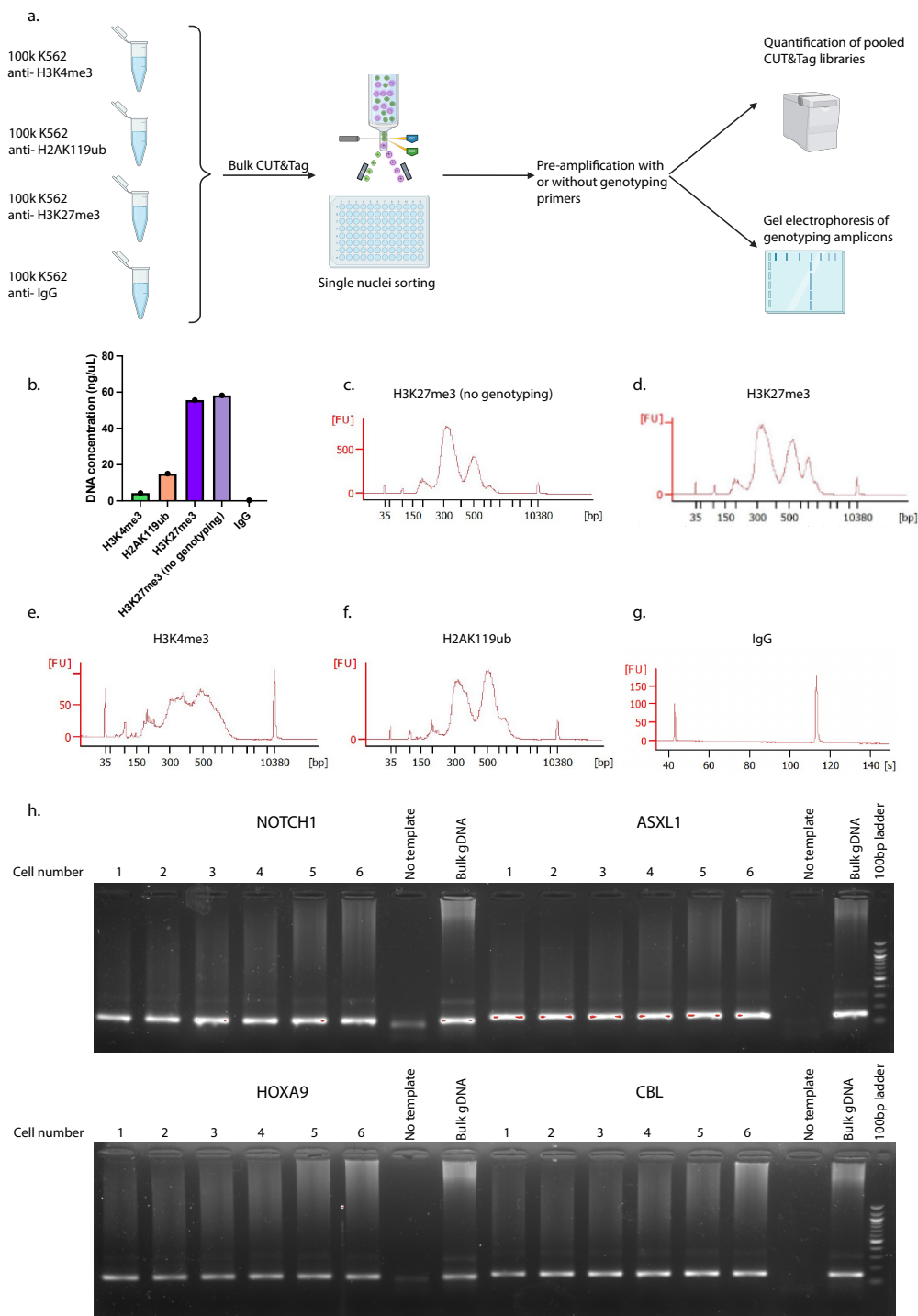


Figure 6.4: GCUT single-cell testing in human cell lines. **a)** Experimental overview for testing GCUT in single cells. Following pre-amplification, the product in each well is split to assess genotyping amplicon formation, and CUT&Tag library formation. **b)** DNA concentrations of purified CUT&Tag libraries pooled from 96 single K562 nuclei for each condition. (*continued next page*)

Figure 6.4 (previous page): c)-g) Bioanalyzer fragment size traces for purified CUT&Tag libraries from 96 single K562 nuclei per condition of primary antibody incubation. Positive and negative controls are shown in **c)** (anti-H3K27me3; no spike in genotyping primers) and **d)**(anti-IgG). Test conditions are shown in **e)-g)**. **h)** Gel electrophoresis bands for amplicons of four separate genomic loci amplified during PCR1 from single K562 nuclei. For bands with corresponding cell numbers, the starting gDNA template came from the same single nucleus. Positive control (bulk gDNA template) and negative control (no gDNA template) for each targeted locus is displayed on the right hand side of each condition.

can be generated from single cells, and that each of these single cells can be genotyped for multiple genomic loci.

6.5 Benchmarking of GCUT with a cell mixing sequencing experiment

6.5.1 Overview of cell mixing experiment

Having established that GCUT can produce CUT&Tag libraries from single cells, whilst simultaneously generating multiple targeted genotyping amplicons, a “killer” cell mixing sequencing experiment was designed to examine the quality of the CUT&Tag and genotyping libraries. The aims of this experiment were use GCUT to:

1. Distinguish three separate cell lines based on the genomic distribution of different histone marks in single cells
2. Distinguish three separate cells lines based on targeted single-cell genotyping from those same cells
3. Establish the quality of single-cell CUT&Tag libraries with and without simultaneous genotyping

Overview of the cell mixing sequencing experiment is shown in Figure 7.5. The cell lines K562 (CML), OCI-AML3 (AML), and JURKAT (T cell leukaemia) are derived from different haematopoietic lineages [283, 351, 352]. Therefore, their gene expression profiles will differ. A major assumption of this experiment is that their epigenomic profiles, determining these differences in gene expression, would be

sufficiently distinguishable. Evidence to support this assumption this comes from another cell mixing experiment for GTAC, which I was involved in designing and performing [260]. In this experiment, K562, OCI-AML3, and JURKAT cell lines could be distinguished based on their chromatin accessibility profiles. In the GCUT cell mixing experiment, three separate conditions for each of these cell lines (anti-H3K27me3, anti-H2AK119ub, or anti-H3K4me3 primary antibodies) underwent the modified CUT&Tag protocol, before single nuclei sorting (Figure 6.5).

To address the first two aims, two 384-well plates per histone mark were sorted: 128 nuclei from each cell line were sorted per plate. Two empty wells were left per plate to act as no template controls for genotype calling. The ground truth identity for each cell line was recorded at time of sorting. This acted as a positive control assessing quality of genotyping and cluster formation. All plates then underwent downstream GCUT library preparation steps as described in Chapter 2, including single-cell genotyping. All nuclei were genotyped for 5 genomic loci: *ASXL1* p.Y591, *HOXA9* p.D192N (both mutant in K562), *DNMT3A* p.R882C, *NRAS* p.Q61 (mutant in OCI-AML3), *RUNX1* p.A149 (mutant in JURKAT). Purified genotyping and CUT&Tag libraries were then sequenced and analysed.

A positive result for this experiment would be the identification of three separate clusters formed on the basis of the distribution of each histone mark. These would correspond to each cell line. By highlighting each cell by genotype, the cell line identity of each cluster could be confirmed.

An additional plate of anti-H3K27me3 nuclei from all 3 cell lines was also sorted, but this plate would not be genotyped. These cells would act as a control to establish whether genotyping affects the quality of single-cell CUT&Tag library.

6.5.2 CUT&Tag library preparation of cell lines

Nucleosomal banding patterns were detected for all CUT&Tag libraries from pooled plates of 382 cells (Figure 6.6 a-g). The same shape of fragment distribution for each histone mark was observed as in the single-cell testing experiments. Interestingly, the H3K4me3 libraries displayed skewing to larger fragments. This is surprising as

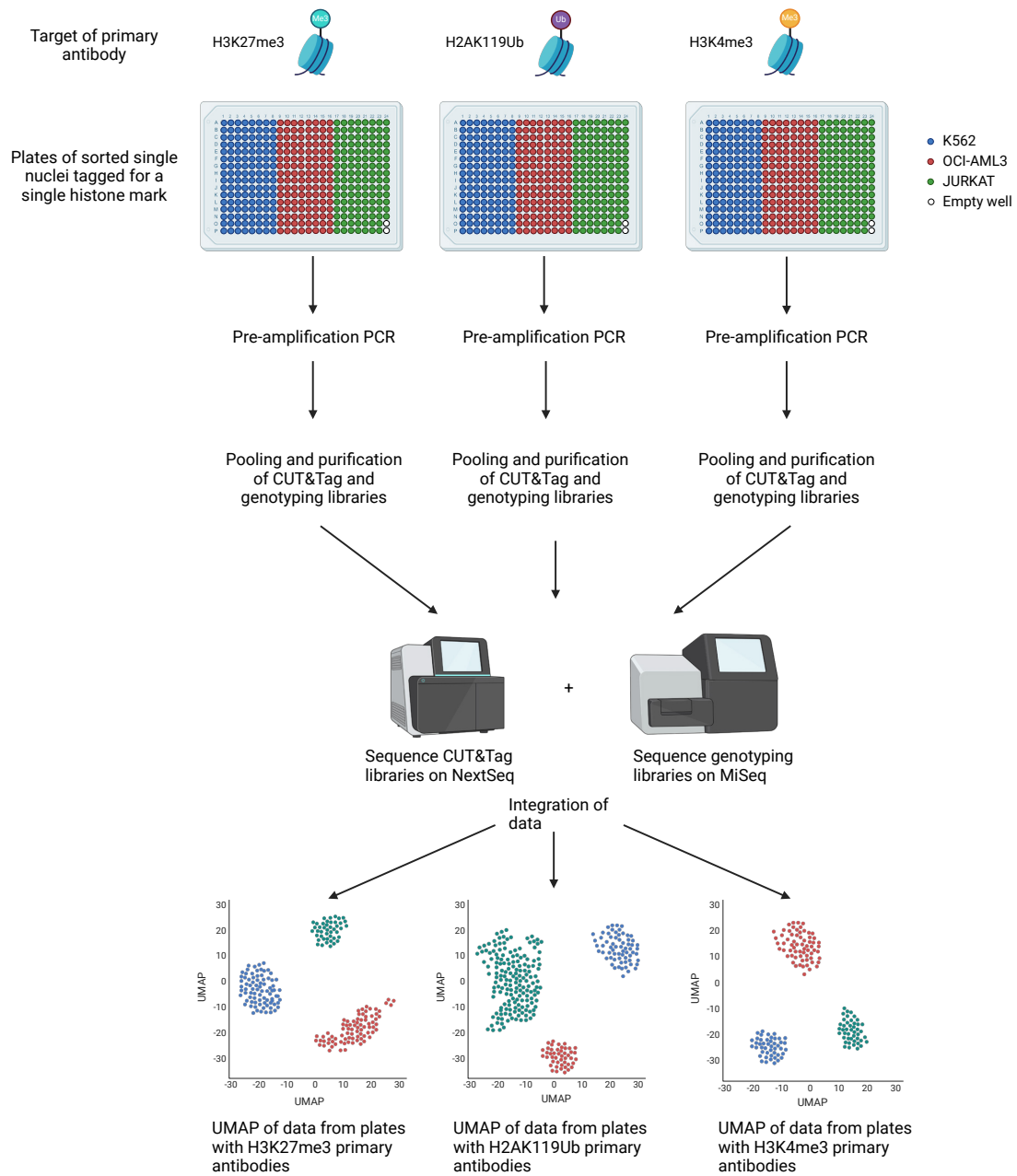


Figure 6.5: GCUT cell mixing experiment overview. Experimental overview for GCUT cell mixing experiment..

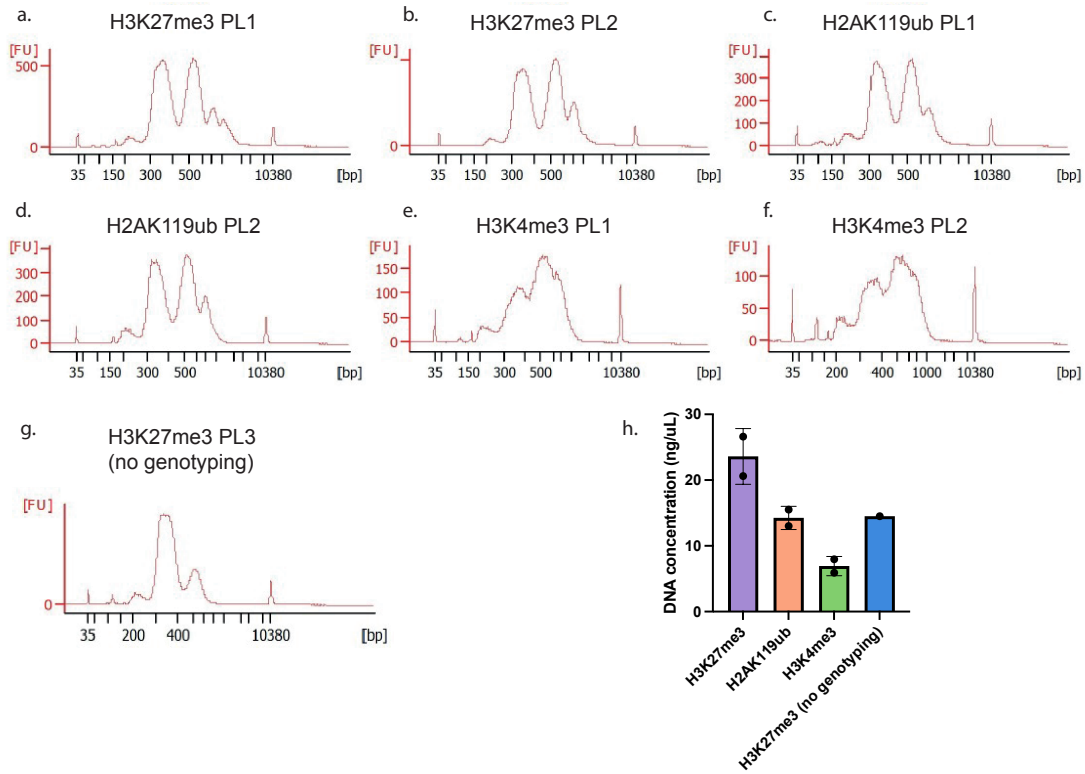


Figure 6.6: Purified CUT&Tag libraries from GCUT cell mixing experiment. a)-g) Bioanalyzer traces of pools of CUT&Tag libraries from 382 single-cells produced using GCUT. Primary antibody condition is indicated above each trace (PL = Plate). h) DNA concentrations (ng/μL) of pools of 382 cells. Each dot represents a pooled plate. Error bars are standard error of the mean.

H3K4me3 is associated with transcriptional activation [101]. Therefore, one would expect these to be enriched for more mononucleosomal fragments, indicative of a more permissive chromatin state. Alternatively, a more permissive chromatin state may mean that there is a greater distance between two separate nucleosomes with H3K4me3. This would increase the distance between transposition sites, and result in the generation of larger fragments. The concentrations of DNA from each library were consistent across replicate plates, and also with the pattern observed across conditions in the single-cell testing (Figure 6.6h).

6.5.3 Genotyping of cell lines

Due to the similarity of the protocol to existing methods, genotyping quality control thresholds were set as for TARGET-Seq and GTAC [45, 232, 260]. Reads

covering the variant nucleotide were detected for all 5 loci, with a mean coverage of 1,968 reads per cell (Figure 6.7a). Of the 2,304 cells that were genotyped, 2,005 cells passed QC for at least one amplicon (number of reads covering the variant nucleotide > 30), indicating efficient capture of each locus by the genotyping amplicons. Importantly, all wells without a sorted nucleus ($n = 12$) failed genotyping QC. Any reads covering the variant nucleotide detected in these wells are likely due to cross-well contamination.

Next, accuracy and sensitivity of genotyping was benchmarked. For the two loci mutant in K562, the scVAF of the mutant reads was < 0.0625 in OCI-AML3 and JURKAT (Figure 6.7b), indicating specific capture of the mutation. The same pattern was also observed for the mutant loci in OCI-AML3 and JURKAT where the mutant scVAF for the non-mutant cell lines was < 0.0715 , and < 0.143 respectively. From scVAF distributions, thresholds for genotype calling were set as described for the TARGET-seq+ genotyping data analysis. Across all five mutant loci, a mean of 74.83% cells were successfully assigned a genotype of either wildtype or mutant for each locus (Figure 6.7c). When inspecting the scVAF distributions, in OCI-AML3 a considerable difference in mean VAF was seen between the two mutant loci: for *DNMT3A* R882C, a mutant scVAF of 0.711 was seen, whereas for *NRAS* Q61, a mutant scVAF of 0.988 was recorded. This is consistent with the notion that OCI-AML3 carries a heterozygous mutation in *DNMT3A* and homozygous mutant for *NRAS*. This indicates that the single-cell genotyping data from GCUT can differentiate zygosity of mutations.

Cells were next assigned to a cell line based on genotype (Figure 6.7d) using the following criteria:

1. A cell was assigned to a cell line if: one or both correct mutant loci for that cell line were genotyped as mutant
2. An “Undetermined” cell line was assigned if that cell was either WT, undetermined, or undetected for all 5 loci

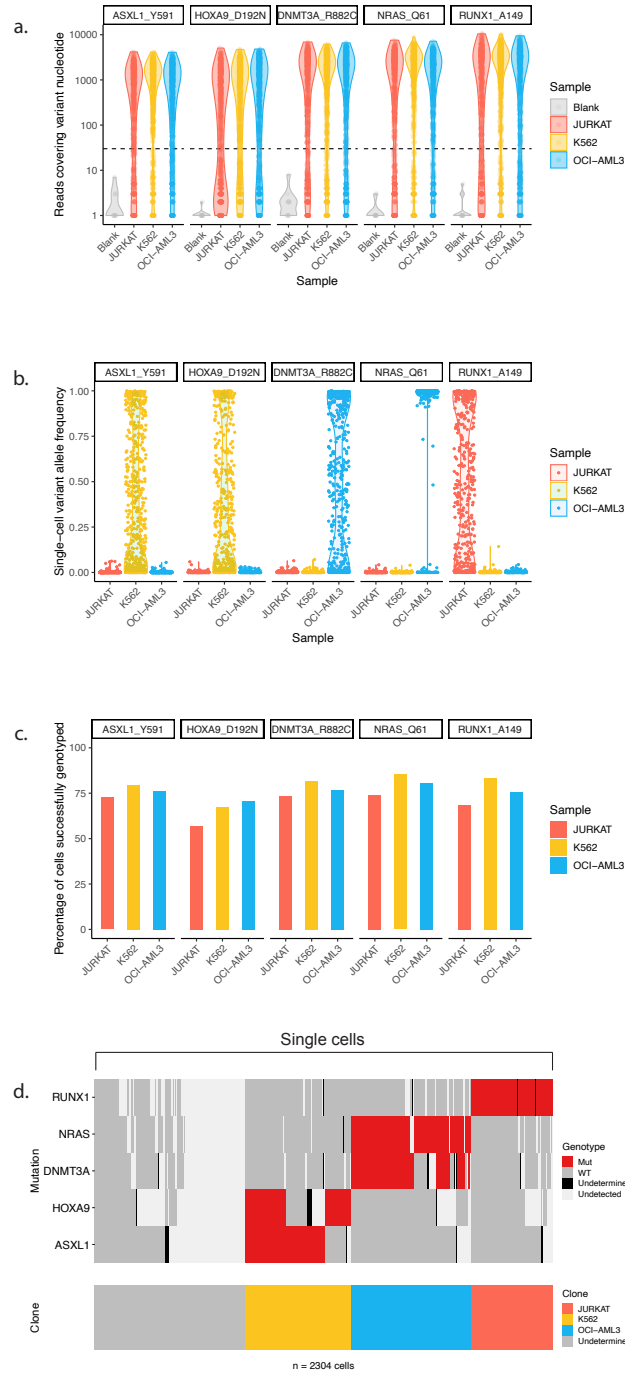


Figure 6.7: GCUT cell mixing experiment genotyping quality control. a) Number of reads covering five targeted genomic loci in single nuclei across three separate cell lines and blank wells (grey). Each dot represents a single cell. Horizontal line represents minimum coverage for single nuclei (30 reads). Percentage of cells failing QC (below line) = 13%. **b)** Single cell VAF distributions for five targeted genomic loci in single nuclei across three separate cell lines. Each dot represents a single cell. Thresholds determining whether a cell is called mutant are individual to each targeted genomic loci, and are not shown. (Continued on next page)

Figure 6.7 (previous page): **c)** Bar plot showing the percentage of each cell line successfully genotyped for each targeted locus. A successful genotyping call was defined as being called as mutant or wildtype. **d)** Top: Raster plot of single-nucleus genotyping results. Each column represents a single nucleus, and each row represents each targeted genomic locus. Colour represents whether each nucleus was called as WT (grey), mutant (red), undetermined (black), undetected (white). Bottom: Assignment of cell type based on genotyping.

Using these criteria, 67.19% of sorted cells could be assigned a cell type based on genotyping. Of these, 67.67% of K562, 78.70% of OCI-AML3, and 54.1% of JURKAT cells were correctly assigned. 32.81% of all sorted cells could not be successfully assigned to a cell line based on genotype. These genotyping success rates are slightly lower ($\approx 10\%$ lower) than average rates recorded in TARGET-seq and GTAC.

Next the allelic dropout (ADO) rates for both mutant and WT alleles across the heterozygous mutations were estimated (Fig 6.8). As previously explained, high ADO rates can result in misassigned genotype, resulting in inaccuracy of downstream analysis. Since the cell lines should be homogenous populations, the rate of homozygous mutation was used as an estimate for ADO for the WT allele, whereas the rate of homozygous WT allele detection was used to estimate the mutant allele ADO rate. Across all five loci, the mutant ADO rate was estimated to be 16.8 - 30.21%, and the WT ADO rate was estimated to be 6.38 - 20.96%. This is higher than other plate-based single-cell multi-omic methods such as TARGET-seq and GTAC, which can achieve ADO rates as low as 1.9% [260]. One reason for this may be the preamplification PCR protocol, where genotyping fragments are amplified simultaneously with CUT&Tag fragments. The protocol includes short cycles of a combined denaturing and annealing step, followed by an elongation step both 10 seconds in length. This PCR protocol is designed to enrich for shorter mononucleosomal fragments of the CUT&Tag library. ADO is likely to occur during the first few cycles of PCR, and any inefficiency of denaturing and annealing during these cycles may explain increased ADO rates.

In summary, GCUT enables high accuracy genotyping of multiple genomic loci in single cells. However, optimisation of the genotyping conditions or protocol could

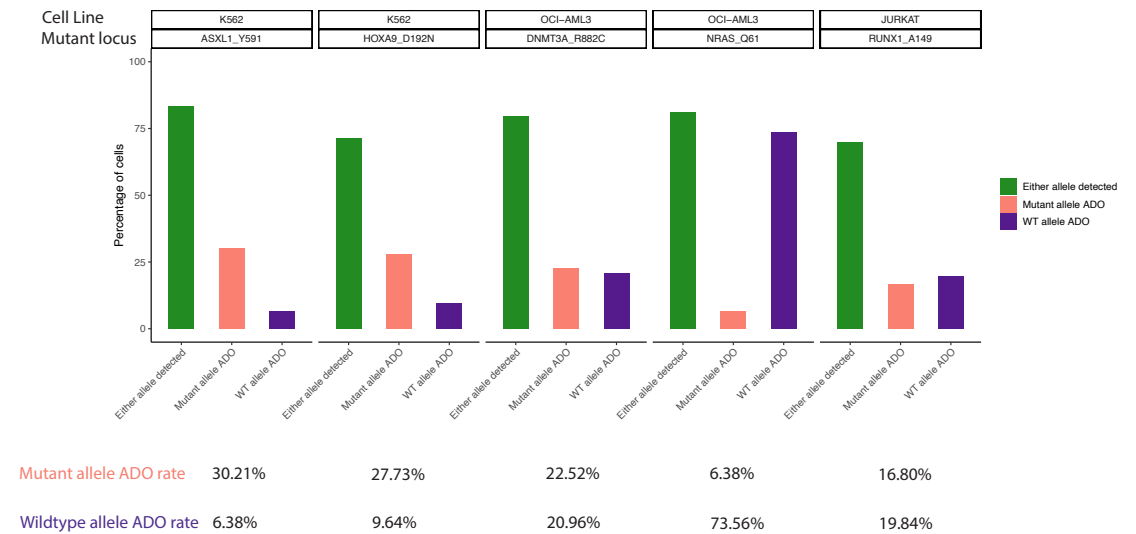


Figure 6.8: Estimation of allelic dropout rate for GCUT genotyping amplicons. Percentage of cells for each amplicon for which: either allele was detected (mutant/WT; Green); Mutant allele ADO was detected (Pink), WT allele ADO was detected (Purple).

increase the confidence in correctly called genotypes.

6.5.4 Chromatin data pre-processing and QC

Next, the quality of data from the single-cell CUT&Tag libraries was assessed. The data was first divided by histone mark (using plate barcodes- each plate contained three cell lines exclusively treated with one type of primary antibody). Across the three histone marks, the mean number of reads per cell was 41,241, of which 79.45% mapped to hg38 (Figure 6.9 a+b). Interestingly, there was a notable difference in library complexity for each cell between histone marks (Figure 6.9c). For H2AK119ub a mean number of unique nuclear fragments was 241, whereas for H3K27me3, the mean number of unique nuclear fragments per cell was 434. This is surprising as the genomic distribution of H2AK119ub and H3K27me3 are largely synonymous [115], so a difference in quantity shouldn't be expected. One possible explanation for this is difference in affinity for respective epitopes of the primary antibodies. Expectedly, H3K4me3- which is found at transcription start sites of active genes [353, 354]- had a lower mean number of unique nuclear fragments per cell (mean = 144). Given the difference between the number of reads and

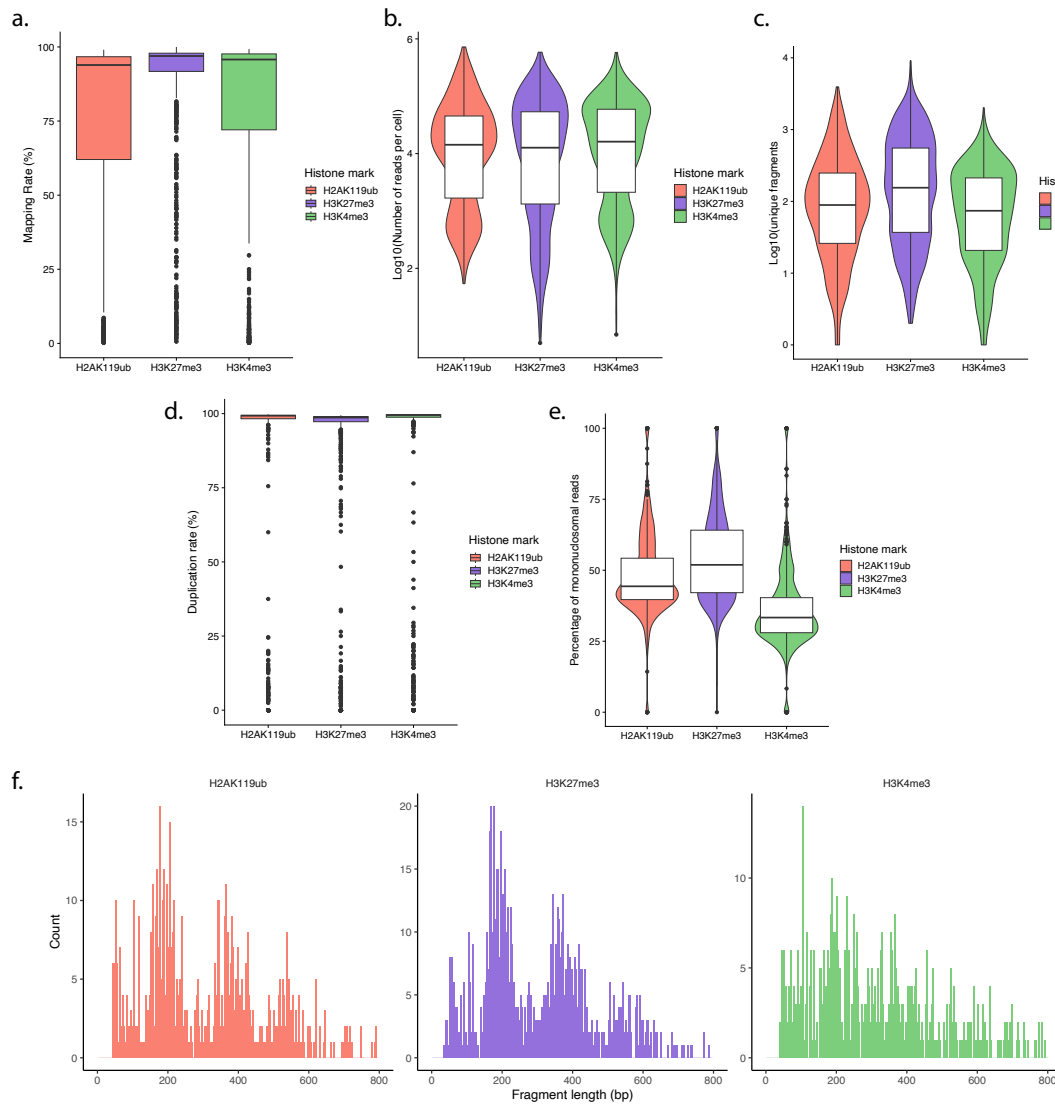


Figure 6.9: Single-cell CUT&Tag library sequencing quality metrics. **a)** Percentage of successfully mapped reads for each cell separated by histone mark. Each dot represents a single cell. **b)** Log10-transformed number of reads per cell separated by histone mark. **c)** Log10-transformed number of unique nuclear fragments per cell separated by histone mark. Mean number of unique nuclear fragments per cell. **d)** Percentage of reads with at least one duplicate for each cell separated by histone mark. Each dot represents a single cell. **e)** Percentage of reads with length between 147bp and 294bp per cell separated by histone mark. Mean percentage of mono-nucleosomal fragments. **f)** Fragment length distribution histogram separated by histone mark.

the number of unique nuclear fragments, it is unsurprising that the duplication rate (the percentage of reads that are duplicates) is very high, with a mean of 82.44% for the entire dataset (Figure 6.9d).

The percentage of reads that are likely to be mononucleosomal (147-294bp

long) is a good indicator of library quality, as mononucleosomal reads give the most accurate location of the epitope of interest (Figure 6.9e). Libraries with very low mononucleosomal percentages are likely to be of low quality. The fragment size distributions based on sequencing data (Figure 6.9f) reflect the overall higher mono-nucleosomal read percentage in cells assaying H3K27me3 distribution (mean = 54.3%), compared to those assaying H3K4me3 (mean = 35.8%).

It is also worth noting that peaks (genomic regions of high numbers of reads) were not called for each histone mark. The decision for this was the lack of library complexity and cell numbers. One possible quality control metric that could have been used is the fraction of reads in peaks (FRiP), and if library complexity can be improved then this would be another metric to use to filter low-quality cells.

To filter low quality cells, two metrics were used. First, cells were filtered that had fewer than 15% of mono-nucleosomal reads. Second, the number of unique nuclear fragments was used to filter low complexity cells. Due to the lack of complexity of the single-cell CUT&Tag libraries, it is challenging to define a threshold that filters data for high quality cells, whilst retaining enough cells to draw meaningful conclusions. Because of this, I decided to test different thresholds for the minimum number of unique nuclear fragments. An alternative strategy would be to establish a quality metric of similarity of each single cell's profile to a bulk CUT&Tag profile of each histone mark in each cell type. One example of this could be an AUCell score.

6.5.5 Dimensionality reduction and clustering

As the number of unique nuclear fragments (UNFs) per cell is a measure of library complexity, low complexity cells could impact the ability to distinguish different cell populations. Previous published studies investigating the use of droplet-based scCUT&Tag have used a threshold of 100 unique nuclear fragments per cell [242, 349, 355]. However, as all of these studies used the high-throughput 10X platform, data from many single cells increased the complexity of the dataset. Only 2,674 nuclei were processed in the GCUT dataset, so the complexity for this dataset was

a lot lower. Therefore, I reasoned that a higher threshold for the minimum number of unique nuclear fragments would be beneficial for analyses.

I decided to test three different thresholds: 10, 100, and 200 unique nuclear fragments per cell. I used dimensionality reduction and subsequent clustering as a method to assess different thresholds for minimum number of unique nuclear fragments per cell. For dimensionality reduction using standard Signac package workflow [263], singular value decomposition (SVD) to calculate latent semantic index (LSI) components. The top 30 LSI components were then used for uniform manifold approximation and projection (UMAP) analysis, and to find clusters using Louvain algorithm.

When applying a liberal threshold of 10 unique nuclear fragments per cell, the number of cells passing QC for the H3K27me3 condition was 93.51% (Figure 6.10a). Four separate clusters were identified following dimensionality reduction, and when highlighted by cell line (plate coordinates of the cell lines were recorded at time of sorting), three of the four clusters corresponded to the cell lines (Figure 6.10b,c). The largest cluster (cluster 1, n=373 cells) was a heterogeneous population of the three cell lines. When a threshold of 100 UNF per cell was applied on the same data, 5 clusters were identified and formed distinct entities corresponding to 3 cell lines (Figure 6.10e,f). A much smaller heterogeneous population was identified this time (cluster 3, n = 139 cells). However, across all three cell lines, the percentage of cells passing QC dropped to 59.16% (Figure 6.10d). A similar pattern was seen when applying a more stringent threshold of 200 UNF per cell (Figure 6.10h,i). However, at this threshold, the percentage of cells passing QC dropped to between 3-10%, depending on the cell line, and histone mark (number of cells for whole dataset passing QC= 936; Figure 6.10g).

In order to formally quantify the difference between UNF thresholds, two quality metrics were established. First, the percentage of cells from each cell lines correctly assigned to each cluster was calculated (Figure 6.11a). Despite no significant difference between 10 and 100 unique nuclear fragments, there was a trend for the numbers of correctly assigned cells to increase with a stricter UNF threshold.

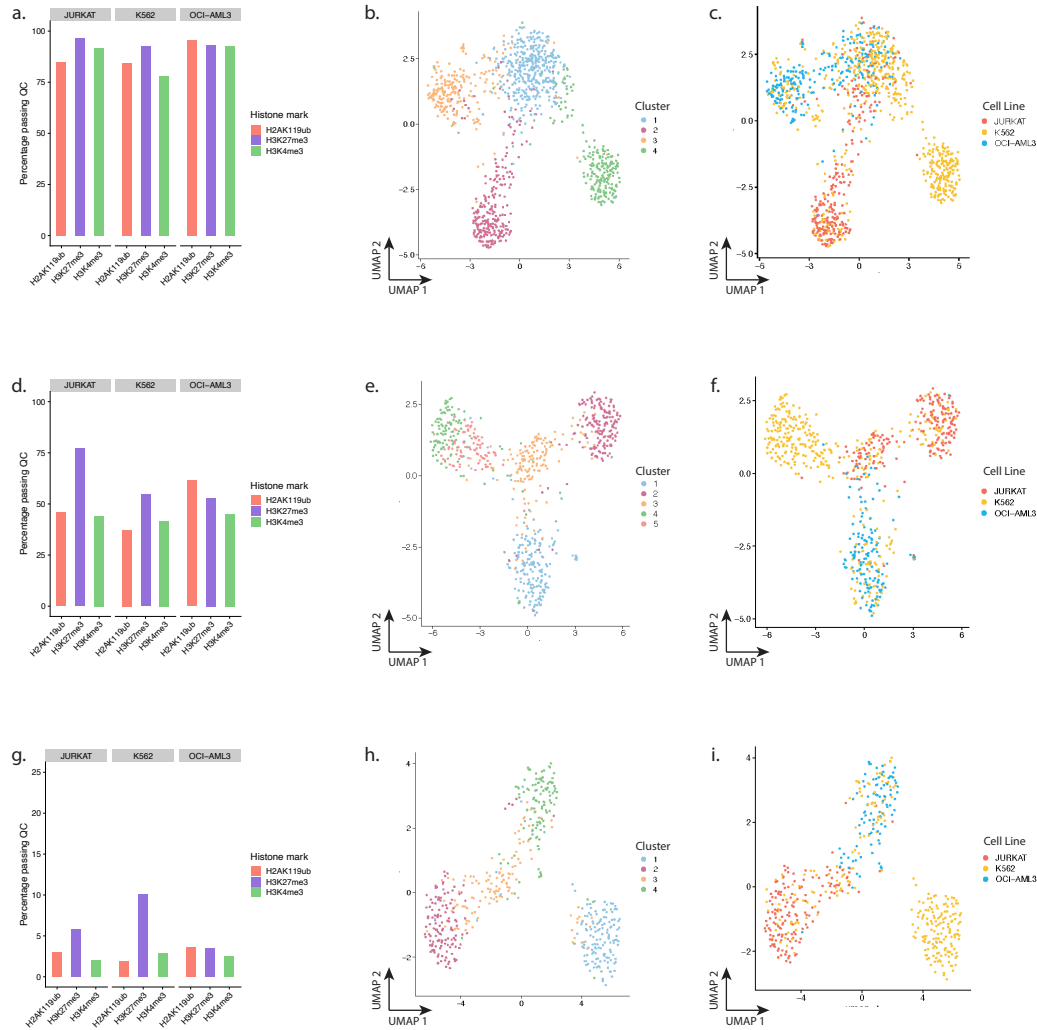


Figure 6.10: Comparison of quality control thresholds for H3K27me3 single-cell CUT&Tag libraries. a) Percentage of cells passing QC in each cell line, divided by histone mark with number of unique nuclear fragments ≥ 10 . b) UMAP of cells based on H3K27me3 distribution with number of unique nuclear fragments ≥ 10 , coloured by Louvain cluster c) UMAP of cells based on H3K27me3 distribution with number of unique nuclear fragments ≥ 10 , coloured by cell type recorded at time of sort. d)-f) As for a) b) and c), but with number of unique nuclear fragments ≥ 100 . e)-h) As for a) b) and c), but with number of unique nuclear fragments ≥ 200 .

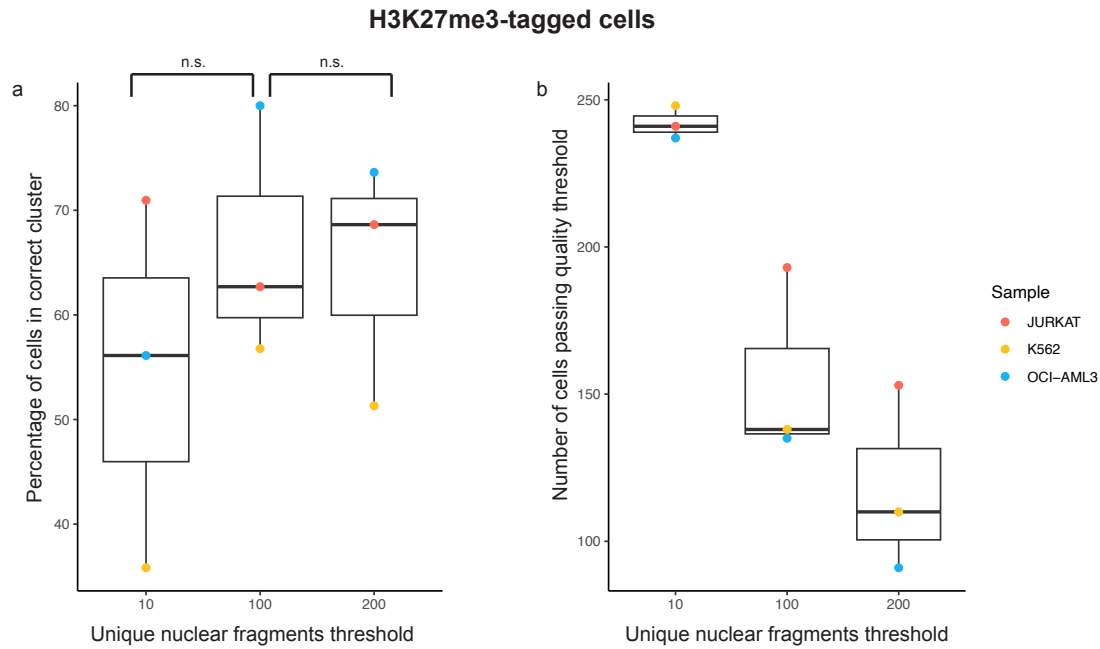


Figure 6.11: Quality metrics of different UNF thresholds for H3K27me3 single-cell CUT&Tag libraries. **a)** Boxplot showing the percentage of cells from each cell line present in the correct cluster at each UNF threshold. Each dot represents a cell line. Student's t-test was used to assess significant difference ($P < 0.05$) between thresholds. **b)** Boxplot showing the number of cells passing quality control at each UNF threshold. Each dot represents a cell line.

Secondly, the number of cells passing quality control at each UNF threshold was calculated (Figure 6.11b). The number of cells passing quality control dramatically decreased between 10 and 100 UNFs (mean difference across cell lines = 95 cells). However this dropped again when the threshold was raised to 200 UNF. The threshold of 100 UNF per cell appeared to strike the right balance of granularity of dataset, whilst retaining a high amount of cells.

The same thresholds were tested for the H2AK119ub and H3K4me3 datasets, however it is worth noting that the mean number of UNFs per cell for these marks was significantly lower than that of H3K27me3 (Figure 6.9c). Applying a more conservative threshold for both H2AK119ub and H3K4me3 did not result in more distinguishable clusters entities corresponding to different cell types (Figure 6.12, 6.13). Furthermore, the percentage of cells passing QC at the higher threshold was noticeably less in both conditions (Figure 6.10 a,g).

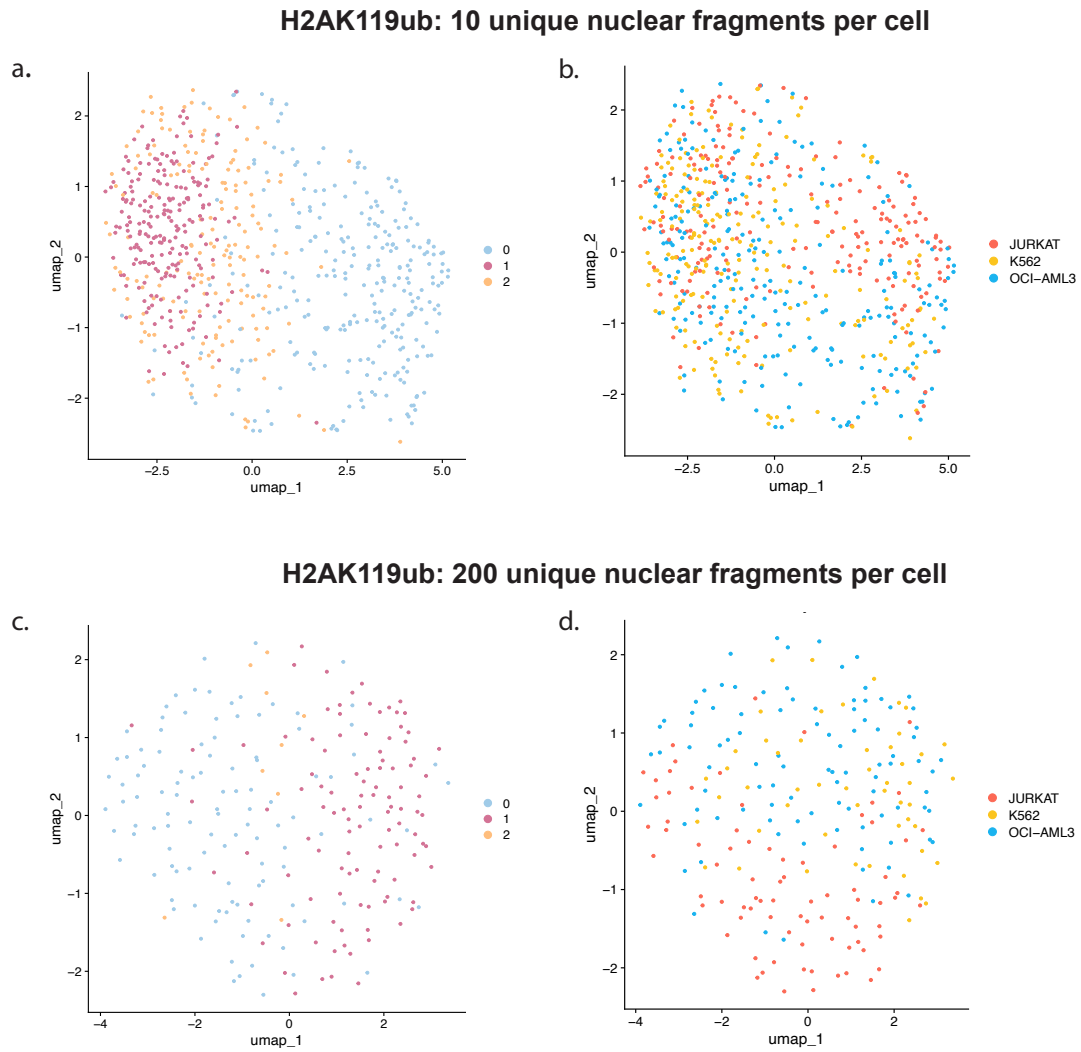


Figure 6.12: Comparisons of different UNF thresholds for H2AK119ub single-cell CUT&Tag libraries. **a)** UMAP of cells based on H2AK119ub distribution with number of unique nuclear fragments ≥ 10 , coloured by Louvain cluster **b)** UMAP of cells based on H2AK119ub distribution with number of unique nuclear fragments ≥ 10 , coloured by cell type recorded at time of sort. **c)+d)** As for **a)+b)** but UNF threshold at ≥ 200 .

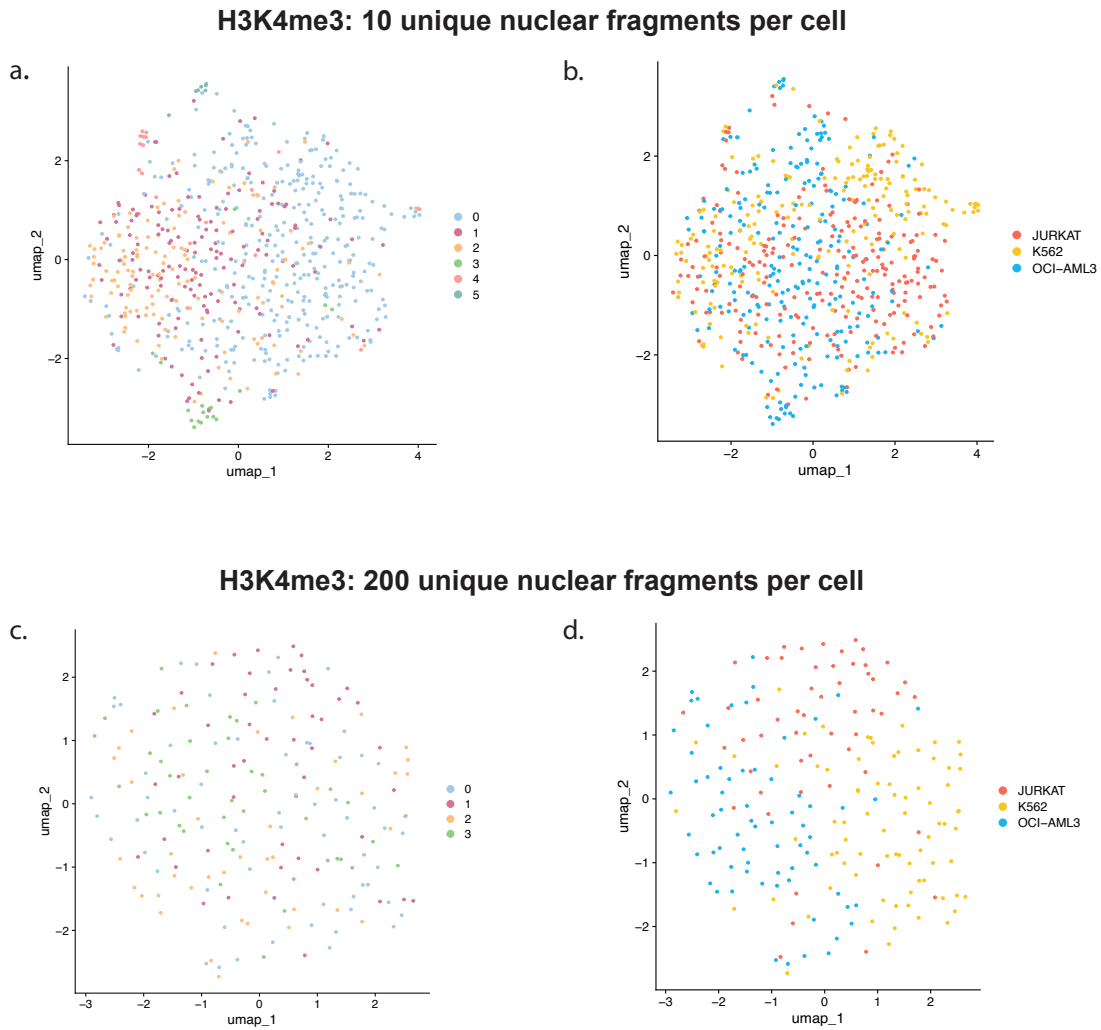


Figure 6.13: Comparisons of different UNF thresholds for H3K4me3 single-cell CUT&Tag libraries. **a)** UMAP of cells based on H3K4me3 distribution with number of unique nuclear fragments ≥ 10 , coloured by Louvain cluster **b)** UMAP of cells based on H3K4me3 distribution with number of unique nuclear fragments ≥ 10 , coloured by cell type recorded at time of sort. **c)+d)** As for **a)+b)** but UNF threshold at ≥ 200 .

Based on these data, I decided to implement a threshold of 100 UNFs per cell for further downstream analyses, striking a balance between library complexity and cell recovery.

6.5.6 CUT&Tag library quality with and without genotyping

Next, I asked whether the inclusion of single-cell genotyping resulted in differences in single-cell CUT&Tag library quality. For this, an extra plate of the three cell lines that had undergone CUT&Tag with anti-H3K27me3 primary antibodies was

sorted, and single-cell CUT&Tag libraries were generated without the inclusion of spike-in genotyping primers. Importantly, these cells were from the same aliquot of each cell line used for GCUT. The sequencing data from this library could then be compared to the H3K27me3 GCUT data (Fig 6.14).

There was no significant difference in library complexity between scCUT&Tag and GCUT in K562 and OCI-AML3 cell lines. However, for JURKAT, the genotyped cells had a significantly higher number of UNFs per cell than the scCUT&Tag condition ($p = 1.34 \times 10^{-5}$; Figure 6.14a). Interestingly, for all three cell lines, the percentage of mononucleosomal fragments was significantly higher in the scCUT&Tag condition (Figure 6.14b). Since the genotyping amplicons are similar in length to mononucleosomal fragments, there may be greater competition for resources during the pre-amplification PCR reaction in the GCUT condition. This would result in a smaller number of mononucleosomal fragments being amplified. There was no clear difference in the number of cells passing QC between the two conditions, and this largely correlated with the number of UNFs per cell (Figure 6.14c). When comparing the fragment distributions, the cells from the scCUT&Tag condition had a fewer number of fragments corresponding to multi-nucleosomal lengths than the GCUT condition (Figure 6.14d), consistent with the pattern for the percentage of mono-nucleosomal reads (Figure 6.14b).

When the K562 cells from the scCUT&Tag condition were mapped onto the UMAP of combined H3K27me3 data, they were mixed heterogeneously with the rest of the K562 H3K27me3 cells, and importantly didn't form any new cell clusters, or map to the clusters corresponding to the other cell lines (Fig 6.14e). This suggests that across the whole genome, there is little difference in the profiles of histone mark distribution generated by the two methods from the same group of cells. However, when inspecting distribution of fragments the gene level, some differences can be identified (Fig 6.14f). When comparing H3K27me3 distribution at the posterior *HOXA* gene cluster in K562 cells, fragments are detected in cells from both conditions as expected. The distributions of these fragments are different, which is unexpected considering that these cells should be homogenous. This may

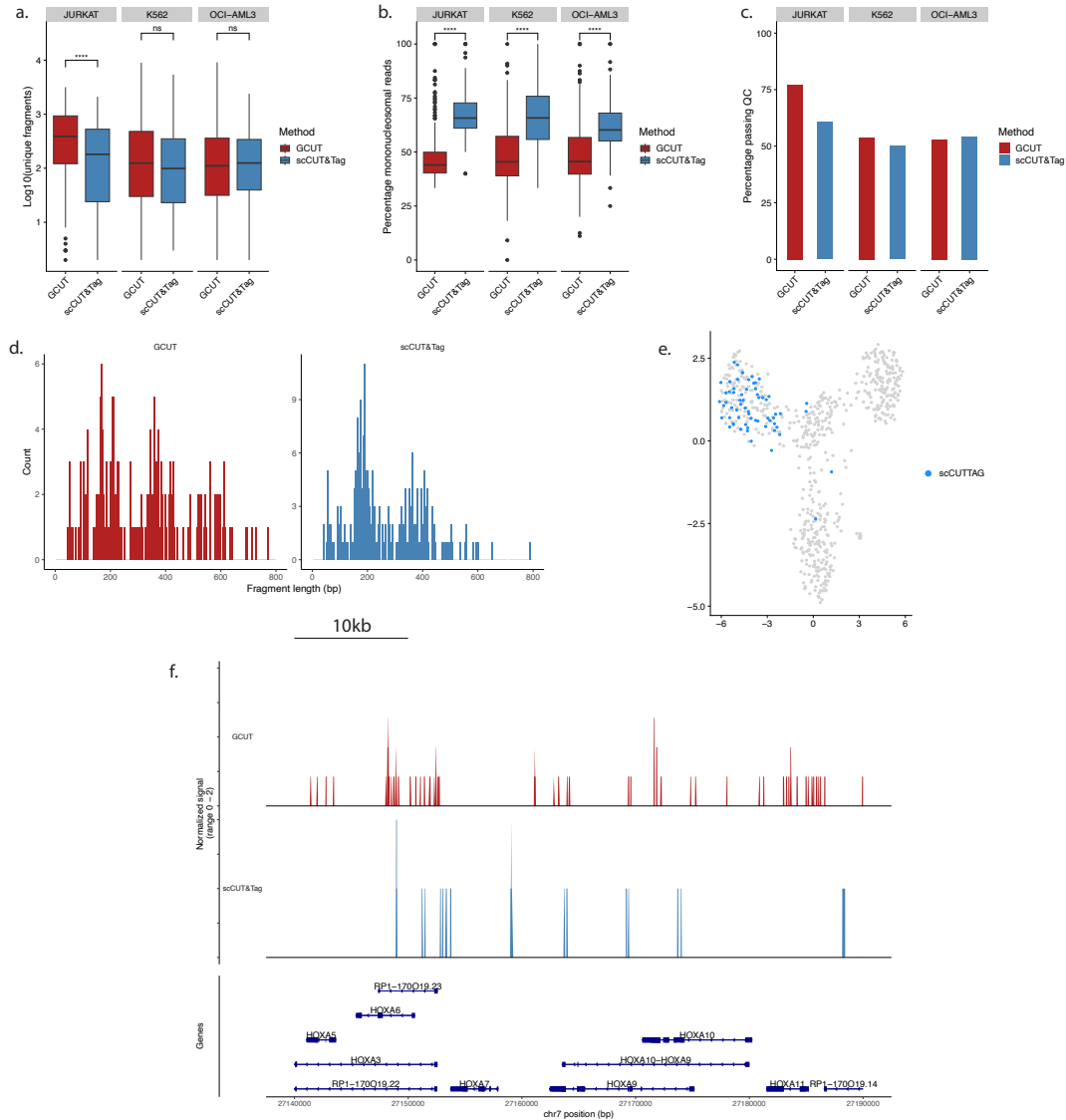


Figure 6.14: Single-cell CUT&Tag library quality sequencing metrics with and without genotyping. **a)** Log10-transformed number of unique nuclear fragments per cell separated by method. Mean number of unique nuclear fragments per cell: H3K27me3 = , H2AK119ub = , H3K4me3 = . Statistics were calculated using student's t-test ($p < 0.05$) **b)** Percentage of reads with length between 147bp and 294bp per cell separated by histone mark. Mean percentage of mono-nucleosomal fragments: H3K27me3 = , H2AK119ub = , H3K4me3 = . Statistics were calculated using student's t-test ($p < 0.05$). **c)** Percentage of cells passing QC in each cell line, divided by method. Cells passed QC when number of unique nuclear fragments ≥ 100 and mono-nucleosomal percentage $\geq 15\%$. **d)** Fragment length distribution histogram separated by histone mark. **e)** UMAP of cells based on H3K27me3 distribution, overlaid with data from K562 cells that were not genotyped (blue). Cells that were not genotyped are distributed heterogeneously with those that were genotyped. **f)** Genome browser tracks from the posterior *HOXA* gene cluster (chr7:27140000-27190000) of H3K27me3 distribution in K562 cells, split by method.

be due to the combination of low library complexity (if the CUT&Tag process has limited resolution), and low cell number (scCUT&Tag $n = 128$; GCUT $n = 256$). Aggregating data from more cells may result in greater similarity in distribution profiles.

Overall, there is little difference in single-cell CUT&Tag library quality with or without genotyping. The difference in percentage of mono-nucleosomal reads between scCUT&Tag and GCUT may be resolved by optimising the preamplification PCR conditions or protocol.

6.5.7 Integration of genotyping and chromatin data

Next, the single-cell genotyping data was integrated with the profiles of histone mark distribution. When the UMAP of cells from the H3K27me3 condition was highlighted with the mutant and wildtype genotypes from the single-cell genotyping data, the mutant cells for each mutant locus were mapped to the cell cluster corresponding to the correct cell line (Figure 6.15). For example, cells mutant for *ASXL1* and *HOXA9*, which are mutant loci in K562, were highlighted in the cluster corresponding to K562 identified by H3K27me3 distribution profiles (Figure 6.15c,d). The same pattern was also true for loci mutant in JURKAT and OCI-AML3 (Figure 6.15e-g). Cluster identities were confirmed by quantifying their composition with each cell line using the plate coordinates recorded during single nuclear sorting (Figure 6.15h).

When the same analysis was done to the H2AK119ub dataset, the mutant cells mapped onto the correct cell line (Figure 6.16 b-e). However, based on H2AK119ub distribution and using the previously discussed QC thresholds, cell lines could not be clearly distinguished into separate clusters (Figure 6.16a, b). Colouring the UMAP by single-cell genotype, identified the correct cell line identities (i.e. recorded at time of sorting), however these also did not distinguish separate clusters (Figure 6.16c,d,e).

Two possible hypotheses to explain this could be: either the primary antibody was not specific enough to bind H2AK119ub leading to reduced complexity of CUT&Tag fragment library, or that the cell lines are not distinguishable based on

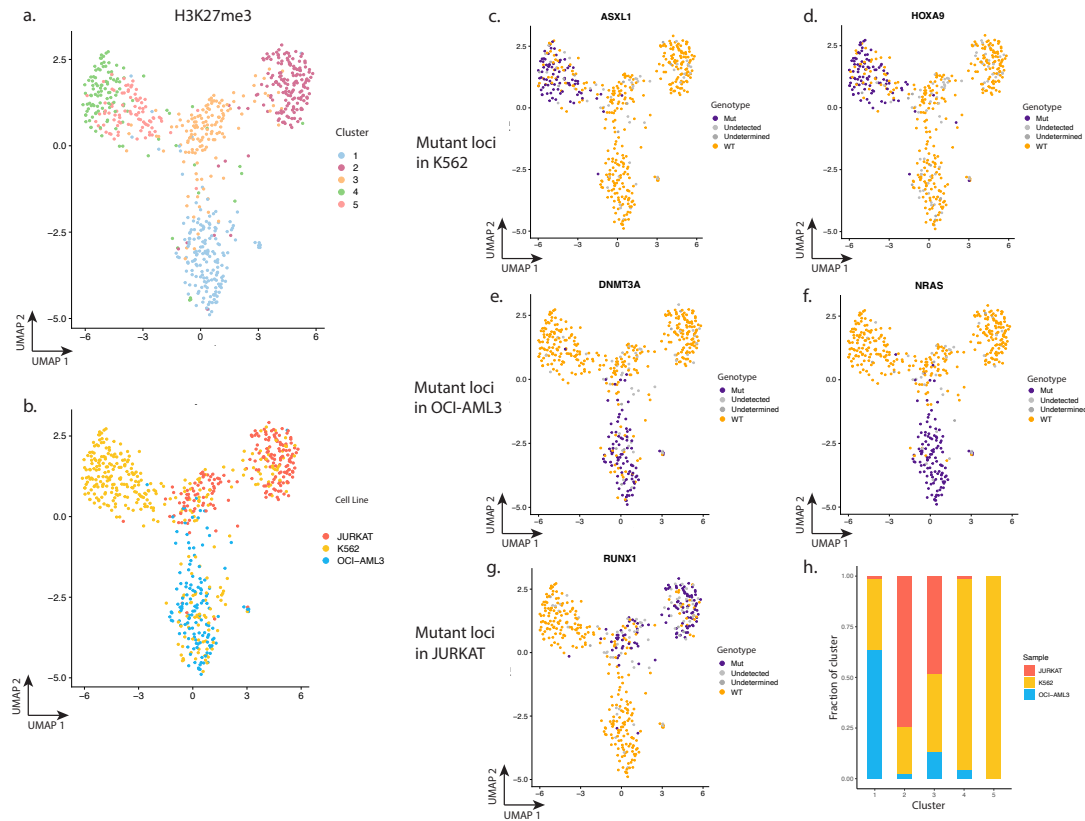


Figure 6.15: Integration of single-cell genotyping and CUT&Tag data for H3K27me3. In a-g) 256 cells from indicated cell lines were subjected to targeted genotyping for *ASXL1*, *HOXA9*, *DNMT3A*, *NRAS*, *RUNX1*. a) UMAP of cells based on H3K27me3 distribution, passing QC at unique nuclear fragments ≥ 100 , coloured by Louvain cluster. b) As in a) but cells are coloured by cell type recorded at time of sort c)-g) The UMAP from a) but coloured by genotyping calls for individual cells. For each indicated locus, mutant cells are coloured in purple, wildtype cells are coloured orange. Cells for which genotype could not be determined or were undetected were coloured grey. h) Composition of each Louvain cluster by recorded cell type

the distribution of H2AK119ub. This represents a major problem in the development of GCUT for use in primary *ASXL1*-mutant CH samples. A bulk CUT&Tag sequencing experiment which trials different primary antibodies against H2AK119ub in the three cell lines would be required to solve this problem and optimise GCUT for this specific application.

Cluster composition was used as a metric to investigate the efficiency of each histone mark to distinguish between cell lines. When manually inspecting each cluster in the H2AK119ub dataset, subtle patterns of enrichment could be observed—for example JURKAT cells (*RUNX1*-mutant cells; Figure 6.16e) were specifically

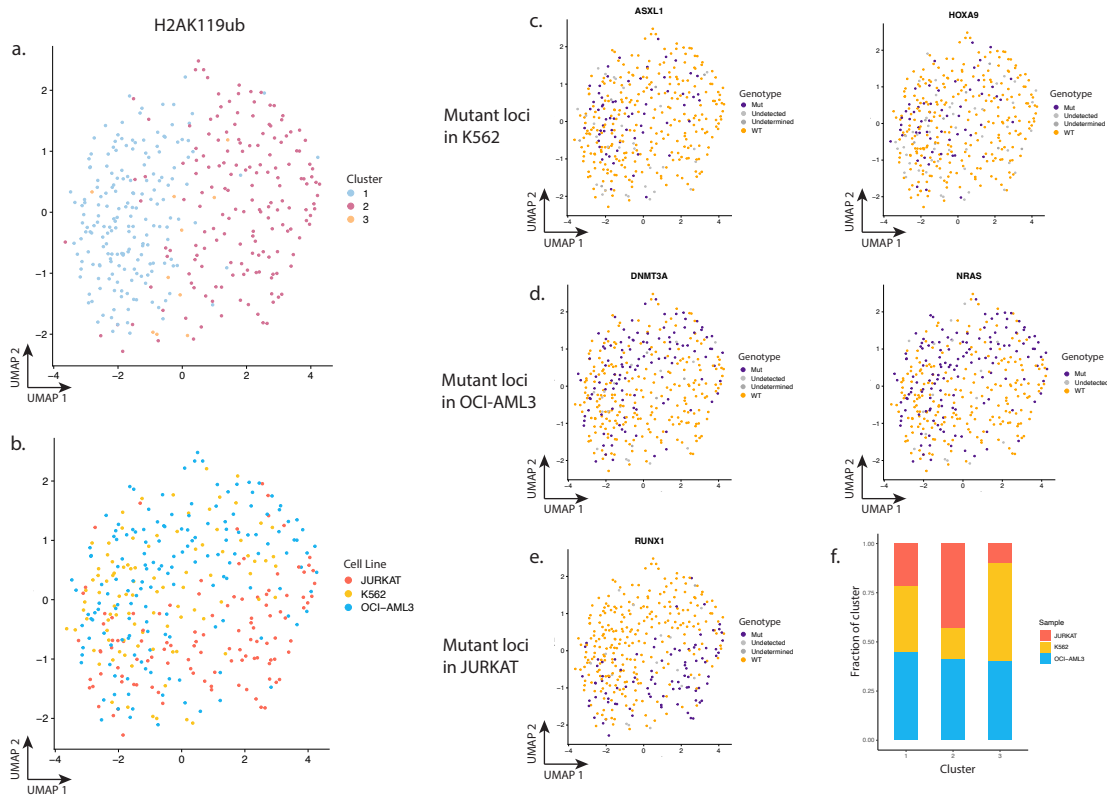


Figure 6.16: Integration of single-cell genotyping and CUT&Tag data for H2AK119ub. In **a-g)** 256 cells from indicated cell lines were subjected to targeted genotyping for *ASXL1*, *HOXA9*, *DNMT3A*, *NRAS*, *RUNX1*. **a)** UMAP of cells based on H2AK119ub distribution, passing QC at unique nuclear fragments ≥ 100 , coloured by Louvain cluster. **b)** As in **a)** but cells are coloured by cell type recorded at time of sort **c-e)** The UMAP from **a)** but coloured by genotyping calls for individual cells. For each indicated locus, mutant cells are coloured in purple, wildtype cells are coloured orange. Cells for which genotype could not be determined or were undetected were coloured grey. **f)** Composition of each Louvain cluster by recorded cell type

enriched in cluster 2. When cluster composition was quantified, JURKAT cells were found to compose 42.85% of cluster 2 (Figure 6.15f). In contrast, the data for H3K27me3 showed that JURKAT cells were enriched in cluster 2, composing 74.69% of the entire cluster (Figure 6.15h). Additionally, cluster 2 in the H3K27me3 dataset was a distinct entity separated from the other clusters. Interestingly, there was also a lack of distinct clusters that formed separate entities in the H3K4me3 dataset (Figure 6.17a), however cells from each individual cell line were grouped together (Figure 6.17b-f). Upon manual inspection, cluster 3 corresponded to JURKAT cells. When cluster composition was quantified, 70.68% of cluster 3 was

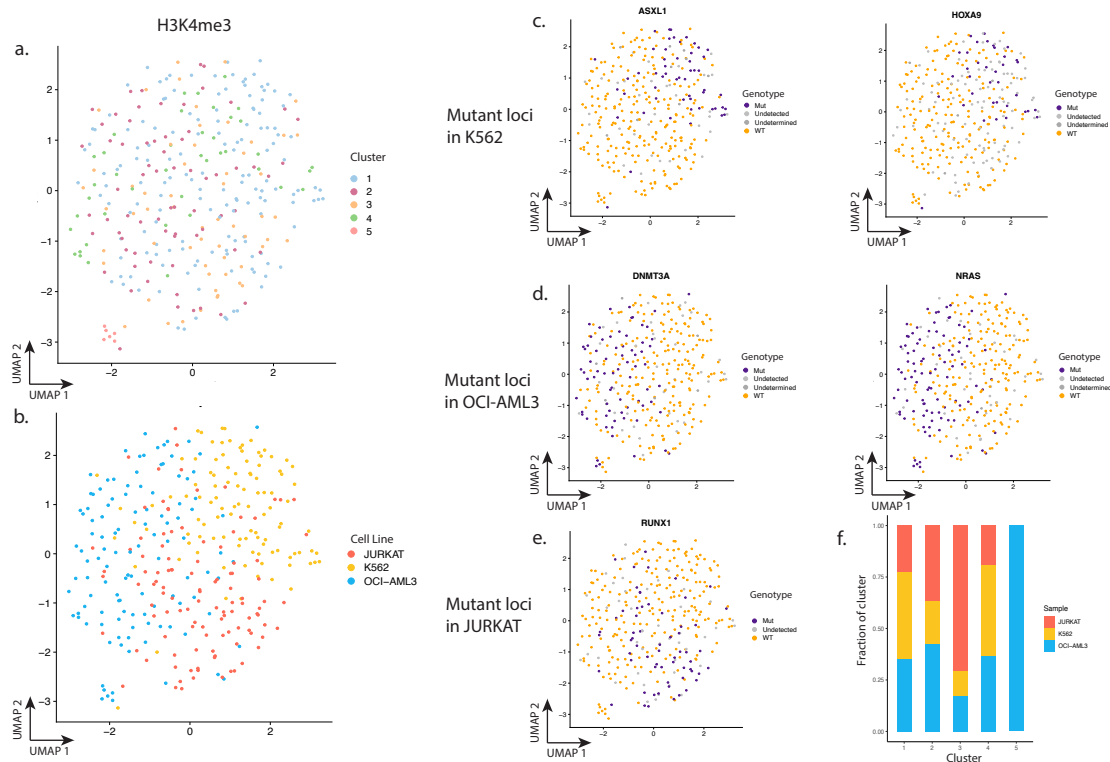


Figure 6.17: Integration of single-cell genotyping and CUT&Tag data for H3K4me3. In a-g) 256 cells from indicated cell lines were subjected to targeted genotyping for *ASXL1*, *HOXA9*, *DNMT3A*, *NRAS*, *RUNX1*. **a)** UMAP of cells based on H3K4me3 distribution, passing QC at unique nuclear fragments ≥ 100 , coloured by Louvain cluster. **b)** As in a) but cells are coloured by cell type recorded at time of sort **c)-e)** The UMAP from **a)** but coloured by genotyping calls for individual cells. For each indicated locus, mutant cells are coloured in purple, wildtype cells are coloured orange. Cells for which genotype could not be determined or were undetected were coloured grey. **f)** Composition of each Louvain cluster by recorded cell type

JURKAT cells. Therefore, H2AK119ub distribution was worse at classifying cell lines in this experiment than H3K27me3 and H3K4me3.

Overall, single-cell genotyping can identify cell line identities of clusters formed based on histone mark distribution, but the distinction of clusters was highly dependent on the histone mark. A correlate with cluster definition was the number of unique fragments per cell. The H3K27me3 dataset clearly displayed the best cluster definition, which also had the highest number of unique nuclear fragments per cell. Improving the number of unique nuclear fragments per cell for the H2AK119ub condition in particular may aid in the definition of distinct clusters.

6.5.8 Genomic distribution of histone marks of from GCUT CUT&Tag data

The analyses to this point have considered the genome-wide view of histone mark distribution in single-cells collected using GCUT. This has provided limited success in distinguishing populations of cells derived from different haematopoietic lineages. One reason for this may be the lack of CUT&Tag library complexity in each single cell. However, pseudobulk histone mark distribution profiles of cell lines from single-cell data may provide insight into the quality and resolution of the data. If pseudobulking the single-cell data cannot distinguish between cell lines, then an alternate strategy to assay histone mark distribution must be considered. To test this, pseudobulk single-cell CUT&Tag profiles were aggregated for each cell line per histone mark. Pseudobulk profiles were then manually inspected at genomic loci of marker genes of each cell lines (Figure 6.18).

GATA2, which should be expressed in K562 [356], was found to be enriched for the transcriptionally activating H3K4me3 mark, and not enriched for the transcriptionally repressive H3K27me3 and H2AK119ub marks. The inverse was observed in OCI-AML3 and JURKAT cell lines. OCI-AML3 was enriched for H3K4me3 at the posterior *HOXA* gene cluster, as expected [357], whereas K562 and JURKAT were not. Similarly, JURKAT was enriched for H3K4me3 at the canonically expressed *GATA3* locus [358], whereas OCI-AML3 and K562 was enriched for H3K27me3 and H2AK119ub.

In summary, pseudobulk CUT&Tag data from single-cells generated using GCUT can distinguish cell lines based on histone mark distribution at marker genes. Quantifying this by correlative analyses with bulk CUT&Tag data or ChIP-seq data for each histone mark in each cell line, would be an important analyses once the CUT&Tag steps of the GCUT protocol have first been optimised.

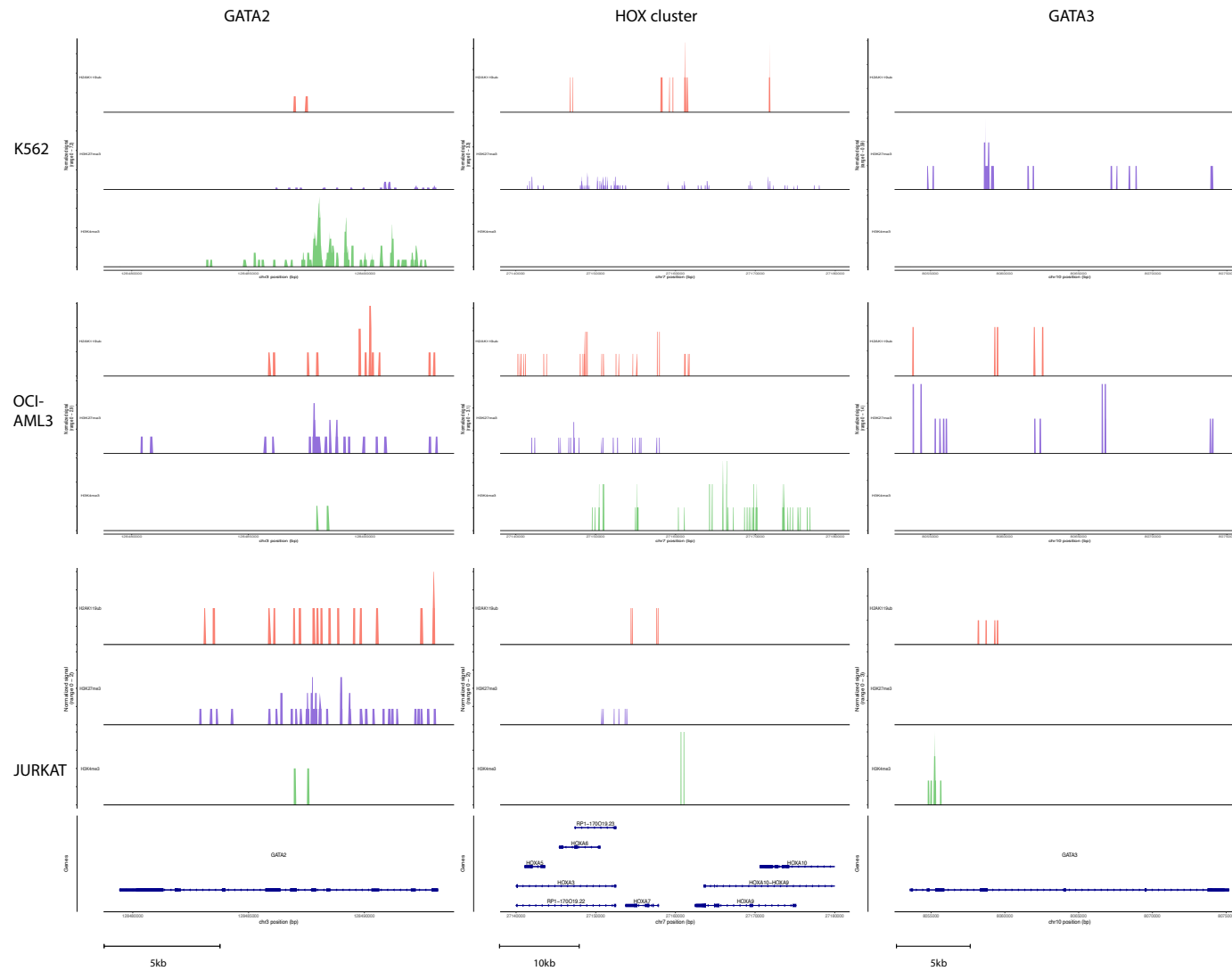


Figure 6.18: Genomic distribution of histone marks of from GCUT CUT&Tag data. Genome browser tracks of pseudobulk single-cell histone mark distributions at three genomic loci that distinguish cell lines. Each track is coloured by histone mark (Pink = H2AK119ub, Purple = H3K27me3, Green = H3K4me3). Each column represents a genomic region: *GATA2* (chr3: 128480000-128500000; expressed in K562), Posterior *HOXA* cluster (chr7:27140000-27180000; expressed in OCI-AML3), *GATA3* (chr10: 8054500-8075500; expressed in JURKAT). Each row represents pseudobulk histone mark distributions from each cell line.

6.6 Concluding remarks

In summary, I have reported for the first time a novel method that can simultaneously profile genomic distribution of histone modifications and genotype multiple genomic loci in single cells. A proof-of-concept experiment has shown that high-accuracy genotyping can identify different cell types based on SNVs, which can be used to identify clusters of cells based on histone mark distributions of single cells. For use in identifying differences in histone mark distribution in rare cell populations between cells of different genotypes, improvements in the accuracy of single-cell genotyping and complexity of CUT&Tag library would need to be made to GCUT. Optimisation of both the CUT&Tag protocol, and the downstream library preparation, and careful selection of reagents would be required to improve quality of genotyping and CUT&Tag libraries in GCUT.

Chapter 7

Discussion

Contents

7.1	The cellular landscape of <i>ASXL1</i> -mutant CH	187
7.2	The impact of <i>ASXL1</i> mutations on the transcriptome of human HSPCs	190
7.3	GCUT: a novel single-cell method to simultaneously assay genotype and histone modification distribution in heterogeneous cell populations	196
7.4	Study limitations and future work	200

A common characterization of ageing is an increase in mutational burden across bodily tissue. The acquisition of fitness-inducing mutations can result in the preferential selection of stem cells. Over the lifetime of an individual, this process will result in clonal mosaicism throughout the body. Clonal mosaicism has been identified in almost all tissues in the human body maintained by adult stem cells [3, 17–20]. In the past 10 years, studies describing this phenomenon in the haematopoietic system of healthy individuals have generated huge amounts of interest due to the association with an increased incidence of blood cancer and other adverse outcomes [24–26]. Despite extensive investigation, it remains unclear how different CH-associated mutations confer clonal advantage. Additionally, the reasons why some clones confer a greater risk of leukaemic transformation, and others do not remain elusive. The risk for progression from CH to malignancy and other adverse outcomes, greatly increases with mutant clone size. Identification of the mechanisms by which mutant clones expand is therefore crucial in understanding the onset of blood cancer.

Mutations in *ASXL1*, a putative epigenetic regulator, are frequently found in CH and a number of myeloid malignancies, including CMML, MDS, MPN, and AML. Notably, *ASXL1* mutations not only confer a greater probability of transformation from CH to AML than for mutations in *DNMT3A* and *TET2*, which account for

$\approx 75\%$ of all CH mutations [26, 53]. Additionally, *ASXL1* mutations are associated with unfavourable outcome in myeloid leukaemias. It should be noted that there is the small possibility that some cases of myeloid malignancy are not preceded by CH, so the association of *ASXL1* mutations and poor prognosis of malignancy may only be relevant if the *ASXL1* mutation is clonal to the transforming mutation. Nevertheless, given these observations it is puzzling why *ASXL1*-mutant clones only account for 5-10% of all CH mutations.

Despite efforts to understand exactly how *ASXL1* mutations affect haematopoiesis *in vitro* and *in vivo*, the exact molecular mechanisms driving the initial clonal expansion remain elusive. Models of mutations in *DNMT3A* and *TET2* *in vivo* and *in vitro* have shown an increased proliferative capacity of mutant HSPCs [203–206]. Curiously, similar models of *ASXL1* mutations display the opposite phenotype- mutant HSPCs have a clonal disadvantage, which is not consistent with clinical observations from humans. Two possibilities may explain this difference.

First, the mechanisms responsible for clonal advantage in *ASXL1*-mutant CH may be distinct from those governing *DNMT3A*- and *TET2*-mutant CH. Functionally, all three genes are epigenetic regulators, however they differ in how this is achieved. *DNMT3A* and *TET2* encode enzymes that regulate DNA methylation, whereas *ASXL1* encodes a crucial co-factor in the PR-DUB complex which regulates the distribution of the histone modification H2AK119ub. Epigenetic control of gene expression is an exquisitely interconnected process, but disruption of different components of epigenetic machinery may result in different phenotypes.

Secondly, models of *ASXL1*-mutations may inaccurately replicate the conditions seen in humans. *In vivo* murine models have accurately replicated the genetics of *ASXL1* mutations, however clonal expansion in humans is thought to occur over the course of decades. During this time, a spectrum of different external stressors may impact how *ASXL1*-mutant clones are selected. This is nearly impossible to replicate in the 2-year life span of a laboratory mouse. It is entirely possible

environmental factors could have profound effects on clonal selection of *ASXL1*-mutant cells. Accurately defining which stimuli *ASXL1*-mutant cells are responding to in humans will likely improve accuracy of *in vivo* models.

From current literature, two outstanding questions remained about *ASXL1* mutations. First, which molecular mechanisms are disrupted by *ASXL1* mutations, and which might confer selective advantage? Secondly, is the selection of *ASXL1*-mutant clones consistent throughout clonal haematopoiesis? In this study, I aimed to characterise the cellular and molecular changes driving selection of *ASXL1*-mutant clones in healthy human patients with clonal haematopoiesis. To gain insight into the cellular and transcriptional changes in *ASXL1*-mutant cells, I performed TARGET-seq+ on cells from 4 *ASXL1*-mutant CH bone marrow samples and 5 age-matched control healthy bone marrow samples. Additionally, I presented for the first time, proof-of-concept results of a novel single-cell multiomic method which simultaneously characterises epigenomic profile and genotype of single cells.

7.1 The cellular landscape of *ASXL1*-mutant CH

One problem faced when investigating clonal haematopoiesis is the availability of primary human samples. Compounding this, the lower prevalence of *ASXL1* mutant patients ($\approx 5 - 10\%$ of CH cases) makes investigation in human tissue challenging. In this study, I utilised samples from the MARCH study, which was established to investigate clonal haematopoiesis. In this cohort of 195 patients, 85 were found to have a CH clone (43.5%), and of these 14 patients (16.5%) were identified with an *ASXL1*-mutant clone with VAF > 0.01 . The percentage of *ASXL1* mutations is higher than what has been recorded in other cohorts, and may be due to the prevalence of smoking amongst the cohort. Smoking has previously been linked to a higher incidence of *ASXL1* mutations in CH [53, 195], and of note, 8/14 patients with an *ASXL1* mutation in this cohort were either smokers or ex-smokers.

To characterise the cellular and transcriptomic consequences of *ASXL1* mutations across haematopoiesis, a single-cell multi-omic method called TARGET-seq+ was

performed on 11,400 cells isolated from bone marrow of 4 patients with *ASXL1*-mutant CH, and 5 age-matched healthy control patients. Of the 8,740 cells that were genotyped, 92.62% were successfully genotyped as either mutant or wildtype for *ASXL1*. This success rate is comparable to other studies using TARGET-seq or other plate-based single-cell genotyping protocols [45, 232, 260, 359].

However, the success rate for each genotyping amplicon was heavily dependent on the efficiency of the primers designed for each locus. In particular, the *ASXL1* p.G646fs locus- an insertion into a homopolymer of 8 guanines which is found recurrently mutated in myeloid malignancies- was the most challenging locus to accurately genotype. This mutation has previously been postulated to be an artifact of polymerase slippage during PCR amplification of the homopolymer, but has since been validated as a *bona fide* mutation [267, 269]. When genotyping this amplicon, PCR error cannot be ruled out, so it is likely that the number of mutant reads in all cells is modestly overestimated. Evidence for the existence of PCR error in this study is that mutant reads were detected in cells from wildtype healthy control samples, albeit at much lower counts than mutant cells. To account for this in future experiments, the VAFs from the WT healthy control samples could be used to establish a “baseline” of PCR error. Thresholds for calling mutant cells could then be calculated using this information. One potential problem with this strategy is that it may be difficult to distinguish mutant reads that were a product of *bona fide* PCR error, from those that occurred as a product cross-well contamination of genetic material from CH cells. More sophisticated mutation calling algorithms are required to improve the accuracy of genotyping for this locus.

When single-cell genotyping information was combined with immunophenotype data, several themes emerged. First, across all four samples, the mutant clone was expanded in the immunophenotypic HSC compartment. On average, adult humans possess between 50,000-200,000 HSCs [270], which sit at the apex of the haematopoietic differentiation hierarchy. Therefore, a mutant clone had to originate from a single HSC. If we use the NOC180 sample as an example, which of the four CH samples had the lowest *ASXL1* VAF, the mutant clone was detected in 30.97% of the

immunophenotypic HSC compartment. Using the lower estimate of the number of HSCs, this represents a $\approx 15,500$ -fold expansion of the mutant clone since acquisition of the mutation. This finding was verified when the mutant clone was quantified in the transcriptomically defined HSC/MPP compartment. This is consistent with the notion that clonal haematopoiesis is the product of clonal expansion in the stem and progenitor cell compartment. Secondly, the CD34+ compartment in CH samples was not significantly expanded compared to wildtype controls, consistent with the clinical observation that patients with CH are haematologically healthy.

Thirdly, there was no overt lineage bias of the mutant clone across the immunophenotypic cell populations. Previous studies modelling *ASXL1* mutations have reported a differentiation bias of *ASXL1*-mutant cells towards the myeloid lineage, and megakaryocyte lineages, and depletion of the lymphoid and erythroid lineages [144, 199, 207]. These findings in models make sense when considering the prevalence of *ASXL1* mutations in myeloid malignancies, which are characterised by the expansion of myeloid progenitor and precursor cells. This can be caused by both differentiation bias towards the myeloid lineages, and differentiation block of the lymphoid and erythroid lineages. In *ASXL1*-mutant CH however, there is no significant evidence of differentiation bias. In fact, at the immunophenotypic level, the opposite was seen. Across the four CH samples the *ASXL1*-mutant clone was modestly expanded in the MEP compartment relative to the myeloid progenitor compartments. Similarly, no obvious differentiation block of the lymphoid or erythroid lineages were seen, with the mutant clone reaching maturity in these lineages, albeit at lower frequencies. It may be that *ASXL1* mutations prime cells towards lineage-specific differentiation trajectories, but are insufficient to induce lineage bias. Subsequent acquisition of transforming mutations may then be required for lineage-biased differentiation.

One obvious limitation of this immunophenotypic analysis is that mature cells for each lineage could not be identified. Antibodies targeting markers of mature cells were included in the lineage antibody staining mix. Quantifying the mutant clone in the immunophenotypic mature cell populations may reveal lineage biases at a mature

differentiation stage than progenitor populations. As an alternative strategy, the densities of the mutant clone were quantified in each transcriptomically defined cell compartment. When the clonal densities were normalised relative to the expanded HSC compartment, it was found that there was an inefficiency in terminal maturation of the B and T lymphoid, and erythroid lineages, but not the myeloid lineage. A hypothesis may be that *ASXL1* has a greater effect on haematopoietic differentiation further towards mature cells, rather than the progenitor cells. Evidence supporting this hypothesis is the high frequency of CMML cases with *ASXL1* mutations (40-49%), a myeloid malignancy characterised by expansion of monocytes. Unfortunately, too few monocytes ($n = 35$) were sorted to accurately assess any early evidence of mutant clone expansion in *ASXL1*-mutant CH.

Lastly, there was evidence of sample-to-sample variability in CD34+ compartment composition. One notable example of this was the enrichment of the CD34+CD38+CD10+ B-NKT progenitor cells in the NOC008 and NOC180 samples. In contrast, the same compartment was barely present in the NOC137 compartment. Interestingly, in NOC180 the detected mutant clone size within the B-NKT progenitor compartment was greater than the HSC compartment, suggesting that the mutant clone is preferentially differentiating towards the lymphoid lineage. This is surprising as *ASXL1* mutations are rare in lymphoid malignancies. However, it must also be noted that the same enrichment of B-NKT cells was also observed in some WT healthy control samples. One possible explanation for this could be heightened levels environmental stimuli (e.g. CXCL12, IL7) secreted by bone marrow stromal cells, which could encourage progenitor cells to differentiate to produce more lymphocytes.

7.2 The impact of *ASXL1* mutations on the transcriptome of human HSPCs

Which molecular mechanisms are disrupted by *ASXL1* mutations and which are likely to cause clonal advantage? Given that the best-described function of *ASXL1* is epigenetic regulation, the assumption is that downstream consequences of mutation will converge upon transcription. Single-cell transcriptome data was used to answer

this question. After filtering, high quality transcriptomic data for 13,525 cells from four CH samples and 5 age-matched control samples was used to compare differences between mutant and WT cells. From this data, several broad observations emerged.

Firstly, after dimensionality reduction *ASXL1*-mutant CH samples did not form separate clusters, and were found to mix heterogeneously with cells from WT control samples. This suggests there is minimal difference in broad transcriptomic state between cells from CH and healthy bone marrow samples. Similarly, *ASXL1*-mutant cells also did not form any new clusters, and were heterogeneously mixed with WT cells from both CH and non-CH samples. This suggests that the transcriptomic differences between *ASXL1*-mutant and WT cells are subtle. Notably, this is unlike what has been observed in single-cell RNA-seq data derived from AML patients [360], where mutant cells follow vastly different transcriptomic trajectories to healthy bone marrow cells.

Secondly, the greatest amount of biological variation was observed between individual samples, regardless of whether samples were CH or non-CH. This is useful to consider a wider perspective when investigating CH in humans: humans are a genetically heterogeneous population of organisms, so how tissues adapt to their environment may differ between individuals. Additionally, throughout life individuals are exposed to a unique set of selective environmental pressures. In a group of patients with a CH clone, exactly how haematopoiesis adapts in response to environmental changes may be specific to each individual, however in all of them the mutant clone still has an advantage. It is therefore challenging to decipher exactly what selective pressures the mutation is adapting cells to, as it could be different in each individual.

Thirdly, the transcriptomic differences between mutant and wildtype cells from CH samples were very subtle. Few significantly differentially expressed genes were detected between mutant and WT cells. Upon gene set enrichment analysis, were very few pathways strongly enriched for, suggesting a very mild phenotype. This suggests that *ASXL1* mutations may only confer very modest clonal advantage, the results of which accumulate over a long period of time. This is also consistent

with the idea that CH clones are established over decades of life [68, 69]. When considering *ASXL1* mutations in the context of malignancy, this may seem counter-intuitive given that they are associated with poor prognosis- one might expect a stronger transcriptomic phenotype. A single *ASXL1* mutation clearly is insufficient to induce malignancy, but speculatively, it may be that it enhances the pathogenic effects of more potent transforming mutations to drive a more aggressive disease. Understanding how single *ASXL1* mutations affect genomic architecture may help generate more informed hypotheses.

How then do *ASXL1*-mutant clones confer a selective advantage? The transcriptomic HSC/MPP compartment was identified as the site of the greatest clonal expansion, so it was hypothesised that *ASXL1* mutations had the most obvious effect in the HSC.

When gene expression profiles between HSCs from CH samples and non-CH samples were compared, one clear observation was the enrichment in transcriptomic response to inflammation in CH cells, regardless of genotype. Specifically, there was a greater enrichment in interferon-mediated signalling rather than NF- κ B. This suggests that there is a heightened inflammatory bone marrow microenvironment in CH bone marrow. This has previously been identified in *DNMT3A*- and *TET2*-mutant CH samples [45] and suggests that increased concentration of pro-inflammatory cytokines is a common feature across CH with mutations in different genes. Assays to measure concentrations of cytokines in bone marrow plasma of *ASXL1*-mutant CH patients will be required to confirm this observation. Provided that inflammatory cytokines are selective pressures, it is challenging to know whether CH precedes the inflammation or vice versa. One hypothesis is that CH clones express more pro-inflammatory cytokines relative to WT cells, which could drive clonal selection. However, evidence suggests that CH clones are established decades before detection but are present in negligible numbers to significantly alter the bone marrow milieu. Inflammation is also a hallmark of ageing [213, 214], so a more likely scenario is that throughout most of adult life, the CH clone is selected for

by the age-associated increase in inflammation, but a CH-associated increase in inflammation accelerates the process of selection as the clone grows in size.

Interestingly however, there was no significant differences in transcriptional response to inflammation between *ASXL1*-mutant and WT HSCs from CH samples. This contrasts with what has previously been reported in *DNMT3A*- and *TET2*-mutant samples [45]. In this study, *DNMT3A*- and *TET2*-mutant HSCs reported an attenuated transcriptional response to inflammation, leading the authors to hypothesise that mutant HSC are more resistant to inflammatory stimulus. Since this was not evident in *ASXL1*-mutant HSCs, it is unlikely that inflammation is driving selection in these samples. One reductionist approach to validating this observation would be to first expand HSCs edited with an *ASXL1* mutation, and WT HSCs *in vitro*. These different populations of HSCs could then be cultured independently with different inflammatory cytokines such as interferons α , β , γ and TNF- α . Apoptosis rate, doubling rate, and colony-forming potential could be compared between the mutant and WT conditions. Performing these assays *in vitro* to assess differences in response to inflammatory cytokines is important to limit the effect of confounding extrinsic influences that are present with *in vivo* approaches. One limitation for this approach is that robust protocols to maintain human HSCs in culture need to first be established.

Another common feature between *ASXL1*-, *DNMT3A*-, and *TET2*-mutant CH bone marrow was the enrichment in aged HSC gene signatures. This would be consistent with the idea that raised inflammatory cytokine levels impair HSC function. One surprising observation was the upregulation of genes involved in B cell development in *ASXL1*-mutant and WT HSCs in CH samples. Interestingly, no lineage bias towards B lymphoid differentiation was observed in all CH samples, and may be indicative of response to external cues such as CXCL12 or IL7 which are related to expression of these genes.

Consistent with previous reports [45, 273, 300], separate subpopulations of HSCs were able to be distinguished based on transcriptomic profile. HSC2, an HSC subpopulation, was characterised by enrichment of gene signatures of inflammation.

Interestingly, specific samples- both CH and non-CH- composed this population. One *ASXL1* sample in particular- NOC082- made up the majority of cells, whereas the other *ASXL1*-mutant samples were barely present. Most mutant HSCs were found in the HSC1 subpopulation. Curiously, when comparing the mutant HSCs with WT-CH HSCs in HSC1 and HSC2 subpopulations, they showed opposite enrichments for inflammatory response gene signatures: mutant HSC2 were enriched for these gene signatures. This observation of sample-specific enrichments in different subpopulations might suggest that *ASXL1* mutations confer different mechanisms of clonal advantage dependent on the context. Studying more patients would aid in improving the robustness of this hypothesis.

One clear observation across the 4 CH samples was the transcriptomic shift of mutant HSCs towards a more ST-HSC/MPP-like, less dormant, less quiescent, and differentiation-primed state. This is contrary to what has been reported in mouse models, whereby CD150+ LT-HSCs were enriched for in aged mutant mice [207]. This observation may be counter-intuitive for clonal advantage, as maintenance of a dormant, long-term HSC-like state would be an advantage for HSC survival. Less quiescence may lead to stem cell exhaustion, telomere shortening, and HSC ageing if a more active HSC divides at a greater rate than a dormant one. Mouse models have showed that *Asxl1*-mutant HSPCs show a competitive disadvantage in serial transplantation assays, and stem cell exhaustion may explain this observation [200, 202, 207]. However, it is important to stress that increased cell division can still be a very rare event (e.g. twice a year rather than once a year), but this is impossible to prove using the present dataset. This may provide a plausible explanation for the lack of stem cell exhaustion.

Furthermore, within the HSC1 subpopulation, mutant cells showed negative enrichment of HSC-like gene signatures, and downregulation of HSC marker genes such as *MECOM* and *HLF*. In an *in vitro* model where *ASXL1* mutations were edited into human CD34+ cord blood cells using CRISPR-Cas9, the opposite was observed [144]. The difference in results may be explained by the difference in cell types being assayed. In the present study only HSCs were compared, whereas in

the *in vitro* study, the whole CD34+ fraction was included. Additionally there are likely to be intrinsic differences in transcriptional state between cord blood derived cells and aged bone marrow cells.

An intriguing observation is the upregulation of cytoskeletal genes such as *SDK2*, *PDLIM4*, and *MAGI4* in mutant HSC1. One interpretation for this could be a taxis away from an environmental pressure, for example cells expressing pro-inflammatory cytokines. Alternatively, mutant HSCs could be migrating towards a supportive cell type within the bone marrow niche, for example mesenchymal stromal cells, shown to secrete supportive factors such as SCF1 and CXCL12 [361, 362]. Cells that make up the bone marrow niche have been shown to play a critical supportive role for HSCs in mice [363–366]. Therefore, HSC interactions with supporting cells and architecture are likely to be deterministic factors in clonal selection. There is a paucity of knowledge regarding the role of the human bone marrow niche, and investigative studies are likely to provide valuable insight into CH and the onset of malignancy. A critical first step for this however, is accurately defining the human bone marrow niche. Subsequent experiments to investigate the role of different bone marrow niche cell types in clonal selection could involve their co-culture with gene-edited HSCs/HSPCs.

A prevailing hypothesis for how *ASXL1* mutations confer selective advantage is the dysregulation of polycomb target gene expression through hyperactive PR-DUB complex deubiquitinase activity. In the HSC1 subpopulation, where the mutant clone was enriched, mutant cells were only modestly enriched for polycomb target gene expression. Understanding which of these dysregulated genes correlate with the observed ST-HSC-like transcriptomic phenotype could generate refined hypotheses to explain clonal advantage. One challenge is that the role of the polycomb repressive complexes in healthy haematopoiesis is relatively understudied, despite many polycomb-group associated genes being implicated in blood cancer [367, 368]. Accurately defining which polycomb-regulated genes are importance in the maintenance of haematopoiesis could provide important insights into the mechanisms of development of a sub-group of leukaemias.

Models of *ASXL1* mutations in human cell lines have displayed a genome-wide reduction of H2AK119ub distribution [144, 197, 198, 210]. If enhanced PR-DUB complex activity is indeed responsible for clonal advantage in *ASXL1*-mutant cells, this observation suggests that this mechanism may not be limited to dysregulation of polycomb target gene expression. Furthermore, recent reports have suggested that polycomb target genes are actually repressed in cells with hyperactive PR-DUB complex activity [144, 369]. Genome wide profiling of H2AK119ub distribution of mutant cells from primary human samples is therefore likely to shed some light on the mechanisms driving the observed transcriptional phenotype.

7.3 GCUT: a novel single-cell method to simultaneously assay genotype and histone modification distribution in heterogeneous cell populations

As previously discussed, mutant *ASXL1* causes an enhancement of PR-DUB complex activity, leading to genome wide reduction of H2AK119ub. It is not known however, if the action of hyperactive PR-DUB complex is target-specific or random. Identifying the precise genomic loci at which this occurs may reveal the molecular mechanisms directly dysregulated by mutant *ASXL1*. Several other groups have previously sought to address this using bulk approaches to assay genomic H2AK119ub distribution in both mouse and human tissue [144, 199, 200, 207, 370]. However, a major failing of these approaches is that they do not capture the complex heterogeneous cellular and molecular conditions that characterize age-related pre-leukaemic and leukaemic states.

To address this, I developed a novel plate-based single-cell method, GCUT, that simultaneously captures epigenetic profile as well as genotype from single cells isolated from heterogeneous tissue. GCUT would therefore enable the comparison of epigenetic post-translational modification landscapes in mutant and wildtype cells. The aims addressed in this thesis were to develop and benchmark the GCUT method to assess the suitability for use in primary human *ASXL1*-mutant CH samples.

Through the development of GCUT, it was possible to amplify CUT&Tag libraries from single cells isolated by FACS sorting, whilst simultaneously amplifying up to five separate genomic loci using targeted primers. It is theoretically possible to target greater numbers of genomic loci, enabling the use of GCUT to assay malignancies with high mutational burden such as solid tumours. Genotyping accuracy may suffer however, if targeting more loci causes an increase in allelic dropout (ADO) rate due to increased competition for resources in the PCR reaction. Additionally, CUT&Tag libraries were able to be generated from single cells using primary antibodies targeting three separate histone marks (H2AK119ub, H3K27me3 and H3K4me3). Theoretically, using the appropriate primary antibodies, any genomic component has the potential to be profiled with GCUT.

GCUT was benchmarked using a cell mixing sequencing experiment. One of the most important requirements for GCUT for the use in primary human tissue is the ability to genotype cells with high accuracy and confidence. Across the 5 genomic loci that were genotyped in single cells from 3 cell lines, a mean of 74.83% of cells were successfully genotyped as either mutant or wildtype for each locus. Whilst this is significantly higher than what was recorded for droplet-based single-cell genotyping strategies, it is lower than GTAC and TARGET-seq - two plate-based methods that use the same genotyping strategy [232, 260, 350]. The range of ADO rates for the WT allele across the 5 genomic loci was 6.38 – 20.96%, again slightly worse than other plate-based single-cell genotyping strategies.

One possible reason for the difference in genotyping success and ADO rates could be the preamplification PCR protocol. ADO is most likely to occur within the first couple of cycles of PCR, resulting in the preferential, exponential amplification of one allele. Optimisation of this pre-amplification PCR protocol to efficiently amplify both single-cell genotyping amplicons and CUT&Tag libraries will be required to potentially improve the genotyping success rate. One suggestion may be to lengthen the combined degradation and annealing step to increase the probability of the genotyping primers annealing to both alleles. This would also retain the

short elongation step (10 seconds), which is required to bias the CUT&Tag library towards amplification of mononucleosomal fragments.

One of the major limitations of GCUT is the complexity of the CUT&Tag libraries. The average number of unique nuclear fragments per cell ranged from 144-434, depending on the histone mark. For H2AK119ub, and H3K4me3, these low-complexity libraries were insufficient to distinguish distinct clusters using dimensionality reduction. Whilst distinct clusters could be distinguished when assaying H3K27me3, after filtering low quality cells only 59.16% remained. The number of unique nuclear fragments for all three histone marks is comparable to what was reported for pre-existing single-cell CUT&Tag methods [242, 349]. However, since both these methods were high-throughput droplet-based methods, high cell numbers increased the overall complexity of the dataset and distinct clusters could be identified.

When assaying rare progenitor populations using plate-based single-cell methods, maximising the percentage of high-quality cells from (ideally) a large number of assayed cells is crucial. This enables high-powered comparisons between different cell states. Increasing the number of unique nuclear fragments per cell will be crucial for the viability of using GCUT in primary human tissue. Hypotheses explaining low library complexity could be: low primary antibody specificity; sub-optimal primary antibody incubation conditions (e.g. concentration of antibody, buffer content); or inefficient transposition by the pAG-Tn5 transposase. Since the duplication rate for all conditions was incredibly high ($\approx 99\%$), the PCR protocol is unlikely to be the cause of the inefficient capture of unique fragments. The major limiting factor for library complexity is the specificity of the primary antibody. The pAG-Tn5 transposase will only bind to genomic sites where immunoglobulin proteins are present. Therefore, if a low percentage of the epitope of interest across the genome is bound by primary antibodies due to lack of specificity, a maximum of that percentage can be assayed per cell. Whilst all hypotheses should be tested to optimise the library complexity, the most important first step would be to test different antibodies for specificity.

Similarly, cell recovery needs to be optimised. From starting with $\approx 100,000$ cells per condition, $\approx 15,000$ were recovered from the bulk CUT&Tag process. Single-nuclei are then sorted from this number, which is also a very inefficient process, recovering $\approx 20\%$ of the input. Immunophenotypic HSCs take up roughly 15% of the CD34+ compartment of the *ASXL1*-CH samples from the MARCH cohort. The mutant fraction of the HSC compartment in these samples is an average of 43%. Therefore the inefficiency of cell recovery currently makes it very difficult to perform powerful analyses between mutant and wildtype cells.

One potential reason for poor cell recovery during the CUT&Tag process is the number of centrifuge steps. Currently in GCUT, a reduced number of 7 centrifuge steps are used, which limits the number of cell washing steps. One recent publication by Bartosovic et al. reduced the number of centrifuge steps of single-cell CUT&Tag, recovering up to 68% of cells [355]. This was achieved by replacing the independent secondary antibody incubation and Tn5 tagmentation steps with a single incubation with a nanobody-Tn5 transposase fusion protein. Interestingly, this method also drastically improved the number of unique nuclear fragments per cell. Implementing the upstream modified CUT&Tag steps may be a solution to optimise both cell recovery and library complexity.

Despite the current limitations of GCUT, the potential of the method extends beyond that of assaying epigenetic modifications in CH. Applying GCUT to a range of diseases resulting from genetic alterations could provide insights into epigenetic foundations of disease development and progression. One of the major obstacles for the optimisation of GCUT is the number of cells that can be assayed at one time. Droplet-based single-cell CUT&Tag implementations currently show similar sensitivity to GCUT, whilst assaying tens of thousands of cells. Developing a more sensitive method of droplet-based single-cell gDNA genotyping than those which currently exist [350, 371], and can be incorporated with scCUT&Tag, is therefore likely to be a long-term solution to GCUT.

7.4 Study limitations and future work

This study is the first of its kind to assay the consequences of *ASXL1* mutations upon haematopoiesis within humans without malignancy. However it is not without limitations.

Firstly, the limited number of *ASXL1*-mutant samples studied is likely to hamper accuracy of analysis. This is particularly relevant given the amount of biological variation observed between individuals, even with the same mutation. The analysis of more samples will increase the depth of the dataset and generate more granular hypotheses of how *ASXL1* mutations confer clonal advantage. Additionally, compared to *DNMT3A* and *TET2* mutations, the lower prevalence of *ASXL1* mutations in CH limits the number of samples available. Additionally, the low VAFs observed CH limits the number of samples with *ASXL1* VAF large enough to warrant assaying. Similarly, the number of cells assayed from each compartment limits the number of analyses possible, particularly among the more mature cell types. It is regrettable that more monocytes, and erythroblasts were not sorted, since these are the cell types in which two of the haematological diseases in which *ASXL1* mutations are prevalent: CMML and aplastic anaemia respectively.

The MARCH study also only collects single time-point samples. Assaying these samples therefore enables only a “snapshot” perspective of molecular activity in mutant and wildtype cells. Evidence suggests that clonal selection is a dynamic process which happens over decades of life. It is entirely possible that the inter-sample variation recorded here may be due to a specific event in the patient’s life just before sample collection, for example viral infection. Therefore, the differences between wild-type and mutant cells described here may not be representative of how the clones have behaved as they have developed through life. Longitudinal sampling is therefore important in future studies. Additionally, bone marrow sample collection is a very invasive procedure which subjects cells to a great deal of environmental

stresses. This therefore has the potential to affect gene expression and may not give a true reflection of the transcriptomic profiles *in situ*.

One basic view of clonal haematopoiesis is that the function of the stem cell is changed resulting from somatic mutation. The use of scRNAseq to distinguish different functional change of stem cells is therefore limited. Change in transcription may not always result in change in function, owing to the action of numerous intermediary regulatory mechanisms. Instead, scRNAseq is a powerful tool to identify changes in functional potential, which is important in estimating stem cell properties. As previously discussed, epigenetic machinery plays an important role in control of transcription. Mutations in genes encoding epigenetic regulators are the most common in CH, so profiling epigenetic changes such as chromatin, histone mark distribution, or DNA methylation patterns between different cell states may provide a more detailed analysis of stem cell potential. This was the aim of developing GCUT. The limitations of GCUT have been discussed in the previous section, and needs to be optimised further to warrant use in primary human tissue. However, it may also provide supporting evidence for, or refine, hypotheses generated from TARGET-seq+ data. Alternative approaches to assay functional potential of cells may be to use technologies such as GTAC or GoTChA which assay differences chromatin accessibility between mutant and WT cells [260, 350].

For methods such as TARGET-seq+ and GCUT, the reliance of the genotyping strategy on PCR can impact accuracy of genotyping through factors such as ADO, or polymerase error. This was particularly evident with the *ASXL1*p.G646WfsX12 mutation genotyped in NOC008 and NOC082, which has previously been attributed as polymerase error. Allelic dropout rate could not be estimated for these amplicons, so it is likely the number of mutant cells is overestimated. However, the fact that transcriptomic differences existed between called mutant and WT cells is evidence that the mutation is real, and that mutant cells had been called. The advent of fourth-generation next-generation sequencing technology has opened the possibility of genotyping from single cells with limited use of PCR [372]. The use of long-read NGS technologies such as Oxford Nanopore Technologies' nanopore sequencing

methods could enable mutations to be detected directly from purified nucleic acid libraries. In the case of TARGET-seq+, this would allow genotyping from purified cDNA libraries without the need of additional genotyping PCR steps. This strategy would assume expression of the mutated gene but may provide a more accurate method of genotyping by reducing the probability of ADO.

Ultimately however, these technologies are limited to produce hypotheses rather than develop causative links between mutation and clonal advantage. This study has revealed evidence that *ASXL1* mutations may confer clonal advantage through multiple molecular mechanisms, rather than one. Specifically, in HSCs, these include the activation of a less-dormant, differentiation-primed transcriptomic state, and upregulation of polycomb target genes. The extent to which these mechanisms confer clonal advantage may also differ between cellular compartments. It is entirely possible that these mechanisms are interlinked, but this will only be identified by exploring these mechanisms on a functional level. To test these hypotheses of clonal expansion, it will be necessary to perform functional assays such as CRISPR screens using the genes found to be dysregulated in *ASXL1*-mutant cells. Ideally, primary human cells should also be used for these assays to correctly assess the impact of dysregulation of each gene upon clonal expansion. Only through understanding the human condition can accurate models of *ASXL1*-mutant CH be developed to test reductionist hypotheses of clonal expansion.

Appendices

Chapter A

Appendices

Contents

A.1	<i>ASXL1</i> -mutant samples of the MARCH cohort.	204
A.2	Antibodies	206
A.3	TARGET-seq+ reagents	207
A.4	GCUT reagents	209
A.5	TARGET-seq+ oligonucleotides	212
A.6	GCUT oligonucleotides	214
A.7	Marker genes of HSC characteristics and subpopulations	216

A.1 *ASXL1*-mutant samples of the MARCH cohort.

Sample	Gene	Chromosome	Position	Type	Reference	Alternate	VAF	Variant classification	Accession	Exon	cDNA change	Protein change
NOC008	ASXL1	chr20	32434638	INDEL	A	AG	0.1665	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
NOC008	PTPN11	chr12	112450496	SNP	G	A	0.0712	nonsynonymous_SNV	NM_002834	exon3	c.316G>A	p.D106N
NOC008	GNB1	chr1	1815790	SNP	T	C	0.0134	nonsynonymous_SNV	NM_002074	exon5	c.169A>G	p.K57E
NOC008	CBL	chr11	119278532	SNP	C	G	0.0107	nonsynonymous_SNV	NM_005188	exon9	c.1250C>G	p.P417R
NOC008	TP53	chr17	7674872	SNP	T	C	0.0159	nonsynonymous_SNV	NM_000546	exon6	c.659A>G	p.Y220C
NOC010	ASXL1	chr20	32434638	INDEL	A	AG	0.0363	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
NOC010	DNMT3A	chr2	25248036	SNP	C	G	0.0118	splicing	NM_022552	exon7	c.855+1G>C	p.E285X
NOC016	ASXL1	chr20	32434638	INDEL	A	AG	0.0155	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
NOC017	ASXL1	chr20	32433373	SNP	C	A	0.0162	stopgain	NM_015338	exon11	c.1175C>A	p.S392X
NOC030	ASXL1	chr20	32433315	SNP	C	T	0.0105	stopgain	NM_015338	exon11	c.1117C>T	p.Q373X
NOC030	NF1	chr17	31232164	SNP	G	T	0.0206	stopgain	NM_001042492	exon25	c.3289G>T	p.E1097X
NOC050	ASXL1	chr20	32434638	INDEL	A	AG	0.0229	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
NOC064	ASXL1	chr20	32433739	INDEL	CTG	C	0.012	frameshift_deletion	NM_015338	exon11	c.1542_1543del	p.V515Gfs
NOC064	ASXL1	chr20	32434825	SNP	G	T	0.0324	stopgain	NM_015338	exon12	c.2113G>T	p.E705X
NOC064	DNMT3A	chr2	25247714	SNP	C	A	0.0149	nonsynonymous_SNV	NM_022552	exon8	c.891G>T	p.W297C
NOC064	KMT2E	chr7	105110540	SNP	C	A	0.0492	stopgain	NM_182931	exon25	c.3908C>A	p.S1303X
NOC064	TET2	chr4	105235611	INDEL	CAGCACTA	C	0.0242	frameshift_deletion	NM_001127208	exon3	c.1670_1676del	p.Q557fs
NOC082	ASXL1	chr20	32434638	INDEL	A	AG	0.0804	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs
NOC092	ASXL1	chr20	32434773	SNP	T	A	0.0337	stopgain	NM_015338	exon12	c.2061T>A	p.C687X
NOC092	DNMT3A	chr2	25235749	SNP	A	C	0.0221	nonsynonymous_SNV	NM_022552	exon22	c.2555T>G	p.M852R
NOC096	ASXL1	chr20	32434636	INDEL	GGAGGGGGGGGT	G	0.0194	frameshift_deletion	NM_015338	exon12	c.1925_1935del	p.G642fs
NOC096	SMC1A	chrX	53414817	SNP	T	A	0.021	nonsynonymous_SNV	NM_006306	exon3	c.352A>T	p.S118C
NOC137	ASXL1	chr20	32435605	SNP	C	T	0.1413	stopgain	NM_015338	exon12	c.2893C>T	p.R965X
NOC137	TET2	chr4	105243621	SNP	C	T	0.0405	stopgain	NM_001127208	exon6	c.3646C>T	p.R1216X
NOC151	ASXL1	chr20	32436833	INDEL	T	TG	0.0161	frameshift_insertion	NM_015338	exon12	c.4122dupG	p.V1374fs
NOC151	JAK2	chr9	5073770	SNP	G	T	0.0106	nonsynonymous_SNV	NM_004972	exon14	c.1849G>T	p.V617F
NOC180	ASXL1	chr20	32433562	INDEL	AGACT	A	0.0764	frameshift_deletion	NM_015338	exon11	c.1365_1368del	p.K455fs

Table A.1: *ASXL1*-mutant samples of the MARCH cohort. List of variants detected by the targeted sequencing panel of BMMNCs in the 13 *ASXL1*-mutant CH samples. Genomic positions are as annotated by HG38.

A.2 Antibodies

Antibody	Fluorochrome	Clone	Isotype	Manufacturer	Catalogue No.	Dilution
Anti-human CD45RA	BB515	HI100	Mouse IgG2b, κ	BD	564552	1:40
Anti-human CD90	PE/Cy7	5E10	Mouse IgG1, κ	BioLegend	328124	1:20
Anti-human CD123	PE	6H6	Mouse IgG1, κ	BioLegend	306006	1:40
Anti-human CD49f	PE/Dazzle 594	GoH3	Rat IgG2a, κ	BioLegend	313626	1:160
Anti-human CD38	BV421	HIT2	Mouse IgG1, κ	BioLegend	303526	1:20
Anti-human CD10	BV605	HI10a	Mouse IgG1, κ	BioLegend	312222	1:40
Anti-human CD34	APC	581	Mouse IgG1, κ	BioLegend	343510	1:160
Anti-human CD117 c-kit	BV785	104D2	Mouse IgG1, κ	BioLegend	313238	1:40
Anti-human CD41	BV510	HIP8	Mouse IgG1, κ	BioLegend	303736	1:40
Lineage master mix						
Anti-human CD2	PE/Cy5	RPA-2.10	Mouse IgG1, κ	BioLegend	300210	1:160
Anti-human CD3	PE/Cy5	HIT3a	Mouse IgG2a, κ	BioLegend	300310	1:320
Anti-human CD4	PE/Cy5	RPA-T4	Mouse IgG1, κ	BioLegend	300510	1:160
Anti-human CD8a	PE/Cy5	RPA-T8	Mouse IgG1, κ	BioLegend	301010	1:320
Anti-human CD11b	PE/Cy5	ICRF44	Mouse IgG1, κ	BioLegend	301308	1:160
Anti-human CD14	PE/Cy5	61D3	Mouse IgG1, κ	eBioscience	15-0149-42	1:160
Anti-human CD19	PE/Cy5	HIB19	Mouse IgG1, κ	BioLegend	302210	1:160
Anti-human CD20	PE/Cy5	2H7	Mouse IgG2b, κ	BioLegend	302308	1:160
Anti-human CD56	PE/Cy5	MEM188	Mouse IgG2a, κ	BioLegend	304608	1:80
Anti-human CD235ab	PE/Cy5	HIR2	Mouse IgG2b, κ	BioLegend	306606	1:320

Table A.2: Antibodies used for immunophenotypic analysis and FACS sorting of BMMNCs.

A.3 TARGET-seq+ reagents

Reagent	Reaction concentration	1 cell (μl)
Poly-ethylene Glycol 8000 (40% solution)	0.05	0.5
Triton X-100 (10% solution)	0.1%	0.03
ERCC 1:4e5	-	0.015
RNAse Inhibitor (40u/uL)	0.5u/uL	0.04
dNTPs (10mM/each)	0.5mM/each	0.2
Protease (1.09 AU/mL)	2.7E-05 AU/μL	0.075
Nuclease Free Water		1.94
Total		2.8
OligodT barcode (10uM)	0.5uM	0.2
Total with barcode		3

Table A.3: Composition of lysis buffer used for TARGET-seq+.

RT mix	Reaction concentration	1 cell (μl)
Tris-HCl pH 8.3 (1M)	25mM	0.1
NaCl (1M)	30mM	0.12
MgCl ₂ (100mM)	2.5mM	0.1
GTP (100mM)	1mM	0.04
DTT (100mM)	8mM	0.32
cDNA primers (25uM)	70nM	0.112
Nuclease-free water	n/a	Variable
RNAse Inhibitor (40u/μL)	0.5u/μL	0.05
SS2 TSO-LNA (100μM)	2uM	0.08
Maxima H-minus RT enzyme (200U/μL)	2u/μL	0.04
RT mix		1.00
TOTAL (Cumulative)		4.00

Table A.4: Composition of reverse transcription reaction used for capture of single-cell whole-transcriptome libraries.

PCR mix	Reaction concentration	1 cell (μl)
Kapa HiFi HotStart buffer (5X)	1X	2.00
dNTPs (10mM/each)	0.3mM/each	0.30
MgCl ₂ (100mM)	0.5mM	0.05
ISPCR Primer (10uM)	0.05uM	0.050
Polymerase (1U/uL)	0.1U/uL	0.200
cDNA primers (25uM)	28nM	0.01
gDNA primers	400nM	0.04
Nuclease-free water	n/a	variable
TOTAL (PCR mix + primers + water)		6.00
TOTAL (Cumulative)		10.00

Table A.5: Composition of preamplification reaction of whole-transcriptome libraries and genotyping amplicons from single cells.

A.4 GCUT reagents

Reagent	Catalogue number	Manufacturer
HEPES pH7.6 1M	15630-106	Gibco
NaCl 5M	AM9760G	invitrogen
Spermidine 100X	27287S	Cell signalling technologies
Protease Inhibitor tablets	11873580001	Merck
BSA 30%	A8577-50ML	Sigma-Aldrich
MgCl ₂ 1M	AM9530G	invitrogen
Digitonin powder	300410-250MG	Sigma-Aldrich
NP40 10%	85124	Thermofisher
EDTA 0.5M	AM9260G	invitrogen
Rabbit anti-H3K27me3 primary Ab (C36B11) (1:50)	9733T	Cell signalling technologies
Rabbit anti-H2AK119Ub primary Ab (D27C4) (1:50)	8240S	Cell signalling technologies
Rabbit anti-H3K4me3 primary Ab (C42D8) (1:50)	9751T	Cell signalling technologies
Rabbit anti-IgG primary Ab (DA1E) (1:10)	3900S	Cell signalling technologies
Guinea-pig anti-rabbit IgG Secondary Ab (1:100)	NBP1-72763	Novus
pAG-Tn5	15-107	EpiCypher
DAPI solution (1mg)	564907	BD Biosciences
Q5 High-Fidelity 2X Master Mix	M0492L	New England Biolabs

Table A.6: List of reagents used for the CUT&Tag protocol in GCUT.

Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Nuclease-free water	n/a	n/a	3738
HEPES pH7.6	1M	0.02M	84
NaCl	5M	0.15M	126
Spermidine	100M	1M	42
Protease Inhibitor	20X	1X	210
		Total volume (uL)	4200

Washing Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Buffer	n/a	n/a	1867
BSA	30%	2%	133.3
		Total volume (uL)	2000

Dig 300 Wash Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Buffer	n/a	n/a	1162
Digitonin	5%	0.01%	1.2
NP40	10%	0.01%	1.2
NaCl	5M	0.15M	36
		Total volume (uL)	1200

Antibody Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Washing Buffer	n/a	n/a	596
EDTA	0.5M	0.002M	2.4
Digitonin	5%	0.01%	0.6
NP40	10%	0.01%	0.6
		Total volume (uL)	600

Tagmentation Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Dig 300 Wash Buffer	n/a	n/a	396
MgCl2	1M	0.01M	4
		Total volume (uL)	400

Dig Washing Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Washing Buffer	n/a	n/a	798
Digitonin	5%	0.01%	0.8
NP40	10%	0.01%	0.8
		Total volume (uL)	800

Stop Buffer			
Reagent	Initial concentration	Final concentration	Volume (uL)
Dig 300 Wash Buffer	n/a	n/a	190
EDTA	0.5M	0.025M	10
		Total volume (uL)	200

Table A.7: Composition of buffers used in the CUT&Tag reactions for GCUT.

Reagent	Initial concentration	Final concentration	1 cell (uL)
Tris-HCl pH 8	1M	0.05M	0.2
NaCl	5M	0.05M	0.04
SDS 10%	10%	0.20%	0.08
Nuclease-free water	n/a	n/a	1.68
Barcoded i5 primer	20uM	5uM	1
Barcoded i7 primer	20uM	5uM	1
Total			4

Table A.8: Reagents used in lysis buffer for GCUT.

Reagent	Initial concentration	Final concentration	1 reaction (uL)
NEB Q5 mastermix	2X	1X	8
Water	n/a	n/a	variable
gDNA primers	10uM	100nM	0.167
Total (PCR mix + primers)	n/a	n/a	10
Sample (with Tween-20)	n/a		6
Total (Cumulative)			16

Table A.9: Reagents used in pre-amplification reaction in GCUT.

A.5 TARGET-seq+ oligonucleotides

N.B. All oligonucleotides listed below are read in the 5'-3' direction.

Oligo name	Oligo type	Sequence	Purification
TSO-LNA	RNA LNA	AAGCAGTGGTATCAACGCAGAGTACATrGrG+G	Rnase free HPLC
ISPCR primer	DNA	AAGCAGTGGTATCAACGCAGAGT	HPLC
i7 index primer (Nextera)	DNA	CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG	HPLC
i5-index primer (custom)	DNA	AATGATACGGCGACCACCGAGATCTACAC[i5]GCCTGTCCGCG GAAGCAGTGGTATCAACGCAGAGT*T*G	HPLC
P5-SEQ	DNA	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTTGC*T	PAGE
i5-SEQ	DNA	AGCAACTCTGCGTTGATACCACTGCTCCGCGGACAGG*C	PAGE
LCS1	DNA	GGCGACCACCGAGATCTACACTGACGACATGGTTCTACA	HPLC
CS2	DNA LNA	T+AC+GGT+AGCAGAGACTTGGTCT	HPLC
CS2rc	DNA LNA	A+GAC+CA+AGTCTCTGCTACCGTA	HPLC

Table A.10: List of oligonucleotides used throughout TARGET-seq+ experiments.

Sample	Gene	Exon	Target	Forward primer sequence	Reverse primer sequence
NOC 008	ASXL1	13	gDNA	AGCCCACTTACCAGATATGCC	TCCTCTCTCCTAGCTCTGGAC
	PTPN11	3	gDNA	TTGCCTCCCTTTCCAATGGAC	ACACAGACCGTCATGCATTTTC
	PTPN11	3	cDNA	GTGTGGAGGCAGAAAACTAC	TGCTCTCCCTTTGTGCATCAC
	GNB1	5	gDNA	ACACAGGTTCTGATGACCAG	GGAATATGCACCTGTGAACGC
	GNB1	5	cDNA	GAAGGTGGCATTGAAGAGCAC	GGCATATGCACAGGTCATGAC
	CBL	9	gDNA	TCTTACATCCTGGCAGGTACG	GCACCAGCCAATTCCTTCATC
	TP53	6	gDNA	GTGCCCTGACTTTCAACTCTG	ACTGACAACCACCCTTAACCC
NOC082	ASXL1	13	gDNA	AGCCCACTTACCAGATATGCC	TCCTCTCTCCTAGCTCTGGAC
NOC137	ASXL1	13	gDNA	AGCCCAAACCAGAATCCAGAG	ATCCTTGTCCACCGAAGATCC
	TET2	6	cDNA	AGGCAAAAGTTCTCAGGGATGT	GGTCATCCCCAAGCAGCTTA
	TET2	6	gDNA	TACCCAGATCAAAGACTCATGC	AGTGCACGCTGAACTCTCTTC
NOC180	ASXL1	12	gDNA	AGGCAGTTGGAGCTATGAGTC	AGGACTGACGATCTTCAAGCC

Table A.11: List of genotyping primers used in the RT-PCR and pre-amplification reactions of TARGET-seq+.

Table A.12: List of Barcoded primers used in the PCR1 genotyping reactions of TARGET-seq+. CS1 (forward primer) and CS2 (reverse primer) sequences are located at the 5' end. Plate-specific barcodes are in lower case. Locus-specific sequences are at the 3' end.

A.6 GCUT oligonucleotides

Primer ID	Barcode	P5 sequence	i5 adapter	Full Sequence
GCUT_i5_index_501	TAGATCGC	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACTAGATCGC TCGTCGGCAGCGTC
GCUT_i5_index_502	CTCTCTAT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACCTCTCTAT TCGTCGGCAGCGTC
GCUT_i5_index_503	TATCCTCT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACTATCCTCT TCGTCGGCAGCGTC
GCUT_i5_index_504	AGAGTAGA	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACAGAGTAGA TCGTCGGCAGCGTC
GCUT_i5_index_505	GTAAGGAG	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACGTAAGGAG TCGTCGGCAGCGTC
GCUT_i5_index_506	ACTGCATA	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACACTGCATA TCGTCGGCAGCGTC
GCUT_i5_index_507	AAGGAGTA	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACAAGGAGTA TCGTCGGCAGCGTC
GCUT_i5_index_508	CTAAGCCT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACCTAAGCCT TCGTCGGCAGCGTC
GCUT_i5_index_510	CGTCTAAT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACCGTCTAAT TCGTCGGCAGCGTC
GCUT_i5_index_511	TCTCTCCG	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACTCTCTCCG TCGTCGGCAGCGTC
GCUT_i5_index_513	TCGACTAG	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACTCGACTAG TCGTCGGCAGCGTC
GCUT_i5_index_515	TTCTAGCT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACTTCTAGCT TCGTCGGCAGCGTC
GCUT_i5_index_516	CCTAGAGT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACCCTAGAGT TCGTCGGCAGCGTC
GCUT_i5_index_517	GCCTAAGA	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACGCTAAGA TCGTCGGCAGCGTC
GCUT_i5_index_518	CTATTAAG	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACCTATTAAG TCGTCGGCAGCGTC
GCUT_i5_index_520	AAGGCTAT	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACAAGGCTAT TCGTCGGCAGCGTC
GCUT_i5_index_521	GAGCCTTA	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACGAGCCTTA TCGTCGGCAGCGTC
GCUT_i5_index_522	TTATGCGA	AATGATACGGCGACCAACCGAGATCTACAC	TCGTCGGCAGCGTC	AATGATACGGCGACCAACCGAGATCTACACTTATGCGA TCGTCGGCAGCGTC

Table A.13: List of custom i5-indexed primers included in the lysis buffer for GCUT.

Cell line	Gene	Exon	Target	Forward primer sequence	Reverse primer sequence
K562	ASXL1	12	qDNA	ACTTTCACGTATCAAACCAACC	CGGGCTTTAATGTCTGCGAG
K562	HOXA9	1	qDNA	TGCTTGTGTTCTCCTCCAG	CCAACTCGAGAGGCTTCCAG
OCI-AML3	NRAS	3	qDNA	CCACACCCCCAGGATTCCTTAC	TGCTCCTAGTACCTGTAGAGG
OCI-AML3	DNMT3A	23	qDNA	CCCTCTCTGCTTTTCTCCC	tgTCGCTACCTCAGTTTGCC
JURKAT	RUNX1	5	qDNA	GGGATGTTCCAGATGGCACT	TGGGTTTGTGCCATGAAACG

Table A.14: List of genotyping primers used in the pre-amplification reactions of GCUT cell mixing experiment.

Cell line	Gene	Exon	Target	Forward primer sequence	Reverse primer sequence
K562	ASXL1	12	gDNA	ACACTGACGACATGGTTCTACAaactgcCCCTGGGTGGTTAAAGGTCAG	TACGGTAGCAGAGACTTGGTCTaaagaccCAGGAGGACTCCGTGGTGG
				ACACTGACGACATGGTTCTACAacclcaaCCCTGGGTGGTTAAAGGTCAG	TACGGTAGCAGAGACTTGGTCTccagttCAGGAGGACTCCGTGGTGG
				ACACTGACGACATGGTTCTACAagactgCCCTGGGTGGTTAAAGGTCAG	TACGGTAGCAGAGACTTGGTCTtgacaaCAGGAGGACTCCGTGGTGG
				ACACTGACGACATGGTTCTACAagttgaCCCTGGGTGGTTAAAGGTCAG	TACGGTAGCAGAGACTTGGTCTgaacagCAGGAGGACTCCGTGGTGG
				ACACTGACGACATGGTTCTACAaactccCCCTGGGTGGTTAAAGGTCAG	TACGGTAGCAGAGACTTGGTCTacttccCAGGAGGACTCCGTGGTGG
				ACACTGACGACATGGTTCTACAagtaagCCCTGGGTGGTTAAAGGTCAG	TACGGTAGCAGAGACTTGGTCTtcttgCAGGAGGACTCCGTGGTGG
K562	HOXA9	1	gDNA	ACACTGACGACATGGTTCTACAaactccCGAAGGCGCCTTCTCTGAAA	TACGGTAGCAGAGACTTGGTCTaaagaccCGGCGCGGATTGGAAG
				ACACTGACGACATGGTTCTACAacclcaaCGAAGGCGCCTTCTCTGAAA	TACGGTAGCAGAGACTTGGTCTccagttCGGCGCGGATTGGAAG
				ACACTGACGACATGGTTCTACAagactgCGAAGGCGCCTTCTCTGAAA	TACGGTAGCAGAGACTTGGTCTtgacaaCGGCGCGGATTGGAAG
				ACACTGACGACATGGTTCTACAagttgaCGAAGGCGCCTTCTCTGAAA	TACGGTAGCAGAGACTTGGTCTgaacagCGGCGCGGATTGGAAG
				ACACTGACGACATGGTTCTACAaactccCGAAGGCGCCTTCTCTGAAA	TACGGTAGCAGAGACTTGGTCTacttccCGGCGCGGATTGGAAG
				ACACTGACGACATGGTTCTACAagtaagCGAAGGCGCCTTCTCTGAAA	TACGGTAGCAGAGACTTGGTCTtcttgCGGCGCGGATTGGAAG
OCI-AML3	NRAS	3	gDNA	ACACTGACGACATGGTTCTACAaactgcACCTGTTTGTGGACATACTGGA	TACGGTAGCAGAGACTTGGTCTaaagaccTACACAGAGGAAGCCTTCGC
				ACACTGACGACATGGTTCTACAacclcaaACCTGTTTGTGGACATACTGGA	TACGGTAGCAGAGACTTGGTCTccagttTACACAGAGGAAGCCTTCGC
				ACACTGACGACATGGTTCTACAagactgACCTGTTTGTGGACATACTGGA	TACGGTAGCAGAGACTTGGTCTtgacaaTACACAGAGGAAGCCTTCGC
				ACACTGACGACATGGTTCTACAagttgaACCTGTTTGTGGACATACTGGA	TACGGTAGCAGAGACTTGGTCTgaacagTACACAGAGGAAGCCTTCGC
				ACACTGACGACATGGTTCTACAaactccACCTGTTTGTGGACATACTGGA	TACGGTAGCAGAGACTTGGTCTacttccTACACAGAGGAAGCCTTCGC
				ACACTGACGACATGGTTCTACAagtaagACCTGTTTGTGGACATACTGGA	TACGGTAGCAGAGACTTGGTCTtcttgTACACAGAGGAAGCCTTCGC
OCI-AML3	DNMT3A	23	gDNA	ACACTGACGACATGGTTCTACAaactgcTCTCCCCAGGGTATTTGGT	TACGGTAGCAGAGACTTGGTCTaaagaccTTTGCCCCCATGTCCCTTAC
				ACACTGACGACATGGTTCTACAacclcaaTCTCCCCAGGGTATTTGGT	TACGGTAGCAGAGACTTGGTCTccagttTTTGCCCCCATGTCCCTTAC
				ACACTGACGACATGGTTCTACAagactgTCTCCCCAGGGTATTTGGT	TACGGTAGCAGAGACTTGGTCTtgacaaTTTGCCCCCATGTCCCTTAC
				ACACTGACGACATGGTTCTACAagttgaTCTCCCCAGGGTATTTGGT	TACGGTAGCAGAGACTTGGTCTgaacagTTTGCCCCCATGTCCCTTAC
				ACACTGACGACATGGTTCTACAaactccTCTCCCCAGGGTATTTGGT	TACGGTAGCAGAGACTTGGTCTacttccTTTGCCCCCATGTCCCTTAC
				ACACTGACGACATGGTTCTACAagtaagTCTCCCCAGGGTATTTGGT	TACGGTAGCAGAGACTTGGTCTtcttgTTTGCCCCCATGTCCCTTAC
JURKAT	RUNX1	5	gDNA	ACACTGACGACATGGTTCTACAaactgcTCACTGTGATGGCTGGCAAT	TACGGTAGCAGAGACTTGGTCTaaagaccTCCACTTCGACCGACAAACC
				ACACTGACGACATGGTTCTACAacclcaaTCACTGTGATGGCTGGCAAT	TACGGTAGCAGAGACTTGGTCTccagttTCCACTTCGACCGACAAACC
				ACACTGACGACATGGTTCTACAagactgTCACTGTGATGGCTGGCAAT	TACGGTAGCAGAGACTTGGTCTtgacaaTCCACTTCGACCGACAAACC
				ACACTGACGACATGGTTCTACAagttgaTCACTGTGATGGCTGGCAAT	TACGGTAGCAGAGACTTGGTCTgaacagTCCACTTCGACCGACAAACC
				ACACTGACGACATGGTTCTACAaactccTCACTGTGATGGCTGGCAAT	TACGGTAGCAGAGACTTGGTCTacttccTCCACTTCGACCGACAAACC
				ACACTGACGACATGGTTCTACAagtaagTCACTGTGATGGCTGGCAAT	TACGGTAGCAGAGACTTGGTCTtcttgTCCACTTCGACCGACAAACC

Table A.15: List of Barcoded primers used in the PCR1 genotyping reactions of the GCUT cell mixing experiment. CS1 (forward primer) and CS2 (reverse primer) sequences are located at the 5' end. Plate-specific barcodes are in the middle of the sequence. Locus-specific sequences are at the 3' end.

A.7 Marker genes of HSC characteristics and subpopulations

Bone Marrow LT-HSC	Bone Marrow ST-HSC	Dormant BM HSC	Quiescent HSC	Activated HSC	HSC Cycle Primed	HSC vs Progenitor	Aged HSC	Inflammatory HSC
SCARA3	CLC	NBL1	AREG	ASPM	SLC25A5	CLEC14A	NEDD4L	CDKN1A
GASK1B	CPA3	ID3	JUNB	AURKB	SRSF3	ZFPM2	AKAP12	FTH1
EMCN	EPX	RPL11	TAMALIN	KIF4A	NCL	HOXB-AS3	MAST4	MAFF
HLF	AC244502.1	SH3BGRL3	PELI1	MAD2L1	NUCB2	CDH4	PRR16	FGFR1
THY1	LGALS1	CD52	CLEC14A	TXNDC17	STMN1	AVP	ACTB	SKI
CD4	ADCYAP1	JUN	MAP3K8	PCNA	XRCC5	MECOM	HSPA5	PPP1R16B
ABI3BP	HPGDS	GADD45A	STK17B	GINS2	DUT	PREX2	GRAMD1B	LRRFIP1
AVP	AMHR2	AC103591.3	RBM38	RAD51	RBM3	NFIB	INSIG1	FOSL2
LAG3	FCER1A	S100A10	CSF3R	NEIL3	EEF2	DMKN	CALR	MIDN
EMP2	BLNK	S100A6	NFIL3	TYMS	RAN	HLF	STARD4	DNAJB6
KLF2	CSF2RB	LMNA	CSNK1E	DHFR	ENO1	RPSAP52	GLIS3	ZNF331
TNFRSF1B	AL157895.1	ATF3	HMGAA2	MCM10	CDK6	TAMALIN	ARHGAP21	SLC7A5
ABCA13	DHRS9	RHOB	IRGQ	BRCA1	HMGB2	EPPK1	SLC41A2	SLC2A3
CDH7	P2RX5	PPP1CB	CSRNP1	KIFC1	HSPD1	FOXO1	VEGFA	BHLHE40
TCEAL2	SLC24A3	YPEL5	NR4A1	C14orf119	HSP90AA1	TIE1	MIR222HG	ID2
APBB2	PLIN2	NR4A2	SLC25A27	NDUFB3	DEK	ARHGAP20	SKIL	CHN2
RAI14	ACSM1	LRRFIP1	MAP3K3	CIP2A	H2AZ1	HIF3A	ARHGAP31	SMOX
CNTN1	RAB44	SH3BP5	TSPYL2	ACAT2	RPS6	CLEC9A	SLC44A1	CSRNP1
GATA3	TFR2	ANKRD28	MDK	NEK2	DDX5	OXT	HSP90B1	ARSI
ACOX2	CST7	MLF1	ARRDC2	FOXM1	NAP1L1	IQCA1	PKIA	PPARD
KCND1	PLD1	HES1	IRF1	POLQ	SUB1	FXYD6	SOX5	SRGN
ANK3	SLC45A3	RPL9	ELF1	CDC25C	NOP58	SH3GL2	SV2C	DDIT4
SHE	FYB1	HOPX	KLF3	CDC6	RBM39	LIMCH1	SLC8A3	NEAT1
PRSS2	MITF	AREG	ZBTB21	DEPDC1	TPR	HOXB5	PPFIBP1	GADD45A
PTGS2	SPON2	RPL34	ST3GAL1	MYBL2	CAT	SPAG17	PECAM1	EZR
CEACAM1	RASGRP3	MAP9	TSC22D3	HMGB2	SERBP1	RASSF6	EGR1	PPP1R3B
RRAS	ABTB2	CRHBP	PER1	CDCA3	PNN	EMCN	CXCL8	RILPL2
NRBP2	CLEC12A	CAST	MIDN	ABI3	C1QBP	NKX2-3	DGKH	CSNK1D
CLDN7	KCNK5	DPYSL3	ATG16L2	RAD54L	SRSF7	STX3	ITGA2	TIPARP
MAGI2	ANGPT2	CD74	SERTAD2	SDHD	CCT8	AC026369.3	FLT1	PRR5
TMEM200A	NIBAN3	DUSP1	MEF2D	TMEM97	PRPF40A	CAVIN1	EMP1	RET
MORC4	NTRK1	HIST1H1C	RILPL2	PARBP	GPX1	SPP1	AC092069.1	TAMALIN
SH3BP5	MICAL2	HIST1H2BC	FAM117A	KIF23	PTPRC	FBLN5	CYTOR	PIM3
NAALAD2	LTBP1	HIST1H2AE	OSER1	FBXO43	PTGES3	DLGAP2	PLAUR	SYTL3
NLRP12	IGLL1	HLA-E	TAPBP	IDH1	RSL24D1	SELENOP	FNDC3A	UBC
GPAT3	ARPP21	HLA-C	GSTM2	MKI67	XRN2	MFAP4	FAM160A1	SH3BGRL3
LILRB2	HES6	HSPA1A	GAB2	CTSC	VAPA	LOXL1-AS1	SLC12A6	AREG
PREX2	PTH2R	HSPA1B	FCMR	CENPW	PNISR	SKAP1	PCBP3	ARID5A
PTGER4	TNFRSF19	HLA-DRA	JUND	SMC2	HSP90AB1	EHD2	KDM6B	ARRDC2
RUND3B	TLE1	HLA-DRB5	VPS37B	FEN1	SERPINB1	CNMD	LXN	CD109
C11orf45	HBD	HLA-DRB1	SDE2	UNG	UBXN4	AC009041.2	FOCAD	GXYLT2
IQCA1	IL5RA	HLA-DQA1	AVP	SKA1	ANP32B	CRHBP	ACVR1	LAPTM4A
EPPK1	VPREB1	HLA-DQB1	CCNL1	NEK6	PCM1	XXYL1-AS2	AFDN	USP12
AJ009632.2	TRPM2	HLA-DMA	SRGN	FABP5P7	TMED2	ROBO4	PFKFB3	GNA15
ALS2CL	STAR	HLA-DPA1	FOXO1	SKA3	LCP1	MLLT3	KIF21A	PHF20
EPHX2	CDK15	HLA-DPB1	GBP4	TADA1	HNRNPDL	BEX5	BMP6	USP36
FLT1	MME	RPS18	SEMA4C	ICMT	JMJD1C	MYCT1	PTMA	KLF13
LSR	HPGD	RPS12	MBOAT7	PGAM1	PARP1	MEG3	AC234772.2	SKIL
STOX2	CEBPE	EZR	AMPD2	MRPL17	SNRPD1	HMGAA2	MEF2C-AS2	RANBP2
SNX21	TESC	ARL4A	IGF2BP2-AS1	DOLK	ATRX	GATA3	MIR4435-2HG	SSH1

Table A.16: List of genes used to define different HSC characteristics.

HSC1	HSC2	HSC3	HSC4
VPREB1	ECEL1	OVCH2	AC026150.8
DLEU7	ACKR1	DIRAS3	RP11-1000B6.7
AC006994.2	CABP7	LINC02716	SFTPA1
TDRD5	LINC00313	PABPC1L2A	RP11-612A1.1
VWC2L	CAMK2A	RP11-466P24.6	RP1-239B22.6
RP11-114G22.1	DIO3OS	RP11-476H24.1	RP11-536K17.2
CTD-2126E3.3	ARSI	PRSS16	TTY2B
MAN1C1	RP1-140K8.1	C6orf141	RP11-34A14.3
FGF20	FUT3	AC023481.1	RP4-777L9.2
ST8SIA1	RP11-350N15.4	GIMAP8	RP11-356B19.11
LCN10	RET	LDHD	RP11-19B4.2
SIGLEC15	DIO3	RP11-240D10.4	CTC-244M17.1
LINC00996	SOCS3-DT	HIF3A	RP11-506H21.5
CTC-251H24.14	CYP1A1	TCEAL6	AP002856.4
NR3C2	HBZ	CTD-3193K9.4	LINC02817
GALM	INHBA	CLEC3B	SLC5A11
AC010084.1	SMOX	FBP1	MRM2
WIF1	PLAT	PCDH17	AC135178.7
RP11-700H6.1	SIPA1L2	NR3C2	RP11-40A13.1
RP11-439A17.10	CDKN1A	CNTN1	LINC01702
SLC16A9	ARHGAP8	DSC2	PRKG2
TEX14	AC024592.9	LINC00891	AJ006998.2
DLGAP2	UBE2QL1	ACVR1C	NKAIN2
SNHG28	PDE2A	RNASE4	RP11-669C19.1
LINC00891	ID2	FAM47E	RP11-863P13.4
STEAP2	NCMAP	TRAM1L1	TAT
MFAP3L	ADAMTS1	FALEC	
GALNTL6	PPARD	RP11-439A17.10	
FAM198B-AS1	FGF9	DLK1	
CACNG8	OXCT2	RRAS	
RPPH1	SLCO4A1	NUPR2	
FOXI2	CCNJL	HLA-DQB2	
CDH7	PDCD1	PERP	
PLCL1	RP11-66H6.4	LINC01504	
AC107218.3	AXL	GIMAP1	
RP5-934G17.7	SHISA2	SCN7A	
ST6GALNAC5	LINC01433	CD4	
MECOM	HR	ABI3BP	
ARHGAP20	PDLIM4	TTC12	
THBS3-AS1	CLEC14A	ZFYVE21	
CUEDC1	PIWIL3	CADPS2	
NRXN3	RP11-488L18.8	CDH7	
RNU12	PDGFA	ECHDC2	
HSD17B3	TKTL1	CLU	
NOS1AP	SPSB1	PDZRN4	
CH17-353B19.2	FTH1	PLAG1	
CEROX1	SCUBE1	CTA-243E7.1	
CLVS1	LRRC32	RNF180	
H4C1	PRRX2	DMKN	
RARG	CRIP2	WIF1	

Table A.17: List of top 50 upregulated genes from each HSC subcluster that was defined by the SAM algorithm. Marker genes are ordered in descending order by log2FC. All upregulated genes are FDR < 0.05.

Gene	log2FC	FDR
RP11-41917.1	1.47352015	1.77E-08
MAGI1	0.9491038	6.67E-07
PYGO1	0.56382504	6.44E-05
PNMA3	0.64116008	6.65E-05
KCNQ5	1.29295414	1.19E-04
PLEKHG4B	0.90742759	1.33E-04
NINL	0.70445382	1.71E-04
PTPRK	0.72999614	2.04E-04
PWRN1	0.68434332	2.46E-04
MARCKS	0.61811541	2.76E-04
CACNA1C	0.80555326	3.02E-04
GALNT14	1.05599874	3.69E-04
CD9	0.93136757	3.91E-04
RP11-527D7.1	0.70167859	4.92E-04
AUTS2	0.8077225	4.92E-04
ALPK3	0.54798264	6.29E-04
PXDN	0.70030986	9.11E-04
TRBC2	0.82455872	9.34E-04
CDH4	1.11523543	0.00119466
UBE2E2	0.96404176	0.00140032
SFMBT2	0.97108826	0.00144695
CLMN	0.64749032	0.00144695
ADAMTS9-AS2	0.54334976	0.00174339
TNNC2	0.79887374	0.00200173
C9orf139	0.71514254	0.00226832
LPAR2	0.52701688	0.00252766
IGHM	0.78068878	0.00439325
MEG3	0.98152972	0.00440947
PSD3	0.8177155	0.00442852
RARRES2	0.57001608	0.00442852
ARMH1	0.88652534	0.00463626
GRIP1	0.86427412	0.00590269
FMNL2	0.72797473	0.00665343
FAM184A	0.61119038	0.00750515
VWF	0.78767447	0.01053972
CARD11	0.91878754	0.01175586
AATK	0.58282903	0.01262875
LAT2	0.80508005	0.01503287
ZNF704	0.74334993	0.01581058
TMEM130	0.58050563	0.01914669
PCBP3	0.76693412	0.01947738
GOLM1	0.57129233	0.02255083
MIR181A1HG	0.68474181	0.02316105
PTPN14	0.51193788	0.02334075
PLIN2	0.67108809	0.02358902
UCHL1	0.8514327	0.02380693
SLC7A5	0.59970108	0.02382796
ARHGAP10	0.59032035	0.02863639
SLC24A3	0.81723926	0.02877881
RP11-264E23.4	0.76857786	0.03374892
RNF130	0.65404263	0.03507691
GPC6	0.69355993	0.03827856
ISG20	0.6756305	0.04087757
GDF11	0.5191965	0.04105933
PRKCA	0.65643957	0.04608736
PARP12	0.51355546	0.04971096

Table A.18: List of upregulated genes in *ASXL1*-mutant HSC1 compared to WT HSC1. Genes are ordered in descending order by log2FC. All upregulated genes are $FDR < 0.05$.

Gene	log2FC	FDR
CDH4	1.060489	2.97E-05
RP11-112.1	1.17237648	1.02E-04
SDK2	1.03781755	1.60E-04
LITAF	0.95964393	3.69E-04
RARRES2	0.78748619	4.29E-04
EPCAM	0.74835014	4.77E-04
PLIN2	0.78408259	5.61E-04
PDLIM4	0.62016978	0.00140529
GSE1	0.74127136	0.00261353
FBLN7	0.45452929	0.00414354
RP11-41917.1	0.80325661	0.00513118
HSPG2	0.75972805	0.00513118
INAFM2	0.80386147	0.00524324
MAGI1	0.65368263	0.00597558
MAP3K15	0.45100062	0.00849351
ATP10A	0.48251172	0.00849351
MSI2	0.28743647	0.00889869
PDE4D	0.64424267	0.00912395
PPP1CB	0.35770329	0.0102083
BTBD11	0.6584129	0.01089148
BTG2	0.60383455	0.01335182
KSR1	0.74524232	0.01662018
KCNIP3	0.38129391	0.02145828
ZNF385D	0.49991818	0.02588314
COL6A2	0.38939922	0.02588314
ARMH1	0.60098165	0.02628263
ALPK3	0.41805982	0.02712926
ARMCX1	0.65273788	0.02862777
TNNC2	0.58001216	0.03020299
RIN3	0.44654501	0.03102764
ST3GAL4	0.58093607	0.03153755
LRMDA	0.44079452	0.03318265
NFKB1	0.60545347	0.03448552
DAP	0.61624729	0.03512157
KDM4B	0.6413939	0.03743096
AUTS2	0.54046141	0.03995738
UCHL1	0.65126902	0.04047811
WVOX	0.72375127	0.04221161
RP11-783N9.1	0.29244695	0.04428288
UBE2H	0.35949481	0.04514693
GTSF1	0.55704206	0.04653327
RETREG1	0.31212173	0.04871524

Table A.19: List of upregulated genes in *ASXL1*-mutant HSC2 compared to WT HSC2. Genes are ordered in descending order by log2FC. All upregulated genes are FDR < 0.05.

References

- [1] F Blokzijl *et al.* “Tissue-specific mutation accumulation in human adult stem cells during life”. In: *Nature* 538 (7624 2016), pp. 260–264. DOI: 10.1038/nature19768.
- [2] M L Hoang *et al.* “Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing”. In: *Proc Natl Acad Sci U S A* 113 (35 2016), pp. 9846–9851. DOI: 10.1073/pnas.1607794113.
- [3] I Martincorena and P J Campbell. “Somatic mutation in cancer and normal cells”. In: *Science* 349 (6255 2015), pp. 1483–1489. DOI: 10.1126/science.aab4082.
- [4] L B Alexandrov *et al.* “Mutational signatures associated with tobacco smoking in human cancer”. In: *Science* 354 (6312 2016), pp. 618–622. DOI: 10.1126/science.aag0299.
- [5] L Moore *et al.* “The mutational landscape of human somatic and germline cells”. In: *Nature* 597 (7876 2021), pp. 381–386. DOI: 10.1038/s41586-021-03822-7.
- [6] L B Alexandrov *et al.* “Clock-like mutational processes in human somatic cells”. In: *Nat Genet* 47 (12 2015), pp. 1402–1407. DOI: 10.1038/ng.3441.
- [7] L B Alexandrov *et al.* “The repertoire of mutational signatures in human cancer”. In: *Nature* 578 (7793 2020), pp. 94–101. DOI: 10.1038/s41586-020-1943-3.
- [8] K Rodgers and M McVey. “Error-Prone Repair of DNA Double-Strand Breaks”. In: *J Cell Physiol* 231 (1 2016), pp. 15–24. DOI: 10.1002/jcp.25053.
- [9] R Saiki *et al.* “Combined landscape of single-nucleotide variants and copy number alterations in clonal hematopoiesis”. In: *Nat Med* 27 (7 2021), pp. 1239–1249. DOI: 10.1038/s41591-021-01411-9.
- [10] K Takahashi *et al.* “Copy number alterations detected as clonal hematopoiesis of indeterminate potential”. In: *Blood Adv* 1 (15 2017), pp. 1031–1036. DOI: 10.1182/bloodadvances.2017007922.
- [11] K Yizhak *et al.* “RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues”. In: *Science* 364 (6444 2019). DOI: 10.1126/science.aaw0726.
- [12] Mel Greaves and Carlo C Maley. “Clonal evolution in cancer”. In: *Nature* 481 (7381 2012), pp. 306–313. DOI: 10.1038/nature10762. URL: <https://doi.org/10.1038/nature10762>.
- [13] Michael S Lawrence *et al.* “Mutational heterogeneity in cancer and the search for new cancer-associated genes”. In: *Nature* 499 (7457 2013), pp. 214–218. DOI: 10.1038/nature12213. URL: <https://doi.org/10.1038/nature12213>.
- [14] Iñigo Martincorena *et al.* “Universal Patterns of Selection in Cancer and Somatic Tissues”. In: *Cell* 171 (5 2017), 1029–1041.e21. DOI: <https://doi.org/10.1016/j.cell.2017.09.042>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867417311364>.

- [15] Matthew H Bailey *et al.* “Comprehensive Characterization of Cancer Driver Genes and Mutations”. In: *Cell* 173 (2 2018), 371–385.e18. DOI: <https://doi.org/10.1016/j.cell.2018.02.060>. URL: <https://www.sciencedirect.com/science/article/pii/S009286741830237X>.
- [16] Peter J Campbell *et al.* “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2020), pp. 82–93. DOI: 10.1038/s41586-020-1969-6. URL: <https://doi.org/10.1038/s41586-020-1969-6>.
- [17] Iñigo Martincorena *et al.* “Somatic mutant clones colonize the human esophagus with age”. In: *Science* 362 (6417 2018), pp. 911–917. DOI: doi:10.1126/science.aau3879. URL: <https://www.science.org/doi/abs/10.1126/science.aau3879>.
- [18] Henry Lee-Six *et al.* “The landscape of somatic mutation in normal colorectal epithelial cells”. In: *Nature* 574 (7779 2019), pp. 532–537. DOI: 10.1038/s41586-019-1672-7. URL: <https://doi.org/10.1038/s41586-019-1672-7>.
- [19] Kosaku Nanki *et al.* “Somatic inflammatory gene mutations in human ulcerative colitis epithelium”. In: *Nature* 577 (7789 2020), pp. 254–259. DOI: 10.1038/s41586-019-1844-5. URL: <https://doi.org/10.1038/s41586-019-1844-5>.
- [20] Kenichi Yoshida *et al.* “Tobacco smoking and somatic mutations in human bronchial epithelium”. In: *Nature* 578 (7794 2020), pp. 266–272. DOI: 10.1038/s41586-020-1961-1. URL: <https://doi.org/10.1038/s41586-020-1961-1>.
- [21] J D Rowley. “Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining”. In: *Nature* 243 (5405 1973), pp. 290–293. DOI: 10.1038/243290a0.
- [22] M F Fey *et al.* “Clonality and X-inactivation patterns in hematopoietic cell populations detected by the highly informative M27 beta DNA probe”. In: *Blood* 83 (4 1994), pp. 931–938.
- [23] Lambert Busque *et al.* “Nonrandom X-Inactivation Patterns in Normal Females: Lyonization Ratios Vary With Age”. In: *Blood* 88.1 (1996), pp. 59–65. URL: <https://www.sciencedirect.com/science/article/pii/S0006497120646043>.
- [24] M Xie *et al.* “Age-related mutations associated with clonal hematopoietic expansion and malignancies”. In: *Nat Med* 20 (12 2014), pp. 1472–1478. DOI: 10.1038/nm.3733. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25326804>.
- [25] Giulio Genovese *et al.* “Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence”. In: *New England Journal of Medicine* 371 (26 2014), pp. 2477–2487. DOI: 10.1056/NEJMoa1409405.
- [26] Siddhartha Jaiswal *et al.* “Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes”. In: *New England Journal of Medicine* 371 (26 2014), pp. 2488–2498. DOI: 10.1056/NEJMoa1408617. URL: <https://www.nejm.org/doi/full/10.1056/NEJMoa1408617>.
- [27] G J Spangrude, S Heimfeld, and I L Weissman. “Purification and characterization of mouse hematopoietic stem cells”. In: *Science* 241 (4861 1988), pp. 58–62. DOI: 10.1126/science.2898810.

- [28] Mary F Lyon. “Sex chromatin and gene action in the mammalian X-chromosome”. In: *American journal of human genetics* 14 (2 1962), p. 135.
- [29] Rosemary E. Gale *et al.* “Tissue Specificity of X-Chromosome Inactivation Patterns”. In: *Blood* 83.10 (1994), pp. 2899–2905. DOI: <https://doi.org/10.1182/blood.V83.10.2899.2899>. URL: <https://www.sciencedirect.com/science/article/pii/S0006497120692060>.
- [30] L Busque *et al.* “Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis”. In: *Nat Genet* 44 (11 2012), pp. 1179–1181. DOI: 10.1038/ng.2413.
- [31] F Delhommeau *et al.* “Mutation in TET2 in myeloid cancers”. In: *N Engl J Med* 360 (22 2009), pp. 2289–2301. DOI: 10.1056/NEJMoa0810069.
- [32] T Miyamoto, I L Weissman, and K Akashi. “AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation”. In: *Proc Natl Acad Sci U S A* 97 (13 2000), pp. 7521–7526. DOI: 10.1073/pnas.97.13.7521.
- [33] J S Welch *et al.* “The origin and evolution of mutations in acute myeloid leukemia”. In: *Cell* 150 (2 2012), pp. 264–278. DOI: 10.1016/j.cell.2012.06.023.
- [34] Petter S Woll *et al.* “Myelodysplastic Syndromes Are Propagated by Rare and Distinct Human Cancer Stem Cells *In Vivo*”. In: *Cancer Cell* 25 (6 2014), pp. 794–808. DOI: 10.1016/j.ccr.2014.03.036. URL: <https://doi.org/10.1016/j.ccr.2014.03.036>.
- [35] Nicolas Goardon *et al.* “Coexistence of LMPP-like and GMP-like Leukemia Stem Cells in Acute Myeloid Leukemia”. In: *Cancer Cell* 19 (1 2011), pp. 138–152. DOI: 10.1016/j.ccr.2010.12.012. URL: <https://doi.org/10.1016/j.ccr.2010.12.012>.
- [36] Lynn Quek *et al.* “Genetically distinct leukemic stem cells in human CD34⁺ acute myeloid leukemia are arrested at a hemopoietic precursor-like stage”. In: *Journal of Experimental Medicine* 213 (8 2016), pp. 1513–1535. DOI: 10.1084/jem.20151775. URL: <https://doi.org/10.1084/jem.20151775>.
- [37] M Jan *et al.* “Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia”. In: *Sci Transl Med* 4 (149 2012), 149ra118. DOI: 10.1126/scitranslmed.3004315.
- [38] L I Shlush *et al.* “Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia”. In: *Nature* 506 (7488 2014), pp. 328–333. DOI: 10.1038/nature13038.
- [39] M R Corces-Zimmerman *et al.* “Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission”. In: *Proc Natl Acad Sci U S A* 111 (7 2014), pp. 2548–2553. DOI: 10.1073/pnas.1324297111.
- [40] A L Young *et al.* “Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults”. In: *Nat Commun* 7 (2016), p. 12484. DOI: 10.1038/ncomms12484.

- [41] C C Coombs *et al.* “Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes”. In: *Cell Stem Cell* 21 (3 2017), 374–382.e4. DOI: 10.1016/j.stem.2017.07.010.
- [42] Kelly L Bolton *et al.* “Cancer therapy shapes the fitness landscape of clonal hematopoiesis”. In: *Nature Genetics* 52 (11 2020), pp. 1219–1226. DOI: 10.1038/s41588-020-00710-0. URL: <https://doi.org/10.1038/s41588-020-00710-0>.
- [43] T H Mouhieddine *et al.* “Clonal hematopoiesis is associated with adverse outcomes in multiple myeloma patients undergoing transplant”. In: *Nat Commun* 11 (1 2020), p. 2996. DOI: 10.1038/s41467-020-16805-5.
- [44] D Midic *et al.* “Prevalence and dynamics of clonal hematopoiesis caused by leukemia-associated mutations in elderly individuals without hematologic disorders”. In: *Leukemia* 34 (8 2020), pp. 2198–2205. DOI: 10.1038/s41375-020-0869-y.
- [45] Niels Asger Jakobsen *et al.* “Selective advantage of mutant stem cells in human clonal hematopoiesis is associated with attenuated response to inflammation and aging”. In: *Cell Stem Cell* 31.8 (2024), 1127–1144.e17. DOI: 10.1016/j.stem.2024.05.010. URL: <https://doi.org/10.1016/j.stem.2024.05.010>.
- [46] Oriol Pich *et al.* “Discovering the drivers of clonal hematopoiesis”. In: *Nature Communications* 13 (1 2022), p. 4267. DOI: 10.1038/s41467-022-31878-0. URL: <https://doi.org/10.1038/s41467-022-31878-0>.
- [47] F Zink *et al.* “Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly”. In: *Blood* 130 (6 2017), pp. 742–752. DOI: 10.1182/blood-2017-02-769869.
- [48] Nicholas Bernstein *et al.* “Analysis of somatic mutations in whole blood from 200,618 individuals identifies pervasive positive selection and novel drivers of clonal hematopoiesis”. In: *Nature Genetics* 56 (6 2024), pp. 1147–1155. DOI: 10.1038/s41588-024-01755-1. URL: <https://doi.org/10.1038/s41588-024-01755-1>.
- [49] Masaki Okano, Shaoping Xie, and En Li. “Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases”. In: *Nature Genetics* 19 (3 1998), pp. 219–220. DOI: 10.1038/890. URL: <https://doi.org/10.1038/890>.
- [50] K D Rasmussen and K Helin. “Role of TET enzymes in DNA methylation, development, and cancer”. In: *Genes Dev* 30 (7 2016), pp. 733–750. DOI: 10.1101/gad.276568.115.
- [51] C Fisher Berger J. Randazzo F. Brock H.W. “A human homolog of Additional sex combs, ADDITIONAL SEX COMBS-LIKE 1, maps to chromosome 20q11”. In: *Gene* 306 (2003), pp. 115–126. DOI: 10.1016/s0378-1119(03)00430-x.
- [52] Michael D Kessler *et al.* “Common and rare variant associations with clonal haematopoiesis phenotypes”. In: *Nature* 612 (7939 2022), pp. 301–309. DOI: 10.1038/s41586-022-05448-9. URL: <https://doi.org/10.1038/s41586-022-05448-9>.

- [53] Siddhartha P Kar *et al.* “Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis”. In: *Nature Genetics* 54 (8 2022), pp. 1155–1166. DOI: 10.1038/s41588-022-01121-z. URL: <https://doi.org/10.1038/s41588-022-01121-z>.
- [54] Joshua S. Weinstock *et al.* “Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis”. In: *Nature* 616.7958 (2023), pp. 755–763. DOI: 10.1038/s41586-023-05806-1. URL: <https://doi.org/10.1038/s41586-023-05806-1>.
- [55] L. Virgilio *et al.* “Identification of the TCL1 gene involved in T-cell malignancies”. In: *Proc Natl Acad Sci U S A* 91.26 (1994), pp. 12530–4. DOI: 10.1073/pnas.91.26.12530.
- [56] A. Schrader *et al.* “Actionable perturbations of damage responses by TCL1/ATM and epigenetic lesions form the basis of T-PLL”. In: *Nat Commun* 9.1 (2018), p. 697. DOI: 10.1038/s41467-017-02688-6.
- [57] J. Laine *et al.* “The protooncogene TCL1 is an Akt kinase coactivator”. In: *Mol Cell* 6.2 (2000), pp. 395–407. DOI: 10.1016/s1097-2765(00)00039-3.
- [58] Astrid Eijkelenboom and Boudewijn M. T. Burgering. “FOXOs: signalling integrators for homeostasis maintenance”. In: *Nature Reviews Molecular Cell Biology* 14.2 (2013), pp. 83–97. DOI: 10.1038/nrm3507. URL: <https://doi.org/10.1038/nrm3507>.
- [59] P R Loh *et al.* “Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations”. In: *Nature* 559 (7714 2018), pp. 350–355. DOI: 10.1038/s41586-018-0321-x.
- [60] C Terao *et al.* “Chromosomal alterations among age-related haematopoietic clones in Japan”. In: *Nature* 584 (7819 2020), pp. 130–135. DOI: 10.1038/s41586-020-2426-2.
- [61] P R Loh, G Genovese, and S A McCarroll. “Monogenic and polygenic inheritance become instruments for clonal selection”. In: *Nature* 584 (7819 2020), pp. 136–141. DOI: 10.1038/s41586-020-2430-6.
- [62] T Gao *et al.* “Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis”. In: *Nat Commun* 12 (1 2021), p. 338. DOI: 10.1038/s41467-020-20565-7.
- [63] K B Jacobs *et al.* “Detectable clonal mosaicism and its relationship to aging and cancer”. In: *Nat Genet* 44 (6 2012), pp. 651–658. DOI: 10.1038/ng.2270.
- [64] M J Machiela *et al.* “Characterization of large structural genetic mosaicism in human autosomes”. In: *Am J Hum Genet* 96 (3 2015), pp. 487–497. DOI: 10.1016/j.ajhg.2015.01.011.
- [65] C C Laurie *et al.* “Detectable clonal mosaicism from birth to old age and its relationship to cancer”. In: *Nat Genet* 44 (6 2012), pp. 642–650. DOI: 10.1038/ng.2271.
- [66] C J Watson *et al.* “The evolutionary dynamics and fitness landscape of clonal hematopoiesis”. In: *Science* 367 (6485 2020), pp. 1449–1454. DOI: 10.1126/science.aay9333.

- [67] G Y P Poon *et al.* “Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues”. In: *Nat Genet* 53 (11 2021), pp. 1597–1605. DOI: 10.1038/s41588-021-00957-1.
- [68] Margarete A Fabre *et al.* “The longitudinal dynamics and natural history of clonal haematopoiesis”. In: *Nature* 606 (7913 2022), pp. 335–342. DOI: 10.1038/s41586-022-04785-z. URL: <https://doi.org/10.1038/s41586-022-04785-z>.
- [69] E Mitchell *et al.* “Clonal dynamics of haematopoiesis across the human lifespan”. In: *Nature* 606 (7913 2022), pp. 343–350. DOI: 10.1038/s41586-022-04786-y.
- [70] Verena Körber *et al.* “Detecting and quantifying clonal selection in somatic stem cells”. In: *bioRxiv* (2023), p. 2021.12.15.472780. DOI: 10.1101/2021.12.15.472780. URL: <https://www.biorxiv.org/content/biorxiv/early/2023/10/10/2021.12.15.472780.full.pdf>.
- [71] Neil A Robertson *et al.* “Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects”. In: *Nature Medicine* 28 (7 2022), pp. 1439–1446. DOI: 10.1038/s41591-022-01883-3. URL: <https://doi.org/10.1038/s41591-022-01883-3>.
- [72] Marianna Rossi *et al.* “Clinical relevance of clonal hematopoiesis in persons aged 80 years”. In: *Blood* 138 (21 2021), pp. 2093–2105. DOI: <https://doi.org/10.1182/blood.2021011320>. URL: <https://www.sciencedirect.com/science/article/pii/S0006497121012350>.
- [73] S Jaiswal *et al.* “Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease”. In: *N Engl J Med* 377 (2 2017), pp. 111–121. DOI: 10.1056/NEJMoa1701719. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28636844>.
- [74] José J Fuster *et al.* “Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice”. In: *Science* 355 (6327 2017), pp. 842–847.
- [75] Waihay J Wong *et al.* “Clonal haematopoiesis and risk of chronic liver disease”. In: *Nature* 616 (7958 2023), pp. 747–754. DOI: 10.1038/s41586-023-05857-4. URL: <https://doi.org/10.1038/s41586-023-05857-4>.
- [76] Peter G Miller *et al.* “Association of clonal hematopoiesis with chronic obstructive pulmonary disease”. In: *Blood* 139 (3 2022), pp. 357–368. DOI: <https://doi.org/10.1182/blood.2021013531>. URL: <https://www.sciencedirect.com/science/article/pii/S0006497121018759>.
- [77] Peter Geon Kim *et al.* “Dnmt3a-mutated clonal hematopoiesis promotes osteoporosis”. In: *Journal of Experimental Medicine* 218 (12 2021), e20211872.
- [78] Ahmed A Z Dawoud *et al.* “Clonal myelopoiesis promotes adverse outcomes in chronic kidney disease”. In: *Leukemia* 36 (2 2022), pp. 507–515. DOI: 10.1038/s41375-021-01382-3. URL: <https://doi.org/10.1038/s41375-021-01382-3>.
- [79] Mridul Agrawal *et al.* “TET2-mutant clonal hematopoiesis and risk of gout”. In: *Blood, The Journal of the American Society of Hematology* 140 (10 2022), pp. 1094–1103.

- [80] José J Fuster *et al.* “TET2-loss-of-function-driven clonal hematopoiesis exacerbates experimental insulin resistance in aging and obesity”. In: *Cell Reports* 33 (4 2020).
- [81] Seishi Ogawa. “Clonal hematopoiesis in acquired aplastic anemia”. In: *Blood* 128.3 (July 2016), pp. 337–347. DOI: 10.1182/blood-2016-01-636381. eprint: <https://ashpublications.org/blood/article-pdf/128/3/337/1358967/337.pdf>. URL: <https://doi.org/10.1182/blood-2016-01-636381>.
- [82] Seyedeh M Zekavat *et al.* “Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection”. In: *Nature Medicine* 27 (6 2021), pp. 1012–1024. DOI: 10.1038/s41591-021-01371-0. URL: <https://doi.org/10.1038/s41591-021-01371-0>.
- [83] Lachelle D Weeks *et al.* “Age-related diseases of inflammation in myelodysplastic syndrome and chronic myelomonocytic leukemia”. In: *Blood* 139 (8 2022), pp. 1246–1250. DOI: <https://doi.org/10.1182/blood.2021014418>. URL: <https://www.sciencedirect.com/science/article/pii/S0006497121021005>.
- [84] Hind Bouzid *et al.* “Clonal hematopoiesis is associated with protection from Alzheimer’s disease”. In: *Nature Medicine* 29 (7 2023), pp. 1662–1670. DOI: 10.1038/s41591-023-02397-2. URL: <https://doi.org/10.1038/s41591-023-02397-2>.
- [85] D Bron *et al.* “Aging and blood disorders: new perspectives, new challenges”. In: *Haematologica* 100 (4 2015), pp. 415–417. DOI: 10.3324/haematol.2015.126771.
- [86] David P Steensma *et al.* “Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes”. In: *Blood* 126 (1 2015), pp. 9–16. DOI: 10.1182/blood-2015-03-631747. URL: <https://doi.org/10.1182/blood-2015-03-631747>.
- [87] S Abelson *et al.* “Prediction of acute myeloid leukaemia risk in healthy individuals”. In: *Nature* 559 (7714 2018), pp. 400–404. DOI: 10.1038/s41586-018-0317-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29988082>.
- [88] P Desai *et al.* “Somatic mutations precede acute myeloid leukemia years before diagnosis”. In: *Nat Med* 24 (7 2018), pp. 1015–1023. DOI: 10.1038/s41591-018-0081-z. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29988143>.
- [89] Lachelle D Weeks *et al.* “Prediction of risk for myeloid malignancy in clonal hematopoiesis”. In: *NEJM evidence* 2 (5 2023), EVIDoA2200310.
- [90] Daniel A Arber *et al.* “International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data”. In: *Blood* 140 (11 2022), pp. 1200–1228. DOI: <https://doi.org/10.1182/blood.2022015850>. URL: <https://www.sciencedirect.com/science/article/pii/S0006497122008369>.
- [91] Luca Malcovati *et al.* “Clinical significance of somatic mutation in unexplained blood cytopenia”. In: *Blood* 129 (25 2017), pp. 3371–3378. DOI: 10.1182/blood-2017-01-763425. URL: <https://doi.org/10.1182/blood-2017-01-763425>.

- [92] Humaira Gowher and Albert Jeltsch. “Mammalian DNA methyltransferases: new discoveries and open questions”. In: *Biochemical Society Transactions* 46.5 (2018), pp. 1191–1202. DOI: 10.1042/bst20170574. URL: <https://doi.org/10.1042/BST20170574>.
- [93] S Saxonov, P Berg, and D L Brutlag. “A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters”. In: *Proc Natl Acad Sci U S A* 103 (5 2006), pp. 1412–1417. DOI: 10.1073/pnas.0510310103.
- [94] M. Ko *et al.* “Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2”. In: *Nature* 468.7325 (2010), pp. 839–43. DOI: 10.1038/nature09586.
- [95] Andrew J. Bannister and Tony Kouzarides. “Regulation of chromatin by histone modifications”. In: *Cell Research* 21.3 (2011), pp. 381–395. DOI: 10.1038/cr.2011.22. URL: <https://doi.org/10.1038/cr.2011.22>.
- [96] T Fujino and T Kitamura. “ASXL1 mutation in clonal hematopoiesis”. In: *Exp Hematol* 83 (2020), pp. 74–84. DOI: 10.1016/j.exphem.2020.01.002. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31945396>.
- [97] Bernd Schuettengruber *et al.* “Genome Regulation by Polycomb and Trithorax: 70 Years and Counting”. In: *Cell* 171.1 (2017), pp. 34–57. DOI: <https://doi.org/10.1016/j.cell.2017.08.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867417308905>.
- [98] E B Lewis. “A gene complex controlling segmentation in *Drosophila*”. In: *Nature* 276 (5688 1978), pp. 565–570. DOI: 10.1038/276565a0.
- [99] Philip Ingham and Robert Whittle. “Trithorax: A new homoeotic mutation of *Drosophila melanogaster* causing transformations of abdominal and thoracic imaginal segments: I. Putative role during embryogenesis”. In: *Molecular and General Genetics MGG* 179 (1980), pp. 607–614.
- [100] B Schuettengruber *et al.* “Trithorax group proteins: switching genes on and keeping them active”. In: *Nat Rev Mol Cell Biol* 12 (12 2011), pp. 799–814. DOI: 10.1038/nrm3230.
- [101] A Shilatifard. “The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis”. In: *Annu Rev Biochem* 81 (2012), pp. 65–95. DOI: 10.1146/annurev-biochem-051710-134100.
- [102] R A Alharbi *et al.* “The role of HOX genes in normal hematopoiesis and acute leukemia”. In: *Leukemia* 27 (5 2013), pp. 1000–1008. DOI: 10.1038/leu.2012.356.
- [103] A Kuzmichev *et al.* “Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein”. In: *Genes Dev* 16 (22 2002), pp. 2893–2905. DOI: 10.1101/gad.1035902.
- [104] S S Levine *et al.* “The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans”. In: *Mol Cell Biol* 22 (17 2002), pp. 6070–6078. DOI: 10.1128/mcb.22.17.6070-6078.2002.
- [105] L Wang *et al.* “Hierarchical recruitment of polycomb group silencing complexes”. In: *Mol Cell* 14 (5 2004), pp. 637–646. DOI: 10.1016/j.molcel.2004.05.009.

- [106] R Cao *et al.* “Role of histone H3 lysine 27 methylation in Polycomb-group silencing”. In: *Science* 298 (5595 2002), pp. 1039–1043. DOI: 10.1126/science.1076997.
- [107] N P Blackledge and R J Klose. “The molecular principles of gene regulation by Polycomb repressive complexes”. In: *Nat Rev Mol Cell Biol* 22 (12 2021), pp. 815–833. DOI: 10.1038/s41580-021-00398-y.
- [108] Z Gao *et al.* “PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes”. In: *Mol Cell* 45 (3 2012), pp. 344–356. DOI: 10.1016/j.molcel.2012.01.002.
- [109] Simon Hauri *et al.* “A High-Density Map for Navigating the Human Polycomb Complexome”. In: *Cell Reports* 17 (2 2016), pp. 583–595. DOI: 10.1016/j.celrep.2016.08.096. URL: <https://doi.org/10.1016/j.celrep.2016.08.096>.
- [110] L Tavares *et al.* “RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3”. In: *Cell* 148 (4 2012), pp. 664–678. DOI: 10.1016/j.cell.2011.12.029.
- [111] A. M. Farcas *et al.* “KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands”. In: *Elife* 1 (2012), e00205. DOI: 10.7554/eLife.00205.
- [112] B. Stielow *et al.* “MGA, L3MBTL2 and E2F6 determine genomic binding of the non-canonical Polycomb repressive complex PRC1.6”. In: *PLoS Genet* 14.1 (2018), e1007193. DOI: 10.1371/journal.pgen.1007193.
- [113] A. Scelfo *et al.* “Functional Landscape of PCGF Proteins Reveals Both RING1A/B-Dependent-and RING1A/B-Independent-Specific Activities”. In: *Mol Cell* 74.5 (2019), 1037–1052.e7. DOI: 10.1016/j.molcel.2019.04.002.
- [114] S. Cooper *et al.* “Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2”. In: *Nat Commun* 7 (2016), p. 13661. DOI: 10.1038/ncomms13661.
- [115] S Tamburri *et al.* “Histone H2AK119 Mono-Ubiquitination Is Essential for Polycomb-Mediated Transcriptional Repression”. In: *Mol Cell* 77 (4 2020), 840–856 e5. DOI: 10.1016/j.molcel.2019.11.021. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31883952>.
- [116] L A Boyer *et al.* “Polycomb complexes repress developmental regulators in murine embryonic stem cells”. In: *Nature* 441 (7091 2006), pp. 349–353. DOI: 10.1038/nature04733.
- [117] L Aloia, B Di Stefano, and L Di Croce. “Polycomb complexes in stem cells and embryonic development”. In: *Development* 140 (12 2013), pp. 2525–2534. DOI: 10.1242/dev.091553.
- [118] A Rizo *et al.* “Long-term maintenance of human hematopoietic stem/progenitor cells by expression of BMI1”. In: *Blood* 111 (5 2008), pp. 2621–2630. DOI: 10.1182/blood-2007-08-106666.
- [119] A Iwama *et al.* “Enhanced self-renewal of hematopoietic stem cells mediated by the polycomb gene product Bmi-1”. In: *Immunity* 21 (6 2004), pp. 843–851. DOI: 10.1016/j.immuni.2004.11.004.

- [120] T Ikawa *et al.* “Conversion of T cells to B cells by inactivation of polycomb-mediated epigenetic suppression of the B-lineage program”. In: *Genes Dev* 30 (22 2016), pp. 2475–2485. DOI: 10.1101/gad.290593.116.
- [121] K. Ross *et al.* “Polycomb group ring finger 1 cooperates with Runx1 in regulating differentiation and self-renewal of hematopoietic cells”. In: *Blood* 119.18 (2012), pp. 4152–61. DOI: 10.1182/blood-2011-09-382390.
- [122] V. van den Boom *et al.* “Nonredundant and locus-specific gene repression functions of PRC1 paralog family members in human hematopoietic stem/progenitor cells”. In: *Blood* 121.13 (2013), pp. 2452–61. DOI: 10.1182/blood-2012-08-451666.
- [123] M Mochizuki-Kashio *et al.* “Ezh2 loss in hematopoietic stem cells predisposes mice to develop heterogeneous malignancies in an Ezh1-dependent manner”. In: *Blood* 126 (10 2015), pp. 1172–1183. DOI: 10.1182/blood-2015-03-634428.
- [124] S C Lee *et al.* “Polycomb repressive complex 2 component Suz12 is required for hematopoietic stem cell function and lymphopoiesis”. In: *Blood* 126 (2 2015), pp. 167–175. DOI: 10.1182/blood-2014-12-615898.
- [125] H Xie *et al.* “Polycomb repressive complex 2 regulates normal hematopoietic stem cell function in a developmental-stage-specific manner”. In: *Cell Stem Cell* 14 (1 2014), pp. 68–80. DOI: 10.1016/j.stem.2013.10.001.
- [126] C D Jude *et al.* “Unique and independent roles for MLL in adult hematopoietic stem cells and progenitors”. In: *Cell Stem Cell* 1 (3 2007), pp. 324–337. DOI: 10.1016/j.stem.2007.05.019.
- [127] K A McMahon *et al.* “Mll has a critical role in fetal and adult hematopoietic stem cell self-renewal”. In: *Cell Stem Cell* 1 (3 2007), pp. 338–345. DOI: 10.1016/j.stem.2007.07.002.
- [128] B D Yu *et al.* “Altered Hox expression and segmental identity in Mll-mutant mice”. In: *Nature* 378 (6556 1995), pp. 505–508. DOI: 10.1038/378505a0.
- [129] P Ernst *et al.* “Definitive hematopoiesis requires the mixed-lineage leukemia gene”. In: *Dev Cell* 6 (3 2004), pp. 437–443. DOI: 10.1016/s1534-5807(04)00061-9.
- [130] C Meyer *et al.* “The MLL recombinome of acute leukemias in 2017”. In: *Leukemia* 32 (2 2018), pp. 273–284. DOI: 10.1038/leu.2017.213.
- [131] S Ziemmin-van der Poel *et al.* “Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias”. In: *Proc Natl Acad Sci U S A* 88 (23 1991), pp. 10735–10739. DOI: 10.1073/pnas.88.23.10735.
- [132] Y. M. Soto-Feliciano *et al.* “A Molecular Switch between Mammalian MLL Complexes Dictates Response to Menin-MLL Inhibition”. In: *Cancer Discov* 13.1 (2023), pp. 146–169. DOI: 10.1158/2159-8290.Cd-22-0416.
- [133] T. A. Milne *et al.* “Menin and MLL cooperatively regulate expression of cyclin-dependent kinase inhibitors”. In: *Proc Natl Acad Sci U S A* 102.3 (2005), pp. 749–54. DOI: 10.1073/pnas.0408836102.
- [134] D A Sinclair *et al.* “Genetic analysis of the additional sex combs locus of *Drosophila melanogaster*”. In: *Genetics* 130 (4 1992), pp. 817–825.

- [135] T A Milne, D A R Sinclair, and H W Brock. “The Additional sex combs gene of *Drosophila* is required for activation and repression of homeotic loci, and interacts specifically with Polycomb and super sex combs”. In: *Molecular and General Genetics MGG* 261 (4 1999), pp. 753–761. DOI: 10.1007/s004380050018. URL: <https://doi.org/10.1007/s004380050018>.
- [136] C L Fisher *et al.* “Characterization of Asxl1, a murine homolog of Additional sex combs, and analysis of the Asx-like gene family”. In: *Gene* 369 (2006), pp. 109–118. DOI: 10.1016/j.gene.2005.10.033.
- [137] A P Szczepanski and L Wang. “Emerging multifaceted roles of BAP1 complexes in biological processes”. In: *Cell Death Discov* 7 (1 2021), p. 20. DOI: 10.1038/s41420-021-00406-2.
- [138] L Aravind and L M Iyer. “The HARE-HTH and associated domains: novel modules in the coordination of epigenetic DNA and protein modifications”. In: *Cell Cycle* 11 (1 2012), pp. 119–131. DOI: 10.4161/cc.11.1.18475.
- [139] K Jain *et al.* “Characterization of the plant homeodomain (PHD) reader family for their histone tail interactions”. In: *Epigenetics Chromatin* 13 (1 2020), p. 3. DOI: 10.1186/s13072-020-0328-z.
- [140] U H Park *et al.* “Additional sex comb-like (ASXL) proteins 1 and 2 play opposite roles in adipogenesis via reciprocal regulation of peroxisome proliferator-activated receptor {gamma}”. In: *J Biol Chem* 286 (2 2011), pp. 1354–1363. DOI: 10.1074/jbc.M110.177816.
- [141] Y S Cho *et al.* “Additional sex comb-like 1 (ASXL1), in cooperation with SRC-1, acts as a ligand-dependent coactivator for retinoic acid receptor”. In: *J Biol Chem* 281 (26 2006), pp. 17588–17598. DOI: 10.1074/jbc.M512616200.
- [142] S Daou *et al.* “Monoubiquitination of ASXLs controls the deubiquitinase activity of the tumor suppressor BAP1”. In: *Nat Commun* 9 (1 2018), p. 4385. DOI: 10.1038/s41467-018-06854-2.
- [143] D D Sahtoe *et al.* “BAP1/ASXL1 recruitment and activation for H2A deubiquitination”. In: *Nat Commun* 7 (2016), p. 10292. DOI: 10.1038/ncomms10292.
- [144] Thomas Köhnke *et al.* “Human ASXL1-Mutant Hematopoiesis Is Driven by a Truncated Protein Associated with Aberrant Deubiquitination of H2AK119”. In: *Blood Cancer Discovery* 5.3 (2024), pp. 202–223. DOI: 10.1158/2643-3230.Bcd-23-0235. URL: <https://doi.org/10.1158/2643-3230.BCD-23-0235>.
- [145] J C Scheuermann *et al.* “Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB”. In: *Nature* 465 (7295 2010), pp. 243–247. DOI: 10.1038/nature08966. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20436459>.
- [146] O Abdel-Wahab *et al.* “ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression”. In: *Cancer Cell* 22 (2 2012), pp. 180–193. DOI: 10.1016/j.ccr.2012.06.032.

- [147] Omar Abdel-Wahab *et al.* “Deletion of *Asxl1* results in myelodysplasia and severe developmental defects in vivo”. In: *Journal of Experimental Medicine* 210 (12 2013), pp. 2641–2659. DOI: 10.1084/jem.20131141. URL: <https://doi.org/10.1084/jem.20131141>.
- [148] D. Inoue *et al.* “A novel ASXL1-OGT axis plays roles in H3K4 methylation and tumor suppression in myeloid malignancies”. In: *Leukemia* 32.6 (2018), pp. 1327–1337. DOI: 10.1038/s41375-018-0083-3.
- [149] P Zhou *et al.* “Mixed lineage leukemia 5 (MLL5) protein regulates cell cycle progression and E2F1-responsive gene expression via association with host cell factor-1 (HCF-1)”. In: *J Biol Chem* 288 (24 2013), pp. 17532–17543. DOI: 10.1074/jbc.M112.439729.
- [150] J Wang *et al.* “Loss of *Asxl1* leads to myelodysplastic syndrome-like disease in mice”. In: *Blood* 123 (4 2014), pp. 541–553. DOI: 10.1182/blood-2013-05-500272. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24255920>.
- [151] H Shi *et al.* “ASXL1 plays an important role in erythropoiesis”. In: *Sci Rep* 6 (2016), p. 28789. DOI: 10.1038/srep28789.
- [152] Antoine Campagne *et al.* “BAP1 complex promotes transcription by opposing PRC1-mediated H2A ubiquitylation”. In: *Nature Communications* 10.1 (2019), p. 348. DOI: 10.1038/s41467-018-08255-x. URL: <https://doi.org/10.1038/s41467-018-08255-x>.
- [153] H Yu *et al.* “The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression”. In: *Mol Cell Biol* 30 (21 2010), pp. 5071–5085. DOI: 10.1128/mcb.00396-10.
- [154] A Dey *et al.* “Loss of the tumor suppressor BAP1 causes myeloid transformation”. In: *Science* 337 (6101 2012), pp. 1541–1546. DOI: 10.1126/science.1221711.
- [155] Y Okino *et al.* “BRCA1-associated protein 1 (BAP1) deubiquitinase antagonizes the ubiquitin-mediated activation of FoxK2 target genes”. In: *J Biol Chem* 290 (3 2015), pp. 1580–1591. DOI: 10.1074/jbc.M114.609834.
- [156] H I Baymaz *et al.* “MBD5 and MBD6 interact with the human PR-DUB complex through their methyl-CpG-binding domain”. In: *Proteomics* 14 (19 2014), pp. 2179–2189. DOI: 10.1002/pmic.201400013.
- [157] Weiran Ge *et al.* “Basis of the H2AK119 specificity of the Polycomb repressive deubiquitinase”. In: *Nature* 616 (7955 2023), pp. 176–182. DOI: 10.1038/s41586-023-05841-y. URL: <https://doi.org/10.1038/s41586-023-05841-y>.
- [158] A Bohring *et al.* “Severe end of Opitz trigonocephaly (C) syndrome or new syndrome?” In: *Am J Med Genet* 85 (5 1999), pp. 438–446. DOI: 10.1002/(sici)1096-8628(19990827)85:5<438::aid-ajmg2>3.0.co;2-a.
- [159] L Zhao *et al.* “Transcription factor MEF2D is required for the maintenance of MLL-rearranged acute myeloid leukemia”. In: *Blood Adv* 5 (22 2021), pp. 4727–4740. DOI: 10.1182/bloodadvances.2021004469. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34597364>.

- [160] C L Fisher *et al.* “Loss-of-function Additional sex combs like 1 mutations disrupt hematopoiesis but do not cause severe myelodysplasia or leukemia”. In: *Blood* 115 (1 2010), pp. 38–46. DOI: 10.1182/blood-2009-07-230698. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19861679>.
- [161] Ben Kinnersley *et al.* “Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology”. In: *Nature Genetics* (2024). DOI: 10.1038/s41588-024-01785-9. URL: <https://doi.org/10.1038/s41588-024-01785-9>.
- [162] J B Micol and O Abdel-Wahab. “The Role of Additional Sex Combs-Like Proteins in Cancer”. In: *Cold Spring Harb Perspect Med* 6 (10 2016). DOI: 10.1101/cshperspect.a026526.
- [163] V Gelsi-Boyer *et al.* “ASXL1 mutation is associated with poor prognosis and acute transformation in chronic myelomonocytic leukaemia”. In: *Br J Haematol* 151 (4 2010), pp. 365–375. DOI: 10.1111/j.1365-2141.2010.08381.x.
- [164] R Itzykson *et al.* “Prognostic score including gene mutations in chronic myelomonocytic leukemia”. In: *J Clin Oncol* 31 (19 2013), pp. 2428–2436. DOI: 10.1200/jco.2012.47.3314.
- [165] V. Gelsi-Boyer *et al.* “Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia”. In: *Br J Haematol* 145.6 (2009), pp. 788–800. DOI: 10.1111/j.1365-2141.2009.07697.x.
- [166] T Haferlach *et al.* “Landscape of genetic lesions in 944 patients with myelodysplastic syndromes”. In: *Leukemia* 28 (2 2014), pp. 241–247. DOI: 10.1038/leu.2013.336.
- [167] J Boulton *et al.* “Frequent mutation of the polycomb-associated gene ASXL1 in the myelodysplastic syndromes and in acute myeloid leukemia”. In: *Leukemia* 24 (5 2010), pp. 1062–1065. DOI: 10.1038/leu.2010.20.
- [168] F Thol *et al.* “Prognostic significance of ASXL1 mutations in patients with myelodysplastic syndromes”. In: *J Clin Oncol* 29 (18 2011), pp. 2499–2506. DOI: 10.1200/jco.2010.33.4938.
- [169] A M Vannucchi *et al.* “Mutations and prognosis in primary myelofibrosis”. In: *Leukemia* 27 (9 2013), pp. 1861–1869. DOI: 10.1038/leu.2013.119.
- [170] P Lundberg *et al.* “Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms”. In: *Blood* 123 (14 2014), pp. 2220–2228. DOI: 10.1182/blood-2013-11-537167.
- [171] N Carbuccia *et al.* “Mutations of ASXL1 gene in myeloproliferative neoplasms”. In: *Leukemia* 23 (11 2009), pp. 2183–2186. DOI: 10.1038/leu.2009.141.
- [172] E Stieglitz *et al.* “The genomic landscape of juvenile myelomonocytic leukemia”. In: *Nat Genet* 47 (11 2015), pp. 1326–1333. DOI: 10.1038/ng.3400.
- [173] A Caye *et al.* “Juvenile myelomonocytic leukemia displays mutations in components of the RAS pathway and the PRC2 network”. In: *Nat Genet* 47 (11 2015), pp. 1334–1340. DOI: 10.1038/ng.3420.
- [174] W C Chou *et al.* “Distinct clinical and biological features of de novo acute myeloid leukemia with additional sex comb-like 1 (ASXL1) mutations”. In: *Blood* 116 (20 2010), pp. 4086–4094. DOI: 10.1182/blood-2010-05-283291.

- [175] J P Patel *et al.* “Prognostic relevance of integrated genetic profiling in acute myeloid leukemia”. In: *N Engl J Med* 366 (12 2012), pp. 1079–1089. DOI: 10.1056/NEJMoa1112304.
- [176] E Papaemmanuil *et al.* “Genomic Classification and Prognosis in Acute Myeloid Leukemia”. In: *N Engl J Med* 374 (23 2016), pp. 2209–2221. DOI: 10.1056/NEJMoa1516192.
- [177] G Marcucci, T Haferlach, and H Döhner. “Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications”. In: *J Clin Oncol* 29 (5 2011), pp. 475–486. DOI: 10.1200/jco.2010.30.2554.
- [178] J Huang *et al.* “Mutations of ASXL1 and TET2 in aplastic anemia”. In: *Haematologica* 100 (5 2015), e172–5. DOI: 10.3324/haematol.2014.120931.
- [179] T Yoshizato *et al.* “Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia”. In: *N Engl J Med* 373 (1 2015), pp. 35–47. DOI: 10.1056/NEJMoa1414799.
- [180] T C Chen *et al.* “Dynamics of ASXL1 mutation and other associated genetic alterations during disease progression in patients with primary myelodysplastic syndrome”. In: *Blood Cancer Journal* 4 (1 2014), e177–e177. DOI: 10.1038/bcj.2013.74. URL: <https://doi.org/10.1038/bcj.2013.74>.
- [181] Steven M Johnson *et al.* “A reappraisal of ASXL1 mutation sites and the cohesin-binding motif in myeloid disease”. In: *Blood Cancer Journal* 13 (1 2023), p. 96. DOI: 10.1038/s41408-023-00876-w. URL: <https://doi.org/10.1038/s41408-023-00876-w>.
- [182] J W Tyner *et al.* “Functional genomic landscape of acute myeloid leukaemia”. In: *Nature* 562 (7728 2018), pp. 526–531. DOI: 10.1038/s41586-018-0623-z.
- [183] S Schnittger *et al.* “ASXL1 exon 12 mutations are frequent in AML with intermediate risk karyotype and are independently associated with an adverse outcome”. In: *Leukemia* 27 (1 2013), pp. 82–91. DOI: 10.1038/leu.2012.262.
- [184] A Bidikian *et al.* “Prognostic impact of ASXL1 mutations in chronic phase chronic myeloid leukemia”. In: *Blood Cancer J* 12 (10 2022), p. 144. DOI: 10.1038/s41408-022-00742-1.
- [185] David A Sallman *et al.* “ASXL1 frameshift mutations drive inferior outcomes in CMML without negative impact in MDS”. In: *Blood Cancer Journal* 7 (12 2017), p. 633. DOI: 10.1038/s41408-017-0004-0. URL: <https://doi.org/10.1038/s41408-017-0004-0>.
- [186] A G Kulasekararaj *et al.* “Somatic mutations identify a subgroup of aplastic anemia patients who progress to myelodysplastic syndrome”. In: *Blood* 124 (17 2014), pp. 2698–2704. DOI: 10.1182/blood-2014-05-574889.
- [187] M Pratcorona *et al.* “Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value”. In: *Haematologica* 97 (3 2012), pp. 388–392. DOI: 10.3324/haematol.2011.051532.
- [188] H Döhner *et al.* “Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN”. In: *Blood* 140 (12 2022), pp. 1345–1377. DOI: 10.1182/blood.2022016867.

- [189] Y Guo *et al.* “ASXL1 alteration cooperates with JAK2V617F to accelerate myelofibrosis”. In: *Leukemia* 33 (5 2019), pp. 1287–1291. DOI: 10.1038/s41375-018-0347-y.
- [190] P Zhang *et al.* “Chromatin regulator Asxl1 loss and Nf1 haploinsufficiency cooperate to accelerate myeloid malignancy”. In: *J Clin Invest* 128 (12 2018), pp. 5383–5398. DOI: 10.1172/jci121366.
- [191] Véronique Gelsi-Boyer *et al.* “Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases”. In: *Journal of Hematology Oncology* 5 (1 2012), p. 12. DOI: 10.1186/1756-8722-5-12. URL: <https://doi.org/10.1186/1756-8722-5-12>.
- [192] D Inoue *et al.* “Truncation mutants of ASXL1 observed in myeloid malignancies are expressed at detectable protein levels”. In: *Exp Hematol* 44 (3 2016), 172–6 e1. DOI: 10.1016/j.exphem.2015.11.011. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26700326>.
- [193] M M Patnaik *et al.* “DNMT3A mutations are associated with inferior overall and leukemia-free survival in chronic myelomonocytic leukemia”. In: *Am J Hematol* 92 (1 2017), pp. 56–61. DOI: 10.1002/ajh.24581.
- [194] X Gao *et al.* “Role of ASXL1 in hematopoiesis and myeloid diseases”. In: *Exp Hematol* 115 (2022), pp. 14–19. DOI: 10.1016/j.exphem.2022.09.003.
- [195] A A Z Dawoud, W J Tapper, and N C P Cross. “Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking”. In: *Leukemia* 34 (10 2020), pp. 2660–2672. DOI: 10.1038/s41375-020-0896-8.
- [196] A P Bracken *et al.* “The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells”. In: *Genes Dev* 21 (5 2007), pp. 525–530. DOI: 10.1101/gad.415507.
- [197] S Asada *et al.* “Mutant ASXL1 cooperates with BAP1 to promote myeloid leukaemogenesis”. In: *Nat Commun* 9 (1 2018), p. 2733. DOI: 10.1038/s41467-018-05085-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30013160>.
- [198] A Balasubramani *et al.* “Cancer-associated ASXL1 mutations may act as gain-of-function mutations of the ASXL1-BAP1 complex”. In: *Nat Commun* 6 (2015), p. 7307. DOI: 10.1038/ncomms8307. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26095772>.
- [199] M Uni *et al.* “Modeling ASXL1 mutation revealed impaired hematopoiesis caused by derepression of p16Ink4a through aberrant PRC1-mediated histone modification”. In: *Leukemia* 33 (1 2019), pp. 191–204. DOI: 10.1038/s41375-018-0198-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29967380>.
- [200] R Nagase *et al.* “Expression of mutant Asxl1 perturbs hematopoiesis and promotes susceptibility to leukemic transformation”. In: *J Exp Med* 215 (6 2018), pp. 1729–1747. DOI: 10.1084/jem.20171151. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29643185>.

- [201] Hui Yang *et al.* “Gain of function of ASXL1 truncating protein in the pathogenesis of myeloid malignancies”. In: *Blood* 131 (3 2018), pp. 328–341. DOI: 10.1182/blood-2017-06-789669. URL: <https://doi.org/10.1182/blood-2017-06-789669>.
- [202] Y C Hsu *et al.* “The distinct biological implications of Asxl1 mutation and its roles in leukemogenesis revealed by a knock-in mouse model”. In: *J Hematol Oncol* 10 (1 2017), p. 139. DOI: 10.1186/s13045-017-0508-x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28697759>.
- [203] Grant A Challen *et al.* “Dnmt3a is essential for hematopoietic stem cell differentiation”. In: *Nature Genetics* 44 (1 2012), pp. 23–31.
- [204] Grant A Challen *et al.* “Dnmt3a and Dnmt3b Have Overlapping and Distinct Functions in Hematopoietic Stem Cells”. In: *Cell Stem Cell* 15 (3 2014), pp. 350–364. DOI: 10.1016/j.stem.2014.06.018. URL: <https://doi.org/10.1016/j.stem.2014.06.018>.
- [205] Cyril Quivoron *et al.* “TET2 Inactivation Results in Pleiotropic Hematopoietic Abnormalities in Mouse and Isxa0;a Recurrent Event during Human Lymphomagenesis”. In: *Cancer Cell* 20 (1 2011), pp. 25–38. DOI: 10.1016/j.ccr.2011.06.003. URL: <https://doi.org/10.1016/j.ccr.2011.06.003>.
- [206] K Moran-Crusio *et al.* “Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation”. In: *Cancer Cell* 20 (1 2011), pp. 11–24. DOI: 10.1016/j.ccr.2011.06.001. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21723200>.
- [207] T Fujino *et al.* “Mutant ASXL1 induces age-related expansion of phenotypic hematopoietic stem cells through activation of Akt/mTOR pathway”. In: *Nat Commun* 12 (1 2021), p. 1826. DOI: 10.1038/s41467-021-22053-y. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33758188>.
- [208] F Notta *et al.* “Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment”. In: *Science* 333 (6039 2011), pp. 218–221. DOI: 10.1126/science.1201219.
- [209] M. J. Kiel, G. L. Radice, and S. J. Morrison. “Lack of evidence that hematopoietic stem cells depend on N-cadherin-mediated adhesion to osteoblasts for their maintenance”. In: *Cell Stem Cell* 1.2 (2007), pp. 204–17. DOI: 10.1016/j.stem.2007.06.001.
- [210] L Wang *et al.* “Epigenetic targeted therapy of stabilized BAP1 in ASXL1 gain-of-function mutated leukemia”. In: *Nat Cancer* 2 (5 2021), pp. 515–526. DOI: 10.1038/s43018-021-00199-4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35122023>.
- [211] O H Yilmaz *et al.* “Pten dependence distinguishes haematopoietic stem cells from leukaemia-initiating cells”. In: *Nature* 441 (7092 2006), pp. 475–482. DOI: 10.1038/nature04703.
- [212] J Zhang *et al.* “PTEN maintains haematopoietic stem cells and acts in lineage choice and leukaemia prevention”. In: *Nature* 441 (7092 2006), pp. 518–522. DOI: 10.1038/nature04747.

- [213] L Ferrucci *et al.* “The origins of age-related proinflammatory state”. In: *Blood* 105 (6 2005), pp. 2294–2299. DOI: 10.1182/blood-2004-07-2599.
- [214] L Ferrucci and E Fabbri. “Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty”. In: *Nat Rev Cardiol* 15 (9 2018), pp. 505–522. DOI: 10.1038/s41569-018-0064-2.
- [215] Eric M Pietras *et al.* “Chronic interleukin-1 exposure drives haematopoietic stem cells towards precocious myeloid differentiation at the expense of self-renewal”. In: *Nature Cell Biology* 18 (6 2016), pp. 607–618. DOI: 10.1038/ncb3346. URL: <https://doi.org/10.1038/ncb3346>.
- [216] Katie A Matatall *et al.* “Chronic Infection Depletes Hematopoietic Stem Cells through Stress-Induced Terminal Differentiation”. In: *Cell Reports* 17 (10 2016), pp. 2584–2595. DOI: 10.1016/j.celrep.2016.11.031. URL: <https://doi.org/10.1016/j.celrep.2016.11.031>.
- [217] Ruzhica Bogeska *et al.* “Inflammatory exposure drives long-lived impairment of hematopoietic stem cell self-renewal activity and accelerated aging”. In: *Cell Stem Cell* 29 (8 2022), 1273–1284.e8. DOI: 10.1016/j.stem.2022.06.012. URL: <https://doi.org/10.1016/j.stem.2022.06.012>.
- [218] Carl A Mitchell *et al.* “Stromal niche inflammation mediated by IL-1 signalling is a targetable driver of haematopoietic ageing”. In: *Nature Cell Biology* 25 (1 2023), pp. 30–41. DOI: 10.1038/s41556-022-01053-0. URL: <https://doi.org/10.1038/s41556-022-01053-0>.
- [219] Daniel Hormaechea-Agulla *et al.* “Chronic infection drives Dnmt3a-loss-of-function clonal hematopoiesis via IFN γ 3b3; signaling”. In: *Cell Stem Cell* 28 (8 2021), 1428–1442.e6. DOI: 10.1016/j.stem.2021.03.002. URL: <https://doi.org/10.1016/j.stem.2021.03.002>.
- [220] Jennifer M SanMiguel *et al.* “Distinct Tumor Necrosis Factor Alpha Receptors Dictate Stem Cell Fitness versus Lineage Output in Dnmt3a-Mutant Clonal Hematopoiesis”. In: *Cancer Discovery* 12 (12 2022), pp. 2763–2773. DOI: 10.1158/2159-8290.Cd-22-0086. URL: <https://doi.org/10.1158/2159-8290.CD-22-0086>.
- [221] Zhigang Cai *et al.* “Inhibition of Inflammatory Signaling in Tet2^{-/-} Mutant Preleukemic Cells Mitigates Stress-Induced Abnormalities and Clonal Hematopoiesis”. In: *Cell Stem Cell* 23 (6 2018), 833–849.e5. DOI: 10.1016/j.stem.2018.10.013. URL: <https://doi.org/10.1016/j.stem.2018.10.013>.
- [222] Marlies Meisel *et al.* “Microbial signals drive pre-leukaemic myeloproliferation in a Tet2-deficient host”. In: *Nature* 557 (7706 2018), pp. 580–584. DOI: 10.1038/s41586-018-0125-z. URL: <https://doi.org/10.1038/s41586-018-0125-z>.
- [223] Francisco Caiado *et al.* “Aging drives Tet2^{+/+} clonal hematopoiesis via IL-1 signaling”. In: *Blood* 141 (8 2023), pp. 886–903. DOI: 10.1182/blood.2022016835. URL: <https://doi.org/10.1182/blood.2022016835>.

- [224] S Avagyan *et al.* “Resistance to inflammation underlies enhanced fitness in clonal hematopoiesis”. In: *Science* 374 (6568 2021), pp. 768–772. DOI: doi:10.1126/science.aba9304. URL: <https://www.science.org/doi/abs/10.1126/science.aba9304>.
- [225] K D Min *et al.* “Experimental ASXL1-Mediated Clonal Hematopoiesis Promotes Inflammation and Accelerates Heart Failure”. In: *J Am Heart Assoc* 11 (19 2022), e026154. DOI: 10.1161/jaha.122.026154.
- [226] A G Bick *et al.* “Inherited causes of clonal haematopoiesis in 97,691 whole genomes”. In: *Nature* 586 (7831 2020), pp. 763–768. DOI: 10.1038/s41586-020-2819-2.
- [227] Anna S Nam *et al.* “Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation”. In: *bioRxiv* (2022), p. 2022.01.14.476225. DOI: 10.1101/2022.01.14.476225. URL: <https://www.biorxiv.org/content/biorxiv/early/2022/01/16/2022.01.14.476225.full.pdf>.
- [228] Zhongwu Lai *et al.* “VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research”. In: *Nucleic Acids Research* 44 (11 2016), e108–e108. DOI: 10.1093/nar/gkw227. URL: <https://doi.org/10.1093/nar/gkw227>.
- [229] Kristian Cibulskis *et al.* “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples”. In: *Nature Biotechnology* 31 (3 2013), pp. 213–219. DOI: 10.1038/nbt.2514. URL: <https://doi.org/10.1038/nbt.2514>.
- [230] A Untergasser *et al.* “Primer3—new capabilities and interfaces”. In: *Nucleic Acids Res* 40 (15 2012), e115. DOI: 10.1093/nar/gks596.
- [231] Jian Ye *et al.* “Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction”. In: *BMC Bioinformatics* 13 (1 2012), p. 134. DOI: 10.1186/1471-2105-13-134. URL: <https://doi.org/10.1186/1471-2105-13-134>.
- [232] A Rodriguez-Meira *et al.* “Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing”. In: *Mol Cell* 73 (6 2019), 1292–1305 e8. DOI: 10.1016/j.molcel.2019.01.009. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30765193>.
- [233] Michael Hagemann-Jensen *et al.* “Single-cell RNA counting at allele and isoform resolution using Smart-seq3”. In: *Nature Biotechnology* 38 (6 2020), pp. 708–714. DOI: 10.1038/s41587-020-0497-0. URL: <https://doi.org/10.1038/s41587-020-0497-0>.
- [234] Robert A Amezcua *et al.* “Orchestrating single-cell analysis with Bioconductor”. In: *Nature Methods* 17 (2 2020), pp. 137–145. DOI: 10.1038/s41592-019-0654-x. URL: <https://doi.org/10.1038/s41592-019-0654-x>.
- [235] Yuhao Hao *et al.* “Dictionary learning for integrative, multimodal and scalable single-cell analysis”. In: *Nature Biotechnology* 42 (2 2024), pp. 293–304. DOI: 10.1038/s41587-023-01767-y. URL: <https://doi.org/10.1038/s41587-023-01767-y>.

- [236] Aaron T L. Lun, Karsten Bach, and John C Marioni. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. In: *Genome Biology* 17 (1 2016), p. 75. DOI: 10.1186/s13059-016-0947-7. URL: <https://doi.org/10.1186/s13059-016-0947-7>.
- [237] Tallulah S Andrews and Martin Hemberg. “M3Drop: dropout-based feature selection for scRNASeq”. In: *Bioinformatics* 35 (16 2018), pp. 2865–2867. DOI: 10.1093/bioinformatics/bty1044. URL: <https://doi.org/10.1093/bioinformatics/bty1044>.
- [238] Philip Brenneke *et al.* “Accounting for technical noise in single-cell RNA-seq experiments”. In: *Nature Methods* 10 (11 2013), pp. 1093–1095. DOI: 10.1038/nmeth.2645. URL: <https://doi.org/10.1038/nmeth.2645>.
- [239] Itay Tirosh *et al.* “Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq”. In: *Science* 352 (6282 2016), pp. 189–196. DOI: doi:10.1126/science.aad0501. URL: <https://www.science.org/doi/abs/10.1126/science.aad0501>.
- [240] Ilya Korsunsky *et al.* “Fast, sensitive and accurate integration of single-cell data with Harmony”. In: *Nature Methods* 16 (12 2019), pp. 1289–1296. DOI: 10.1038/s41592-019-0619-0. URL: <https://doi.org/10.1038/s41592-019-0619-0>.
- [241] Gennady Korotkevich *et al.* “Fast gene set enrichment analysis”. In: *bioRxiv* (2021), p. 60012. DOI: 10.1101/060012. URL: <https://www.biorxiv.org/content/biorxiv/early/2021/02/01/060012.full.pdf>.
- [242] Bingjie Zhang *et al.* “Characterizing cellular heterogeneity in chromatin state with scCUTTag-pro”. In: *Nature Biotechnology* 40 (8 2022), pp. 1220–1230. DOI: 10.1038/s41587-022-01250-0. URL: <https://doi.org/10.1038/s41587-022-01250-0>.
- [243] Lars Velten *et al.* “Human haematopoietic stem cell lineage commitment is a continuous process”. In: *Nature Cell Biology* 19 (4 2017), pp. 271–281. DOI: 10.1038/ncb3493. URL: <https://doi.org/10.1038/ncb3493>.
- [244] Noa Novershtern *et al.* “Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis”. In: *Cell* 144 (2 2011), pp. 296–309. DOI: <https://doi.org/10.1016/j.cell.2011.01.004>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867411000055>.
- [245] Roy Drissen *et al.* “Identification of two distinct pathways of human myelopoiesis”. In: *Science Immunology* 4 (35 2019), eaau7148. DOI: doi:10.1126/sciimmunol.aau7148. URL: <https://www.science.org/doi/abs/10.1126/sciimmunol.aau7148>.
- [246] Anindita Roy *et al.* “Transitions in lineage specification and gene regulatory networks in hematopoietic stem/progenitor cells over human development”. In: *Cell Reports* 36 (11 2021), p. 109698. DOI: <https://doi.org/10.1016/j.celrep.2021.109698>. URL: <https://www.sciencedirect.com/science/article/pii/S2211124721011451>.
- [247] S Z Xie *et al.* “Sphingosine-1-phosphate receptor 3 potentiates inflammatory programs in normal and leukemia stem cells to promote differentiation”. In: *Blood Cancer Discov* 2 (1 2021), pp. 32–53. DOI: 10.1158/2643-3230.Bcd-20-0155.

- [248] Dimitris Karamitros *et al.* “Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells”. In: *Nature Immunology* 19 (1 2018), pp. 85–97. DOI: 10.1038/s41590-017-0001-2. URL: <https://doi.org/10.1038/s41590-017-0001-2>.
- [249] Lu Chen *et al.* “Transcriptional diversity during lineage commitment of human blood progenitors”. In: *Science* 345 (6204 2014), p. 1251033. DOI: [doi:10.1126/science.1251033](https://doi.org/10.1126/science.1251033). URL: <https://www.science.org/doi/abs/10.1126/science.1251033>.
- [250] Joyce B Kang *et al.* “Efficient and precise single-cell reference atlas mapping with Symphony”. In: *Nature Communications* 12 (1 2021), p. 5890. DOI: 10.1038/s41467-021-25957-x. URL: <https://doi.org/10.1038/s41467-021-25957-x>.
- [251] Tal Galili. “dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering”. In: *Bioinformatics* 31 (22 2015), pp. 3718–3720. DOI: 10.1093/bioinformatics/btv428. URL: <https://doi.org/10.1093/bioinformatics/btv428>.
- [252] Daniel B Burkhardt *et al.* “Quantifying the effect of experimental perturbations at single-cell resolution”. In: *Nature Biotechnology* 39 (5 2021), pp. 619–629. DOI: 10.1038/s41587-020-00803-5. URL: <https://doi.org/10.1038/s41587-020-00803-5>.
- [253] Gabriel E Hoffman and Panos Roussos. “Dream: powerful differential expression analysis for repeated measures designs”. In: *Bioinformatics* 37 (2 2020), pp. 192–201. DOI: 10.1093/bioinformatics/btaa687. URL: <https://doi.org/10.1093/bioinformatics/btaa687>.
- [254] Matthew E Ritchie *et al.* “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43 (7 2015), e47–e47. DOI: 10.1093/nar/gkv007. URL: <https://doi.org/10.1093/nar/gkv007>.
- [255] Davis J McCarthy *et al.* “Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R”. In: *Bioinformatics* 33 (8 2017), pp. 1179–1186. DOI: 10.1093/bioinformatics/btw777. URL: <https://doi.org/10.1093/bioinformatics/btw777>.
- [256] M D Robinson, D J McCarthy, and G K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26 (1 2010), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.
- [257] Alexander J Tarashansky *et al.* “Self-assembling manifolds in single-cell RNA sequencing data”. In: *eLife* 8 (2019). Ed. by Alex K Shalek, Naama Barkai, and Itai Yanai, e48994. DOI: 10.7554/eLife.48994. URL: <https://doi.org/10.7554/eLife.48994>.
- [258] Sara Aibar *et al.* “SCENIC: single-cell regulatory network inference and clustering”. In: *Nature Methods* 14 (11 2017), pp. 1083–1086. DOI: 10.1038/nmeth.4463. URL: <https://doi.org/10.1038/nmeth.4463>.

- [259] Samuel A Lambert *et al.* “The Human Transcription Factors”. In: *Cell* 172 (4 2018), pp. 650–665. DOI: <https://doi.org/10.1016/j.cell.2018.01.029>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867418301065>.
- [260] Sven Turkalj *et al.* “GTAC enables parallel genotyping of multiple genomic loci with chromatin accessibility profiling in single cells”. In: *Cell Stem Cell* 30.5 (2023), 722–740.e11. DOI: <https://doi.org/10.1016/j.stem.2023.04.012>. URL: <https://www.sciencedirect.com/science/article/pii/S1934590923001303>.
- [261] Wei Xu *et al.* “A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility”. In: *Nature Protocols* 16 (8 2021), pp. 4084–4107. DOI: 10.1038/s41596-021-00583-5. URL: <https://doi.org/10.1038/s41596-021-00583-5>.
- [262] Xi Chen *et al.* “A rapid and robust method for single cell chromatin accessibility profiling”. In: *Nature Communications* 9 (1 2018), p. 5345. DOI: 10.1038/s41467-018-07771-0. URL: <https://doi.org/10.1038/s41467-018-07771-0>.
- [263] Tim Stuart *et al.* “Single-cell chromatin state analysis with Signac”. In: *Nature Methods* 18 (11 2021), pp. 1333–1341. DOI: 10.1038/s41592-021-01282-5. URL: <https://doi.org/10.1038/s41592-021-01282-5>.
- [264] Darren A Cusanovich *et al.* “Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing”. In: *Science* 348 (6237 2015), pp. 910–914. DOI: doi:10.1126/science.aab1601. URL: <https://www.science.org/doi/abs/10.1126/science.aab1601>.
- [265] A S Nam *et al.* “Somatic mutations and cell identity linked by Genotyping of Transcriptomes”. In: *Nature* 571 (7765 2019), pp. 355–360. DOI: 10.1038/s41586-019-1367-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31270458>.
- [266] Dominique Bonnet and John E Dick. “Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell”. In: *Nature Medicine* 3 (7 1997), pp. 730–737. DOI: 10.1038/nm0797-730. URL: <https://doi.org/10.1038/nm0797-730>.
- [267] O Abdel-Wahab *et al.* “The most commonly reported variant in ASXL1 (c.1934dupG;p.Gly646TrpfsX12) is not a somatic alteration”. In: *Leukemia* 24 (9 2010), pp. 1656–1657. DOI: 10.1038/leu.2010.144.
- [268] E Viguera, D Canceill, and S D Ehrlich. “Replication slippage involves DNA polymerase pausing and dissociation”. In: *Embo j* 20 (10 2001), pp. 2587–2595. DOI: 10.1093/emboj/20.10.2587.
- [269] C K Yannakou *et al.* “ASXL1 c.1934dup;p.Gly646Trpfs*12-a true somatic alteration requiring a new approach”. In: *Blood Cancer J* 7 (12 2017), p. 656. DOI: 10.1038/s41408-017-0025-8.
- [270] H Lee-Six *et al.* “Population dynamics of normal human blood inferred from somatic mutations”. In: *Nature* 561 (7724 2018), pp. 473–478. DOI: 10.1038/s41586-018-0497-0.

- [271] S L Berger and H L Cooper. “Very short-lived and stable mRNAs from resting human lymphocytes”. In: *Proceedings of the National Academy of Sciences* 72 (10 1975), pp. 3873–3877. DOI: doi:10.1073/pnas.72.10.3873. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.72.10.3873>.
- [272] M S Kowalczyk *et al.* “Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells”. In: *Genome Res* 25 (12 2015), pp. 1860–1872. DOI: 10.1101/gr.192237.115.
- [273] L García-Prat *et al.* “TFEB-mediated endolysosomal activity controls human hematopoietic stem cell fate”. In: *Cell Stem Cell* 28 (10 2021), 1838–1850.e10. DOI: 10.1016/j.stem.2021.07.003.
- [274] Bernhard Lehnertz *et al.* “HLF expression defines the human hematopoietic stem cell state”. In: *Blood* 138 (25 2021), pp. 2642–2654. DOI: 10.1182/blood.2021010745. URL: <https://doi.org/10.1182/blood.2021010745>.
- [275] Amir Giladi *et al.* “Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis”. In: *Nature Cell Biology* 20 (7 2018), pp. 836–846. DOI: 10.1038/s41556-018-0121-4. URL: <https://doi.org/10.1038/s41556-018-0121-4>.
- [276] Pia Sommerkamp *et al.* “The long non-coding RNA Meg3 is dispensable for hematopoietic stem cells”. In: *Scientific Reports* 9 (1 2019), p. 2110. DOI: 10.1038/s41598-019-38605-8. URL: <https://doi.org/10.1038/s41598-019-38605-8>.
- [277] M Cazzola. “Introduction to a review series: the 2016 revision of the WHO classification of tumors of hematopoietic and lymphoid tissues”. In: *Blood* 127 (20 2016), pp. 2361–2364. DOI: 10.1182/blood-2016-03-657379.
- [278] J Sulek *et al.* “Increased expression of the collagen internalization receptor uPARAP/Endo180 in the stroma of head and neck cancer”. In: *J Histochem Cytochem* 55 (4 2007), pp. 347–353. DOI: 10.1369/jhc.6A7133.2006.
- [279] D Wienke *et al.* “The collagen receptor Endo180 (CD280) Is expressed on basal-like breast tumor cells and promotes tumor growth in vivo”. In: *Cancer Res* 67 (21 2007), pp. 10230–10240. DOI: 10.1158/0008-5472.Can-06-3496.
- [280] X Gai *et al.* “MRC2 expression correlates with TGF β 1 and survival in hepatocellular carcinoma”. In: *Int J Mol Sci* 15 (9 2014), pp. 15011–15025. DOI: 10.3390/ijms150915011.
- [281] Yanli Jin *et al.* “Gas6/AXL Signaling Regulates Self-Renewal of Chronic Myelogenous Leukemia Stem Cells by Stabilizing β -Catenin”. In: *Clinical Cancer Research* 23 (11 2017), pp. 2842–2855. DOI: 10.1158/1078-0432.Ccr-16-1298. URL: <https://doi.org/10.1158/1078-0432.CCR-16-1298>.
- [282] Ana Catarina Menezes *et al.* “Increased expression of RUNX3 inhibits normal human myeloid development”. In: *Leukemia* 36 (7 2022), pp. 1769–1780. DOI: 10.1038/s41375-022-01577-2. URL: <https://doi.org/10.1038/s41375-022-01577-2>.
- [283] B Zhou *et al.* “Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562”. In: *Genome Res* 29 (3 2019), pp. 472–484. DOI: 10.1101/gr.234948.118.

- [284] Pernilla Eliasson *et al.* “Hypoxia mediates low cell-cycle activity and increases the proportion of long-term-reconstituting hematopoietic stem cells during in vitro culture”. In: *Experimental Hematology* 38 (4 2010), 301–310.e2. DOI: <https://doi.org/10.1016/j.exphem.2010.01.005>. URL: <https://www.sciencedirect.com/science/article/pii/S0301472X10000135>.
- [285] Robert Nechanitzky *et al.* “Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells”. In: *Nature Immunology* 14 (8 2013), pp. 867–875. DOI: 10.1038/ni.2641. URL: <https://doi.org/10.1038/ni.2641>.
- [286] S O’Byrne *et al.* “Discovery of a CD10-negative B-progenitor in human fetal life identifies unique ontogeny-related developmental programs”. In: *Blood* 134 (13 2019), pp. 1059–1071. DOI: 10.1182/blood.2019001289.
- [287] Robin D Lee *et al.* “Single-cell analysis identifies dynamic gene expression networks that govern B cell development and transformation”. In: *Nature Communications* 12 (1 2021), p. 6843. DOI: 10.1038/s41467-021-27232-5. URL: <https://doi.org/10.1038/s41467-021-27232-5>.
- [288] Leo D Wang and Marcus R Clark. “B-cell antigen-receptor signalling in lymphocyte development”. In: *Immunology* 110 (4 2003), pp. 411–420. DOI: <https://doi.org/10.1111/j.1365-2567.2003.01756.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2567.2003.01756.x>.
- [289] T Nagasawa. “The chemokine CXCL12 and regulation of HSC and B lymphocyte development in the bone marrow niche”. In: *Adv Exp Med Biol* 602 (2007), pp. 69–75. DOI: 10.1007/978-0-387-72009-8_9.
- [290] Y K Parrish *et al.* “IL-7 Dependence in human B lymphopoiesis increases during progression of ontogeny from cord blood to bone marrow”. In: *J Immunol* 182 (7 2009), pp. 4255–4266. DOI: 10.4049/jimmunol.0800489.
- [291] L A Freihat *et al.* “IRAK3 modulates downstream innate immune signalling through its guanylate cyclase activity”. In: *Scientific Reports* 9 (1 2019), p. 15468. DOI: 10.1038/s41598-019-51913-3. URL: <https://doi.org/10.1038/s41598-019-51913-3>.
- [292] Anthony S Zannas *et al.* “Epigenetic upregulation of FKBP5 by aging and stress contributes to NF- κ B-driven inflammation and cardiovascular risk”. In: *Proceedings of the National Academy of Sciences* 116 (23 2019), pp. 11370–11379. DOI: [doi:10.1073/pnas.1816847116](https://doi.org/10.1073/pnas.1816847116). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1816847116>.
- [293] M Kolesnichenko *et al.* “Transcriptional repression of NFKBIA triggers constitutive IKK- and proteasome-independent p53/RelA activation in senescence”. In: *Embo j* 40 (6 2021), e104296. DOI: 10.15252/embj.2019104296.
- [294] V Wittamer *et al.* “Specific recruitment of antigen-presenting cells by chemerin, a novel processed ligand from human inflammatory fluids”. In: *J Exp Med* 198 (7 2003), pp. 977–985. DOI: 10.1084/jem.20030382.
- [295] H Mumme *et al.* “Single-cell analysis reveals altered tumor microenvironments of relapse- and remission-associated pediatric acute myeloid leukemia”. In: *Nat Commun* 14 (1 2023), p. 6209. DOI: 10.1038/s41467-023-41994-0.

- [296] Luciano Nicosia *et al.* “Pharmacological inhibition of LSD1 triggers myeloid differentiation by targeting GSE1 oncogenic functions in AML”. In: *Oncogene* 41 (6 2022), pp. 878–894. DOI: 10.1038/s41388-021-02123-7. URL: <https://doi.org/10.1038/s41388-021-02123-7>.
- [297] Farzaneh Atashrazm and Sarah Ellis. “The polarity protein PARD3 and cancer”. In: *Oncogene* 40 (25 2021), pp. 4245–4262. DOI: 10.1038/s41388-021-01813-6. URL: <https://doi.org/10.1038/s41388-021-01813-6>.
- [298] Baoming Zhao *et al.* “Annexin A1 translocates to nucleus and promotes the expression of pro-inflammatory cytokines in a PKC-dependent manner after OGD/R”. In: *Scientific Reports* 6 (1 2016), p. 27028. DOI: 10.1038/srep27028. URL: <https://doi.org/10.1038/srep27028>.
- [299] L Thanh Nha Uyen *et al.* “PCDH17 functions as a common tumor suppressor gene in acute leukemia and its transcriptional downregulation is mediated primarily by aberrant histone acetylation, not DNA methylation”. In: *Int J Hematol* 111 (3 2020), pp. 451–462. DOI: 10.1007/s12185-019-02799-4.
- [300] Stephanie Z Xie *et al.* “Sphingolipid Modulation Activates Proteostasis Programs to Govern Human Hematopoietic Stem Cell Self-Renewal”. In: *Cell Stem Cell* 25 (5 2019), 639–653.e7. DOI: 10.1016/j.stem.2019.09.008. URL: <https://doi.org/10.1016/j.stem.2019.09.008>.
- [301] Xiaoren Tang *et al.* “LPS-induced TNF- α factor (LITAF)-deficient mice express reduced LPS-induced cytokine: Evidence for LITAF-dependent LPS signaling pathways”. In: *Proceedings of the National Academy of Sciences* 103 (37 2006), pp. 13777–13782. DOI: doi:10.1073/pnas.0605988103. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.0605988103>.
- [302] Max Karlsson *et al.* “A single-cell type transcriptomics map of human tissues”. In: *Science Advances* 7.31 (2021), eabh2169. DOI: 10.1126/sciadv.abh2169. URL: <https://doi.org/10.1126/sciadv.abh2169>.
- [303] A Fontalba, O Gutierrez, and J L Fernandez-Luna. “NLRP2, an inhibitor of the NF-kappaB pathway, is transcriptionally activated by NF-kappaB and exhibits a nonfunctional allelic variant”. In: *J Immunol* 179 (12 2007), pp. 8519–8524. DOI: 10.4049/jimmunol.179.12.8519.
- [304] K Yang *et al.* “TRIM21 is essential to sustain IFN regulatory factor 3 activation during antiviral response”. In: *J Immunol* 182 (6 2009), pp. 3782–3792. DOI: 10.4049/jimmunol.0803126.
- [305] P H Driggers *et al.* “An interferon gamma-regulated protein that binds the interferon-inducible enhancer element of major histocompatibility complex class I genes”. In: *Proc Natl Acad Sci U S A* 87 (10 1990), pp. 3743–3747. DOI: 10.1073/pnas.87.10.3743.
- [306] Andy G X Zeng *et al.* “Identification of a human hematopoietic stem cell subset that retains memory of inflammatory stress”. In: *bioRxiv* (2023), p. 2023.09.11.557271. DOI: 10.1101/2023.09.11.557271. URL: <https://www.biorxiv.org/content/biorxiv/early/2023/12/18/2023.09.11.557271.full.pdf>.

- [307] Ting Liu *et al.* “NF- κ B signaling in inflammation”. In: *Signal Transduction and Targeted Therapy* 2 (1 2017), p. 17023. DOI: 10.1038/sigtrans.2017.23. URL: <https://doi.org/10.1038/sigtrans.2017.23>.
- [308] Douglas E Wright *et al.* “Physiological Migration of Hematopoietic Stem and Progenitor Cells”. In: *Science* 294 (5548 2001), pp. 1933–1936. DOI: doi:10.1126/science.1064081. URL: <https://www.science.org/doi/abs/10.1126/science.1064081>.
- [309] J A Cancelas *et al.* “Rac GTPases differentially integrate signals regulating hematopoietic stem cell localization”. In: *Nat Med* 11 (8 2005), pp. 886–891. DOI: 10.1038/nm1274.
- [310] F C Yang *et al.* “Rac and Cdc42 GTPases control hematopoietic stem cell shape, adhesion, migration, and mobilization”. In: *Proc Natl Acad Sci U S A* 98 (10 2001), pp. 5614–5618. DOI: 10.1073/pnas.101546898.
- [311] Y Gu *et al.* “Hematopoietic cell regulation by Rac1 and Rac2 guanosine triphosphatases”. In: *Science* 302 (5644 2003), pp. 445–449. DOI: 10.1126/science.1088485.
- [312] Y Liu *et al.* “CD9, a potential leukemia stem cell marker, regulates drug resistance and leukemia development in acute myeloid leukemia”. In: *Stem Cell Res Ther* 12 (1 2021), p. 86. DOI: 10.1186/s13287-021-02155-6.
- [313] Richard A. Voit *et al.* “A genetic disorder reveals a hematopoietic stem cell regulatory network co-opted in leukemia”. In: *Nature Immunology* 24.1 (2023), pp. 69–83. DOI: 10.1038/s41590-022-01370-4. URL: <https://doi.org/10.1038/s41590-022-01370-4>.
- [314] Karolina Komorowska *et al.* “Hepatic Leukemia Factor Maintains Quiescence of Hematopoietic Stem Cells and Protects the Stem Cell Pool during Regeneration”. In: *Cell Reports* 21 (12 2017), pp. 3514–3523. DOI: <https://doi.org/10.1016/j.celrep.2017.11.084>. URL: <https://www.sciencedirect.com/science/article/pii/S2211124717317515>.
- [315] A Grover *et al.* “Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells”. In: *Nat Commun* 7 (2016), p. 11075. DOI: 10.1038/ncomms11075.
- [316] K C MacNamara *et al.* “Transient activation of hematopoietic stem and progenitor cells by IFN γ during acute bacterial infection”. In: *PLoS One* 6 (12 2011), e28669. DOI: 10.1371/journal.pone.0028669.
- [317] C R C Zhang *et al.* “Inflammatory cytokines promote clonal hematopoiesis with specific mutations in ulcerative colitis patients”. In: *Exp Hematol* 80 (2019), 36–41.e3. DOI: 10.1016/j.exphem.2019.11.008.
- [318] V van den Boom *et al.* “Non-canonical PRC1.1 Targets Active Genes Independent of H3K27me3 and Is Essential for Leukemogenesis”. In: *Cell Rep* 14 (2 2016), pp. 332–346. DOI: 10.1016/j.celrep.2015.12.034.
- [319] I Ben-Porath *et al.* “An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors”. In: *Nat Genet* 40 (5 2008), pp. 499–507. DOI: 10.1038/ng.127.

- [320] Zhonghua Gao *et al.* “An AUTS2–Polycomb complex activates gene expression in the CNS”. In: *Nature* 516 (7531 2014), pp. 349–354. DOI: 10.1038/nature13921. URL: <https://doi.org/10.1038/nature13921>.
- [321] B Sun *et al.* “Sox4 is required for the survival of pro-B cells”. In: *J Immunol* 190 (5 2013), pp. 2080–2089. DOI: 10.4049/jimmunol.1202736.
- [322] C Wang *et al.* “A critical role of RUNX1 in governing megakaryocyte-primed hematopoietic stem cell differentiation”. In: *Blood Adv* 7 (11 2023), pp. 2590–2605. DOI: 10.1182/bloodadvances.2022008591.
- [323] Daisuke Kurotaki *et al.* “IRF8 inhibits C/EBP α activity to restrain mononuclear phagocyte progenitors from differentiating into neutrophils”. In: *Nature Communications* 5 (1 2014), p. 4978. DOI: 10.1038/ncomms5978. URL: <https://doi.org/10.1038/ncomms5978>.
- [324] N Ahmed *et al.* “Blood stem cell PU.1 upregulation is a consequence of differentiation without fast autoregulation”. In: *J Exp Med* 219 (1 2022). DOI: 10.1084/jem.20202490.
- [325] Prashant Trikha *et al.* “E2f1–3 Are Critical for Myeloid Development*”. In: *Journal of Biological Chemistry* 286 (6 2011), pp. 4783–4795. DOI: <https://doi.org/10.1074/jbc.M110.182733>. URL: <https://www.sciencedirect.com/science/article/pii/S0021925820561802>.
- [326] Peiman Shooshtarizadeh *et al.* “Gfi1b regulates the level of Wnt/ β -catenin signaling in hematopoietic stem cells and megakaryocytes”. In: *Nature Communications* 10 (1 2019), p. 1270. DOI: 10.1038/s41467-019-09273-z. URL: <https://doi.org/10.1038/s41467-019-09273-z>.
- [327] P Kumar *et al.* “HMGA2 promotes long-term engraftment and myeloerythroid differentiation of human hematopoietic stem and progenitor cells”. In: *Blood Adv* 3 (4 2019), pp. 681–691. DOI: 10.1182/bloodadvances.2018023986.
- [328] M M Nakagawa, H Chen, and C V Rathinam. “Constitutive Activation of NF- κ B Pathway in Hematopoietic Stem Cells Causes Loss of Quiescence and Deregulated Transcription Factor Networks”. In: *Front Cell Dev Biol* 6 (2018), p. 143. DOI: 10.3389/fcell.2018.00143.
- [329] Roberto Avellino and Ruud Delwel. “Expression and regulation of C/EBP α in normal myelopoiesis and in malignant transformation”. In: *Blood* 129 (15 2017), pp. 2083–2091. DOI: 10.1182/blood-2016-09-687822. URL: <https://doi.org/10.1182/blood-2016-09-687822>.
- [330] R P DeKoter, M B Kamath, and I B Houston. “Analysis of concentration-dependent functions of PU.1 in hematopoiesis using mouse models”. In: *Blood Cells Mol Dis* 39 (3 2007), pp. 316–320. DOI: 10.1016/j.bcmd.2007.06.004.
- [331] H D Lacorazza *et al.* “The transcription factor MEF/ELF4 regulates the quiescence of primitive hematopoietic cells”. In: *Cancer Cell* 9 (3 2006), pp. 175–187. DOI: 10.1016/j.ccr.2006.02.017.
- [332] Namrata Rastogi *et al.* “Nuclear factor I-C overexpression promotes monocytic development and cell survival in acute myeloid leukemia”. In: *Leukemia* 37 (2 2023), pp. 276–287. DOI: 10.1038/s41375-022-01801-z. URL: <https://doi.org/10.1038/s41375-022-01801-z>.

- [333] M Shafarenko *et al.* “Fos modulates myeloid cell survival and differentiation and partially abrogates the c-Myc block in terminal myeloid differentiation”. In: *Blood* 103 (11 2004), pp. 4259–4267. DOI: 10.1182/blood-2002-09-2704.
- [334] A Tamir *et al.* “Fli-1, an Ets-related transcription factor, regulates erythropoietin-induced erythroid proliferation and differentiation: evidence for direct transcriptional repression of the Rb gene during differentiation”. In: *Mol Cell Biol* 19 (6 1999), pp. 4452–4464. DOI: 10.1128/mcb.19.6.4452.
- [335] Fong-Ying Tsai and Stuart H Orkin. “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”. In: *Blood* 89 (10 1997), pp. 3636–3643. DOI: 10.1182/blood.V89.10.3636. URL: <https://doi.org/10.1182/blood.V89.10.3636>.
- [336] A Vegiopoulos *et al.* “Coordination of erythropoiesis by the transcription factor c-Myb”. In: *Blood* 107 (12 2006), pp. 4703–4710. DOI: 10.1182/blood-2005-07-2968.
- [337] H C Masuoka and T M Townes. “Targeted disruption of the activating transcription factor 4 gene results in severe fetal anemia in mice”. In: *Blood* 99 (3 2002), pp. 736–745. DOI: 10.1182/blood.v99.3.736.
- [338] Zhaofeng Zheng *et al.* “The ATF4-RPS19BP1 axis modulates ribosome biogenesis to promote erythropoiesis”. In: *Blood* (2024). DOI: 10.1182/blood.2023021901. URL: <https://doi.org/10.1182/blood.2023021901>.
- [339] J. D. Blais *et al.* “Activating transcription factor 4 is translationally regulated by hypoxic stress”. In: *Mol Cell Biol* 24.17 (2004), pp. 7469–82. DOI: 10.1128/mcb.24.17.7469-7482.2004.
- [340] G. O. Nwosu, J. A. Powell, and S. M. Pitson. “Targeting the integrated stress response in hematologic malignancies”. In: *Exp Hematol Oncol* 11.1 (2022), p. 94. DOI: 10.1186/s40164-022-00348-0.
- [341] P. van Galen *et al.* “Integrated Stress Response Activity Marks Stem Cells in Normal Hematopoiesis and Leukemia”. In: *Cell Rep* 25.5 (2018), 1109–1117.e5. DOI: 10.1016/j.celrep.2018.10.021.
- [342] Boris Bartholdy and Patrick Matthias. “Transcriptional control of B cell development and function”. In: *Gene* 327 (1 2004), pp. 1–23. DOI: <https://doi.org/10.1016/j.gene.2003.11.008>. URL: <https://www.sciencedirect.com/science/article/pii/S0378111903010746>.
- [343] D. Soodgupta *et al.* “RAG-Mediated DNA Breaks Attenuate PU.1 Activity in Early B Cells through Activation of a SPIC-BCLAF1 Complex”. In: *Cell Rep* 29.4 (2019), 829–843.e5. DOI: 10.1016/j.celrep.2019.09.026.
- [344] E Laurenti *et al.* “The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment”. In: *Nat Immunol* 14 (7 2013), pp. 756–763. DOI: 10.1038/ni.2615.
- [345] Leonid Teytelman *et al.* “Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins”. In: *Proceedings of the National Academy of Sciences* 110 (46 2013), pp. 18602–18607. DOI: doi:10.1073/pnas.1316064110. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1316064110>.

- [346] Hatice S Kaya-Okur *et al.* “CUTTag for efficient epigenomic profiling of small samples and single cells”. In: *Nature Communications* 10 (1 2019), p. 1930. DOI: 10.1038/s41467-019-09982-5. URL: <https://doi.org/10.1038/s41467-019-09982-5>.
- [347] Peter J Skene and Steven Henikoff. “An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites”. In: *eLife* 6 (2017). Ed. by Danny Reinberg, e21856. DOI: 10.7554/eLife.21856. URL: <https://doi.org/10.7554/eLife.21856>.
- [348] Peter J Skene, Jorja G Henikoff, and Steven Henikoff. “Targeted in situ genome-wide profiling with high efficiency for low cell numbers”. In: *Nature Protocols* 13 (5 2018), pp. 1006–1019. DOI: 10.1038/nprot.2018.015. URL: <https://doi.org/10.1038/nprot.2018.015>.
- [349] Marek Bartosovic, Mukund Kabbe, and Gonalo Castelo-Branco. “Single-cell CUTTag profiles histone modifications and transcription factors in complex tissues”. In: *Nature Biotechnology* 39 (7 2021), pp. 825–835. DOI: 10.1038/s41587-021-00869-9. URL: <https://doi.org/10.1038/s41587-021-00869-9>.
- [350] Franco Izzo *et al.* “Mapping genotypes to chromatin accessibility profiles in single cells”. In: *Nature* 629 (8014 2024), pp. 1149–1157. DOI: 10.1038/s41586-024-07388-y. URL: <https://doi.org/10.1038/s41586-024-07388-y>.
- [351] H Quentmeier *et al.* “Cell line OCI/AML3 bears exon-12 NPM gene mutation-A and cytoplasmic expression of nucleophosmin”. In: *Leukemia* 19 (10 2005), pp. 1760–1767. DOI: 10.1038/sj.leu.2403899.
- [352] Robert T Abraham and Arthur Weiss. “Jurkat T cells and development of the T-cell receptor signalling paradigm”. In: *Nature Reviews Immunology* 4 (4 2004), pp. 301–308. DOI: 10.1038/nri1330. URL: <https://doi.org/10.1038/nri1330>.
- [353] B E Bernstein *et al.* “Genomic maps and comparative analysis of histone modifications in human and mouse”. In: *Cell* 120 (2 2005), pp. 169–181. DOI: 10.1016/j.cell.2005.01.001.
- [354] N D Heintzman *et al.* “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”. In: *Nat Genet* 39 (3 2007), pp. 311–318. DOI: 10.1038/ng1966.
- [355] Marek Bartosovic and Gonalo Castelo-Branco. “Multimodal chromatin profiling using nanobody-based single-cell CUTTag”. In: *Nature Biotechnology* 41 (6 2023), pp. 794–805. DOI: 10.1038/s41587-022-01535-4. URL: <https://doi.org/10.1038/s41587-022-01535-4>.
- [356] P Ikonomi *et al.* “Overexpression of GATA-2 inhibits erythroid and promotes megakaryocyte differentiation”. In: *Exp Hematol* 28 (12 2000), pp. 1423–1431. DOI: 10.1016/s0301-472x(00)00553-1.
- [357] M W K hn *et al.* “Targeting Chromatin Regulators Inhibits Leukemogenic Gene Expression in NPM1 Mutant Leukemia”. In: *Cancer Discov* 6 (10 2016), pp. 1166–1181. DOI: 10.1158/2159-8290.Cd-16-0237.

- [358] Jean-Marc Grégoire and Paul-Henri Roméo. “T-cell Expression of the Human *GATA-3* Gene Is Regulated by a Non-lineage-specific Silencer *”. In: *Journal of Biological Chemistry* 274 (10 1999), pp. 6567–6578. DOI: 10.1074/jbc.274.10.6567. URL: <https://doi.org/10.1074/jbc.274.10.6567>.
- [359] Alba Rodriguez-Meira *et al.* “Single-cell multi-omics identifies chronic inflammation as a driver of TP53-mutant leukemic evolution”. In: *Nature Genetics* 55 (9 2023), pp. 1531–1541. DOI: 10.1038/s41588-023-01480-1. URL: <https://doi.org/10.1038/s41588-023-01480-1>.
- [360] A G X Zeng *et al.* “Precise single-cell transcriptomic mapping of normal and leukemic cell states reveals unconventional lineage priming in acute myeloid leukemia”. In: *bioRxiv* (2023). DOI: 10.1101/2023.12.26.573390.
- [361] Z. Cao *et al.* “Hepatocyte growth factor (HGF) and stem cell factor (SCF) maintained the stemness of human bone marrow mesenchymal stem cells (hBMSCs) during long-term expansion by preserving mitochondrial function via the PI3K/AKT, ERK1/2, and STAT3 signaling pathways”. In: *Stem Cell Res Ther* 11.1 (2020), p. 329. DOI: 10.1186/s13287-020-01830-4.
- [362] Adam Greenbaum *et al.* “CXCL12 in early mesenchymal progenitors is required for haematopoietic stem-cell maintenance”. In: *Nature* 495.7440 (2013), pp. 227–230. DOI: 10.1038/nature11926. URL: <https://doi.org/10.1038/nature11926>.
- [363] S J Morrison and D T Scadden. “The bone marrow niche for haematopoietic stem cells”. In: *Nature* 505 (7483 2014), pp. 327–334. DOI: 10.1038/nature12984.
- [364] T Sugiyama *et al.* “Maintenance of the hematopoietic stem cell pool by CXCL12-CXCR4 chemokine signaling in bone marrow stromal cell niches”. In: *Immunity* 25 (6 2006), pp. 977–988. DOI: 10.1016/j.immuni.2006.10.016.
- [365] W Li *et al.* “Hematopoietic stem cell repopulating ability can be maintained in vitro by some primary endothelial cells”. In: *Exp Hematol* 32 (12 2004), pp. 1226–1237. DOI: 10.1016/j.exphem.2004.09.001.
- [366] B O Zhou *et al.* “Bone marrow adipocytes promote the regeneration of stem cells and haematopoiesis by secreting SCF”. In: *Nat Cell Biol* 19 (8 2017), pp. 891–903. DOI: 10.1038/ncb3570.
- [367] T. Bhattacharyya and J. Bond. “Polycomb Alterations in Acute Myeloid Leukaemia: From Structure to Function”. In: *Cancers (Basel)* 15.6 (2023). DOI: 10.3390/cancers15061693.
- [368] V. Di Carlo, I. Mocavini, and L. Di Croce. “Polycomb complexes in normal and malignant hematopoiesis”. In: *J Cell Biol* 218.1 (2019), pp. 55–69. DOI: 10.1083/jcb.201808028.
- [369] N A Fursova *et al.* “BAP1 constrains pervasive H2AK119ub1 to control the transcriptional potential of the genome”. In: *Genes Dev* 35 (9-10 2021), pp. 749–770. DOI: 10.1101/gad.347005.120.

-
- [370] Moritz Binder *et al.* “Oncogenic gene expression and epigenetic remodeling of cis-regulatory elements in ASXL1-mutant chronic myelomonocytic leukemia”. In: *Nature Communications* 13 (1 2022), p. 1434. DOI: 10.1038/s41467-022-29142-6. URL: <https://doi.org/10.1038/s41467-022-29142-6>.
- [371] A S Nam, R Chaligne, and D A Landau. “Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics”. In: *Nat Rev Genet* 22 (1 2021), pp. 3–18. DOI: 10.1038/s41576-020-0265-5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32807900>.
- [372] Yanxiao Feng *et al.* “Nanopore-based Fourth-generation DNA Sequencing Technology”. In: *Genomics, Proteomics Bioinformatics* 13 (1 2015), pp. 4–16. DOI: <https://doi.org/10.1016/j.gpb.2015.01.009>. URL: <https://www.sciencedirect.com/science/article/pii/S1672022915000133>.