Characterising the Cellular and Molecular Basis of Age-Related Clonal Haematopoiesis



Asger Jakobsen

Corpus Christi College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Acknowledgements

I would like to thank my supervisor Prof. Paresh Vyas for his support and scientific mentorship throughout my DPhil and always encouraging me to follow my curiosity. Many members of the Vyas lab have been extremely helpful in discussing scientific ideas, and giving me advice and feedback. I would particularly like to thank Marlen Metzner and Bilyana Stoilova for teaching me a lot of molecular biology when I started at the lab, Rachel Moore and Alison Kennedy who helped me with targeted DNA sequencing experiments, and Batchimeg Usukhbayar, Mirian Salazar and Jessica Doondeea for their help with banking the many bone marrow samples. Thank you to Grigore Gafencu, Supat Thongjuea, Thomas Höfer and Naser Ansari-Pour for their advice on bioinformatic analysis, and to Vasiliki Symeonidou, Susann Rahmig, David Cruz Hernandez, Connor Sweeney, Sarah Gooding, Hedia Chagraoui and Lynn Quek. A special thanks to Sven Turkalj who joined the lab in the midst of the pandemic in 2020, and helped me with the experiments in Chapter 5. He has always been cheerful, eager to help with late-night FACS sorts and discussing ideas. Thanks also to Jennifer O'Sullivan and Alba Rodriguez-Meira for their advice and help when I was starting my TARGET-seq experiments.

At the Botnar Research Centre and the Nuffield Orthopaedic Hospital, I would particularly like to thank Kim Wheway, Bridget Watkins and Debra Beazley for all their help with getting the MARCH Study set up and recruiting patients, without whom this project would not have been feasible. Thank you also to Prof. Andrew Carr for supervising the study, and to Simon Newman, Benjamin Kendrick, Adrian Taylor, Rasheed Afinowi-Luitz and Roger Gundle for collecting samples.

Thank you also to my co-supervisor Prof. Doug Higgs, who has been a wonderful mentor, as well as the other members of my Thesis Committee, Professors Claus Nerlov and Richard Cornall.

I would like to acknowledge the Medical Research Council and Leuka for their funding which enabled me to do my DPhil.

Most importantly, I would like to thank my family and friends who have been a huge support both before and during my DPhil. Thank you especially to my wife Hannah, who has been incredibly patient and understanding every time my experiments ran late and encouraged me through all the hard moments. She and Walter have kept me centred and provided much-needed distraction at home and on our walks. Thank you to my dad, whose inspiration led me to pursue science, and to my mum for her unwavering patience and love. I am grateful to my friends, especially Claire, Tim, Francis, Simon and Kipp who are just amazing.

Finally, I would like to thank all the patients who have donated their samples for this study. Without their generosity, none of this work would have been possible.

Abstract

Throughout life, tissue stem cells accumulate somatic mutations. Occasionally, a mutation may occur that confers a selective advantage leading to growth of an expanded clone. When this occurs in haematopoietic stem cells (HSC), this is called clonal haematopoiesis. Clonal haematopoiesis is common in healthy older people, and is associated with increased risk of haematological malignancy and other adverse outcomes. Surprisingly, most cases of clonal haematopoiesis are associated with mutations in just two genes, *DNMT3A* and *TET2*, which regulate DNA methylation. Mutations in these genes are often the first genetic event in myeloid and lymphoid malignancies and may be acquired many years before the onset of disease. The mechanisms by which mutations in *DNMT3A* and *TET2* cause clonal expansion are poorly understood and no studies have analysed their consequences in humans with unperturbed clonal haematopoiesis.

This study characterises the cellular and molecular consequences of somatic *DNMT3A* and *TET2* mutations in humans with clonal haematopoiesis. Blood and bone marrow samples were collected from a cohort of 195 individuals undergoing total hip replacement surgery. DNA sequencing identified 85 cases with clonal haematopoiesis with a similar profile of mutations to those in the general population. Most individuals had normal blood counts and bone marrow immunophenotyping showed that clonal haematopoiesis is usually associated with normal stem and progenitor frequencies.

I adapted the TARGET-seq method to develop a more sensitive method for combined single cell genotyping with whole transcriptome sequencing. The new TARGET-seq+ method detected more genes per cell with lower dropout rates and reduced technical variation. I applied TARGET-seq+ to 12,847 single cells from 10 samples with clonal haematopoiesis and 3 controls. *DNMT3A* and *TET2*-mutant haematopoietic stem and progenitor cells did not occupy novel transcriptional states, but intermingled with wild-type cells, showing they follow a common differentiation trajectory. Expansion of *DNMT3A*-mutant clones occurred mainly in HSC and the greatest clone size was reached in early multipotent progenitors. In contrast, *TET2*-mutant clones expanded later in differentiation with a transcriptional bias towards the myeloid lineage. These findings highlight distinct effects of *DNMT3A* and *TET2* mutations on haematopoietic stem and progenitor cells and identify dysregulated pathways that may have a role in expansion of the mutant clone.

Word Count: 35710

Contents

List of Figures viii

List of Tables xi

List of Abbreviations xii

1 Introduction 1

1.1 Clonal haematopoiesis and its role in ageing and disease 2

1.1.1 Somatic clonal expansion in ageing 2

1.1.2 Brief history of clonal haematopoiesis 4

1.1.3 Genetic drivers of age-related clonal haematopoiesis 6

1.1.4 Clonal haematopoiesis and risk of haematological cancer . . 9

1.1.5 Clonal haematopoiesis and non-malignant disease 11

1.1.6 Clonal dynamics across the human lifespan 12

1.2 A brief introduction to DNA methylation and the DNMT and TET enzymes 14

1.3 The function of DNMT3A in haematopoiesis and myeloid malignancy 16

1.3.1 Structure and function of DNMT3A 16

1.3.2 *DNMT3A* mutations in CH and myeloid malignancies 18

1.3.3 The role of DNMT3A in haematopoiesis 19

1.4 The function of TET2 in haematopoiesis and myeloid malignancy . 22

1.4.1 Structure and function of TET2 22

1.4.2 *TET2* mutations in CH and myeloid malignancies 23

1.4.3 The role of TET2 in haematopoiesis 24

1.5 Aims 25

2 Materials and Methods 27

2.1 Study participants 27

2.2 Sample collection and processing 28

2.3 Targeted DNA sequencing 29

2.3.1 Library preparation and sequencing 29

2.3.2 Read pre-processing and mapping for targeted DNA sequencing 30

2.3.3 Somatic variant calling 30

2.3.4 Annotation of pathogenic driver variants 32

2.3.5 Comparison of bone marrow and peripheral blood allele frequencies 33

2.3.6 Statistical analysis of targeted sequencing 33
2.4 Cell culture 34
2.5 Flow cytometry and cell sorting 34
2.6 TARGET-seq+ optimisation 35
2.6.1 Comparison of RT-PCR conditions and DNA polymerases . 36
2.6.2 Comparison of template-switching oligos 36
2.6.3 Comparison of targeted primer concentrations 36
2.7 TARGET-seq library preparation 37
2.8 Design and validation of targeted genotyping primers for TARGET-seq+ 38
2.9 TARGET-seq+ library preparation 40
2.9.1 Lysis buffer preparation 40
2.9.2 Reverse transcription and pre-amplification 41
2.9.3 Whole-transcriptome library preparation and sequencing . . 42
2.9.4 Targeted single-cell genotyping 43
2.10 Single-cell genotyping analysis 44
2.10.1 Mutational calling 44
2.10.2 Allelic dropout estimation 45
2.10.3 Inference of clonal hierarchies 46
2.11 Analysis of FACS index data 47
2.12 Single-cell transcriptome data pre-processing 48
2.12.1 Mapping and transcript counting 48
2.12.2 Transcript detection and dropout frequency calculation . . . 48
2.12.3 Cell-to-cell correlation analysis 49
2.13 Single-cell transcriptome analysis 49
2.13.1 Quality control and normalisation of single-cell transcriptome data 49
2.13.2 Dimensionality reduction and clustering for individual samples 50
2.13.3 Single-cell copy number analysis 50
2.13.4 Data integration and visualisation 51
2.13.5 Cell cycle and lineage signature scores 51
2.13.6 Clustering analysis, marker gene identification, and cell type annotation 51
2.13.7 Differential abundance analysis 52
2.13.8 Pseudotime analysis 53
2.13.9 Differential gene expression analysis 54
2.13.10 Gene set enrichment analysis 54
2.13.11 AUCell analysis 55

3 Collection and characterisation of a clonal haematopoiesis cohort 56
3.1 Establishing the MARCH cohort 56
3.1.1 Introduction to the MARCH Study 56
3.1.2 Participant recruitment 57
3.1.3 BM sample processing and cell yields 58
3.2 Identification of cases with clonal haematopoiesis by targeted DNA sequencing 59
3.2.1 Approach to targeted DNA sequencing 59
3.2.2 Somatic variant calling 60
3.2.3 Genetic landscape of cases with clonal haematopoiesis 62
3.2.4 Association of clonal haematopoiesis with clinical parameters 66
3.2.5 Molecular characteristics of somatic variants 68
3.3 Comparison of somatic variants between BM and PB 70
3.4 Bone marrow immunophenotypic analysis 72
3.4.1 Effect of age on HSPC frequencies 73
3.4.2 Effect of CH on HSPC immunophenotypes 74
3.5 Conclusions: Collection and characterisation of a clonal haematopoiesis cohort 78

4 Development of TARGET-seq+ 80
4.1 Introduction and rationale for the development of TARGET-seq+ . 81
4.1.1 Current methods for single cell genotyping and transcriptome sequencing 81
4.1.2 Review of the TARGET-seq method 84
4.1.3 Review of Smart-seq3 chemistry 86
4.1.4 Aims in developing TARGET-seq+ 87
4.2 Modifications to the method and pilot experiments 87
4.3 Validation of TARGET-seq+ in cell lines and primary haematopoietic cells 91
4.3.1 Targeted genotyping with TARGET-seq+ 91
4.3.2 Transcriptome sequencing with TARGET-seq+ 93
4.4 Mutational calling in single-cell genotyping 98
4.5 Allelic dropout and accuracy of single-cell genotyping 101
4.6 Conclusions: Development of TARGET-seq+ 102

5 Single-cell genomic analysis of *DNMT3A* and *TET2*-mutant clonal haematopoiesis 105
5.1 Approach to single-cell analysis of clonal haematopoiesis samples . . 105
5.1.1 Aims in single-cell analysis 105
5.1.2 Experimental design 108
5.1.3 Sample selection 108
5.2 Single-cell genotyping analysis of clonal haematopoiesis samples . . 110
5.2.1 Data QC and filtering 110
5.2.2 Clonal analysis of immunophenotypic populations for *DNMT3A*-mutant samples 113
5.2.3 Surface marker expression in *DNMT3A*-mutant populations 119
5.2.4 Clonal analysis of immunophenotypic populations for *TET2*-mutant samples 119
5.3 Single-cell transcriptome landscape of normal and clonal haematopoiesis samples 127
5.3.1 Data QC and filtering 127
5.3.2 Copy number variation analysis 129
5.3.3 Dimensionality reduction and clustering 129
5.3.4 Integrated analysis of transcriptomic data for normal and clonal haematopoiesis samples 135
5.3.5 Expression of *DNMT3A* and *TET2* in HSPCs 147
5.4 Transcriptomic analysis of *DNMT3A*-mutant clonal haematopoiesis 149
5.4.1 Analysis of *DNMT3A*-mutant clonal size through haematopoietic differentiation 149
5.4.2 Transcriptional consequences of *DNMT3A* mutations in progenitors 154
5.4.3 Transcriptional consequences of *DNMT3A* mutations in HSC 161
5.5 Transcriptomic analysis of *TET2*-mutant clonal haematopoiesis . . 166
5.5.1 Analysis of *TET2*-mutant clonal size through haematopoietic differentiation 166
5.5.2 Transcriptional consequences of *TET2* mutations in progenitors 170
5.5.3 Transcriptional consequences of *TET2* mutations in HSC . . 173
5.5.4 Consequences of biallelic *TET2* mutations 179

6 Discussion 182
6.1 Establishing a human clonal haematopoiesis cohort 183
6.2 Developing the TARGET-seq+ method 185
6.3 Transcriptional consequences of *DNMT3A* and *TET2* mutations in haematopoietic stem and progenitor cells 187
6.4 Limitations of this study and future work 193

Appendices

A Appendices **196**
A.1 Clinical characteristics of the MARCH cohort 196
A.2 Targeted DNA sequencing 200
A.3 Flow cytometry antibodies 204
A.4 Reagents and conditions used for cDNA synthesis in TARGET-seq . 205
A.5 Reagents and conditions used for cDNA synthesis in TARGET-seq+ 206
A.6 Oligonucleotides used in TARGET-seq+ 207

References **217**

List of Figures

1.1 Neutral and selected somatic mutations in ageing 3
1.2 DNA methylation and demethylation 15
1.3 Structure of DNMT3A 16
1.4 Structure of TET2 23

2.1 Validation of targeted primers for a genotyping locus. 40

3.1 Participant recruitment for the MARCH cohort. 58
3.2 Bone marrow mononuclear cell isolation. 59
3.3 Power to detect a variant with VAF of 0.02 as a function of coverage depth. 60
3.4 Mapping and coverage metrics for targeted sequencing. 61
3.5 Somatic variant calling in BM targeted sequencing. 62
3.6 Landscape of somatic mutations in the MARCH cohort. 64
3.7 Prevalence of mutations with age. 65
3.8 Blood indices in CH cases and controls. 67
3.9 Mutations detected by gene. 68
3.10 Distribution of mutations within common genes. 69
3.11 Variant allele frequencies in commonly mutated genes. 70
3.12 Comparison of variants detected in BM and PB sequencing. 71
3.13 Gating strategy for immunophenotypic analysis. 73
3.14 Effect of age on HSPC frequencies. 75
3.15 Effect of CH on HSPC frequencies. 76
3.16 Effect of CH on CD117 and CD38 expression. 77

4.1 Schematic representation of the 3'-TARGET-seq method. 84
4.2 Testing RT-PCR conditions for TARGET-seq+. 89
4.3 Testing targeted amplification of a genotyping locus with TARGET-seq+. 90
4.4 Structure of 3' transcriptome libraries for the TARGET-seq+ method. 91
4.5 Detection of mutations with TARGET-seq+. 92
4.6 Quality metrics for TARGET-seq+ validation. 94
4.7 Transcript quantification with TARGET-seq and TARGET-seq+. . 95
4.8 PCA and UMAP of cells analysed with TARGET-seq and TARGET-seq+. 96

4.9 Sensitivity of TARGET-seq+ for transcript detection. 97
4.10 Dropout rates and reproducibility with TARGET-seq+. 98
4.11 Setting the minimum coverage in single cell genotyping. 99
4.12 Mutation calling in single cell genotyping. 100
4.13 Measurement of allelic dropout in germline SNPs. 103
4.14 Estimation of allelic dropout in gDNA and cDNA amplicons. 104

5.1 Sorting single cells for TARGET-seq+ analysis 107
5.2 Coverage of gDNA amplicons in plates across FACS sorts 111
5.3 Single-cell genotyping of CH samples 112
5.4 Frequency of *DNMT3A*-mutant cells in NOC108 HSPC compartments 113
5.5 Frequency of *DNMT3A*-mutant cells in NOC132 HSPC compartments 114
5.6 Frequency of *DNMT3A*-mutant cells in NOC171 HSPC compartments 115
5.7 Clonal hierarchy and frequency of mutant cells in NOC131 HSPC compartments 117
5.8 Clonal hierarchy and frequency of mutant cells in NOC062 HSPC compartments 118
5.9 CD34 expression in *DNMT3A*-mutant versus WT cells 119
5.10 Frequency of *TET2*-mutant cells in NOC031 HSPC compartments . 120
5.11 Clonal hierarchy and frequency of mutant cells in NOC117 HSPC compartments 121
5.12 Clonal hierarchy and frequency of mutant cells in NOC002 HSPC compartments 123
5.13 Clonal hierarchy and frequency of mutant cells in NOC115 HSPC compartments 125
5.14 Clonal hierarchy and frequency of mutant cells in NOC137 HSPC compartments 126
5.15 Quality control of RNA sequencing 127
5.16 Library size and genes detected per cell 128
5.17 Copy number analysis of cells from control and clonal haematopoiesis samples 130
5.18 UMAP of individual *DNMT3A*-mutant samples 132
5.19 UMAP of individual *TET2*-mutant samples 134
5.20 Harmony integration of transcriptomic data from control and clonal haematopoiesis samples 136
5.21 Identification of clusters in single cell transcriptome data 137
5.22 Gene set enrichment analysis of cluster marker genes 138
5.23 Differentially gene expression amongst HSC and MPP clusters . . . 140
5.24 Identification of megakaryocytic and erythroid progenitors 141
5.25 Differential gene expression amongst lympho-myeloid progenitors . . 143

5.26 Identification of myeloid progenitors and precursors 144
5.27 Identification of lymphoid progenitors and B cell precursors 145
5.28 Identification of T, NK, plasma and endothelial cells 146
5.29 Frequency of cell types in control and CH samples 147
5.30 *DNMT3A* expression through differentiation 148
5.31 *TET2* expression through differentiation 149
5.32 Abundance of *DNMT3A*-mutant cells across the haematopoietic hierarchy 152
5.33 Pseudotime analysis of *DNMT3A*-mutant cells 153
5.34 Pseudotime for individual *DNMT3A*-mutant samples 154
5.35 Differential gene expression analysis in *DNMT3A*-mutant LMPP . . 156
5.36 Differential gene expression analysis in *DNMT3A*-mutant erythroid-megakaryocyte progenitors 158
5.37 Expression of lineage marker genes in *DNMT3A*-mutant progenitors 159
5.38 Expression of HSC marker genes in *DNMT3A*-mutant HSC and early progenitors 160
5.39 Quiescence and lineage bias in *DNMT3A*-mutant HSC/MPP 163
5.40 Differential gene expression analysis in *DNMT3A*-mutant HSC/MPP 165
5.41 Abundance of *TET2*-mutant cells across the haematopoietic hierarchy 168
5.42 Pseudotime analysis of *TET2*-mutant cells 169
5.43 Pseudotime for individual *TET2*-mutant samples 170
5.44 Differential gene expression analysis in *TET2*-mutant GMP 172
5.45 Differential gene expression analysis in *TET2*-mutant LMPP 173
5.46 Quiescence and lineage bias in *TET2*-mutant HSC/MPP 175
5.47 Differential gene expression analysis in *TET2*-mutant HSC/MPP . 176
5.48 Expression of mitochondrial genes in the myeloid lineage in *TET2*-mutant samples 178
5.49 Differentially expressed transcription factors in *TET2*-mutant samples 179
5.50 Consequences of biallelic *TET2*-mutations in sample NOC002 . . . 181

List of Tables

3.1 Summary of clinical characteristics. 66
3.2 Summary of blood cell parameters in the cohort. 67
3.3 Frequency of HSPC populations in control BM samples. 74

5.1 *DNMT3A*-mutant samples analysed by TARGET-seq+ 109
5.2 *TET2*-mutant samples analysed by TARGET-seq+ 110

A.1 Clinical characteristics of the MARCH cohort. 197
A.2 Coverage of targeted sequencing panel. 200
A.3 Variants detected by targeted BM DNA sequencing. 202
A.4 Antibodies used for immunophenotypic analysis of BM HSPCs . . . 204
A.5 Antibodies used for sorting BM samples for TARGET-seq+ 204
A.6 TARGET-seq lysis buffer 205
A.7 TARGET-seq RT buffer 205
A.8 TARGET-seq pre-amplification PCR mix 205
A.9 TARGET-seq+ lysis buffer 206
A.10 TARGET-seq+ RT buffer 206
A.11 TARGET-seq+ pre-amplification PCR mix 206
A.12 Common oligonucleotides used for TARGET-seq+ 207
A.13 Custom i5 transcriptome index primers 207
A.14 Barcoded OligodT-ISPCR primers. 208
A.15 Targeted pre-amplification genotyping primers 214
A.16 Genotyping barcoding primers 215

List of Abbreviations

5hmC 5-hydroxymethylcytosine

5mC 5-methylcytosine

ADO . . . . . Allelic dropout

α-KG . . . . . α-ketoglutarate

AML . . . . . Acute myeloid leukaemia

bp . . . . . Base pair

BM . . . . . Bone marrow

B-NK . . . . . B-natural killer cell progenitor

cDNA . . . . . Complementary DNA

CH . . . . . Clonal haematopoiesis

CHIP . . . . . Clonal haematopoiesis of indeterminate potential

CLP . . . . . Common lymphoid progenitor

CMML . . . . Chronic myelomonocytic leukaemia

CMP . . . . . Common myeloid progenitor

COSMIC . . . Catalogue of Somatic Mutations in Cancer

DC . . . . . Dendritic cell

DMSO . . . . . Dimethyl sulfoxide

DNA . . . . . Deoxyribonucleic acid

DNMT . . . . DNA methyltransferase

DTT 1,4-dithiothreitol

EDTA . . . . . Ethylenediaminetetraacetic

EMPP . . . . . Erythroid/megakaryocyte-primed multipotent progenitor

FACS . . . . . Fluorescence-activated cell sorting

FBS . . . . . Fetal bovine serum

FC . . . . . Fold change

FDR . . . . . False discovery rate

FMO . . . . . Fluorescence-minus-one

gDNA . . . . . Genomic DNA

G-CSF . . . . Granulocyte colony stimulating factor

GMP . . . . . Granulocyte-monocyte progenitor

GSEA . . . . . Gene set enrichment analysis

GTP . . . . . Guanosine triphosphate

HSC . . . . . Haematopoietic stem cell

HSPC . . . . . Haematopoietic stem and progenitor cell

IFN-γ . . . . . Interferon gamma

IGV . . . . . Integrative Genomics Viewer

IL . . . . . Interleukin

IMDM . . . . Iscove's Modified Dulbecco's Medium

Indel . . . . . Insertion or deletion

LMPP . . . . . Lymphoid-primed multipotent progenitor

LNA . . . . . Locked nucleic acid

lncRNA . . . . Long non-coding RNA

Lin . . . . . Lineage

LT-HSC . . . . Long-term haematopoietic stem cell

mCA . . . . . Mosaic chromosomal alteration

MDS . . . . . Myelodysplastic syndrome

MEP . . . . . Megakaryocyte-erythroid progenitor

MFI . . . . . Median fluorescence intensity

MkP . . . . . Megakaryocyte progenitor

MLP . . . . . Multi-lymphoid progenitor

MNC . . . . . Mononuclear cell

MPP . . . . . Multipotent progenitor

mRNA . . . . Messenger ribonucleic acid (RNA)

NOC . . . . . Nuffield Orthopaedic Centre

NK cell . . . . Natural killer cell

P-adj . . . . . Adjusted P value

PB . . . . . Peripheral blood

PBMC . . . . Peripheral blood mononuclear cell

PBS . . . . . Phosphate-buffered saline

PC . . . . . Principal component

PCR . . . . . Polymerase chain reaction

pDC . . . . . Plasmacytoid dendritic cell

polyA . . . . . Polyadenylated

PEG . . . . . Polyethylene glycol

RNA . . . . . Ribonucleic acid

RNA-Seq . . . RNA sequencing

RPMI-1640 . Roswell Park Memorial Institute medium

RT . . . . . Reverse transcription

RT-PCR . . . Reverse transcription polymerase chain reaction

scRNA-Seq . . Single cell RNA sequencing

SD . . . . . Standard deviation

SEM . . . . . Standard error of the mean

SNP . . . . . Single nucleotide polymorphism

SNV . . . . . Single nucleotide variant

ST-HSC . . . . Short-term haematopoietic stem cell

TET2 . . . . . Tet methylcytosine dioxygenase 2

THR . . . . . Total hip replacement

TNFα . . . . . Tumour necrosis factor alpha

TS . . . . . TARGET-seq

TS+ . . . . . TARGET-seq+

TSO . . . . . Template-switching oligo

UMAP . . . . Uniform manifold approximation and projection

UMI . . . . . Unique molecular identifier

VAF . . . . . Variant allele frequency

VUS . . . . . Variant of unknown significance

WT . . . . . Wild type

Chapter 1

Introduction

Contents

1.1 Clonal haematopoiesis and its role in ageing and disease 2
1.1.1 Somatic clonal expansion in ageing 2
1.1.2 Brief history of clonal haematopoiesis 4
1.1.3 Genetic drivers of age-related clonal haematopoiesis . . 6
1.1.4 Clonal haematopoiesis and risk of haematological cancer 9
1.1.5 Clonal haematopoiesis and non-malignant disease 11
1.1.6 Clonal dynamics across the human lifespan 12
1.2 A brief introduction to DNA methylation and the DNMT and TET enzymes 14
1.3 The function of DNMT3A in haematopoiesis and myeloid malignancy 16
1.3.1 Structure and function of DNMT3A 16
1.3.2 *DNMT3A* mutations in CH and myeloid malignancies . 18
1.3.3 The role of DNMT3A in haematopoiesis 19
1.4 The function of TET2 in haematopoiesis and myeloid malignancy 22
1.4.1 Structure and function of TET2 22
1.4.2 *TET2* mutations in CH and myeloid malignancies . . . 23
1.4.3 The role of TET2 in haematopoiesis 24
1.5 Aims 25

1.1 Clonal haematopoiesis and its role in ageing and disease

1.1.1 Somatic clonal expansion in ageing

As organisms age, their tissues accumulate cellular damage, including genetic mutations, telomere attrition, epigenetic changes, mitochondrial dysfunction and stem cell exhaustion [1–10]. Somatic mutations that occur in long-lived tissue stem cells may provide a substrate for clonal selection [1, 11, 12]. The majority of somatically-acquired mutations have a neutral effect on cellular survival and fitness, but rarely a mutation may occur that confers a selective growth advantage.

Through Darwinian selection, descendants of that cell will gain a fitness advantage and over time may grow into an expanded clone (Figure 1.1). Such clones have the potential to alter tissue function and may in some cases be the first step towards the development of cancer.

Cancer has long been thought of in evolutionary terms, as a clonal disorder that evolves from a single ancestor cell through multiple stages of mutation acquisition and clonal selection. Since the identification of the first cancerous somatic point mutation in the *HRAS* gene in 1982, there has been major progress in understanding the genetic basis of cancer [13, 14]. Advances in DNA sequencing now makes it possible to catalogue genetic drivers from thousands of cancer whole genomes [15–18]. These studies have confirmed the genetic diversity within individual cancers and identified clonal structures that provide insight into their origin and evolutionary history [1, 19]. However, little is known about the early stages of cancer development due to the difficulty of identifying individuals before the onset of disease and obtaining pre-cancerous tissue for scientific study. Understanding the cellular and molecular processes that cause pre-cancerous clones to expand, evade normal homeostatic controls, and eventually progress to malignancy will be key to identifying individuals at risk of disease, predicting which clones will progress, and perhaps even preventing the development of cancer.

The presence of clonal expansions in healthy human tissues was first described in blood [20–26] and has since been observed in skin [27–30], oesophagus [31–33], large intestine [33–36], bronchus [33, 37], liver [33, 38, 39], endometrium [40] and bladder [41, 42]. Frequently, these clones harbour cancer-associated mutations, suggesting a strong link to cancer development and have in several cases been to shown to be the cancerous cell of origin [25, 43–45].

Amongst these tissues, blood is unique. First, the migratory nature of blood stem and progenitor cells means that clones are not spatially confined [46] and, given enough time, a clone with a selective advantage could potentially grow to an extraordinary size, or even displace the entire system. Second, haematopoietic stem and progenitor cells (HSPCs) give rise to a great diversity of mature cell types, and so

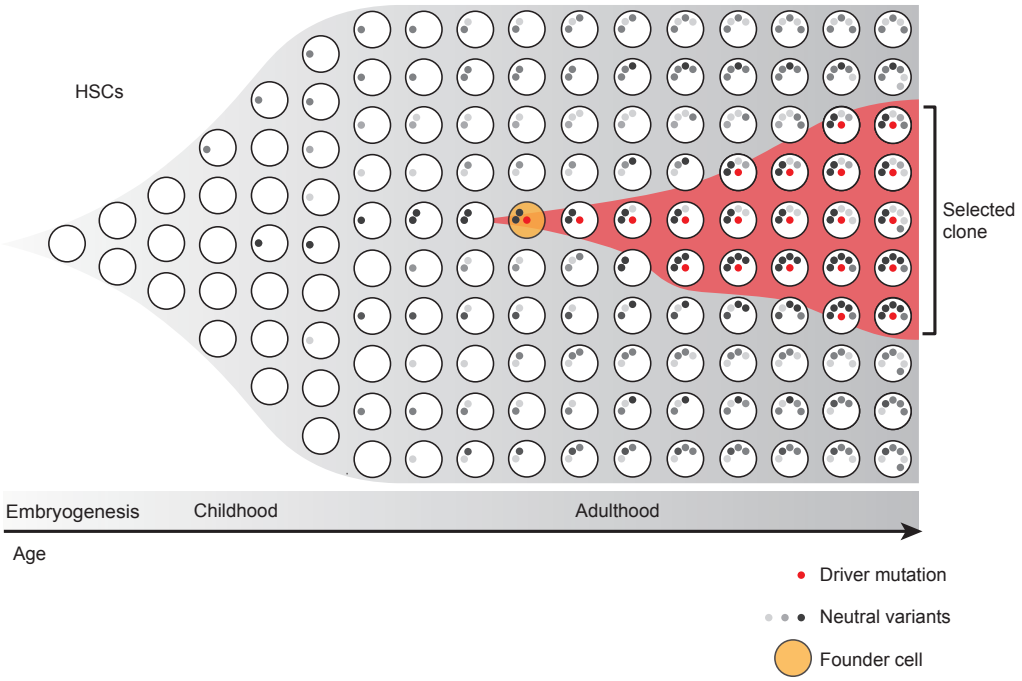


Figure 1.1: Neutral and selected somatic mutations in ageing. Somatic mutations are acquired in all tissues throughout life. When they occur in long-lived tissue stem cells, such as haematopoietic stem cells (HSC), they may be retained within the stem cell population and become substrates for clonal selection. The gradual accumulation of mutations over time leads to increased genetic diversity in older age. Most mutations (or genetic variants) are neutral and have no effect on stem cell fitness (represented by grey dots). The frequency of any such mutation in the population may expand and contract over time by neutral drift as the descendants of the cell in which the mutation was acquired divide, die or differentiate. Rarely, a mutation will confer a selective advantage (red dots) to the cell in which it occurs. Over time, the descendants of that cell, collectively referred to as a "clone", will out-compete their neighbours, leading to clonal expansion over time. In the blood system, this is known as clonal haematopoiesis.

clones might alter the function of not only HSPCs but also circulating erythrocytes, platelets and immune cells descended from the mutant clone. Finally, because many of the immune cells produced by HSPCs will in turn migrate to tissues throughout the body, haematopoietic clones have the potential to influence the function of almost any tissue in the body. This makes clonal haematopoiesis of prominent interest, not only within haematological malignancies but in medical science in general.

1.1.2 Brief history of clonal haematopoiesis

The first evidence of clonal haematopoiesis came from studies of X-chromosome inactivation patterns in women without evidence of haematological malignancies [20–22]. During female embryogenesis, one X-chromosome is randomly inactivated in all cells, and the progeny of each cell maintain the same X-inactivation pattern. Thus in adult tissues, each X-chromosome should be inactivated in approximately half of the cells. A proportion of women were found to have skewed X-inactivation ratios, suggestive of clonality. The incidence of skewing increased with age, and was particularly high in the myeloid compartment [21, 22]. However, this in itself is not a proof of positive selection because clones may expand by a process of neutral drift especially following a tissue insult that creates a genetic bottleneck, such as after chemotherapy treatment, inflammation or infection. In 2012, mutations in the *TET2* gene were detected in approximately 5% of women with skewed X-chromosome inactivation [23]. Since *TET2* is commonly mutated in myeloid malignancies, this observation suggested a link between clonal haematopoiesis and malignancy, and that other somatic mutations might be identified that lead to clonal expansion in healthy individuals.

Another source of evidence came from studies of "pre-leukaemic" clones in patients who were in remission from acute myeloid leukaemia (AML). The *AML1-ETO* fusion transcript, which is present in most AML patients with a t(8;21) translocation, had been identified in a fraction of phenotypically normal HSCs in remission samples [47]. Similar observations were made in later studies, where stem cell clones harbouring early pre-leukaemic mutations but not late mutations were identified, and it was shown that these could give rise to normal multilineage engraftment when transplanted into immunodeficient mice [43–45]. These studies also identified distinct patterns of mutation acquisition. Pre-leukaemic mutations often occurred in genes encoding epigenetic regulators or splicing factors, whereas late transforming mutations often affected the *NPM1* gene, or genes that encode signal transduction molecules such as *FLT3*, *NRAS* and *KRAS*. In myelodysplastic syndromes, early mutations are acquired in the stem cell compartment, but in cases

that transform to AML, later mutations are sometimes only identified in progenitor cells, consistent with the observation that AML leukaemia stem cells frequently have a progenitor phenotype [48–50]. These findings supported a multi-step model of leukaemic evolution, where early mutations occurring in HSCs give rise to a clone that has selective advantage, while late cooperating mutations cause differentiation arrest and disease propagated by progenitor-like cells.

In 2014, three landmark studies analysing blood DNA exome sequencing datasets from thousands of individuals without haematological malignancies were published. One study analysed sequencing data from 2,728 patients with non-haematological cancer enrolled in The Cancer Genome Atlas study [26]. The second study analysed whole-exome sequencing data from 17,182 people enrolled either in the Jackson Heart Study or one of 22 type 2 diabetes association studies [24]. The final study examined a Swedish cohort of 12,380 people with schizophrenia, bipolar disease, or no psychiatric diagnosis [25]. Together, these studies made three key observations. First, clonal haematopoiesis was rare in individuals under the age of 40, but the prevalence increased with age and was identified in 10–20% of individuals over the age of 70. Second, they found that clonal haematopoiesis was associated with mutations in a surprisingly restricted set of genes. Nearly two-thirds of cases were associated with mutations in just two genes, *DNMT3A* and *TET2*, which are enzymes involved in DNA methylation. Finally, they showed that clonal haematopoiesis increases the risk of subsequent haematological cancer as well as cardiovascular disease and overall mortality.

In these original studies, the mutated clones were often large; for example in the study by Jaiswal *et al.* [24], mutated clones contributed on average to 18% of blood cells. This was largely due to the limited sensitivity afforded by low-depth exome sequencing. Subsequent studies using more sensitive sequencing methods, including ultra-deep error-corrected sequencing, have identified a much higher prevalence of CH, where most clones identified are small ($< 2\%$ variant allele frequency) [51–53]. It now appears likely that mutations would be detected

in almost all older adults if one sequences deeply enough, though the clinical relevance of these small clones is unclear.

1.1.3 Genetic drivers of age-related clonal haematopoiesis

The majority of somatic mutations identified in clonal haematopoiesis are cytosine-to-thymine (C→T) transitions occurring in CpG contexts, a characteristic age-related signature resulting from spontaneous deamination of 5-methylcytosine [54, 55]. Since the original three publications in 2014, the genetic drivers of CH have been validated in multiple subsequent studies [24–26, 51, 56–59]. The vast majority of mutations occur in genes involved in epigenetic regulation or RNA splicing, with only a small minority occurring in classical oncogenes (e.g. *JAK2*, *CBL*, *GNAS*, *GNB1*) or tumour suppressor genes (e.g. *TP53*). After *DNMT3A* and *TET2*, the third most commonly mutated gene is *ASXL1*, an epigenetic regulator involved in multiple histone modifications. The splicing factors *SF3B1*, *SRSF2*, *PRPF8* and *U2AF1* are also frequently mutated. Amongst the commonly mutated genes, *DNMT3A* mutations tend to be more common in younger individuals whereas splicing factor mutations are enriched in older individuals [51, 60]. Mutations in genes involved in DNA damage response, including *TP53*, *PPM1D* and *CHEK2*, are particularly common in patients who have received cytotoxic chemotherapy or radiotherapy for cancer [61, 62]. It is notable that mutations are absent or very rare in many other genes that are recurrently mutated in haematopoietic malignancies, such as transcription factors (e.g. *RUNX1*, *GATA2*, *NPM1*, *CEBPA*) and signalling pathways (e.g. *FLT3*, *RAS* family members).

In addition to point mutations and small indels, CH can be detected as large-scale mosaic chromosomal alterations (mCAs), such as amplifications, deletions, and copy-neutral loss of heterozygosity [63–68]. These studies identified common deletions in chromosomes 13q, 2p and 4q. Notably, the latter two deletions contain *DNMT3A* and *TET2* respectively, suggesting haploinsufficiency of these genes might be the cause of selective advantage in these cases. Indeed, deletions of chromosome 4q are significantly enriched in patients that have a co-existing *TET2* point mutation,

suggesting biallelic loss of function confers an additional selective advantage to monoallelic loss [68]. However, many of the other recurrent mCAs do not overlap known driver genes and it is not clear why or whether they are positively selected.

While the majority of studies have relied on prior knowledge of genes linked to haematological malignancies to identify clonal haematopoiesis resulting from point mutations and small indels, a few studies have used an unbiased approach to identify evidence of clonality based on the presence of genetic variants present at low variant allele frequency in peripheral blood [25, 57, 69]. Holstege *et al.* [69] first used this approach to report an extreme case of oligoclonal haematopoiesis without candidate driver mutations in a healthy 115-year-old woman whose peripheral blood cells were found to be the progeny of only two HSC clones. Genovese *et al.* [25] used a similar analysis of exome sequencing data to identify putative somatic mutations based on unusual allelic fractions, and then categorised these cases of clonal haematopoiesis into those with candidate driver mutations and those with unknown drivers based on a list of recurrent mutations associated with haematological malignancies. A similar approach was used in an analysis of whole genome sequencing (WGS) data from an Icelandic cohort, showing that this method captures many more cases of clonal haematopoiesis than analyses limited to known driver genes; candidate driver mutations associated with haematological malignancies were only identified in approximately 40% of cases [57]. Clonal haematopoiesis in the absence of known driver mutations could be explained by rare mutations in as-yet unidentified driver genes, variants in non-coding regions, clonally-inherited epigenetic changes, or stochastic age-related clonal expansion by neutral drift. In fact, computational modelling by Zink *et al.* [57] suggested that some cases of clonal haematopoiesis could arise merely as a result of neutral drift acting on a small population of active HSCs over time. Distinguishing clonal expansion that arises by neutral drift versus positive selection remains an active area of research [70, 71].

Recently, germline genetic variants have been identified that predispose to clonal haematopoiesis. The most common of these are single nucleotide variants (SNVs) at the *TERT* locus, which encodes telomerase reverse transcriptase [57, 65, 72].

Other rarer variants include a loss-of-function mutation in *CHEK2* and a missense variant in *ATM*, both of which are key cell cycle checkpoint regulators [60, 72]. Notably, variants at the *TCL1A* locus have opposing effects across CH drivers. Three non-coding SNPs were associated with increased risk of *DNMT3A*-mutant CH but a reduced risk of CH with *TET2* and other mutations [60, 72]. TCL1A is a transcriptional regulator that physically interacts with DNMT3A, inhibiting its function [73]. It is often a target of chromosomal translocations in T-cell leukaemias where it is over-expressed. Another noteworthy variant was a non-coding SNP near *TET2* identified specifically in individuals of African ancestry. Interestingly, this SNP was found to lie within a *TET2* enhancer where it disrupts a GATA binding motif, leading to reduced *TET2* expression [72].

1.1.4 Clonal haematopoiesis and risk of haematological cancer

One of the key findings of the studies by Jaiswal *et al.* [24] and Genovese *et al.* [25] was that clonal haematopoiesis confers a ∼10 fold increased risk of subsequent haematological cancer. In those studies, myeloid malignancy was most common, but some individuals with CH developed lymphoid malignancies, which is perhaps unsurprising given that *DNMT3A* and *TET2* mutations frequently occur in some subtypes of T-cell and B-cell lymphoma [74, 75]. The absolute incidence of haematological cancer within CH carriers remains low; in the study by Jaiswal *et al.* [24], it translated to a risk of approximately 0.5–1.0% per year. Nonetheless, the idea that individuals at risk of blood cancer might be identified by the presence of mutations in peripheral blood has generated enormous interest. Reflecting this, the term "clonal haematopoiesis of indeterminate potential" (CHIP) was introduced to describe the presence of a cancer-associated mutation with a variant allele frequency greater than 0.02 in individuals with no history of haematological malignancy [76]. Provisional guidelines for the management of patients identified with CHIP have been proposed [77], and there is a lot of effort to improve risk stratification [59].

Multiple factors have been shown to influence the risk of malignancy. The risk is greater in those with multiple mutations, higher variant allele frequency (i.e.

larger clones) and mutations in specific genes, including *TP53*, *U2AF1*, *SF3B1* and *IDH1* [58, 59]. In one study, all participants with *TP53* and *IDH1/2* mutations subsequently developed AML [58]. mCAs are also associated with elevated risk of blood cancer, and co-occurrence of gene mutations with mCAs further increases this risk [65–68].

An important distinction is between individuals with CH and normal blood counts versus those with low blood counts (cytopenia). Individuals with unexplained cytopenia and a cancer-associated mutation (also known as clonal cytopenia of undetermined significance, CCUS) have much greater risk of malignant transformation than those without identifiable mutations [78]. In these individuals, the 5-year cumulative probability of progression to myeloid malignancy was 82%, much higher than for individuals with CHIP and no cytopenia. However, minor reductions in monocyte, platelet, red blood cell and white blood cell counts that would not meet the threshold for clinically-relevant cytopenias also correlate with risk of leukaemia, as does an elevated red blood cell distribution width (RDW) [59].

Myelodysplastic syndromes (MDS) and acute myeloid leukaemia (AML) are clonal diseases involving stem, progenitor and precursor cells of the haematopoietic system. MDS is characterised by morphological dysplasia, ineffective haematopoiesis leading to cytopenias, and risk of transformation to AML. The clinical course is variable, and while some patients live for many years without treatment, approximately a third progress to AML [79]. AML is an aggressive leukaemia, characterised by infiltration of the bone marrow, blood and other tissues by highly proliferative cells with abnormal differentiation. It is amongst the most rapidly progressive of all human cancers and the prognosis remains poor, particularly in older patients where only 5–15% of patients over the age of 65 survive longer than 5 years and most die within a year [80, 81]. While so-called *de novo* AML arises without an antecedant exposure or prodromal haematological disorder, secondary AML is preceded by MDS or a myeloproliferative neoplasm (MPN). Secondary AML tends to affect older patients and is characterised by a distinct mutational profile [82]. Many of the mutations seen in CH are more commonly seen in MDS and secondary AML than

de novo AML, including *TET2*, *ASXL1* and splicing factor mutations. Note that *DNMT3A* is not one of these. While most individuals with CH have a single somatic mutation, MDS is more genetically complex, with most patients having at least two mutations in driver genes, whereas in AML the average number of recurrent mutations is five [1, 79, 83, 84]. In disease remission, CH mutations are often retained with a substantial clone size. It has been hypothesised that these pre-leukaemic clones might be a reservoir for relapse [43, 44], but often *DNMT3A*, *TET2* and *ASXL1* mutations persist in remission with no evidence of recurrent disease [85–89]. CHIP, MDS and secondary AML can therefore be seen as a continuum, characterised by increasing mutational burden and disrupted haematopoietic differentiation.

1.1.5 Clonal haematopoiesis and non-malignant disease

Clonal haematopoiesis has also been associated with non-malignant diseases. One of the most surprising findings by Jaiswal *et al.* [24] was that CHIP conferred an increased risk of overall mortality, and this was explained not by the risk of blood cancer, but by elevated rates of cardiovascular disease (CVD), including coronary heart disease and ischaemic stroke. In a subsequent study, they validated this finding and showed that the risk of coronary heart disease is 1.9 times higher in CHIP carriers than in non-carriers, making CHIP a greater risk factor than smoking, high blood pressure or elevated low-density lipoprotein cholesterol levels [90, 91]. This raised the obvious question of whether CVD was a consequence of mutant clones or merely an association due to shared aetiological factors. By transplanting homozygous or heterozygous *Tet2*-knockout bone marrow cells into atherosclerosis-prone $Ldlr^{-/-}$ mice, they and another group showed that Tet2 loss of function in myeloid cells accelerated atherosclerosis [90, 92]. Macrophages lacking Tet2 overexpressed inflammatory chemokines and cytokines, suggesting a direct causal link that might be relevant to other immune-mediated diseases of ageing. Interestingly, a recent study has also suggested the reverse may be true — presence of atherosclerosis in mice and human patients was associated with increased proliferation in HSPCs, and *Tet2*-knockout bone marrow cells showed greater clonal

expansion in mice with atherosclerosis [93]. However, it was not clear whether this was due to expansion at the level of stem cells or increased myeloid differentiation.

Numerous other conditions have now been associated with clonal haematopoiesis. Rates of CH are higher in smokers and individuals with type II diabetes, osteoarthritis, HIV, aplastic anaemia and other autoimmune diseases [24, 94–97]. Furthermore, CH has been associated with increased predisposition to diverse infections and severe COVID-19 [98, 99]. Recently, it has even been linked to improved outcomes in Alzheimer's disease, where *DNMT3A* and *TET2* mutations were identified in brain microglia isolated from CH carriers [100]. Causal links for most of these associations are yet to be established but it appears highly probable that important interactions between somatic age-related mutations and immune function will be identified in the near future.

1.1.6 Clonal dynamics across the human lifespan

Taken together, the population studies indicate that acquiring driver mutations in CH-associated genes is almost inevitable during the average human lifespan. Only some of these clones expand to a size that is detectable by conventional next-generation sequencing with a variant allele frequency of ∼1–2%, corresponding to 2–4% of blood cells. In turn, a small minority of these acquire a combination of cooperating mutations that lead to malignancy, either MDS, AML or other blood cancers. Given that the clone size and clonal complexity is an important predictor of leukaemia risk, accurate prediction of the dynamics of clonal expansion is likely to be important in risk stratification.

This raises a number of questions. First, how long does it take for clones to expand to a detectable size from the time they are acquired? Do all clones expand, and for those that do, are the expansion rates equal across different mutations? Does the growth rate determine risk of progression to malignancy? What are the phenotypic consequences of mutations in stem and progenitor cells and what are the mechanisms of clonal expansion?

Recent studies have begun to shed some light on the first three questions. By performing whole genome sequencing (WGS) on colonies derived from single HSPCs and using somatic mutations as naturally-occurring clonal markers, Lee-Six *et al.* [101] showed that haematopoiesis in a middle-aged human individual was clonally diverse, estimating that there were a total of 50,000–200,000 HSCs. Using the same method across 10 individuals aged 0–81 years, they found that there is a dramatic decline in clonal diversity in individuals over the age of 75, with 12–18 clones contributing to the majority of blood production, most of which had no known driver mutations [102]. They estimated that most clones arose before the age of 40. Another study performed longitudinal sampling of 697 CH clones over an average of 13 years and found that most clones, particularly those with mutations in *DNMT3A* and *TET2*, expand at a fixed rate during old age [103]. The growth rate differed between clones with mutations in different genes; *DNMT3A*-mutant clones expanded slowly at ∼5% per year, *TET2*-mutant clones expanded at ∼10% per year, whereas *SRSF2*, *PTPN11* and *U2AF1*-mutant clones expanded the fastest at 15–20% per year. However, when they compared this with lifelong growth rates inferred from single-cell phylogenetic trees, they found that the growth rate of *DNMT3A*-mutant clones was too slow to account for the observed clone size, and they must therefore grow at a faster rate early in life and decelerate in older age. Importantly, a higher growth rate was associated with increased risk of AML. These findings are largely consistent with modelling of variant allele frequency distributions from human population-level targeted sequencing data, where the average growth rates of clones with the most commonly observed *DNMT3A* and *TET2* variants are estimated to be 4–15% per year [104].

If the above estimates are correct, CH clones represent a dramatic clonal expansion. Even if we take the lower estimate of HSC number, a clone with variant allele frequency of 0.02 (1 in 25 blood cells) represents a 2000-fold expansion from a single stem cell (1 in 50,000 HSC). For some of the genetic drivers of clonal haematopoiesis (e.g. *TP53* and signalling pathway genes), the mechanisms of clonal expansion are fairly clear, but they are less clear for mutations in the two

most common genes, *DNMT3A* and *TET2*. As regulators of DNA methylation, they are thought to have genome-wide effects on gene regulation. It is especially intriguing that they result in a similar phenotype, even though the two enzymes have biochemically opposing effects — whereas DNMT3A methylates DNA, TET2 catalyses a modification that is the first step towards DNA demethylation. For these reasons, I decided to focus on these two genes in this study, hypothesising that there might be interesting contrasts between the two.

1.2 A brief introduction to DNA methylation and the DNMT and TET enzymes

DNA methylation is an epigenetic modification that is essential for embryonic development, genomic imprinting and repression of repetitive elements, endogenous retroviruses and other retrotransposons [105–108]. The DNA methyltransferases catalyse the conversion of cytosine to 5-methylcytosine (5mC) using S-adenosyl methionine (SAM) as a methyl donor, usually in the context of CpG dinucleotides. Most CpG dinucleotides (60–80%) in the human genome are methylated in a tissue and cell-type specific pattern with the exception of CpG dense clusters, called CpG islands, which are usually unmethylated. CpG islands tend to occur in gene regulatory regions including ∼70% of gene promoters in humans.

There are three known mammalian DNA methyltransferases, DNMT1, DNMT3A and DNMT3B [109, 110]. DNMT1 is a maintenance methyltransferase present at DNA replication forks where it copies the methylation pattern from the parental DNA strand onto the newly synthesised daughter strand during cell division [106, 111]. In contrast, DNMT3A and DNMT3B are primarily *de novo* methyltransferases, able to methylate unmethylated DNA [106, 111]. In addition, germ cells and embryonic tissues express the regulatory protein DNMT3L which modulates the activity of DNMT3A but itself lacks enzymatic activity.

The TET family of dioxygenases catalyse the conversion of 5-methylcytosine to 5-hydroxymethylcytosine (5hmC) and other oxidised derivatives, 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [113–115]. This reaction requires oxygen,

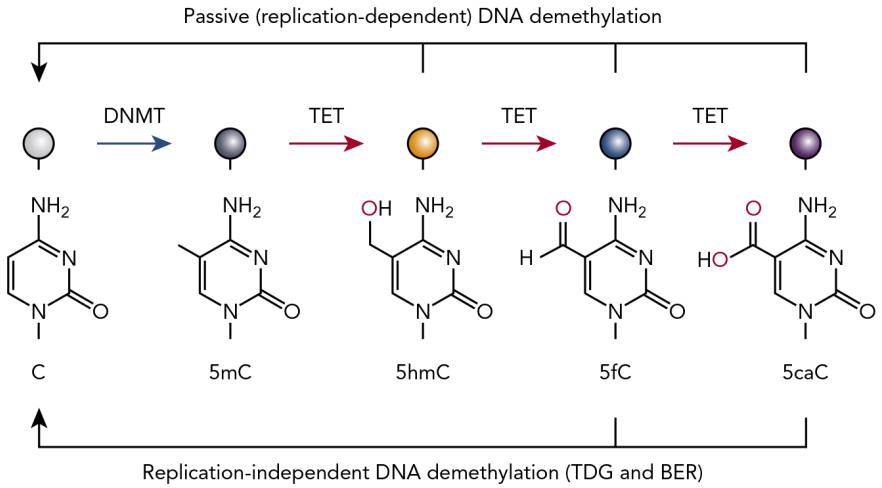


Figure 1.2: DNA methylation and demethylation. Cytosines in DNA CpG dinucleotides are methylated by DNMTs to produce 5mC. TET enzymes oxidise 5mC to 5hmC and other oxidised derivatives, 5fC and 5caC. 5hmC is the most abundant of the oxidised methylcytosine modifications (1% to 10% of 5mC in most somatic cell types). During DNA replication, DNMT1 normally copies the methylation pattern from the parent DNA strand to the daughter strand. 5hmC, 5fC and 5caC present on the parent strand are not recognised by DNMT1; this prevents the restoration of symmetrical methylation on the daughter strand leading to passive (replication-dependent) DNA demethylation (top arrow). TET enzymes can also facilitate DNA demethylation independently of DNA replication (bottom arrow), because 5fC and 5caC can be excised by thymine DNA glycosylase (TDG) and replaced with unmodified cytosine through base-excision repair (BER). Reprinted from Lio, Yuita, and Rao [112], with permission from Elsevier.

α-ketoglutarate (α-KG, also called 2-oxoglutarate), which is a product of the tricarboxylic acid (TCA) cycle, and Fe^{2+}. 5hmC, 5fC and 5caC are intermediates in DNA demethylation. Since DNMT1 does not recognise any of these marks, the unmethylated cytosine on the daughter strand is not methylated when DNA is replicated, resulting in passive loss of DNA methylation when a cell divides [116, 117] (Figure 1.2). Additionally, 5fC and 5caC can be excised by thymine DNA glycosylase, and are then replaced by unmethylated cytosine by the base excision repair system; this is called active DNA demethylation [112, 116]. However, it has recently been recognised that 5hmC can be a stable epigenetic mark that may have a regulatory role itself [118, 119]. 5hmC is enriched in highly expressed genes, and also at active enhancers defined by high levels of H3K4me1 and H3K27Ac [112, 119, 120]. It is also worth noting that the DNMT and TET proteins may be epistatic, in that hydroxymethylation by TET2 may depend on the presence of

5mC marks introduced by DNMTs. Indeed, a study in epidermal stem cells has shown that Dnmt3a is recruited to the centre of active enhancers by interaction with H3K36me3, where it promotes DNA hydroxymethylation in a Tet2-dependent manner. This is in contrast to Dnmt3b, which has a broader binding pattern over enhancers and promotes DNA methylation [121].

The role of DNA methylation in somatic differentiation and gene regulation is debated [122, 123]. While DNA methylation at promoters generally correlates with transcriptional repression and some transcription factors are sensitive to DNA methylation, methylation changes often follow transcription factor binding rather than precede them [123–127].

1.3 The function of DNMT3A in haematopoiesis and myeloid malignancy

1.3.1 Structure and function of DNMT3A

DNMT3A is encoded by 23 exons on human chromosome 2p. The protein is highly conserved, with 98% homology between human and murine forms. The protein is expressed in two major forms: the long DNMT3A1 isoform is 912 amino acids in length, whereas the shorter DNMT3A2 isoform lacks the first 6 exons of the N-terminus (Figure 1.3). DNMT3A2, which is most extensively studied, is expressed in embryonic stem cells (ESC), germ cells, thymus and spleen and is largely silenced in adult tissues but expression has been reported in normal cord blood and AML [128–130]. In contrast, the long isoform is widely expressed [131].

DNMT3A contains three main functional domains. The Pro-Trp-Trp-Pro (PWWP) domain binds DNA and histone H3 that is marked by H3K36me3 and H3K36me2, which recruit DNMT3A to intergenic regions [121, 132–136]. The ATRX-DNMT3-DNMT3L (ADD) domain also binds H3 peptides, mediates protein interactions and modulates the methyltransferase activity through an allosteric mechanism. The C-terminal methyltransferase domain contains SAM and DNA binding regions. DNMT3A forms oligomers in the nucleus through binding interfaces located on the methyltransferase domain [137]. Both DNA

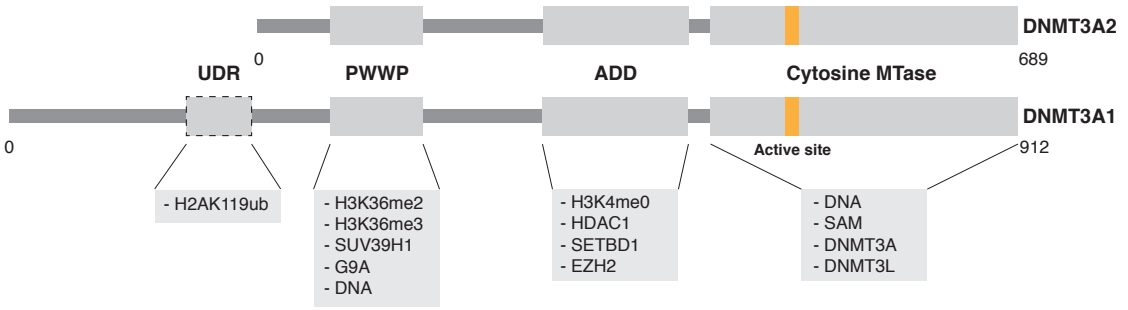


Figure 1.3: Structure of DNMT3A. Schematic of the domain architecture of DNMT3A. The two main isoforms, DNMT3A1 and DNMT3A2, and the number of amino acids in each polypeptide are shown. The main domains are the putative ubiquitin-dependent recruitment region (UDR), the Pro-Trp-Trp-Pro (PWWP) domain, the ATRX-DNMT3-DNMT3L (ADD) domain, and the methyltransferase (Cytosine MTase) domain. The features or interaction partners bound by each domain are shown below.

binding and processivity of the enzyme depend on its ability to form tetramers, and mutations affecting residues on the tetramer interface, such as R882, disrupt processive methylation [138].

Recently the regulation of DNMT3A localisation and activity by chromatin has begun to be elucidated. When not bound to chromatin, the ADD domain forms an auto-inhibitory loop that blocks the methyltransferase binding to DNA [139, 140]. Association with unmethylated H3K4me0, releases the auto-inhibition. This mechanism probably explains the negative correlation between H3K4me3 and DNA methylation, and prevents methylation at CpG islands and promoters of actively expressed genes which are rich in H3K4me3 [139, 141]. Meanwhile, the PWWP domain reads methylation of H3K36, recruiting DNMT3A to H3K36me2-enriched intergenic regions and to H3K36me3-enriched gene bodies [133, 135]. Recently, a putative ubiquitin-dependent recruitment region (UDR) was identified which interacts with H2AK119ub, a mark catalysed by Polycomb repressive complex 1 (PRC1) [136]. This region lies in the N-terminal domain of DNMT3A1 and is not found in the DNMT3A2 isoform. Additionally, DNMT3A interacts with a number of chromatin modifiers involved in gene repression, mostly via the ADD and PWWP domains, including histone deacetylase 1 (HDAC1) and the histone lysine N-methyltransferases EZH2, SETDB1, SUV39H1 and G9A [142–145]. However, the

manner in which DNMT3A is regulated remains poorly understood, particularly how its methylation activity is targeted to specific sites and how interaction with other protein partners modifies its function.

Germline loss-of-function mutations in *DNMT3A* cause an overgrowth syndrome, Tatton-Brown-Rahman syndrome, characterised by distinctive facial appearance, intellectual disability and increased height [146]. Some of these are missense mutations that abolish H3K36me2 binding of the PWWP domain [135]. Conversely, gain-of-function mutations in the PWWP lead to microcephalic dwarfism and aberrant DNA methylation of Polycomb-associated regions encompassing developmental genes [134, 136].

1.3.2 *DNMT3A* mutations in CH and myeloid malignancies

Mutations in *DNMT3A* are found in approximately 50% of cases of clonal haematopoiesis and are reported in approximately 25% of adult AML [83, 84, 147–149], as well as in a significant number of cases of MDS [150, 151], chronic myelomonocytic leukaemia (CMML) and adult T-cell leukaemia/lymphoma [152–155]. It is notable that *DNMT3A* mutations are almost exclusively observed in adult myeloid and T-cell malignancies with very few cases reported in paediatric, B-cell or non-haematological malignancies.

In AML, *DNMT3A* mutations are predominantly heterozygous missense mutations in the methyltransferase domain [83, 154]. Approximately 60% are missense mutations in the R882 residue [147–149, 156], which are thought to exert a dominant-negative effect by preventing tetramerisation of the DNMT3A protein complex [138, 157, 158], although aberrant interaction with PRC1 has also been proposed as a mechanism [159]. Interestingly, R882 mutations also alter the flanking sequence preference of DNMT3A, leading to aberrant methylation patterns [160, 161]. Other mutations are distributed throughout the other functional domains, and include missense, nonsense, frameshift, and splice site mutations. Recently it was shown that many of the missense mutations cause protein instability [162]. *DNMT3A* mutations co-occur with *NPM1* mutations in 60–80% cases [83, 156] and also

associate significantly with mutations in *FLT3* and *IDH2*, whereas they are mutually exclusive with *PML-RARA*, *MYH11-CBFB*, and *MLL*-containing fusion genes [84, 163]. In studies of younger adults with AML, patients with *DNMT3A* mutations were older, had higher white blood cell counts and more frequently had a normal karyotype and myelomonocytic or monocytic morphology (FAB types M4 and M5) [148, 156, 164]. The profile of mutations is somewhat different in the other haematological malignancies; R882 mutations are less predominant in MDS, CMML and T cell malignancies, and a relatively high frequency of biallelic mutations have been reported in T-cell acute lymphoblastic leukaemia [153, 154].

The profile of mutations observed in CH is similar to the malignancies, except that R882 mutations are found in only ∼20% of cases [24–26, 56–59]. This has led some to hypothesise that *DNMT3A* haploinsufficiency confers a selective advantage but the dominant-negative effect of R882 mutations confers an additional leukaemogenic effect [165]. Interestingly, R882-mutant CH clones do not have faster growth rates than other *DNMT3A* mutations [103].

1.3.3 The role of DNMT3A in haematopoiesis

Mouse models have begun to define the function of DNMT3A in haematopoiesis and the phenotype of *DNMT3A*-mutant clones.

Complete knockout of *Dnmt3a* in HSCs biases them slightly towards self-renewal rather than differentiation [166]. This becomes more pronounced with serial transplantation, eventually leading to impaired differentiation and expansion within the HSC compartment [166, 167]. Consequently, *Dnmt3a*$^{-/-}$ HSCs gain a competitive advantage over WT HSCs and have a pre-disposition to malignant transformation in the presence of cooperating mutations [166, 168–171].

Loss of Dnmt3a is associated with upregulation of multipotency and self-renewal genes, including *Runx1*, *Gata3*, *Pbx1* and *Cdkn1a*, as well as enrichment of TGFβ and MAPK signalling [166, 172]. Furthermore, differentiated progeny of *Dnmt3a*$^{-/-}$ HSCs fail to repress stem-cell associated genes. Downregulated

genes were associated with cell cycle, niche interactions and haematopoietic lineage-specification. Some HSC-related genes were associated with DNA hypomethylation, but overall differences in gene expression were better explained by histone marks than DNA methylation changes. Decreased methylation is observed particularly at the at the edges of large undermethylated regions (known as methylation "canyons" or "valleys") [173]. These are evolutionarily-conserved regions that are often regulated by Polycomb and associated with developmental genes, many of which are aberrantly expressed in leukaemia [173, 174].

Complete loss of DNMT3A is rarely seen in patients with myeloid malignancies or CH. More realistic mouse models have since shown similar but more subtle phenotypes. Heterozygous loss of *Dnmt3a* leads to a modest clonal advantage after 12 months in competitive transplantation that was equal across progenitor compartments and associated with a slight myeloid bias and very few gene expression changes [175]. Knock-in of the mouse homolog of $DNMT3A^{\mathrm{R882H}}$ ($Dnmt3a^{\mathrm{R878H}}$) in steady-state haematopoiesis leads to increased frequency of short-term HSCs and granulocyte-monocyte progenitors (GMP), and impaired erythroid output [176]. $Dnmt3a^{\mathrm{R878H}}$ progenitors upregulate HSC-associated genes such as *Hoxa* genes and downregulate cell cycle and lineage-associated genes, though there are substantial differences in the results between studies [159, 176, 177].

The association of CH with age, smoking and inflammatory conditions have led some to hypothesise that inflammation might promote expansion of mutant clones. In the normal haematopoietic system, chronic inflammation leads to HSC depletion due to a loss of quiescence and eventual exhaustion [178–180]. Indeed, exposure to chronic infection or IFN-γ-signalling increases the competitive advantage of $Dnmt3a^{-/-}$ HSCs and is associated with relative suppression of pro-differentiation factors, including *Fos* and *Jun*, which are normally upregulated in response to inflammation [181]. This might account for the high prevalence of CH in individuals with inflammatory disorders, but the idea that *DNMT3A*-mutant cells have a selective advantage specifically in the aged bone marrow microenvironment goes against the observations that individuals with *DNMT3A*-mutant CH are younger

than the average and expansion of *DNMT3A*-mutant clones actually slows down later in life [103].

Several studies have shown impaired or biased differentiation in the setting of *Dnmt3a* loss-of-function [159, 166, 176]. Loss of Dnmt3a methylation activity at lineage-specific enhancers has been suggested as a mechanism for this [170, 171, 182]. Using single-cell multi-omics, Izzo *et al.* [183] found HSPCs in *Dnmt3a*$^{-/-}$ mice were skewed towards the megakaryocyte-erythroid lineage and away from the myelomonocytic lineage. As a potential mechanism, they found that erythroid-related transcription factors (TFs) had higher CpG content in their binding motifs than monocytic TFs, making them more sensitive to the global hypomethylation observed in *Dnmt3a*$^{-/-}$ HSPCs. CpG-rich motifs of the erythroid TFs Tal1 and Klf1 were more accessible in *Dnmt3a*$^{-/-}$ HSCs, and combining single-cell RNA-seq with methylation profiling showed a correlation between enhancer methylation and erythroid transcriptional priming. Interestingly, Gata1 motif accessibility was similarly affected, even though it is CpG poor.

All of the above mouse models are limited by their use of transplantation assays or *Dnmt3a* ablation (or mutation) in all haematopoietic cells, neither of which faithfully represent the human situation where a mutation is acquired in a single mutant cell and a clone slowly expands over many years. Recently, findings from human individuals with somatic *DNMT3A* R882 mutations were reported. This study analysed mobilised HSPCs from 4 individuals in remission from multiple myeloma who had a *DNMT3A* R882 CH clone [184]. Combining single-cell RNA-seq with genotyping they found increased megakaryocyte-erythroid transcriptional priming within immature progenitors, similar to the observations in *Dnmt3a*$^{-/-}$ mice. Pro-inflammatory genes and MYC and MAX targets were upregulated. Hypomethylated regions were enriched for Polycomb repressive complex 2 targets and a specific sequence motif that overlaps with the MYC/MAX binding motif. Combining DNA methylation with chromatin accessibility profiling, they identified a correlation between hypomethylation of the motif and accessibility, suggesting enhanced MYC/MAX binding is linked to loss of methylation.

While differences in DNA methylation sensitivity of TFs leading to differentiation bias is a plausible hypothesis, there are likely to be other factors involved. The degree to which transcription factor binding is influenced by chemical modifications of DNA or chromatin accessibility remains an open question. A systematic analysis of methylation sensitivity for 542 transcription factors showed that DNA methylation reduced DNA binding for 23% of TFs and increased binding for 33% [126]. Furthermore, it has been observed that enhancer activity tracks closely with chromatin accessibility, whereas DNA methylation changes occur later, suggesting methylation loss is secondary to transcription factor binding [127].

In summary, *DNMT3A* loss-of-function appears to alter HSC function such that cell divisions are more likely to lead to self-renewal than differentiation. Changes in DNA methylation and gene expression are subtle even in the setting of serial transplantation, but may partially be explained by altered transcription factor binding at methylation-sensitive motifs. However, the consequences of somatic *DNMT3A* mutations in human haematopoietic stem and progenitor cells remain poorly understood.

1.4 The function of TET2 in haematopoiesis and myeloid malignancy

1.4.1 Structure and function of TET2

TET2 is encoded by 11 exons of chromosome 4 and it is expressed in at least 3 isoforms [185]. The predominant isoform is a protein of 2002 amino acids, with a C-terminal catalytic domain that comprises a cysteine-rich domain and a double-stranded β-helix domain (DSBH) (Figure 1.4). The DSBH domain contains the metal-binding residues for Fe^{2+} and α-KG while the Cys-rich domain stabilises DNA binding [186].

TET2 specifically recognises 5mC in CpG dinucleotides and has no preference for the flanking sequences [186, 187]. Instead, it is recruited to DNA by transcription factors, including PU.1, C/EBPα, TFCP2L1, KLF4, WT1, PML and SNIP1 [188–192]. SNIP1 in turn interacts with c-MYC. In addition, TET2 activity is regulated

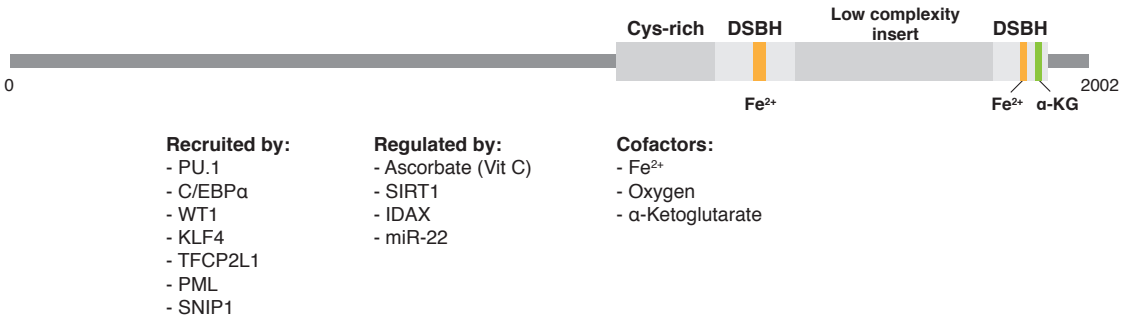


Figure 1.4: Structure of TET2. Schematic of the domain architecture of TET2. The C-terminal catalytic domain consists of a cysteine-rich domain (Cys-rich) and a double-stranded β-helix domain (DSBH). The Cys-rich domain modulates the chromatin targeting. The DSBH domain harbours catalytic motifs, which interact with Fe^{2+} and α-ketoglutarate (α-KG, also called 2-oxoglutarate). A large low-complexity insert is found within the DSBH domain, but its function is yet to be defined. The known interaction partners are listed below.

by ascorbate (Vitamin C), probably through its action of maintaining Fe^{2+} in its reduced state [193]. The regulatory role of ascorbate on TET enzymes and Jumonji histone demethylases has generated interest in the use of ascorbate in myeloid malignancies [194, 195]. TET2 catalytic activity is also enhanced by SIRT1, which de-acetylates TET2 at conserved lysine residues in its catalytic domain [196]. Expression of TET2 is negatively regulated by miR-22 and at the protein level by IDAX [197], a CXXC domain protein that binds TET2 and promotes its degradation by caspases [198].

In addition to its activity on DNA, TET2 oxidation of mRNA and tRNA has recently been reported [199–201]. Hydroxymethylation of RNA by TET enzymes reduces the stability of transcripts in mouse ES cells [202]. In bone marrow cells, Tet2 was specifically shown to repress *Socs3* expression, where loss of 5mC allows Adar1 to bind and destabilise the transcript [200]. This is therefore another regulatory mechanism by which *TET2* mutations might have their effect.

1.4.2 *TET2* mutations in CH and myeloid malignancies

Mutations in *TET2* are found in approximately 20% of cases of clonal haematopoiesis, 13% of myeloproliferative neoplasms (MPNs), 25% of MDS, 50% of CMML, and 23% of AML [79, 83, 84, 150, 152, 203, 204]. Additionally, they are commonly found

in peripheral T-cell lymphomas, where they frequently co-occur with mutations in *RHOA*, and in approximately 6% of diffuse large B cell lymphoma [74, 205–208].

In all cases, *TET2* mutations are missense or truncating variants distributed across the coding region. Most of the leukaemia-associated missense mutations are inactivating mutations that inhibit or abolish the catalytic activity, either by impairing interaction with Fe^{2+} and α-KG at the active site or by affecting the structural integrity of the catalytic core domain [209]. Biallelic mutations occur in approximately 50% of TET2-mutated AML cases and are particularly common in CMML [152, 210, 211]. In AML, *TET2* mutations are mutually exclusive with mutations in the isocitrate dehydrogenase genes *IDH1/IHD2*, and the *PML-RARA* and *MYH11-CBFB* fusion genes [84, 163]. AML patients with *TET2* mutations tend to be older, have higher white blood cell counts and lower platelet counts than those without *TET2* mutations [210].

1.4.3 The role of TET2 in haematopoiesis

As for DNMT3A, mouse models have helped to define the function of TET2 in haematopoiesis. Homozygous loss of Tet2 leads to a phenotype with some similarities to Dnmt3a knockout. *Tet2*$^{-/-}$ HSC have a competitive advantage in transplantation experiments and increased serial colony forming capacity *in vitro*, suggestive of increased self-renewal [212–215]. However, *Tet2*$^{-/-}$ mice also show expansion of downstream common myeloid and granulocyte-macrophage progenitors, and with age some mice develop a myeloproliferative disorder, reminiscent of CMML [213–215]. Heterozygous loss of Tet2 leads to a similar phenotype and also confers a competitive advantage but with longer latency [213, 214].

Interestingly, a recent study showed that *Tet2*$^{-/-}$ HSCs exhaust at the same rate as WT HSCs in serial transplantation experiments, despite an initial competitive advantage [216]. This is in contrast to *Dnmt3a*$^{-/-}$ HSCs, which can be transplanted for many generations without exhaustion [167]. Notably, *Tet2*$^{-/-}$ HSCs show more pronounced transcriptional differences than *Dnmt3a*$^{-/-}$ HSCs. In support of an aberrant progenitor phenotype, knockdown of TET2 in human CD34+ cells leads to

monocytic differentiation bias, and hyperproliferation and impaired differentiation of erythroid progenitors that depend on stem cell factor for their growth [217, 218]. Myeloid lineage bias was also observed in the single-cell study by Izzo *et al.* [183], where *Tet2*$^{-/-}$ mice showed expansion of HSC and myelomonocytic progenitors and reduced frequency of megakaryocyte, erythroid and lymphoid progenitors. *Tet2*$^{-/-}$ HSCs downregulated DNA replication genes and upregulated myelomonocytic lineage genes. As discussed earlier, the authors hypothesised that the lineage bias might be due to differential sensitivity of myelomonocytic and erythroid transcription factors to DNA methylation owing to differences in CpG content of their binding motifs. As expected, *Tet2*$^{-/-}$ HSCs had decreased accessibility of erythroid TF binding motifs, which was associated with a modest increase in enhancer methylation, whereas myelomonocytic motifs were less affected. Therefore they suggest the monocytic differentiation bias results from reduced binding of erythroid TFs at hypermethylated motifs.

However, this is unlikely to account for the entire phenotype. Notably, knockout of Tet2 has phenotypic and transcriptional consequences that are distinct from those observed when a catalytically-inactive protein is expressed [219]. This included downregulation of Gata2, which ameliorated the disease phenotype when overexpressed in *Tet2*$^{-/-}$ mice, suggesting TET2 may have regulatory role independent of its enzymatic activity.

Several studies have now shown that inflammation may play a role in the expansion of *TET2*-mutant clones. Exposure to acute inflammation leads to increased expansion of *Tet2*$^{-/-}$ HSCs and myeloid progenitors [220, 221]. Altered response of mutant cells to inflammatory cytokines, especially IL-6, appears to play a role. *Tet2*$^{-/-}$ HSPCs are resistant to inflammation-induced apoptosis, and *Tet2*$^{-/-}$ granulocyte-monocyte progenitors upregulate surface expression of IL-6Rα, which results in increased IL-6-induced signalling and myeloid proliferation. Interestingly, the production of IL-6, along with other cytokines such as IL-1b, CCL2 and TNF-α, is increased in mature *Tet2*$^{-/-}$ myeloid cells. This may generate

a positive feedback loop, where mature *TET2*-mutant immune cells create an environment that favours expansion of the mutant clone.

1.5 Aims

Studies of human populations have shown that in our lifetime we are all likely to acquire somatic mutations in our haematopoietic stem cells that give rise to expanded clones. Some of these expand to detectable levels, and a small minority progress to cause disease. The mechanisms that cause clonal expansion are poorly understood, particularly for the two most common drivers, mutations in *DNMT3A* and *TET2*, and no studies have yet analysed the consequences of these mutations in humans with unperturbed clonal haematopoiesis. This study therefore aimed to characterise the cellular and molecular consequences of somatic *DNMT3A* and *TET2* mutations on haematopoietic stem and progenitor cells in humans with clonal haematopoiesis.

The overall aims of the study were:

1. To establish a cohort of blood and bone marrow samples from healthy human individuals and identify cases with age-related clonal haematopoiesis.

2. To develop a more sensitive method for single-cell targeted genotyping and whole transcriptome analysis.

3. To identify the patterns of clonal expansion and the effect of mutations on haematopoietic differentiation in cases of *DNMT3A* and *TET2*-mutant clonal haematopoiesis.

4. To determine how gene expression is dysregulated in haematopoietic stem and progenitor cells with somatic *DNMT3A* and *TET2* mutations.

Chapter 2

Materials and Methods

Contents

2.1 Study participants 27
2.2 Sample collection and processing 28
2.3 Targeted DNA sequencing 29
2.3.1 Library preparation and sequencing 29
2.3.2 Read pre-processing and mapping for targeted DNA sequencing 30
2.3.3 Somatic variant calling 30
2.3.4 Annotation of pathogenic driver variants 32
2.3.5 Comparison of bone marrow and peripheral blood allele frequencies 33
2.3.6 Statistical analysis of targeted sequencing 33
2.4 Cell culture 34
2.5 Flow cytometry and cell sorting 34
2.6 TARGET-seq+ optimisation 35
2.6.1 Comparison of RT-PCR conditions and DNA polymerases 36
2.6.2 Comparison of template-switching oligos 36
2.6.3 Comparison of targeted primer concentrations 36
2.7 TARGET-seq library preparation 37
2.8 Design and validation of targeted genotyping primers for TARGET-seq+ 38
2.9 TARGET-seq+ library preparation 40
2.9.1 Lysis buffer preparation 40
2.9.2 Reverse transcription and pre-amplification 41
2.9.3 Whole-transcriptome library preparation and sequencing 42
2.9.4 Targeted single-cell genotyping 43
2.10 Single-cell genotyping analysis 44
2.10.1 Mutational calling 44
2.10.2 Allelic dropout estimation 45
2.10.3 Inference of clonal hierarchies 46
2.11 Analysis of FACS index data 47
2.12 Single-cell transcriptome data pre-processing 48
2.12.1 Mapping and transcript counting 48
2.12.2 Transcript detection and dropout frequency calculation 48
2.12.3 Cell-to-cell correlation analysis 49
2.13 Single-cell transcriptome analysis 49
2.13.1 Quality control and normalisation of single-cell transcriptome data 49

2.13.2 Dimensionality reduction and clustering for individual samples 50
2.13.3 Single-cell copy number analysis 50
2.13.4 Data integration and visualisation 51
2.13.5 Cell cycle and lineage signature scores 51
2.13.6 Clustering analysis, marker gene identification, and cell type annotation 51
2.13.7 Differential abundance analysis 52
2.13.8 Pseudotime analysis 53
2.13.9 Differential gene expression analysis 54
2.13.10 Gene set enrichment analysis 54
2.13.11 AUCell analysis 55

2.1 Study participants

Patient samples were collected with informed consent under the Mechanisms of Age-Related Clonal Haematopoiesis (MARCH) Study. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki. This study was approved by the Yorkshire & The Humber - Bradford Leeds Research Ethics Committee (NHS REC Ref: 17/YH/0382). Participants were recruited from individuals undergoing elective total hip replacement (THR) surgery at the Nuffield Orthopaedic Centre, Oxford. Exclusion criteria were: history of rheumatoid arthritis or other inflammatory arthritis, history of septic arthritis in the limb undergoing surgery, history of haematological cancer, bisphosphonate use, and oral steroid use. Patient characteristics are summarised in Table A.1.

2.2 Sample collection and processing

At the time of surgery, trabecular bone fragments and bone marrow aspirates were obtained from the femoral canal and collected in 10 ml anticoagulated buffer containing acid-citrate-dextrose, heparin sodium and DNase. Samples of peripheral blood were collected in EDTA vacutainers. Hair follicle samples were collected from participants as a germline control and stored at room temperature.

All samples were processed within 24 hours of collection and Ficoll gradient separation used to isolate mononuclear cells (MNCs). Peripheral blood and bone

marrow aspirate samples were diluted 1:2 in RPMI-1640 and filtered through a 70 μm cell strainer. Trabecular bone samples were manually fragmented with scissors and washed thoroughly in RPMI media with DNase to collect trabecular marrow, which was then filtered through a 70 μm cell strainer to obtain a single cell suspension. Peripheral blood and bone marrow suspensions were then layered onto Ficoll-Paque (Sigma-Aldrich), centrifuged at 400 g for 30 minutes at room temperature with no brake and the interphase of MNCs collected. For some samples, bone marrow CD34+ cells were purified using a CD34 MicroBead kit and MACS separation columns (Miltenyi Biotec, Bergisch Gladbach, Germany), according to the manufacturer's instructions. Unseparated MNCs, CD34+ and CD34-deplete fractions were frozen in 90% fetal bovine serum (FBS, Sigma-Aldrich) with 10% dimethyl sulfoxide (DMSO) and stored in liquid nitrogen.

On the day of an experiment, cryopreserved BM sample vials were thawed at 37 °C in a water bath, 1 ml FBS was added, and the suspension then diluted in 8 ml thaw media (Iscove's Modified Dulbecco's Medium (IMDM) (ThermoFisher) with 20% FBS + 100 μg/ml DNAse). The suspension was centrifuged at 400 g for 10 mins, cells were resuspended in a flow cytometry staining medium (IMDM with 10% FBS + 10 μg/ml DNAse) and filtered through a 35 μm cell strainer prior to use in an experiment.

Peripheral blood granulocytic cell pellets isolated by Ficoll density gradient centrifugation were frozen for DNA sequencing analysis of mutations in peripheral blood. For cytokine analysis, peripheral blood and bone marrow plasma samples were obtained by collecting the supernatants after Ficoll density gradient centrifugation, further centrifuged at 500 g for 5 mins to remove cells and frozen at –80 °C.

2.3 Targeted DNA sequencing

2.3.1 Library preparation and sequencing

Targeted DNA sequencing was performed on bone marrow MNCs and peripheral blood granulocytic DNA samples. Genomic DNA was extracted from bone marrow

MNCs and peripheral blood granulocytes using a DNeasy Blood & Tissue Kit (Qiagen) and eluted in EB buffer according the manufacturer's instructions.

Pre-capture DNA libraries were prepared using the KAPA HyperPlus protocol (KK8514, Roche) either by hand or using a Biomek FXp Automated Workstation (Beckman Coulter). 100 ng of genomic DNA was fragmented by enzymatic fragmentation. Following end repair and A-tailing, adapter ligation was performed using KAPA dual-indexed adapters (KK8722, Roche). Library cleanup and double-sided size selection was performed using Agencourt AMPure XP beads (Beckman-Coulter) to obtain fragments of approximately 320 bp. Libraries were amplified by ligation-mediated PCR for 6 cycles using a KAPA HiFi HotStart high-fidelity DNA polymerase (Roche) and purified using AMPure XP beads.

Targeted capture was performed using a custom pool of biotinylated capture probes (SeqCap EZ Prime Choice, Roche) targeting 97 genes recurrently mutated in clonal haematopoiesis (CH) and myeloid malignancies spanning 347 kb (see Table A.2)[24–26, 53, 56, 83, 84]. Amplified DNA libraries were hybridised to the capture probes in pools of 10-12 samples according to the manufacturer's instructions. The captured DNA libraries were amplified by 14 cycles of PCR using a KAPA HiFi HotStart DNA polymerase and purified using AMPure XP Beads.

Post-capture amplified DNA libraries were quantified by Qubit (Life Technologies) and size distribution and quality analysed using a Bioanalyser chip (Agilent Technologies). Libraries were pooled in equimolar concentrations and were sequenced on an Illumina NextSeq 500 using paired-end reads.

2.3.2 Read pre-processing and mapping for targeted DNA sequencing

Sequencing data were analysed in a custom pipeline based on the GATK Best Practices (GATK v4.1.2.0 and Picard v2.21.3). Raw sequencing reads were converted to an unmapped BAM file and adapter sequences soft-clipped using Picard MarkIlluminaAdapters. Following conversion back to a FASTQ file, reads were mapped to the hg38 human reference genome assembly using the Burrows–Wheeler aligner v0.7.17 [222] with alternate contig-aware alignment. Mapped and unmapped BAM

files were merged using MergeBamAlignment and reads from different sequencing lanes were combined. Duplicate reads were marked using Picard MarkDuplicates and base quality scores recalibrated with GATK BaseRecalibrator and ApplyBQSR. Target enrichment metrics and coverage was calculated using Picard CollectAlignmentSummaryMetrics, CollectHsMetrics and custom scripts.

2.3.3 Somatic variant calling

Somatic variant calling was performed on the pre-processed BAM files using VarDictJava v1.7.0 [223] and Mutect2 [224] in tumour-only mode. For VarDict, variants were called with a minimum variant allele frequency of 0.01, minimum base quality score of 25 and minimum supporting reads of 2, with indel realignment and removal of adapter sequences. For Mutect2, a minimum tumour LOD of 2 was used and variants were filtered for sequence context-dependent artefacts using FilterMutectCalls and FilterByOrientationBias. Indels were left-aligned and normalised using bcftools norm (v1.9). Variants were annotated using Annovar [225].

VarDict and Mutect2 variant calls were analysed separately to identify a consensus list of high-confidence variants. Variants were filtered using custom python scripts. The following post-processing filters were applied to VarDict calls to exclude likely sequencing artefacts:

1. Minimum of 5 variant reads for SNVs (with at least 2 reads in forward and reverse directions), or minimum of 10 variant reads for indels (with at least 4 reads in forward and reverse directions).

2. Minimum base quality score 30.

3. Minimum mapping quality score 40, except for variants in *U2AF1;U2AF1L5*, where the mapping quality was ignored. This is because in hg38, there is a duplication of the *U2AF1* gene on chromosome 21 called *U2AF1L5*, which results in reads being flagged as multi-mapped.

4. Maximum strand bias Fisher P-value of 0.0001.

5. No position bias towards beginning or end of reads.

The following post-processing filters were applied to Mutect2 calls:

1. Passed all default Mutect2 filters or only failing the clustered_events filter.

2. Minimum of 5 variant reads for SNVs (with at least 2 reads in forward and reverse directions), or minimum of 10 variant reads for indels (with at least 4 reads in forward and reverse directions).

Variants were flagged as likely germline polymorphisms or sequencing artefacts if any of the following applied:

1. Variant allele with a population allele frequency >1 in 1,000 according to any of three large polymorphism databases (Gnomad, 1000 Genomes Project, ESP6500) that is not a hotspot driver mutation with COSMIC (v88) recurrence >100 or is present in a list of clonal haematopoiesis-associated mutations compiled from five large studies [24, 25, 53, 58, 59].

2. Variant allele frequency (VAF) between 0.4-0.6 or >0.9 unless recurrent in COSMIC >5 times, or previously reported in clonal hematopoiesis.

3. Present in a panel of normal cord blood samples.

4. Recurrent in the cohort unless present at least 5 times in COSMIC or at least 2 times in the clonal haematopoiesis studies.

After filtering, variants were manually inspected using the Integrative Genomics Viewer (IGV) tool (http://software.broadinstitute.org/software/igv/).

2.3.4 Annotation of pathogenic driver variants

Samples were annotated as having CH based on the presence of driver mutations in bone marrow sequencing at a VAF ≥ 0.01. Variants were annotated as pathogenic driver mutations using the following criteria:

1. Truncating mutations (nonsense, splice site or frameshift indel) in the following genes: *DNMT3A*, *TET2*, *ASXL1*, *NF1*, *IKZF1*, *RAD21*, *WT1*, *KMT2D*, *SH2B3*, *TP53*, *CEBPA*, *RUNX1*, *BCOR*, *KDM6A*, *STAG2*, *PHF6*, *KMT2C*, *KMT2E*, *PPM1D*, *ATRX*, *EZH2*, *CREBBP*, *NOTCH1*, *CUX1* and *ZRSR2*.

2. Non-synonymous variants at the following hotspot residues: *CBL* E366, L380, C384, C396, C404 and R420; *DNMT3A* R882; *FLT3* D835; *IDH1* R132; *IDH2* R140 and R172; *JAK2* V617F; *KIT* W557, V559 and D816; *KRAS* G12, G13, Q61 and A146; *MPL* W515; *NRAS* G12, G13 and Q61; *SF3B1* K666 and K700; *SRSF2* P95; *U2AF1* S34, R156 and Q157.

3. Non-synonymous variants occurring within the following residues of *DNMT3A*: p.292-350, p.482-614 and p.634-912; *TET2*: p.1104-1481 and p.1843-2002; or *NOTCH1*: p.1574-1620, p.1671-1721.

4. Truncating variants in *CALR* exon 9.

5. *FLT3* internal tandem duplications.

6. Non-synonymous variants reported at least 10 times in COSMIC with VAF $<$ 0.4.

7. Non-synonymous variants falling within an annotated InterPro domain with VAF $<$ 0.4.

8. Non-synonymous variants reported in COSMIC $>$ 100 times.

If a variant did not meet these criteria, it was annotated as a variant of unknown significance (VUS). Cases with CH were defined by the presence of at least one pathogenic driver mutation at VAF $\geq$ 0.01.

2.3.5 Comparison of bone marrow and peripheral blood allele frequencies

Variants were called in peripheral blood DNA sequencing data as described above, using a minimum variant allele frequency of 0.005. Unfiltered variant calls were intersected with the list of curated bone marrow variants to compare the VAF between bone marrow and blood. For those mutations known to be present in bone marrow which were not called in peripheral blood, raw allele counts were performed directly from the BAM files using bcftools mpileup (with minimum base quality of 30 and minimum mapping quality of 35), and the allele frequency calculated.

2.3.6 Statistical analysis of targeted sequencing

All plots and statistical analyses were performed in the R statistical programming environment. Visualisation of mutational co-occurrence and the position of coding variants in protein domains was performed using the maftools v2.10.0 R package. Fisher's exact test was used to assess the significance of differences in the prevalence of CH between groups. A two-sided Wilcoxon rank-sum test was used to assign significance level for differences in medians. For the analysis comparing the frequency of CH in the MARCH and Abelson *et al.* [59] cohorts, only variants with VAF ≥ 0.01 were included to ensure comparability. From the Abelson study, only samples in the control discovery and validation cohorts (not pre-AML cases) were included, and duplicate samples were excluded where there were multiple timepoints from the same individual.

2.4 Cell culture

JURKAT cells were kindly provided by Prof. Adam Mead and Dr. Nikolaos Sousos. Cells were maintained in culture in RPMI-1640 supplemented with 10% FBS and 1% penicillin-streptomycin. For testing of TARGET-seq methods, an aliquot of the culture was taken and washed in media prior to antibody staining.

2.5 Flow cytometry and cell sorting

Cryopreserved BM samples were thawed as outlined in Section 2.2. Live cells were counted using trypan blue exclusion or using an automated cell counter (NucleoCounter, ChemoMetec). Antibody incubations were carried out in flow cytometry staining buffer (IMDM with 10% FBS + 10 μg/mL DNAse) for 20–30 minutes at 4°C in the dark. Up to 5×10^6 cells were stained in 100 μl volume and scaled up, as appropriate based on cell number. Following antibody incubations, cells were washed with 1 mL flow cytometry staining buffer, centrifuged at 350 g for 5 min and resuspended in flow cytometry staining buffer containing the live/dead stain (Propidium iodide (final concentration 3 μM; Biolegend) or 7-AAD (final concentration 1:100; Biolegend)). For FACS sorting, the cell suspension was filtered through a 35 μm cell strainer and live/dead stain was added immediately before sorting. Unstained, single-stained and fluorescence-minus-one (FMO) controls were used to determine background staining and compensation in each channel. Single-stained controls used the Anti-Mouse Ig, κ/Negative Control Compensation Particle Set and the anti-rat Ig, κ/Negative Control Compensation Particle Set (Becton Dickinson, Oxford UK) and a live/dead-only stained sample.

For flow cytometry analysis of primary BM samples, unseparated BM MNCs were used. Samples were stained with antibodies listed in Table A.4. Gating was kept consistent across all samples to enable quantification of population sizes.

For FACS sorting of BM samples for TARGET-seq/TARGET-seq+ experiments either unseparated BM MNCs or CD34-enriched BM MNCs were used. Samples were stained with antibodies listed in Table A.5 and 7-AAD (Biolegend) for dead cell exclusion. Unseparated BM MNCs were pre-sorted to enrich for Lin–CD34+ cells prior to single-cell sorting into plates containing lysis buffer (see below for details). Doublets and dead cells were excluded. For clonal haematopoiesis and control samples, the following populations were sorted: Live/Lin–CD34+, Live/Lin–CD34+CD38– and a small proportion of Lin+ cells. In addition to the sample

of interest, cells from the NOC153 control sample were sorted onto every plate, making up approximately 10% of wells on each plate, and two wells were left empty as no-template controls. During single-cell sorting, the time to sort each plate was kept under 15 minutes in order to avoid degradation of mRNA. Once completed, each plate was centrifuged and snap frozen on dry ice prior to storage at –80 °C. For index sorting, the following parameters were recorded for all single cells: FSC, BSC (SSC), Lineage/live/dead, CD34, CD38, CD117, CD45RA, CD10, CD90, CD123 and CD49f.

Sorting was performed on a Sony MA900 for 384-well plates or a BD Fusion, and flow cytometry analysis was done on a LSR Fortessa X20. Data analysis was performed using FlowJo v10.8.0 and R.

2.6 TARGET-seq+ optimisation

This section outlines the details of pilot optimisation experiments in the development of TARGET-seq+; the final TARGET-seq+ protocol is described in Section 2.9.

All comparisons were performed on single cells from the JURKAT cell line and primary human HSPCs with 16–48 cells per condition. Cells were sorted into 96-well plates containing 3 µL of lysis buffer consisting of 0.1% Triton X-100 (Sigma-Aldrich), 0.5 mM dNTP (Life Technologies), 5% polyethylene glycol (PEG) 8000 (Sigma-Aldrich), 0.5 U/µl RNase inhibitor (Takara), 2.7×10^{-5} AU/mL protease (Qiagen), 0.5 µM oligodT primer (see Table A.12) and 0.02 µl of 1:2,000,000 diluted ERCC spike-in mix 1 (Life Technologies).

Reverse transcription (RT) reactions were performed using the Smart-seq3 RT buffer containing 25 mM Tris-HCl (Thermo Scientific), 30 mM NaCl (Invitrogen), 2.5 mM MgCl2 (Invitrogen), 1 mM GTP (Thermo Scientific), 8 mM Dithiothreitol (DTT, Thermo Scientific), 0.5 U/µl RNase inhibitor (Takara), 2 µM of different template switching oligos (TSO), and 2 U/µl of Maxima H-minus reverse transcriptase enzyme (Thermo Scientific).

2.6.1 Comparison of RT-PCR conditions and DNA polymerases

Two DNA polymerases, KAPA HiFi HotStart (KAPA Biosystems) and SeqAmp (Clontech), were tested with the Smart-seq3 RT conditions described above. For each, the recommended concentration of enzyme and PCR buffer was used. PCR reaction volumes were maintained at 10 µL by adjusting the volume of H_2O. For all PCR reactions the Smart-seq3 cycling conditions were used, using 22 cycles for JURKAT and 26 for HSPCs. Success of RT-PCR amplification was assessed by fragment analysis of purified single-cell libraries using a High Sensitivity NGS Fragment Analysis Kit (1 bp - 6,000 bp) on a Fragment Analyzer (Agilent).

2.6.2 Comparison of template-switching oligos

Template-switching oligos (TSOs) from the Smart-seq2 (SS2) protocol (containing locked nucleic acid (LNA) modification and ISPCR adapter sequence) and the Smart-seq3 (SS3) protocol (without LNA modification) were compared to determine cDNA yield, which was measured by running purified libraries on a Fragment Analyzer.

2.6.3 Comparison of targeted primer concentrations

Targeted genotyping primers were tested in decreasing concentrations to assess the impact on cDNA amplification and genotyping targeted amplification. For mRNA primers, concentrations of 700 nM (used in the TARGET-seq protocol), 70 nM and 14 nM were tested. For gDNA primers, concentrations of 800 nM (used in the TARGET-seq protocol), 400 nM and 140 nM were tested. No targeted primer controls were included in all tests. Efficiency of cDNA amplification and primer-dimer formation was assessed by fragment analysis and measurement of cDNA concentration. Efficiency of genotyping target amplification was assessed by performing a second PCR reaction on the preamplified cDNA-amplicon mix using nested primers and performing gel electrophoresis (further details below).

2.7 TARGET-seq library preparation

TARGET-seq libraries were generated according to the published protocol [226, 227] for preliminary experiments and the validation experiment for comparison with TARGET-seq+.

Cells were sorted into 96-well plates containing 4.1 µL of lysis buffer. Lysis buffer was prepared as described in Table A.6, consisting of 0.18% Triton X-100 (Sigma-Aldrich), 1.0 mM dNTP (Life Technologies), 1.0 U/µl RNase inhibitor (Takara), 2.7×10^{-5} AU/mL protease (Qiagen), 1.0 µM barcoded oligodT-ISPCR primer (see Table A.14) and 1:8,000,000 diluted ERCC spike-in mix 1 (Life Technologies).

To perform cell lysis and denaturation of the RNA, plates were incubated at 72°C for 15 min and immediately centrifuged and placed on ice afterwards. RT mix was prepared as described in Table A.7. Then 5.6 µL of RT mix, containing SMARTScribe First-Strand Buffer, SMARTScribe enzyme, RNase inhibitor, SS2 TSO-LNA (1 µM final concentration) and targeted mRNA primers (700 nM final concentration), was added to each well. Reverse transcription and template switching were carried out at 42 °C for 90 min followed by 10 cycles of 50 °C for 2 min and 42 °C for 2 min. The reaction was terminated by incubating at 70 °C for 5 min. PCR pre-amplification was performed directly after reverse transcription by adding 15 µL of PCR mix, consisting of SeqAmp PCR Buffer, SeqAmp DNA polymerase, ISPCR primers (50 nM final concentration) and targeted gDNA and cDNA primers (see Table A.8). 20 cycles of amplification were used for JURKAT cells and 24 cycles cycles of amplification for HSPCs.

Single-cell genotyping and 3' transcriptome libraries were prepared as for TARGET-seq+ libraries detailed below.

2.8 Design and validation of targeted genotyping primers for TARGET-seq+

Targeted genotyping primers used in the pre-amplification RT-PCR step were designed to amplify amplicons 180–900 bp long. Where possible, gDNA primers were designed to anneal within intronic regions flanking the mutation of interest, and mRNA primers were designed to anneal to exonic regions outside of the gDNA amplicon, so that independent amplicons would be generated from mRNA and gDNA. Furthermore, when a heterozygous SNP was observed close to the mutation, primers were placed so that this would fall within the amplicon, enabling a direct measurement of allelic dropout. Primers were designed in Primer3Plus [228] and specificity was checked using Primer-BLAST [229]. For each target, at least four primer pairs were tested in a PCR reaction with bulk genomic DNA or cDNA template using the KAPA HiFi HS polymerase and the same reaction conditions as in the pre-amplification PCR reaction. Specificity and efficiency of amplification was assessed using gel electrophoresis. The best pair was taken forward for single-cell validation.

Single-cell validation of pre-amplification genotyping primers was performed on FACS-sorted single Lin–CD34+ HSPCs from a control sample. Testing in primary human single cells was important because it was found that some primer combinations that were successful in cell lines were not successful when used in primary cells. First, to determine whether a given combination of primers was compatible with whole-transcriptome library amplification, RT-PCR reactions were performed using the multiplexed combination of targeted mRNA and gDNA primers intended to be used for a given patient sample. A minimum of 8 single cells were tested for each combination and no-targeted primer controls were used for comparison. The success of cDNA amplification was assessed by fragment analysis of purified libraries using a High Sensitivity NGS Fragment Analysis Kit (1 bp - 6,000 bp) on a Fragment Analyzer (Agilent). If a given combination of targeted

primers compromised cDNA amplification by excessive primer-dimer or concatemer formation, alternative primers were tested.

Targeted primers for use in the genotyping PCR1 step (see Section 2.9.4) were designed to be nested within each of the amplicons generated in the pre-amplification RT-PCR and to generate amplicons of 150–450 bp in length. The use of a second nested PCR increases the efficiency and specificity of target amplification for subsequent sequencing. As for pre-amplification primers, gDNA primers were designed to anneal within intronic regions flanking the mutation, and cDNA primers were designed to anneal to exonic regions outside of the gDNA amplicon, so that independent mutational readouts would be generated from cDNA and gDNA where this was possible. Primer pairs were tested for specificity and efficiency in bulk PCR reactions using the KAPA2G Robust HS polymerase.

Finally, the amplification of genotyping targets was validated in single cells. For each target, an aliquot of the unpurified pre-amplified cDNA-amplicon mix from single-cell testing (above) was used as input, and genotyping PCR1 reactions were performed using the nested primers. At least 8 single cells were tested for each amplicon. Specificity and efficiency of amplification was assessed by gel electrophoresis (an example is shown in Figure 2.1). In the final experiment, nested primers were multiplexed in combinations of similar amplification efficiency in order to generate approximately equimolar amounts of each amplicon for sequencing.

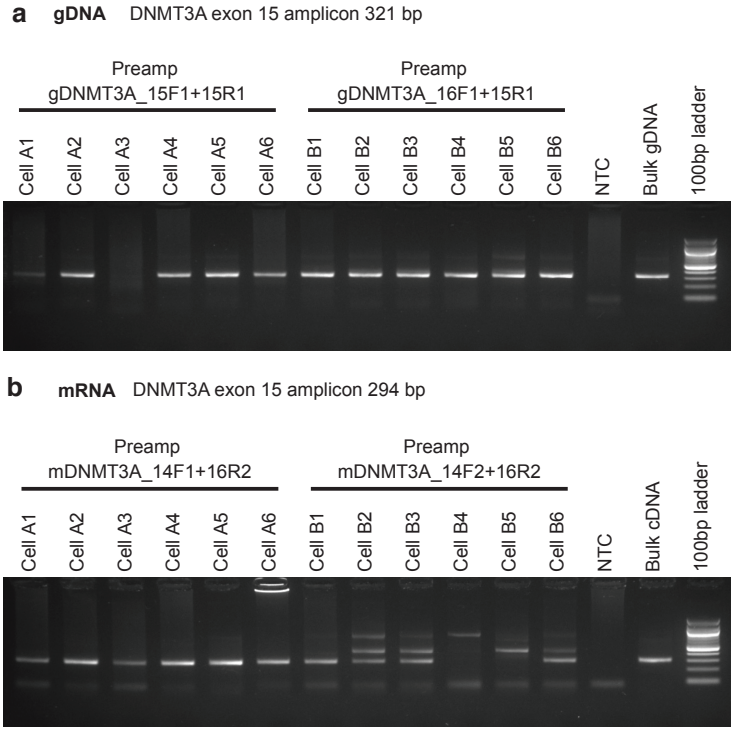


Figure 2.1: Validation of targeted primers for a genotyping locus. Agarose gel electrophoresis showing targeted amplification of *DNMT3A* exon 15 gDNA **(a)** and exon 15 mRNA/cDNA **(b)** from single HSPCs. In this example, two separate combinations of preamplification primers were tested (indicated by the annotation above lanes A1-A6 and B1-B6 respectively), followed by a second PCR using nested primers. For gDNA, amplification is best with the right-hand primer pair. For mRNA/cDNA amplification is best with the left-hand combination, whereas the right-hand combination produces non-specific bands. NTC: no template control, Bulk: positive control using bulk cDNA or gDNA.

2.9 TARGET-seq+ library preparation

2.9.1 Lysis buffer preparation

Lysis buffer was prepared as described in Table A.9, consisting of Triton X-100 (Sigma-Aldrich), dNTP (Life Technologies), PEG 8000 (Sigma-Aldrich), RNase inhibitor (Takara), protease (Qiagen), and 1:8,000,000 diluted ERCC spike-in mix. 25 µL of lysis buffer was dispensed into each well of a 384-well stock plate using a Formulatrix Mantis with high-volume chip, and 1.79 µL of 10 µM barcoded oligodT-ISPCR primer was added to each well using an Integra Viaflow and mixed. 3 µL of barcoded oligodT-lysis buffer mix was then transferred into each well of a 384-well

plate. Plates were sealed and stored at –80 °C until use. After cell sorting was complete, plates were sealed, centrifuged, and snap frozen on dry ice prior to storage at –80 °C. Barcoded oligodT-ISPCR primers were kindly provided by Prof. Adam Mead and Dr. Alba Rodriguez-Meira, and the sequences are listed in Table A.14.

2.9.2 Reverse transcription and pre-amplification

Plates were removed from –80 °C storage, transferred on dry ice and then incubated at 72 °C for 15 min to perform cell lysis, RNA denaturation and protease heat inactivation. RT mix was prepared as described in Table A.10. Immediately after the incubation, plates were centrifuged, placed on a cold block and 1 µL RT mix was added. Plates were then sealed, centrifuged and incubated on a thermocycler at 42 °C for 90 min followed by 10 cycles of 50 °C for 2 min and 42 °C for 2 min to perform reverse transcription and template switching. The reaction was terminated by incubating at 85 °C for 5 min.

Pre-amplification PCR mix with targeted genotyping primers was prepared as described in Table A.11. Pre-amplification PCR was performed directly after reverse transcription by adding 6 µL of PCR mix and incubation on a thermocycler using the following programme: 98 °C for 3 min for initial denaturation, 20–24 cycles of 98 °C for 20 s, 67 °C for 30 s and 72 °C for 6 min. Final elongation was performed at 72 °C for 5 min. 20 cycles of amplification was used for JURKAT cells and 21 cycles of amplification for HSPCs in the main experiments. The sequences of the primers used in the RT-PCR steps for whole transcriptome amplification and targeted genotyping amplification are listed in Table A.15.

After RT-PCR, successful pre-amplified libraries would contain both whole transcriptome cDNA and amplicons spanning each mutation that was targeted with genotyping primers. An aliquot of this cDNA-amplicon mix was used for whole transcriptome library preparation and another aliquot for single-cell genotyping library preparation. 1 µL of cDNA-amplicon mix was pooled per well to create cDNA pools from 192 uniquely barcoded single-cell libraries. Each cDNA pool was purified twice using Ampure XP beads with 0.6:1 beads to cDNA ratio. The

quality of pooled cDNA libraries was checked using a High Sensitivity DNA Kit on a Bioanalyzer (Agilent) or a High Sensitivity NGS Fragment Analysis Kit (1 bp - 6,000 bp) on a Fragment Analyzer (Agilent) and concentrations quantified by Qubit dsDNA HS Assay (Life Technologies). These pools were used to generate 3' biased whole-transcriptome libraries. The remainder of the cDNA-amplicon mix was diluted 1:2 with water and stored at –20 °C for use in single-cell genotyping.

2.9.3 Whole-transcriptome library preparation and sequencing

Bead-purified pooled cDNA libraries were used for tagmentation-based library preparation with a Nextera XT DNA Library Preparation Kit (Illumina) using a custom PCR amplification strategy to generate 3' biased libraries containing oligodT cell barcodes as described in Rodriguez-Meira *et al.* [226] with some modifications. Pooled cDNA libraries were diluted to 800 pg/µL and a total of 4 ng (5 µL) from each 192-cell pool was used in the tagmentation reaction with 10 µL tagmentation buffer (TD) and 5 µL ATM enzyme. The tagmentation reaction was incubated at 55 °C for 10 min, followed by the addition of 5 µL 0.2% SDS to release Tn5 from the DNA. Library amplification was performed using 5 µL Nextera XT i7 forward index primer (Illumina) and 5 µL custom i5 index primers (2 µM) adapted from the P5-index adapter in Rodriguez-Meira *et al.* [226] (see Tables A.12 and A.13 for sequences). The custom i5 index primer binds the barcoded oligodT-ISPCR adapter, resulting in amplification of the 3' fragments containing the cell barcode. PCR was performed by adding NPM enzyme (Nextera XT DNA Library Preparation Kit, Illumina) and incubation on a thermocycler using the following programme: 30 s at 95 °C; 14 cycles of 10 s at 95 °C; 30s at 55 °C; 30 s at 72 °C; and a final elongation of 5 min at 72 °C.

After tagmentation, each indexed pool was purified twice with Ampure XP beads using a 0.7:1 beads to cDNA ratio. Library quality was checked using a High Sensitivity DNA Kit on a Bioanalyzer and quantified using Qubit dsDNA HS Assay. Equimolar pools were made, and sequenced using custom sequencing primers for Read1 and Index2 (P5-SEQ, I5-SEQ, 300 nM in HT1 buffer, see Table

A.12). For benchmarking experiments, libraries were sequenced on a NextSeq 500/550 High Output v2.5 (75 cycle) kit (Illumina) using the following sequencing configuration: 15 bp R1; 8 bp index read; 69 bp R2. For the main experiments, up to 9,984 single-cell libraries (52 pools of 192 single-cell libraries) were sequenced on a NovaSeq S4 flow cell using the following sequencing configuration: 15 bp R1; 8 bp index read 1; 8 bp index read 2; 200 bp R2.

2.9.4 Targeted single-cell genotyping

To generate Illumina-compatible libraries for single-cell genotyping, two PCR steps were required. As the genotyping amplicons generated by the pre-amplification RT-PCR are un-barcoded, genotyping PCR reactions were carried out separately for each single-cell library.

In the first PCR step (genotyping PCR1), nested target-specific primers containing universal CS1 (forward primer) or CS2 (reverse primer) adapters are used to amplify the target regions of interest. Incorporation of a barcode sequence specific to each plate into these primers enables libraries from different plates to be pooled later on. Target-specific primers were designed to amplify annotated coding regions in the case of cDNA amplicons and at least one intronic region in the case of gDNA amplicons. The primer sequences used for genotyping PCR1 for each sample are described in Table A.16. PCR1 reactions were performed using 3.25 µL of KAPA 2G Robust HS Ready Mix (Sigma-Aldrich), 1.5 µL of diluted cDNA-amplicon mix and 300 nM target-specific primers in a 6.5 µL reaction.

In the second PCR step (genotyping PCR2), Illumina-compatible adapters containing a 10 bp cell barcode are attached to the genotyping PCR1 product by binding to the CS1/CS2 adapters. PCR2 reactions were performed using FastStart High Fidelity polymerase (Sigma-Aldrich) with 1.0 µL of PCR1 product and 1.2 µL of each barcode primer mix (Access Array Barcode Library for Illumina Sequencers-384, Single Direction, Fluidigm) in a 6.2 µL reaction.

Indexed amplicons were pooled using a Mosquito HTS liquid handling platform (TTP Labtech) and purified with Ampure XP beads (0.8:1 beads to PCR product

ratio). Purified pooled libraries were quantified using Qubit dsDNA HS Assay and the quality checked using a Tapestation High Sensitivity D1000 kit (Agilent) to ensure the size distribution of amplicons was appropriate. Each pool was diluted to a final concentration of 4 nM and further diluted to 10 pM in HT1 buffer prior sequencing. Libraries were sequenced on a NextSeq 500/550 Mid Output v2.5 kit (300 cycle) (Illumina) using 150 bp paired-end reads, with 10 bp for the cell barcode index read and custom sequencing primers (Table A.12).

2.10 Single-cell genotyping analysis

2.10.1 Mutational calling

Single-cell genotyping reads were pre-processed using the custom TARGET-seq pipeline (https://github.com/albarmeira/TARGET-seq). First, reads were demultiplexed based on the 384 cell barcodes incorporated in genotyping PCR2 and plate barcodes genotyping PCR1, generating separate FASTQ files for each single cell. Reads were aligned to hg38 using STAR (v2.7.3a) with default settings and cDNA/gDNA amplicons were separated into different bam files, extracting reads matching the primer sequences used for targeted PCR barcoding. This allowed independent mutational information to be obtained from cDNA and gDNA amplicons. Variant calling was performed using mpileup (samtools v1.1, options: `--minBQ 30, --count-orphans, --ignore overlaps`) and results were summarized with the custom pipeline.

Mutational calling on single cells was then performed with custom R scripts, separately for each mutation. First, the total coverage across the mutant locus was calculated. A filtering threshold was applied to remove cells where the amplicon was not detected or coverage was too low for reliable genotyping. Filtering thresholds were set for each amplicon based on the number of reads in no-template controls and in outlying cells at the lower end of the coverage distribution (likely failed libraries). The minimum coverage was 30–100 reads for most gDNA amplicons and 20–50 reads for most cDNA amplicons. In cells with coverage below the threshold, the amplicon was called undetected.

The variant allele frequency (VAF) for the mutation was calculated for each cell as the number of reads with the mutant allele divided by the total coverage. PCR amplification and next-generation sequencing have an inherent error rate, which needs to be accounted for when calling cell genotypes. WT control cells sorted onto every plate enabled the error rate for each mutant allele to be determined. The threshold for calling mutations in each amplicon was set based on the VAF observed in control cells (threshold VAF 0.008–0.03, set individually for each mutation). Additionally a minimum of 5–30 mutant reads were required to call a cell mutant. Cells with a borderline VAF or where allelic dropout was confirmed by analysis of a germline SNP (see Section 2.10.2) were called undetermined.

Genotyping information from gDNA and cDNA amplicons were then combined and a consensus genotype assigned. Consensus genotypes were assigned as follows:

1. If the mutation was identified in either the gDNA or cDNA amplicon, the cell was called mutant.

2. If both amplicons were WT, the cell was called WT.

3. If the gDNA amplicon was WT but the cDNA amplicon was undetected or undetermined, the cell was called WT.

4. If the cDNA amplicon was WT but the gDNA amplicon was undetected or undetermined, the cell was called undetermined, due to the high allelic dropout rate of cDNA amplicons.

2.10.2 Allelic dropout estimation

Allelic dropout (ADO) rates were estimated by two independent methods. First, the allelic detection rate was measured for heterozygous germline SNPs present in the gDNA amplicon for four mutations. A VAF threshold and minimum number of reads to detect the reference or alternate alleles was set for each mutation, and cells were called heterozygous (Het) if the VAF and minimum number of reference and alternate reads fell within that boundary. Otherwise, cells were called homozygous reference (Hom Ref) or homozygous alternate (Hom Alt). SNP alleles were phased

with each mutation based on the positive or negative correlation in VAF. When neither a mutation or the in-phase SNP allele were detected, a cell was assigned an undetermined genotype as it could be either WT or mutant.

Second, the dropout rate was estimated from the frequency with which cells were called homozygous mutant, when the variant should be heterozygous. Homozygous mutant frequencies were calculated for all mutations that had at least 16 mutant cells. For each mutation, the threshold VAF and maximum number of reference reads required to call a cell homozygous were set to be the inverse of that used to call cells mutant (i.e. VAF ≥ 0.99 if VAF ≥ 0.01 was used to call mutant cells). The estimated dropout rate of the mutant allele was calculated as the number of homozygous mutant cells divided by the inferred total number of mutant cells ($2 \times$ number of homozygous mutant cells plus number of heterozygous mutant cells).

2.10.3 Inference of clonal hierarchies

In samples with multiple mutations, the pattern of mutational co-occurrence was used to determine clonal structures and assign a clonal identity to each cell. The principles used were based on those in Quek *et al.* [19]. In samples where mutations were mutually exclusive, such as in samples NOC131 (Figure 5.7), NOC117 (Figure 5.11) and NOC115 (Figure 5.13), it was clear that these belonged to independent clones.

In cases where mutations co-occur in the same cells, a linear or branching clonal structure may be present. For example, with mutations A, B and C, the structure might be $A \rightarrow AB \rightarrow ABC$ (linear) or $A \rightarrow AB + AC$ (branching). The occurrence of ADO means that in some cells, a mutation that is present will not be detected. For example a cell in clone AC might have the detected genotype C. In the case of samples NOC062 (Figure 5.8) and NOC137 (Figure 5.14) the clonal structure was unambiguous. In both cases, the dominant clone harbours the initiating mutation and no additional mutations (*DNMT3A* p.G890fs for NOC062 and *ASXL1* p.R965X for NOC137). This was consistent with the results of bulk DNA sequencing. In both cases, cells with the initiating mutation and no additional mutations were assigned

to the founder clone (A), whereas cells with cooperating mutations were assigned to daughter clones (AB, AC or ABC). In cells where a cooperating mutation (B or C) was detected but the initiating mutation (A) was not detected, it was assumed that ADO had occurred, and the cell was assigned to the appropriate daughter clone. In all cases, this was a rare occurrence consistent with the ADO rate.

In sample NOC002, there was a dominant double-mutant clone. Here the most likely sequence of acquisition was determined based on the following two observations: a) the frequency of cells with only *TET2* p.Q726X was greater than the frequency of cells with only *TET2* p.R1261C; and b) the VAF of *TET2* p.Q726X was higher than *TET2* p.R1261C in bulk DNA sequencing. The frequency of cells with *TET2* p.R1261C but not *TET2* p.Q726X was concordant with the estimated ADO rate.

For all downstream analyses, including differential gene expression, the clone assignment rather than the raw genotype was used to categorise WT and mutant cells.

2.11 Analysis of FACS index data

Flow cytometry index data were recorded for each single cell during FACS sorting for TARGET-seq+ experiments. Fluorescence values were recorded for forward scatter (FSC), back-scatter (BSC; equivalent to side scatter, SSC), Lineage/live/dead, CD34, CD38, CD117, CD45RA, CD10, CD90, CD123 and CD49f. Index data were matched with single-cell identifiers based on the well coordinate and combined with genotyping calls and other metadata into a unified data set. Virtual FACS gating was performed in R based on the strategy used for sorting. Gates were set based on populations that were negative for each marker. Cells were labelled as positive or negative for each surface marker, and Boolean logic used to assign an immunophenotypic population label. For example, cells that were Lin–CD34+CD38–CD10–CD45RA–CD90+ were labelled as immunophenotypic HSC. These categories were used to visualise the mutant cell frequency across HSPC compartments in Chapter 5, for labelling UMAP plots and in identifying HSC/MPP compartments for differential gene expression analysis.

2.12 Single-cell transcriptome data pre-processing

2.12.1 Mapping and transcript counting

Transcriptome sequencing data were demultiplexed into FASTQ files for each plate with a unique i7-i5 index combination using bcl2fastq. These files contained reads from up to 384 cells with a shared plate index. A custom python pipeline was used to further demultiplex and map the sequencing reads. First, reads from each plate were demultiplexed using the 14 bp single-cell barcode sequence in Read1 using cutadapt (v3.4). Concurrently, cDNA reads (Read2) were trimmed for polyA tails, Nextera adapters and low-quality reads using the following parameters: `--nextseq-trim=0.15 -A "A{20}N{80};min_overlap=3" -A AGCAACTCTGCGTTGATACCACTGCTT --minimum-length :25 -e 0.15 --pair-filter=any`. This generated individual FASTQ files with single-ended cDNA reads corresponding to each single-cell barcode.

Reads were then mapped to the hg38/GRCh38 reference genome and ERCC92 transcripts with STAR (v2.7.9a) using the GENCODE v38 reference gene annotation. Counts for each gene were obtained with featureCounts (Subread v2.0.1) using the following options: `-t exon -g gene_id -extraAttributes gene_name -primary -s 1`. Sequencing and mapping quality metrics were calculated with FastQC (v0.11.9), Samtools flagstat (v1.12), MultiQC (v1.11) and the outputs of STAR and featureCounts.

2.12.2 Transcript detection and dropout frequency calculation

For the comparison of transcript detection sensitivity between TARGET-seq and TARGET-seq+ (Figure 4.9), data were first downsampled to a set number of assigned reads per cell to remove differences due to unequal sequencing depth. Cells were filtered to include only those with that number of reads. The number of genes detected per cell was calculated as the sum of genes with at least one assigned read.

For calculation of dropout rates (Figure 4.10), data downsampled to 5×10^5 reads per cell were used. A random sample of 16 cells per chemistry were compared for JURKAT and 20 cells per chemistry for HSPC. The dropout frequency for a given gene was calculated as the percentage of cells in which the gene is not detected (normalised counts < 1). Genes were divided into three groups to compare the dropout rate in: a) all expressed genes, defined as genes detected in at least 2 cells by any method; b) frequently expressed genes, defined as genes detected in >50% of all cells; and c) lowly expressed genes, defined as genes with a mean of 2–10 normalised counts per cell.

2.12.3 Cell-to-cell correlation analysis

Cell-to-cell correlations for JURKAT cells processed with each method (Figure 4.10c) were calculated using pairwise Pearson correlations in libraries downsampled to 5×10^5 reads per cell.

2.13 Single-cell transcriptome analysis

2.13.1 Quality control and normalisation of single-cell transcriptome data

Single-cell transcriptome analysis was performed using the SingCellaR package (v1.2.0, `https://github.com/supatt-lab/SingCellaR`) [230]. Metadata including genotyping and FACS index data were matched with single-cell identifiers based on the plate and well coordinates.

Cells meeting the following filtering criteria were included in the analysis: reads assigned to genes $>$ 20,000; genes detected $>$ 1500; reads assigned to ERCC transcripts $<$ 50%; reads in mitochondrial genes $<$ 10–14% depending on an individual sample. 85–95% of cells per sample passed filtering. Genes expressed in fewer than 5 cells were removed.

Reads were normalised by library size using the pool normalisation method with prior clustering from the scran package [231]. Highly variable genes were identified by fitting a generalised linear model to the relationship between the mean expression

and squared coefficient of variation (CV^2) for the ERCC spike-ins, which can be used to estimate technical noise (using the `BrenneckeGetVariableGenes` function from the M3Drop package) [232, 233]. Genes for which the CV^2 exceeded technical noise (using a false discovery rate (FDR) of 0.1) were considered variable.

2.13.2 Dimensionality reduction and clustering for individual samples

Initially, samples were analysed individually. Limma was used to regress out the plate batch effect (`SingCellaR::remove_unwanted_confounders` function). 3087–8768 variable genes were identified per sample after removing mitochondrial genes. Principal Component Analysis (PCA) was then performed using the highly variable genes (`SingCellaR::runPCA` function). An elbow plot was inspected to determine the number of principal components (PCs) capturing most of the variance. Uniform Manifold Approximation and Projection (UMAP) analysis was performed using the first 30 PCs and 15 nearest neighbours using 'cosine' as the uwot metric parameter (`SingCellaR::runUMAP` function) [234]. Cells were clustered using the `SingCellaR::identifyClusters` function with the first 30 PCs, 15 k-nearest neighbours (KNN) and 'Euclidean' as the distance metric. This method clusters cells using a k-nearest neighbour approach. After nearest neighbours are identified, the weighted graph is created with weight values calculated from normalised shared number of the nearest neighbours. The 'louvain' community detection method from the igraph package is then applied to identify clusters.

2.13.3 Single-cell copy number analysis

Copy number variants at the single-cell level were called with the InferCNV R package (v1.8.1), using the following parameters: `cutoff = 1, denoise = TRUE, HMM = TRUE, analysis_mode='subclusters', hclust_method = 'ward.D2'`. Cells from the 3 control samples were used as the reference cells. Cluster identities were used as the cell annotation. The reference gene positions were taken from GENCODE v38.

2.13.4 Data integration and visualisation

Transcriptome data from all 10 clonal haematopoiesis and 3 control samples were combined into a single dataset using the `SingCellaR::preprocess_integration` function. The combined dataset was normalised using scran and variable genes were identified by modelling technical variation of the ERCC spike-ins as described above. This identified 11,928 variable genes (FDR < 0.05), which were used for PCA analysis.

Data integration was performed using Harmony to correct for donor effects, using the sample identifier as the covariate [235]. UMAP visualisation was performed pre- and post-integration to assess the effectiveness of the integration. Integration and visualisation were compared for a range of 20–200 top PCs and 5–50 nearest neighbours. The landscape was similar across all values tested but the best resolution was obtained by using a high number of PCs and a low number of nearest neighbours. For all subsequent analyses, the top 100 Harmony-adjusted PCs and 15 nearest neighbours were used.

2.13.5 Cell cycle and lineage signature scores

S-phase and G2M-phase cell cycle scores were calculated as the mean expression value of cell cycle gene signatures from Tirosh *et al.* [236] for each cell (`SingCellaR::add_cell_cycle_genes_score` function). Lineage signature gene sets with canonical lineage markers were obtained from Roy *et al.* [230]. Expression of signatures was plotted using the `SingCellaR::plot_umap_label_by_multiple_gene_sets` function.

2.13.6 Clustering analysis, marker gene identification, and cell type annotation

Clustering was performed using the `SingCellaR::identifyClusters` function as described in Section 2.13.2 using the Harmony-adjusted PCs. Marker genes differentially expressed in each cluster were identified with the `findMarkerGenes` function. This uses a non-parametric Wilcoxon test on log-transformed, normalised counts to compare expression level, and Fisher's exact test to compare the frequency

of cells expressing each gene [230]. *P*-values generated from both tests are then combined using Fisher's method and adjusted using the Benjamini-Hochberg (BH) multiple-testing correction. Genes expressed by each individual cluster were compared to all other clusters and differentially expressed genes defined with an absolute log2 fold change (log2FC) ≥ 0.5, adjusted *P*-value < 0.05, and fraction of expressing cells > 0.3. Differentially expressed genes were ranked using *P*-values and log2FC to select the top differential genes per cluster.

To annotate clusters, the `SingCellaR::identifyGSEAPrerankedGenes` function was used to pre-rank genes obtained from differential gene expression analysis comparing each individual cluster with all other clusters. Gene ranking scores were calculated as the log2FC multiplied by –log10 of the adjusted *P*-value. Gene set enrichment analysis (GSEA) was performed using the fgsea package (v1.20.0) [237]. Gene sets for cluster annotation were obtained from 9 publications that have characterised human haematopoiesis [236, 238–245]. Clusters were manually annotated using the gene set enrichment results, immunophenotypic surface marker expression and inspection of canonical marker genes.

2.13.7 Differential abundance analysis

Differential abundance between WT and mutant cells was analysed using the MiloR package (v1.2.0) [246]. This method builds a *k*-nearest neighbour graph and assigns cells to partially overlapping neighbourhoods enabling the abundance of cells in different neighbourhoods to be compared between genotypes. A Milo object was created using the top 100 Harmony-adjusted PCs and the following functions were run with respective parameters: `miloR::buildGraph` with `k = 25, d = 100`; then `miloR::makeNhoods` with `prop = 0.1, k = 25, d = 100, refined = TRUE`; then `miloR::countCells` with `sample = "Sample_clone"`, and `miloR::calcNhoodDistance` with `d = 100`. For differential abundance testing, a design matrix was created with the sample and clone identities. The `miloR::testNhoods` function was run with `design = Sample + clone, norm.method = "logMS"` to test for differences between clones, while accounting for sample

effects. Differential abundance results were plotted on the UMAP using the `miloR::buildNhoodGraph` function and annotated with cluster annotations using `miloR::annotateNhoods`. Neighbourhoods with a FDR < 0.2 were labelled as significant. For the beeswarm plot, neighbourhoods were excluded if they were mixed, with less than 60% of cells belonging to the main cluster. Milo analysis was performed separately for *DNMT3A*-mutant and *TET2*-mutant samples, including WT and single-mutant cells only.

For comparison of mutant cell frequency across clusters, the frequency of mutant cells within each cluster was first calculated for each sample. This frequency was then divided by the mutant cell frequency within the HSC-MPP cluster to generate a normalised frequency that could be compared across samples with different clone sizes. For each cluster, samples were excluded if there were fewer than 8 cells assigned to that cluster.

2.13.8 Pseudotime analysis

Pseudotime trajectories were constructed using the Slingshot R package (v2.2.0) [247] using the top 100 Harmony-adjusted PCs as input. The merged 'HSC-MPP' cluster was used as the start cluster, and 'Late Erythroid', 'MkP', 'GM-precursors', 'pDC' and 'Pre-B' were used as the end clusters. Pseudotime ordering scores were extracted and plotted on the UMAP embedding to visualise each trajectory. Differences in median pseudotime score between WT and mutant cells was tested by fitting a gamma generalised linear mixed model to the pseudotime score, using sample identity as the random effect and clone identity as the fixed effect. P values were obtained by a likelihood ratio test of the full model with the clone effect against the model without the clone effect. For comparisons between WT and mutant cells within individual samples, differences in median pseudotime score were tested using a Wilcoxon rank-sum test with Benjamani-Hochberg correction for testing across multiple lineages. Regions along each trajectory with maximal clonal expansion were determined by manual inspection of the density plot.

2.13.9 Differential gene expression analysis

Differential gene expression comparing WT and mutant cells was complicated by the relatively subtle changes between genotypes in the setting of substantial sample-to-sample differences. In addition, the considerable differences in clonal composition between samples meant that in a WT to mutant comparison, expression differences might be dominated by the sample with most mutant cells. To overcome this, it was necessary to account for sample effects in differential expression testing. The strategy was guided by recent reviews and a systematic analysis of differential expression tools [248, 249].

Differential expression testing was performed using limma-trend (v3.50.0) [250] on log normalised counts, using the scran normalisation size factors as described in Vieth *et al.* [249]. A linear model was fitted to each gene using `limma::lmFit` and differential expression testing was performed using `limma::eBayes` with `trend = TRUE, robust = TRUE`. The clone was used as the test variable and sample identity as the covariate. This effectively performs a paired analysis between WT and mutant cells within each biological replicate (i.e. sample). P values were adjusted for multiple testing with Benjamani-Hochberg correction. Differentially expressed genes were defined as genes expressed in at least 30% of WT or mutant cells within each comparison with FDR < 0.05.

2.13.10 Gene set enrichment analysis

Gene rankings for gene set enrichment analysis (GSEA) were generated by differential gene expression testing using limma-trend as described above. Genes were filtered to include those expressed in at least 1% of WT or mutant cells within each comparison. Genes were ranked by the sign of the fold change multiplied by -log10 of the adjusted P-value using the following formula: `sign(log2FC) * -log10(adjusted P-value)`. To perform GSEA, the `fgseaMultilevel` function was run from the fgsea package [237]. Significantly enriched gene sets were filtered for redundant pathways using the `collapsePathways` function and a FDR as described in each figure.

Haematopoietic sets used for comparisons between WT and mutant cells were obtained from 18 publications of human and murine haematopoiesis, in particular studies of HSC quiescence, dormancy and activation, and studies differentiating stem and progenitor subtypes [50, 236, 238–245, 251–258]. Murine gene annotations were converted to their human orthologues using the biomaRt package (v2.50.1). Haematopoietic gene sets were filtered for genes expressed in this dataset. MSigDB Hallmark and Gene Ontology Biological Process gene signatures were downloaded from the MSigDB database (v7.5.1) `https://www.gsea-msigdb.org/gsea/msigdb`.

2.13.11 AUCell analysis

The AUCell package (v1.16.0) [259] was used to quantify the gene set activity in single cells. AUCell gene-expression rankings were created using the `SingCellaR::Build_AUCell_Rankings` function. AUCell gene signature enrichment was then calculated using the `SingCellaR::Run_AUCell` function with the gene matrix transposed (GMT) file of gene sets. Haematopoietic gene sets were the same as those used for GSEA analysis described above. For plotting AUCell scores on UMAP plots, the `SingCellaR::plot_umap_label_by_AUCell_score` function was used. Differences in mean AUCell scores between WT and mutant cells were tested by a two-way ANOVA, using the sample identity as a covariate. A linear model was fitted to the AUCell score, using clone identity and sample identity as fixed effects. P values were obtained by a likelihood ratio test of the full model with the clone effect against the model without the clone effect.

Chapter 3

Collection and characterisation of a clonal haematopoiesis cohort

Contents

3.1 Establishing the MARCH cohort 56
3.1.1 Introduction to the MARCH Study 56
3.1.2 Participant recruitment 57
3.1.3 BM sample processing and cell yields 58
3.2 Identification of cases with clonal haematopoiesis by targeted DNA sequencing 59
3.2.1 Approach to targeted DNA sequencing 59
3.2.2 Somatic variant calling 60
3.2.3 Genetic landscape of cases with clonal haematopoiesis . 62
3.2.4 Association of clonal haematopoiesis with clinical parameters 66
3.2.5 Molecular characteristics of somatic variants 68
3.3 Comparison of somatic variants between BM and PB 70
3.4 Bone marrow immunophenotypic analysis 72
3.4.1 Effect of age on HSPC frequencies 73
3.4.2 Effect of CH on HSPC immunophenotypes 74
3.5 Conclusions: Collection and characterisation of a clonal haematopoiesis cohort 78

I am grateful to Batchimeg Usukhbayar, Mirian Salazar and Rachel Moore for their help with sample banking, and to Rachel and Marlen Metzner for their help with targeted genotyping experiments described in Chapter 3.

3.1 Establishing the MARCH cohort

3.1.1 Introduction to the MARCH Study

In order to study clonal haematopoiesis in humans, it was essential to obtain haematopoietic stem and progenitor cells (HSPCs) from a sufficiently large cohort of aged individuals. Bone marrow is the most abundant source of HSPCs but samples from individuals without haematological disease are not readily available. One

potential source of bone marrow tissue is from individuals undergoing total hip replacement (THR) surgery, where there is access to the femoral canal during the operation. The Mechanisms of Age-Related Clonal Haemopoiesis (MARCH) Study was therefore established to collect samples from individuals undergoing elective THR surgery. The aim was to collect samples from approximately 200 participants. These were screened for somatic mutations by targeted DNA sequencing to identify cases with CH. Given the prevalence of CH is 10–20% in individuals over the age 65 [58, 59, 91], it was predicted that this approach would identify sufficient cases of CH to take forward for further study.

3.1.2 Participant recruitment

Participants were recruited from adult patients undergoing elective THR at the Nuffield Orthopaedic Centre (NOC) in Oxford. Individuals with a previous history of haematological malignancy, rheumatoid or inflammatory arthritis, or who were receiving medications that could alter haematopoietic cell output such as steroids and bisphosphonates were excluded. Recruitment started in January 2018 and samples were collected from 195 participants over the course of 2 years (Figure 3.1). Samples of peripheral blood (PB), bone marrow (BM) and hair follicles were collected from each participant. The cohort had a median age of 70 years (range 30–92), and 55% were female. Clinical details including comorbidities, concurrent medications and full blood counts were collected for all participants. The full data are set out in Table A.1 and summarised in Table 3.1 later in this chapter. Overall, 43.6% of the participants had a history of cardiovascular disease (mainly hypertension), 16.9% had a history of autoimmune disease, and 12.8% had a history of previous non-haematological cancer.

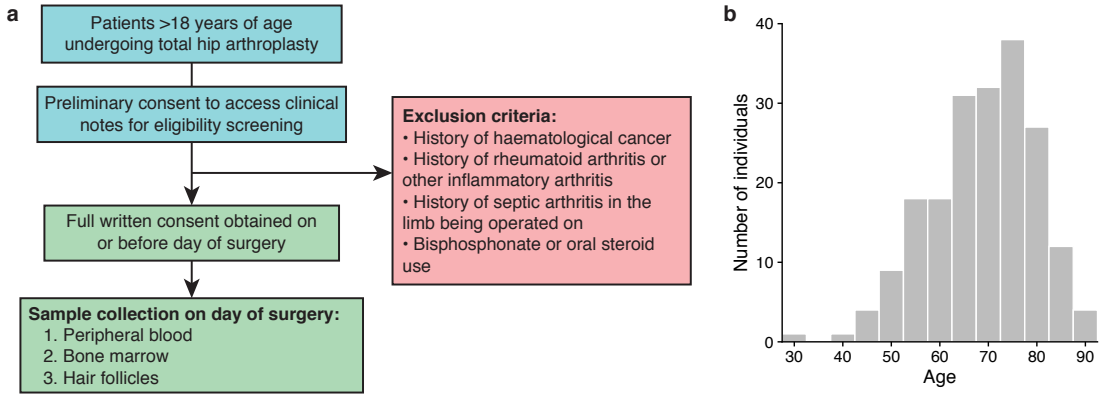


Figure 3.1: Participant recruitment for the MARCH cohort. (a) Participant recruitment procedure. **(b)** Age distribution of the MARCH cohort. 195 individuals were recruited with a median age of 70 years (range 30–92).

3.1.3 BM sample processing and cell yields

As HSPCs constitute a rare population within the normal bone marrow, it was important to obtain good cell yields to make downstream experiments possible. Whereas in diagnostic BM sampling, haematopoietic cells are obtained by needle aspiration usually from the iliac crest, in the THR setting tissue is available from two distinct sources. First, a piece of trabecular bone is removed from the medullary cavity, and second, an aspirate may be taken from the femoral canal. Samples were collected in anticoagulated buffer and processed within 24 hours as shown in Figure 3.2a. It was found that the optimal source of BM mononuclear cells (MNCs) differed between individuals, so by collecting cells from both trabecular marrow and a femoral canal aspirate, the cell yield was improved (Figure 3.2b). Overall, a median of 250×10^6 BM MNCs (range 6–2240×10^6) were obtained and for samples where CD34-enrichment was performed, a median of 6.5×10^6 BM MNCs (range 1–28×10^6) were obtained (Figure 3.2c-d). Cell yields were negatively correlated with age (Figure 3.2e).

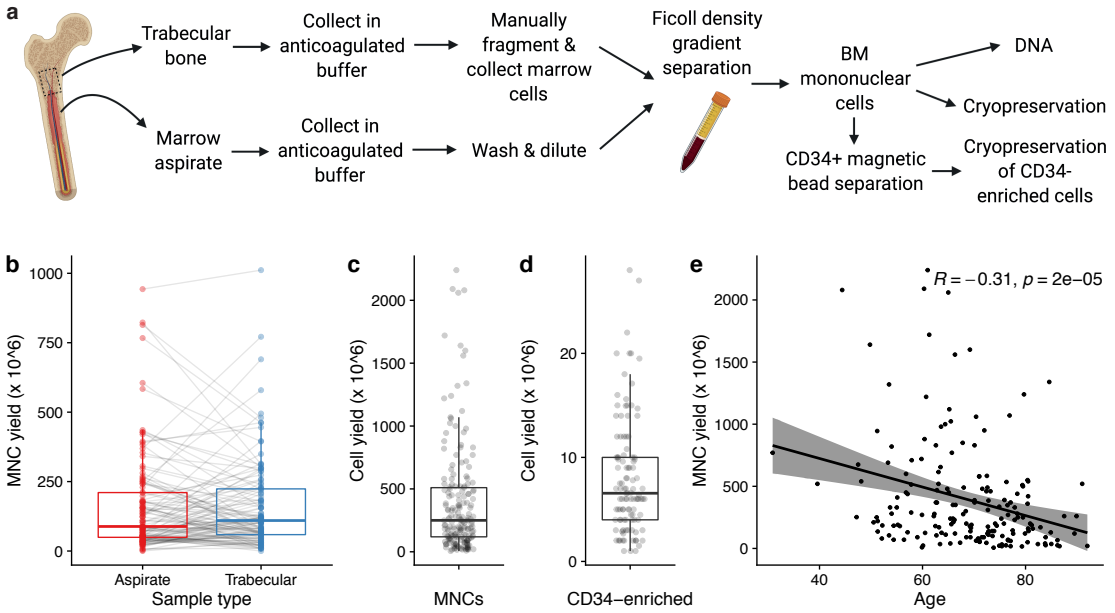


Figure 3.2: Bone marrow mononuclear cell isolation. **(a)** Method for isolation of mononuclear cells from femoral bone marrow. Trabecular bone fragments were manually fragmented and combined with the femoral aspirate after washing. Mononuclear cells were then isolated by Ficoll density gradient centrifugation. MNCs were frozen for DNA extraction and viable cells cryopreserved, in some cases after CD34 magnetic bead separation. *Created with BioRender.com.* **(b)** BM MNC yields from femoral aspirate and trabecular bone samples. Lines connect paired samples from the same individual. **(c)** Total yield of BM MNCs per individual. **(d)** Yield of CD34-enriched MNCs per individual, where this was performed. **(e)** Yield of BM MNCs plotted against the age of the donor. The linear regression line is depicted with the 95% confidence interval, and the Spearman correlation coefficient (R) is noted. BM: Bone marrow, MNCs: Mononuclear cells.

3.2 Identification of cases with clonal haematopoiesis by targeted DNA sequencing

3.2.1 Approach to targeted DNA sequencing

To identify cases with clonal haematopoiesis, a targeted DNA sequencing protocol was established using a hybridisation capture method to enrich for regions of interest. A capture panel was designed to target the coding regions of 97 genes recurrently mutated in CHIP and myeloid malignancies (Table A.2). Sequencing was performed on DNA extracted from BM MNCs because it was hypothesised that somatic mutations might be present at a higher frequency in BM than peripheral blood,

and thus yield a greater sensitivity for detecting CH. To enable the detection of variants with low variant allele frequency (VAF), a minimum sequencing coverage depth of 500X was aimed for. This gives a power of 0.97 to detect variants with a VAF of 0.02 with at least 5 variant reads (Figure 3.3).

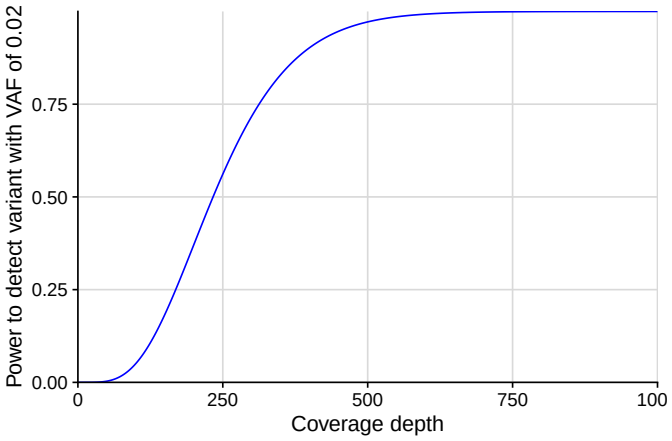


Figure 3.3: Power to detect a variant with VAF of 0.02 as a function of coverage depth. The probability of detecting an allele follows a binomial distribution, which depends on the sequencing depth and number of variant reads required to call a variant. Here the probability of detecting a variant with VAF of 0.02 with at least 5 variant reads is plotted across a range of sequencing depths.

DNA sequencing data were processed by a custom analysis pipeline as described in Chapter 2. Alignment rates were >99% for all samples, with on-target rates of >70% and PCR duplicates <40% for the majority of samples. After removal of poor quality reads and PCR duplicates, the median sequencing coverage was 790X, with 182 of 195 samples achieving >500X coverage (Figure 3.4a-d). Evenness of coverage across genes was good, and above average across the commonly mutated genes, including *DNMT3A*, *TET2*, *ASXL1*, and *TP53* (Figure 3.4e-f).

3.2.2 Somatic variant calling

Somatic variants were called using two variant callers (Mutect2 and VarDict [223, 224]) using a VAF threshold of 0.01. Variants were filtered to exclude sequencing artefacts, germline polymorphisms, and variants present in a panel of normal samples to retain only high confidence somatic variants. Variants detected by each

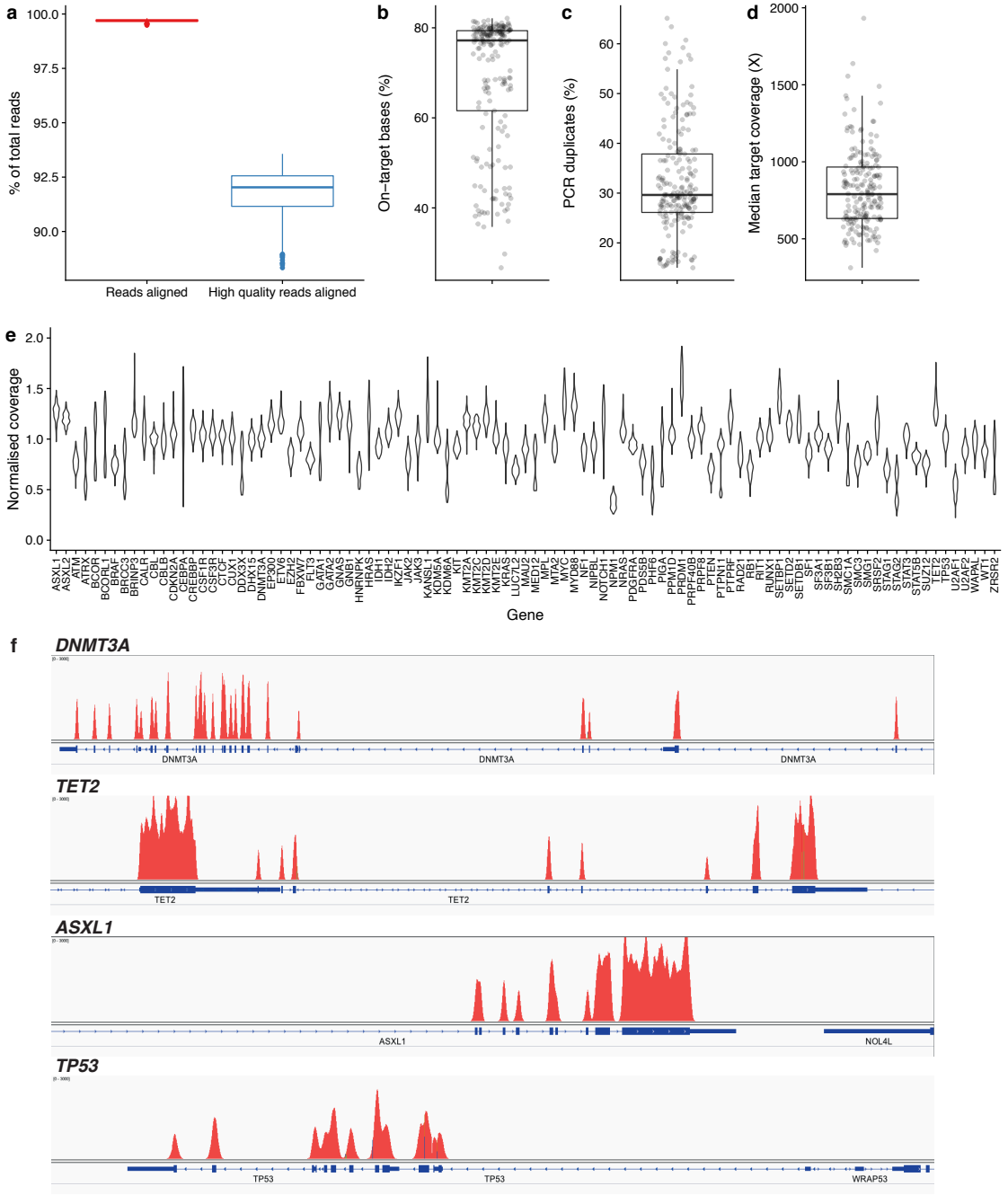


Figure 3.4: Mapping and coverage metrics for targeted sequencing. Each point represents a sample. **(a)** Percentage of reads aligned to the reference sequence, all reads (left) and those with a mapping quality of Q20 or higher (right). **(b)** Percentage of bases located on or near the target capture region. **(c)** Percentage of reads flagged as PCR duplicates. **(d)** The median coverage of a target region. **(e)** The coverage across each gene in the panel, normalised by the coverage per sample. **(f)** Representative IGV coverage tracks are shown for commonly mutated genes for one sample with median coverage of 767X. Coverage is shown in red and exons are indicated by the blue bars in the track below. The scale is 0 - 3000X.

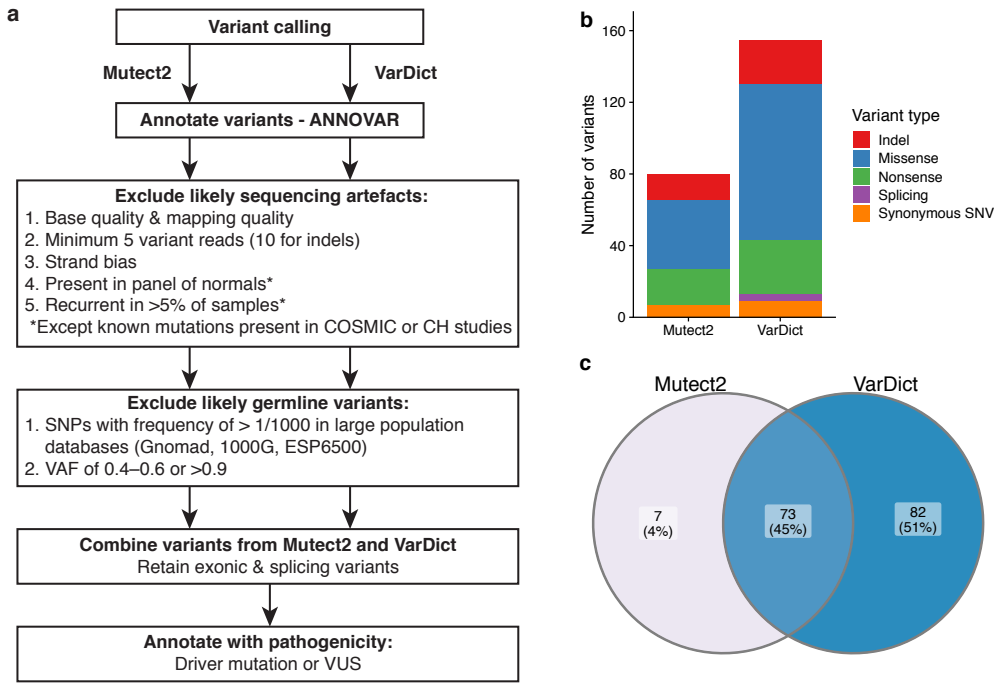


Figure 3.5: Somatic variant calling in BM targeted sequencing. (a) Overview of the variant calling strategy. Variants were called using Mutect2 and VarDict, filtered separately before combining to make a final list. Further details are described in Materials and Methods. **(b)** Number and types of variants called by Mutect2 and VarDict following filtering. **(c)** Overlap between variants called by Mutect2 and VarDict.
COSMIC: Catalogue of Somatic Mutations in Cancer, CH: Clonal haematopoiesis, 1000G: 1000 Genomes Project, VAF: Variant allele frequency, VUS: Variant of unknown significance, SNP: Single nucleotide polymorphism, SNV: Single nucleotide variant.

caller were analysed separately and the union (variants called by either caller) were retained (Figure 3.5a). Of the final filtered variants, more were called by VarDict than Mutect2, and 45% of variants were called by both (Figure 3.5b-c). Finally, variants were annotated as pathogenic driver mutations or variants of unknown significance (VUS) based on their predicted functional consequence and knowledge from the literature (Section 2.3.4). Following filtering, 162 variants were retained, of which 151 (93%) were non-synonymous and 133 (82%) were classified as pathogenic (Table A.3). Only pathogenic mutations were included in the analyses below.

3.2.3 Genetic landscape of cases with clonal haematopoiesis

Of the 195 individuals in the cohort, 85 (43.5%) were found to have CH with at least one pathogenic driver mutation. In a further 10 individuals, a single VUS was

identified and these cases were not considered to have CH (Figure 3.6). The most commonly mutated genes were *DNMT3A* (19% of cases), *TET2* (11% of cases) and *ASXL1* (7% of cases), in concordance with previously published studies of CH in the general population. Amongst SNVs, the most common base-pair change was a C→T transition, which is a somatic mutational signature of ageing that occurs due to spontaneous deamination of 5-methylcytosine [54, 55]. While the majority of individuals with CH had a single driver mutation, 34 (40%) had two or more mutations. This was a larger proportion than that observed in previous studies; for example in Desai *et al.* [58] and Abelson *et al.* [59], 11% and 14% of CH cases had two or more driver mutations respectively. The median VAF was 0.022 (mean 0.043, range 0.010–0.365), as expected from the sensitivity of the sequencing method.

As in previous studies, the prevalence of CH was strongly age-dependent. CH was rare in individuals under 60 years of age but was found in 75% of individuals over the age of 80 (Figure 3.7a). When compared to the control cohort reported by Abelson et al., 2018 which used a comparable sequencing method, the prevalence of CH in the MARCH cohort was found to be significantly higher in individuals over the age of 70 ($P < 0.001$, Fisher's exact test). The proportion of cases with multiple mutations increased with age and was also significantly higher than in the Abelson cohort (Figure 3.7b, $P < 0.001$, Fisher's exact test). Within cases with CH, there was no association between the size of the largest clone (as measured by the VAF) and age, and the two studies were comparable (Figure 3.7c). Clonal complexity was greater in older individuals, and cases with 4 or 5 mutations were exclusively observed in individuals over the age of 80. Interestingly, while *DNMT3A* was the most commonly mutated gene, the increase in CH at older age groups was driven by genes other than *DNMT3A*. The prevalence of mutations in *DNMT3A* did not increase with age, whereas those in *TET2*, *ASXL1* and other genes showed an increase with age (Figure 3.7e).

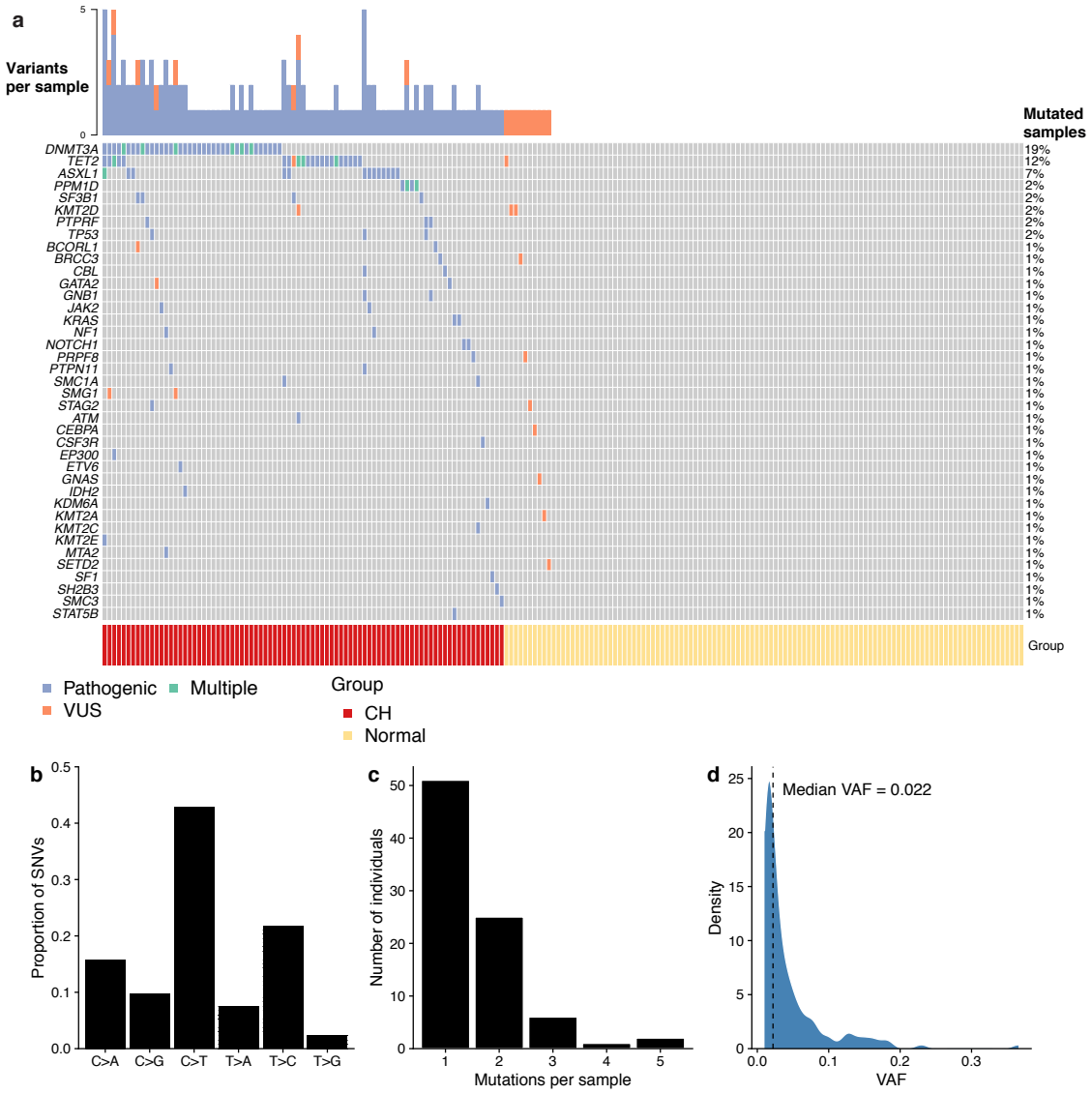


Figure 3.6: Landscape of somatic mutations in the MARCH cohort. **(a)** Spectrum of somatic variants observed in the cohort. Each row represents a gene, and each column corresponds to a participant in the study. Top bar plot indicates the number of mutations per sample. Variants are color-coded and classified as pathogenic or variants of unknown significance (VUS) according to criteria specified in Methods. Samples with >1 variant in any one gene are classified as Multiple. The bottom panel shows whether each sample was categorised as a case of clonal haematopoiesis (CH; red) or normal control (yellow). **(b)** Distribution of the types of single-nucleotide substitutions observed in somatic variants of the cohort. **(c)** Number of individuals with 1, 2, 3, 4, or 5 driver mutations detected. **(d)** Distribution of variant allele frequencies (VAF) observed across the cohort.

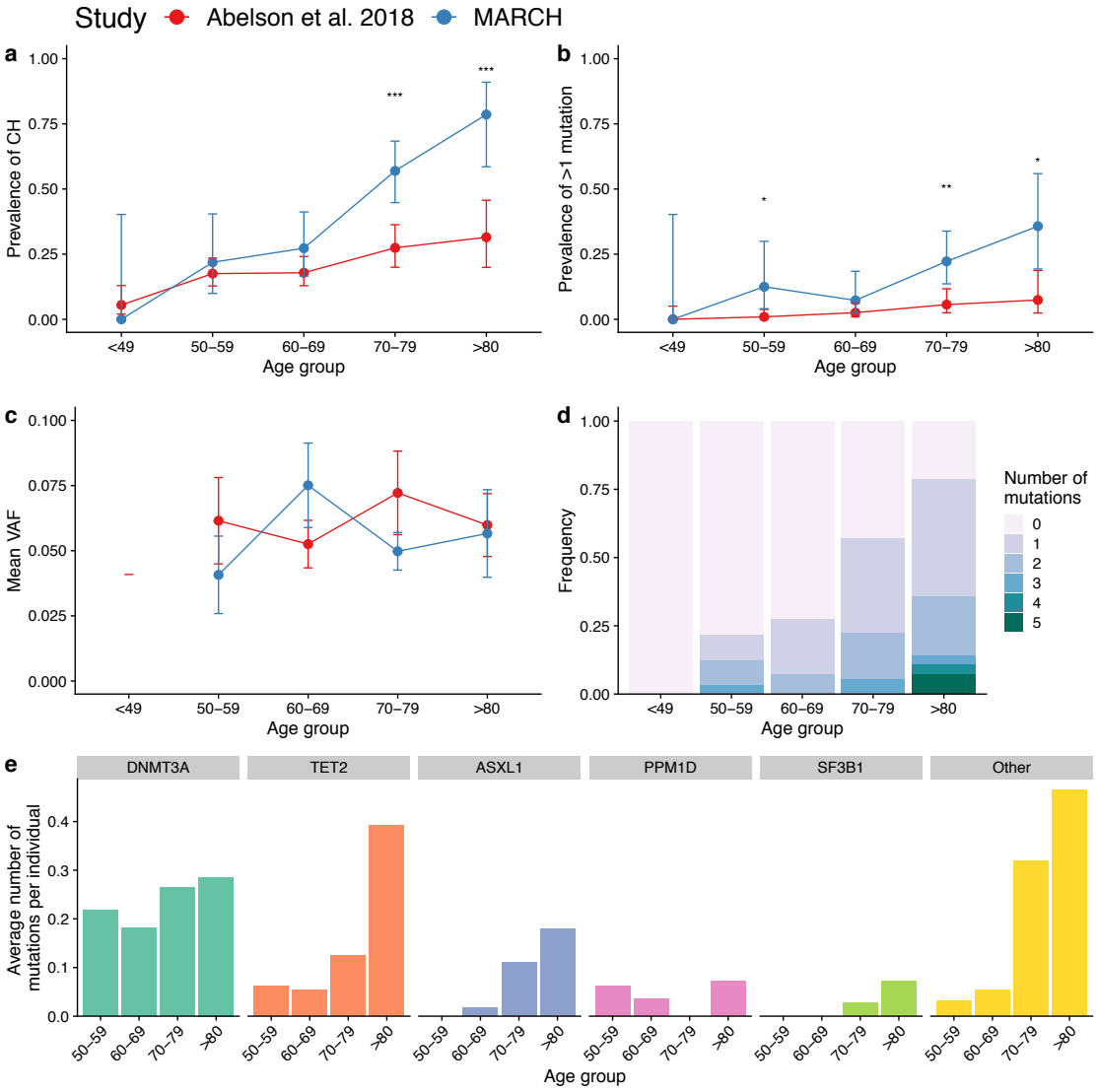


Figure 3.7: Prevalence of mutations with age. (a-c) Comparison of the MARCH cohort (blue) and a cohort of 676 individuals reported by Abelson et al., 2018 [59] (red). **(a)** Incidence of CH with at least one driver mutation (VAF $\geq$ 0.01) with age. **(b)** Incidence of CH with more than one driver mutation (VAF $\geq$ 0.01) with age. Error bars represent 95% confidence intervals. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, two-sided Fisher's exact test with Bonferroni multiple testing correction. **(c)** Mean and standard error of the variant allele frequency (VAF) measured for the largest CH clone with age. Non-significant, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. **(d)** Prevalence of the number of mutations per individual by age. **(e)** Prevalence of mutations with age broken down by gene.

3.2.4 Association of clonal haematopoiesis with clinical parameters

Clonal haematopoiesis has previously been associated with increased prevalence of cardiovascular disease, type 2 diabetes, smoking, chronic obstructive pulmonary disease (COPD) and autoimmune disease.

The clinical characteristics of cases with and without CHIP are shown in Table 3.1. CH cases were significantly more likely to have autoimmune disease (Fisher's

	Number of individuals (%)				
Measure	**Whole cohort**	**Controls without CH**	**Individuals with CH (VAF ≥ 0.01)**	***P***	***P-adj***
All	195 (100)	110 (56.4	85 (43.6)		
Female	108 (55.4)	59 (53.6)	49 (57.6)	0.66	0.66
Male	87 (44.6)	51 (46.4)	36 (42.4)		
Age *median (range)*	70.1 (30-92)	65.4 (30-91)	75.6 (51-92)	$1.5x10^{-10}$	$3.0x10^{-10}$
BMI *median (range)*	28.6 (15.6-65)	28.8 (15.7-46.8	28.3 (15.6-65)	0.55	0.55
Cardiovascular disease	85 (43.6)	43 (39.1)	42 (49.4)	0.19	0.80
Hypertension	72 (36.9)	37 (33.6)	35 (41.2)		
Ischaemic heart disease	13 (6.7)	8 (7.3)	5 (5.9)		
Stroke/TIA	12 (6.2)	6 (5.5)	6 (7.1)		
Cardiac arrhythmia	11 (5.6)	6 (5.5)	5 (5.9)		
Valvular heart disease	6 (3.1)	4 (3.6)	2 (2.4)	0.70	0.91
Type 2 diabetes	17 (8.7)	8 (7.3)	9 (10.6)	0.45	0.82
Hypercholesterolaemia	13 (6.7)	5 (4.5)	8 (9.4)	0.25	0.80
Autoimmune disease	33 (16.9)	12 (10.9)	21 (24.7)	0.013	0.16
Hypothyroidism	18 (9.2)	8 (7.3)	10 (11.8)		
Inflammatory bowel disease	4 (2.1)	2 (1.8)	2 (2.4)		
Other autoimmune disease	12 (6.2)	2 (1.8)	10 (11.8)		
Previous VTE	6 (3.1)	2 (1.8)	4 (4.7)	0.41	0.82
Previous cancer	25 (12.8)	14 (12.7)	11 (12.9)	1.00	1.00
Asthma	22 (11.3)	14 (12.7)	8 (9.4)	0.50	0.82
COPD	9 (4.6)	3 (2.7)	6 (7.1)	0.18	0.80
Chronic kidney disease	9 (4.6)	4 (3.6)	5 (5.9)	0.51	0.82
Previous chemotherapy	3 (1.5)	1 (0.9)	2 (2.4)	0.58	0.84
Previous radiotherapy	9 (4.6)	5 (4.5)	4 (4.7)	1.00	1.00
Current or ex-smoker	101 (51.8)	56 (50.9)	45 (52.9)	0.89	1.00
Medications					
NSAID	51 (26.2)	30 (27.3)	21 (24.7)		
Immunosuppressants	5 (2.6)	3 (2.7)	2 (2.4)		

Table 3.1: Summary of clinical characteristics. Baseline characteristics of samples with and without clonal haematopoiesis (CH) in the MARCH cohort. Summary statistics are shown. Comparison of proportions was performed using Fisher's exact test; comparison of medians was performed with the Wilcoxon rank-sum test; *P*: significance level; *P-adj*: with Benjamini–Hochberg multiple testing correction. TIA: Transient ischaemic attack. COPD: Chronic obstructive pulmonary disease, VTE: Venous thromboembolism, NSAID: Non-steroidal anti-inflammatory drugs.

exact test), which was also identified in a recent study of CH in THR patients [94]. Hypothyroidism and polymyalgia rheumatica were the most common conditions, affecting 10 and 4 of the 21 CH cases with autoimmune disease respectively.

Cardiovascular disease and chronic obstructive pulmonary disease (COPD) was also more common amongst cases with CH, but these did not reach statistical significance.

Data from the most recent full blood count prior to surgery were collected. The blood-cell indices of cases with and without CHIP are shown in Table 3.2. Individuals with CH had lower haemoglobin levels ($P = 0.014$, P-adjusted $= 0.056$, Wilcoxon rank sum test), as has been reported in previous studies (Figure 3.8). Other blood count indices did not show significant differences.

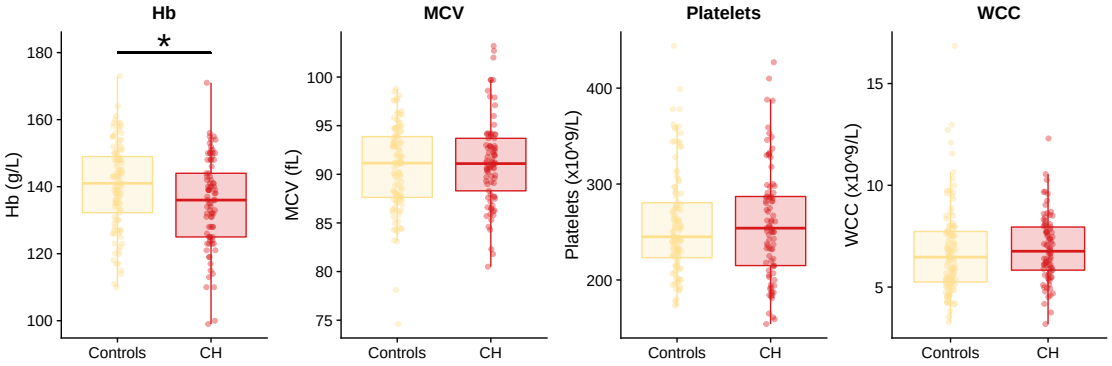


Figure 3.8: Blood indices in CH cases and controls. Haemoglobin (Hb), Mean corpuscular volume (MCV), White cell count (WCC) and Platelet count in individuals with and without CH mutations. For Hb: $P = 0.014$, $P\text{-}adj = 0.056$ (Wilcoxon rank sum test with Benjamini–Hochberg multiple testing correction).

	Median (range)				
Measure	**Whole cohort**	**Controls without CH**	**Individuals with CH (VAF ≥ 1%)**	***P***	***P-adj***
Hb (g/L)	139 (99-173)	141 (110-173)	136 (99-171)	0.014	0.056
MCV (fL)	91.1 (74.6-103.2)	91.2 (74.6-98.8)	91.1 (80.5-103.2)	0.70	0.88
Platelets (x10^9/L)	250 (154-444)	245 (173-444)	254 (154-427)	0.88	0.88
WCC (x10^9/L)	6.6 (3.2-16.9)	6.5 (3.3-16.9)	6.8 (3.2-12.3)	0.21	0.42
Neutrophils (x10^9/L)	4.2 (1.3-13.6)	3.9 (1.3-13.6)	4.4 (2.8-9.4)		
Lymphocytes (x10^9/L)	1.7 (0.5-3.7)	1.7 (0.8-3.7)	1.6 (0.51-3.05)		
Monocytes (x10^9/L)	0.5 (0.2-1.8)	0.5 (0.2-1.8)	0.5 (0.2-1.0)		
Eosinophils (x10^9/L)	0.1 (0-1.1)	0.1 (0-0.8)	0.1 (0-1.1)		
Basophils (x10^9/L)	0.04 (0.01-0.27)	0.05 (0.02-0.13)	0.04 (0.01-0.27)		

Table 3.2: Summary of blood cell parameters in the cohort. Blood cell parameters of individuals with and without clonal haematopoiesis in the MARCH cohort. Measurements are from the most recent blood count prior to surgery. Comparison of medians was performed with the Wilcoxon rank-sum test; P: significance level; $P\text{-}adj$: with Benjamini–Hochberg multiple testing correction.
Hb: Haemoglobin, MCV: Mean corpuscular volume, WCC: White cell count.

3.2.5 Molecular characteristics of somatic variants

As in previous studies of clonal haematopoiesis, most mutations were detected in *DNMT3A*, *TET2* and *ASXL1*. For *DNMT3A*, approximately one third of mutations were predicted to be truncating and the remainder were missense or inframe indels clustered within the structured domains of the protein (Figures 3.9 and 3.10). Only one mutation in the arginine 882 (R882) residue was detected, in contrast to studies of AML where this hotspot constitutes about 60% of mutations. This is consistent with previous studies of CH in non-cancer cohorts [24, 25, 58, 59].

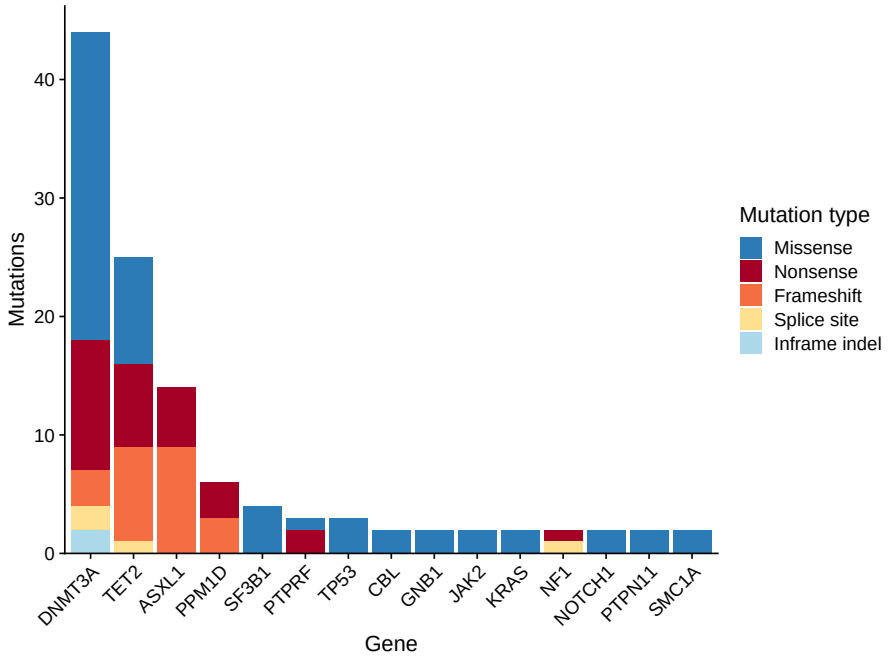


Figure 3.9: Mutations detected by gene. Number of mutations detected in each gene, coloured by the predicted effect on the protein.

Mutations in *TET2* were dispersed across the coding region and predicted to be loss-of-function mutations. Two-thirds of mutations were nonsense or frameshift mutations, and the remainder were missense mutations occurring in the cysteine-rich or double-stranded β-helix domains.

All mutations in *ASXL1* were nonsense or frameshift mutations occurring in the last two exons, predicted to generate a truncated form of the protein lacking the PHD domain. This is consistent with previous studies of CH and AML. The most

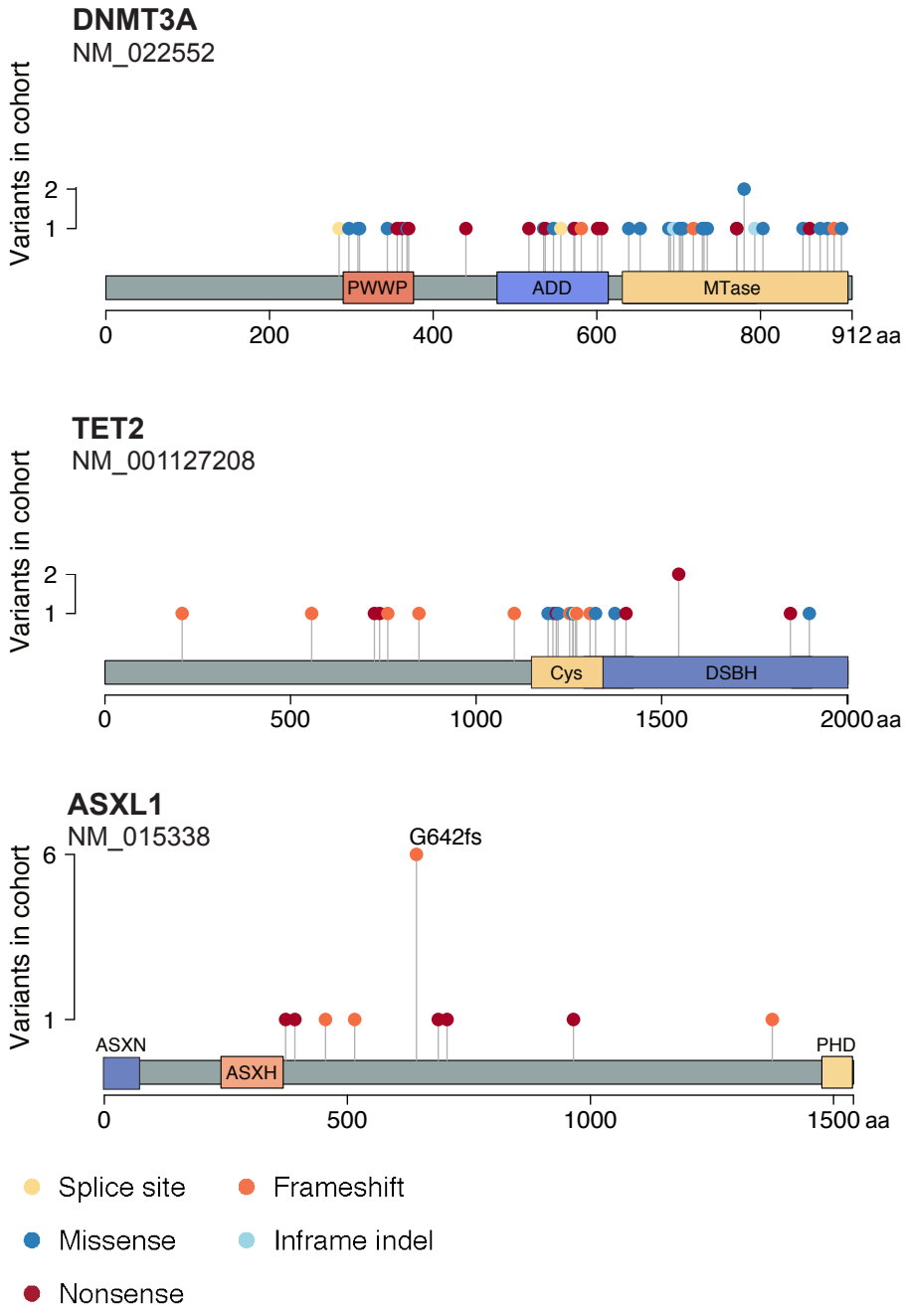


Figure 3.10: Distribution of mutations within common genes. Schematic of the position of coding variants in coding regions of *DNMT3A*, *TET2* and *ASXL1*. Positions of protein domains are shown: PWWP, proline-tryptophan-tryptophan-proline; ADD, ATRX-DNMT3-DNMT3L; MTase, cytosine methyltransferase; Cys, cysteine-rich; DSBH, double-stranded β-helix; ASXH, Additional sex combs homology; PHD, plant homeodomain. Mutations are classified as described above.

common mutation identified was a frameshift insertion at glycine 642 (G642fs), which made up 6 of the 14 (43%) ASXL1 mutations identified. This falls in a homopolymer repeat of eight guanine nucleotides in exon 12.

There was no significant difference in the VAFs observed between the commonly mutated genes (Figure 3.11, Kruskal-Wallis test). Similarly, there was no significant difference in the VAFs of truncating and missense mutations in *DNMT3A* and *TET2* (Wilcoxon rank-sum test).

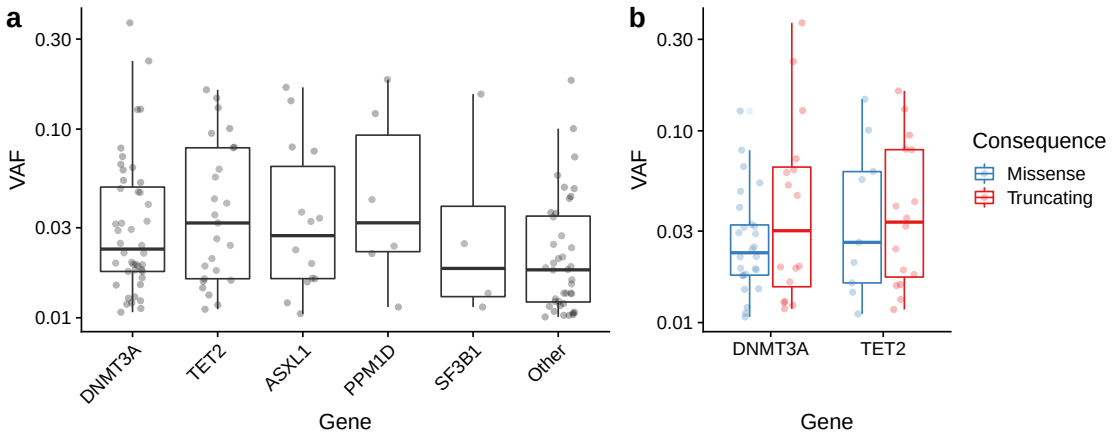


Figure 3.11: Variant allele frequencies in commonly mutated genes. **(a)** Distribution of VAFs observed across the cohort, for the 5 most commonly mutated genes. Distribution of VAFs in all the other genes are binned as "Other". **(b)** Distribution of VAFs for *DNMT3A* and *TET2* comparing missense and truncating variants.

3.3 Comparison of somatic variants between BM and PB

The majority of studies of CH to date have analysed sequencing performed on peripheral blood; only two studies have analysed BM. One of these analysed sequencing of BM from 109 individuals and PB from a separate 91 individuals undergoing THR with no paired samples [94]. The other study analysed paired BM and PB samples from just 17 individuals undergoing THR [260]. To understand how the frequency of somatic variants differed between bone marrow and peripheral blood, re-sequencing of peripheral blood granulocytic DNA was performed on 99 samples (79 with non-synonymous variants and 20 controls). Variant calling was performed

(using a VAF threshold of 0.005), and where variants present in BM were not detected by variant calling, allelic frequencies were counted manually using mpileup.

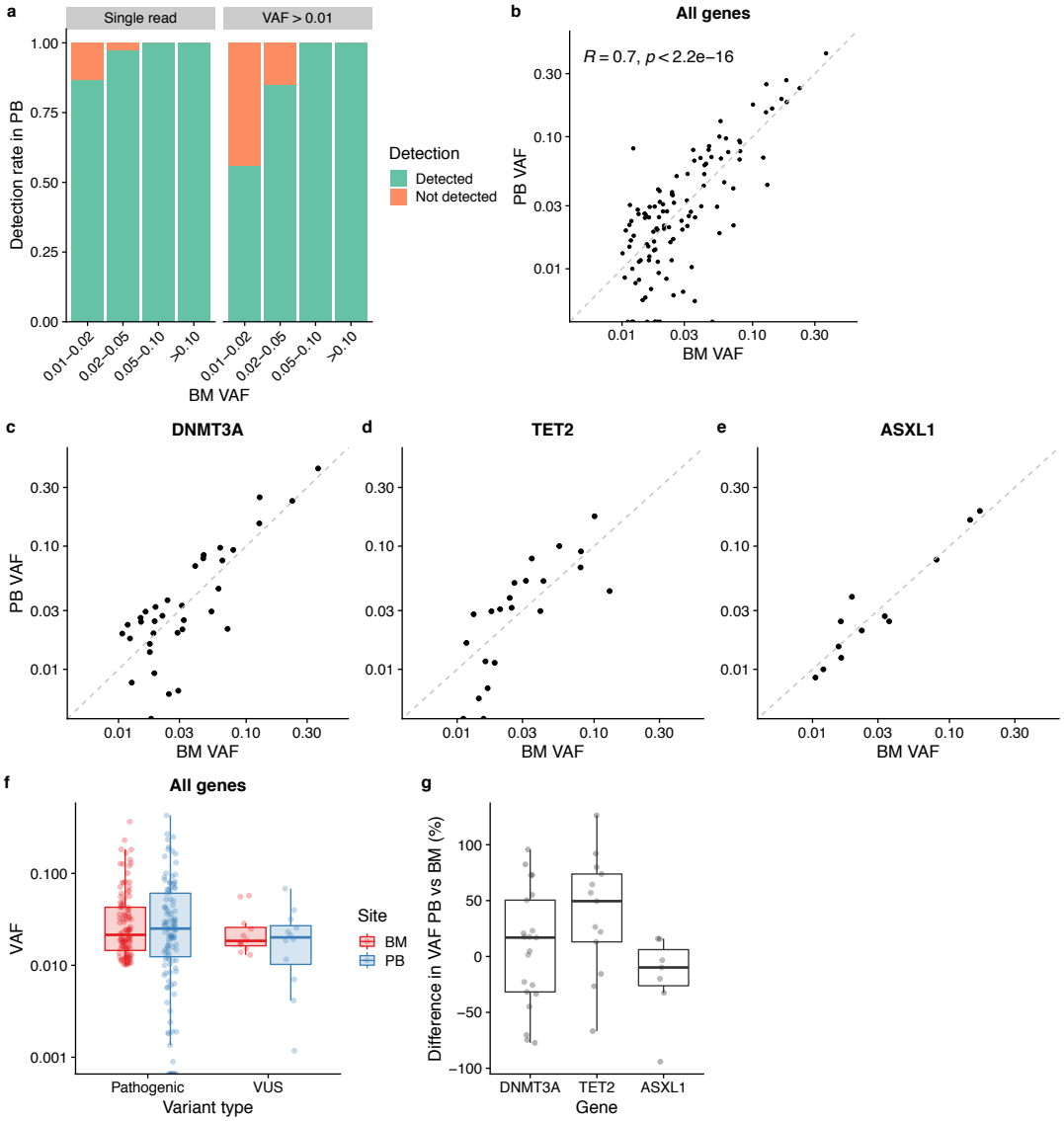


Figure 3.12: Comparison of variants detected in BM and PB sequencing. (a) Detection rate of variants in PB, using at least a single read (left), or a VAF of 0.01 (right) as cutoff. Variants are binned by their VAF measured in BM. Plots **(b-e)** show the correlation of VAFs between BM and PB sequencing for variants in all genes **(b)** and the genes indicated. The dotted line indicates the line of equality where BM VAF = PB VAF. R indicates Spearman correlation coefficient. **(f)** Distribution of VAFs for pathogenic and variants of unknown significance (VUS). Differences are not statistically significant (paired Wilcoxon signed rank test). **(g)** Percentage difference of VAFs between PB and BM for variants in the genes indicated. Positive differences indicate a higher VAF in PB than BM. Variants with VAF < 0.02 in BM were excluded due to the high uncertainty around the VAF measurement at this level. Differences are not statistically significant (paired Wilcoxon signed rank test).

Of 126 variants present in BM in this sample set, 117 (93%) were detected in PB sequencing with at least 1 variant read, and 94 (75%) were detected with a VAF of at least 0.01 (Figure 3.12a). Most of the variants not detected in PB had a VAF below 0.02 in BM but the largest had a VAF of 0.049 (in *KMT2E*). The VAFs of variants across all genes showed a high correlation between BM and PB, and this was also the case in the most commonly mutated genes (Figure 3.12b-e). Overall, there was no significant difference in VAFs between BM and PB. This held true for both variants classified as pathogenic and variants of unknown significance (VUS) (Figure 3.12f). However, in individual cases the VAF in PB could be much higher or lower than in BM (Figure 3.12f-g). Mutations in *TET2* tended towards higher VAFs in PB but this was not statistically significant (paired Wilcoxon signed rank test).

3.4 Bone marrow immunophenotypic analysis

It has previously been shown that the frequency of immunophenotypic haematopoietic stem and progenitor cells (HSPC) changes with age, and in particular an increase in immunophenotypic haematopoietic stem cells (HSC) has been observed [261, 262]. In myeloid malignancies, substantial expansion of myeloid progenitor populations is often seen, and in AML the bone marrow becomes dominated by leukaemic progenitor or precursor-like populations [49, 50]. It is not known whether the presence of CH leads to similar perturbations in the bone marrow composition. Multi-parameter flow cytometric analysis of the lineage negative (Lin–) HSPC compartment in bone marrow samples from cases with CH and controls was performed to compare the frequencies of the classically defined HSPC populations. 27 samples with CH and 20 controls were chosen. The age of control samples ranged from 30 to 89 years of age and were grouped into young (<45), middle-aged (45–65) and old (>65) individuals.

A representative immunophenotypic profile and the gating used to define HSPC populations is shown in Figure 3.13. Table 3.3 summarises the frequency of cells in the populations analysed and compares them to previous data from our laboratory [242]. As shown, Lin–CD34+ HSPCs typically constitute approximately 2% of BM

mononuclear cells, while Lin–CD34–CD117+ precursors constitute approximately 1%. Overall, the size of HSPC populations were consistent with previous results from our laboratory.

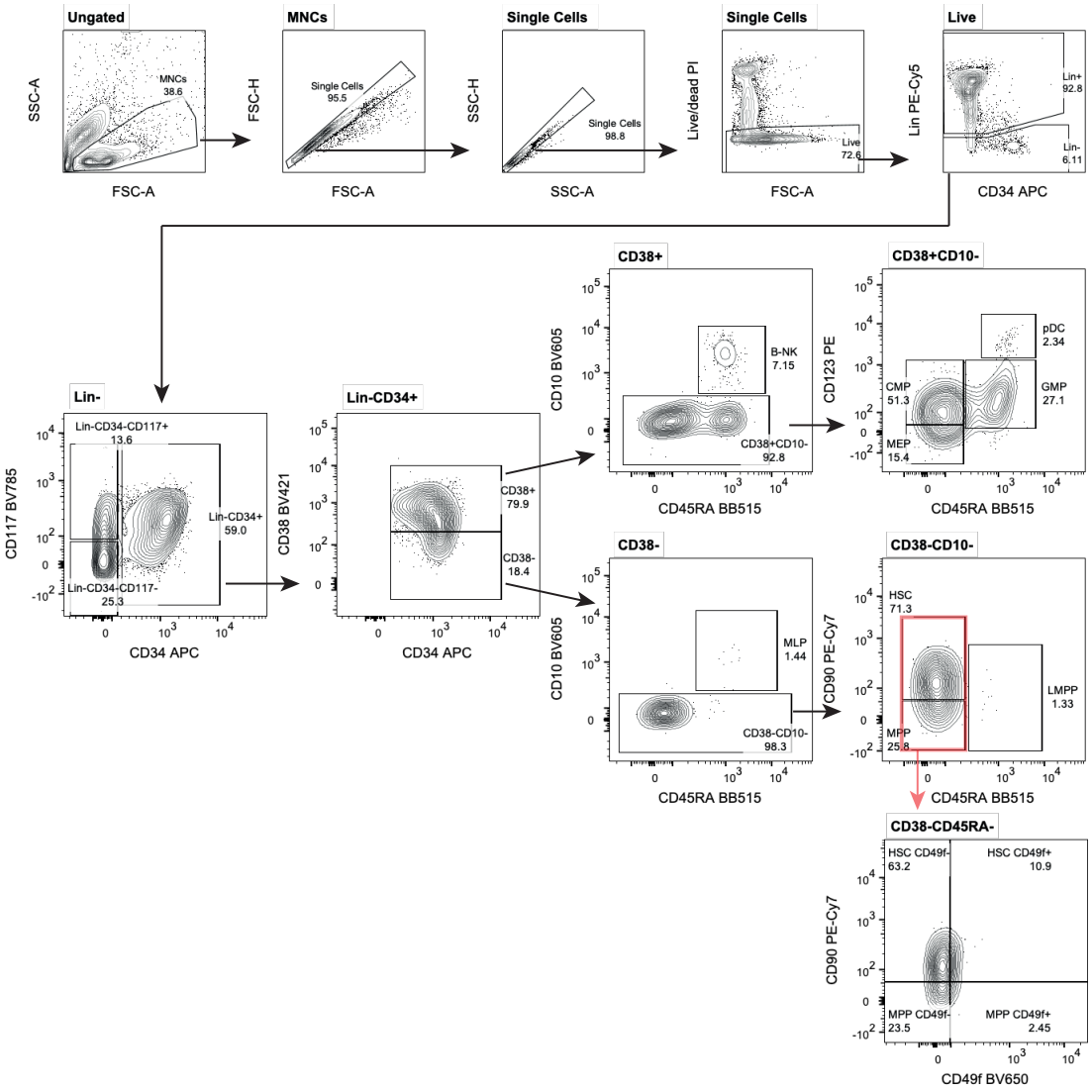


Figure 3.13: Gating strategy for immunophenotypic analysis. Representative flow cytometry plots for a normal BM sample showing the gating used to analyse HSPC population frequencies.

3.4.1 Effect of age on HSPC frequencies

Initially, the effect of age on the frequency of HSPC and precursor frequencies in control samples was analysed. With increasing age, there was an increase in the frequency of CD34–CD117+ precursor cells as a fraction of BM MNCs ($P =$

Population	Immunophenotype	% of Lin-CD34+ (Mean ± SD)	
		Karamitros et al. 2018	Controls; this study
Samples analysed		**7**	**20**
Stem and early progenitors	Lin-CD34+CD38-	16.1 ± 10.3	18.2 ± 7.4
Stem and early progenitors	Lin-CD34+CD38- CD10-	15.8 ± 10.4	17.7 ± 7.3
HSC	Lin-CD34+CD38-CD10-CD90+CD45RA-	11.4 ± 8.5	11.2 ± 4.9
MPP	Lin-CD34+CD38-CD10-CD90-CD45RA-	3.7 ± 3.4	5.7 ± 3.2
LMPP	Lin-CD34+CD38-CD10-CD90-/loCD45RA+	0.2 ± 0.2	0.4 ± 0.2
MLP	Lin-CD34+CD38-CD10+CD90-/loCD45RA+	0.2 ± 0.1	0.2 ± 0.2
Late Progenitors	Lin-CD34+CD38+	83.9 ± 10.3	80.3 ± 7.6
Late Progenitors	Lin-CD34+CD38+CD10-	78.4 ± 8.5	75.3 ± 7
CMP	Lin-CD34+CD38+CD10-CD123midCD45RA-	34.7 ± 6.9	39.3 ± 5.4
GMP	Lin-CD34+CD38+CD10-CD123midCD45RA+	17.8 ± 6.1	17.8 ± 5.1
MEP	Lin-CD34+CD38+CD10-CD123-CD45RA-	14.6 ± 4.9	13.7 ± 4.3
B-NK	Lin-CD34+CD38+CD10+CD45RA+	5.1 ± 7.5	4.6 ± 3.4
		% of BM BMNC (Mean ± SD)	
HSPC	Lin-CD34+	-	2.4 ± 1.1
Precursor	Lin-CD34+CD117+	-	1.1 ± 0.6

Table 3.3: Frequency of HSPC populations in control BM samples. Top section (white) shows the immunophenotypic definition and frequency of HSPC populations as a percentage of BM Lin–CD34+ cells. Data from 20 control samples in the MARCH cohort is compared to previous analysis on 7 adult BM samples from our laboratory published in Karamitros *et al.* [242]. Bottom section (grey) shows the frequency of HSPC and precursor cells as a frequency of total BM MNCs.
HSC: Haematopoietic stem cell, MPP: Multipotent progenitor, LMPP: Lymphoid-primed multipotent progenitor, MLP: Multi-lymphoid progenitor, CMP: Common myeloid progenitor, GMP: Granulocyte-monocyte progenitor, MEP: Megakaryocyte-erythroid progenitor, B-NK: B-natural killer cell progenitor.

0.05, Pearson correlation) but there was no change in total Lin–CD34+ HSPCs (Figure 3.14a-b). Within HSPCs, no significant differences were seen with age (Figure 3.14c-d). Although there was a trend towards higher frequency of HSC (particularly the CD49f+ subset) with age, this was not statistically significant (Figure 3.14e-f), unlike in one previous report [261].

3.4.2 Effect of CH on HSPC immunophenotypes

Next, the effect of CH on HSPC frequencies was analysed. Of the 27 samples with CH analysed, 16 had a single mutation and 11 had more than one. CH samples were grouped according to the gene mutated, and samples with more than one mutation were categorised according to the gene that had the highest VAF, as this is likely to have the greatest impact on the immunophenotype. Of the 27 CH samples, 14 had a *DNMT3A* mutation, 10 had a *TET2* mutation and 3 had an *ASXL1* mutation in

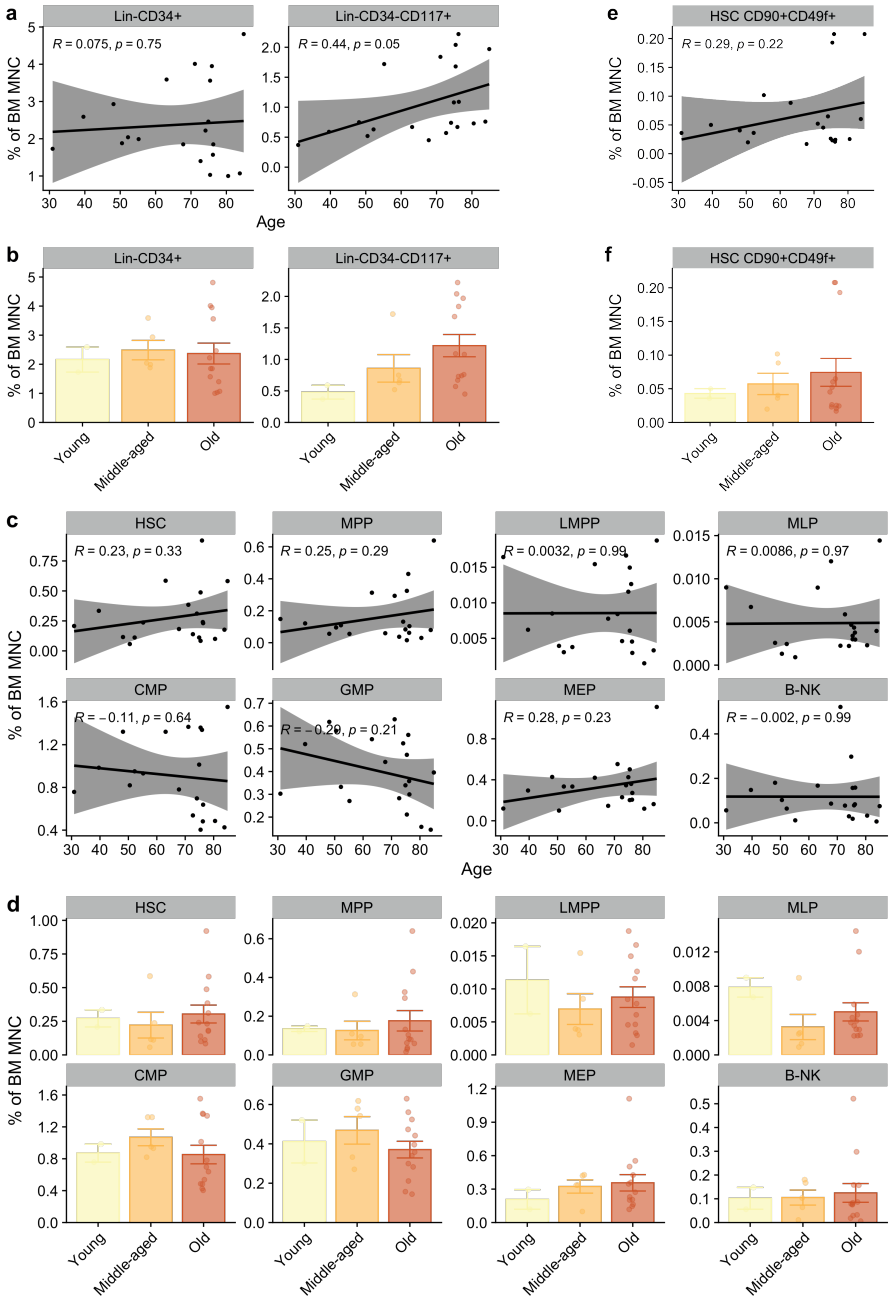


Figure 3.14: Effect of age on HSPC frequencies. (a-b) Frequency of Lin–CD34+ HSPC and Lin–CD34–CD117+ precursor cells within BM MNCs with respect to age. **(c-d)** Frequency of HSPC subpopulations as defined in Table 3.3 within Lin–CD34+ HSPC with respect to age. **(e-f)** Frequency of CD90+CD49f+ HSC within Lin–CD34+ HSPC with respect to age. For dotplots, the linear regression line is shown with the 95% confidence interval. R indicates the Pearson correlation coefficient. For all barplots, samples were grouped into young (<45), middle-aged (45–65) and old (>65) individuals. Bars show the mean frequency and error bars represent the standard error of the mean. Each point represents a sample.

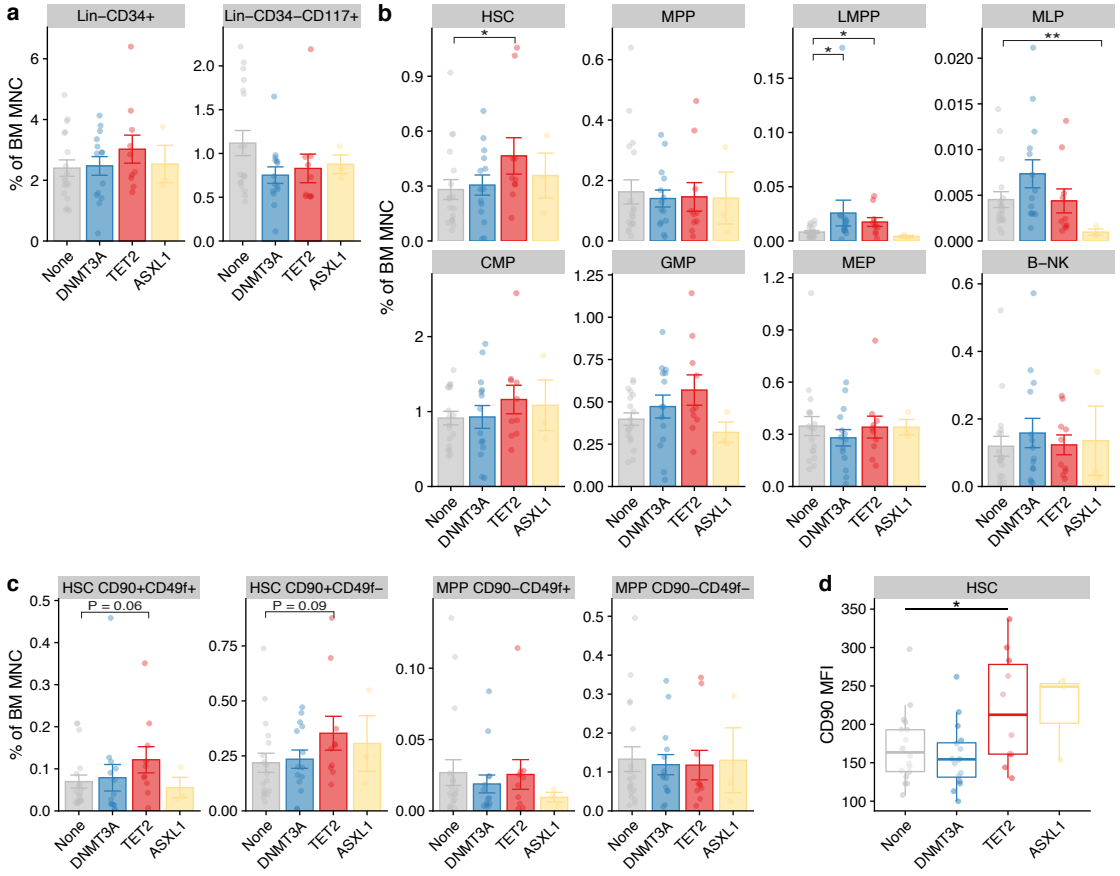


Figure 3.15: Effect of CH on HSPC frequencies. (a) Frequency of Lin–CD34+ HSPC and Lin–CD34–CD117+ precursor cells within BM MNCs with respect to mutation. **(b)** Frequency of HSPC subpopulations as defined in Table 3.3 within Lin–CD34+ HSPC with respect to mutation. **(c)** Frequency of CD90+/–CD49f+/– HSC and MPP within Lin–CD34+ HSPC with respect to mutation. **(d)** Expression of CD90 within the HSC population with respect to mutation. For all plots, CH samples were categorised according to the mutation present in the largest clone (n = 14 *DNMT3A*-mutant, 10 *TET2*-mutant, 3 *ASXL1*-mutant. Bars show the mean frequency and error bars represent the standard error of the mean. Each point represents a sample. * $P < 0.05$, ** $P < 0.01$, Wilcoxon rank-sum test. MFI: Median fluorescence intensity.

the largest clone. The median VAF was 0.076 (equivalent to a clonal fraction of 15.2%). For the control samples, those from young donors were excluded to obtain an age-matched set (median age of controls 74 vs. 72 in CH cases).

There was no significant difference in total Lin–CD34+ HSPC or precursors between cases of CH and controls (Figure 3.15a). Within the Lin–CD34+ HSPC, both *DNMT3A*-mutant and *TET2*-mutant samples had a relative increase in the rare lymphoid-primed multipotent progenitor (LMPP) population, and *ASXL1*-mutant samples had a decrease in the multi-lymphoid progenitor (MLP) (Figure 3.15b).

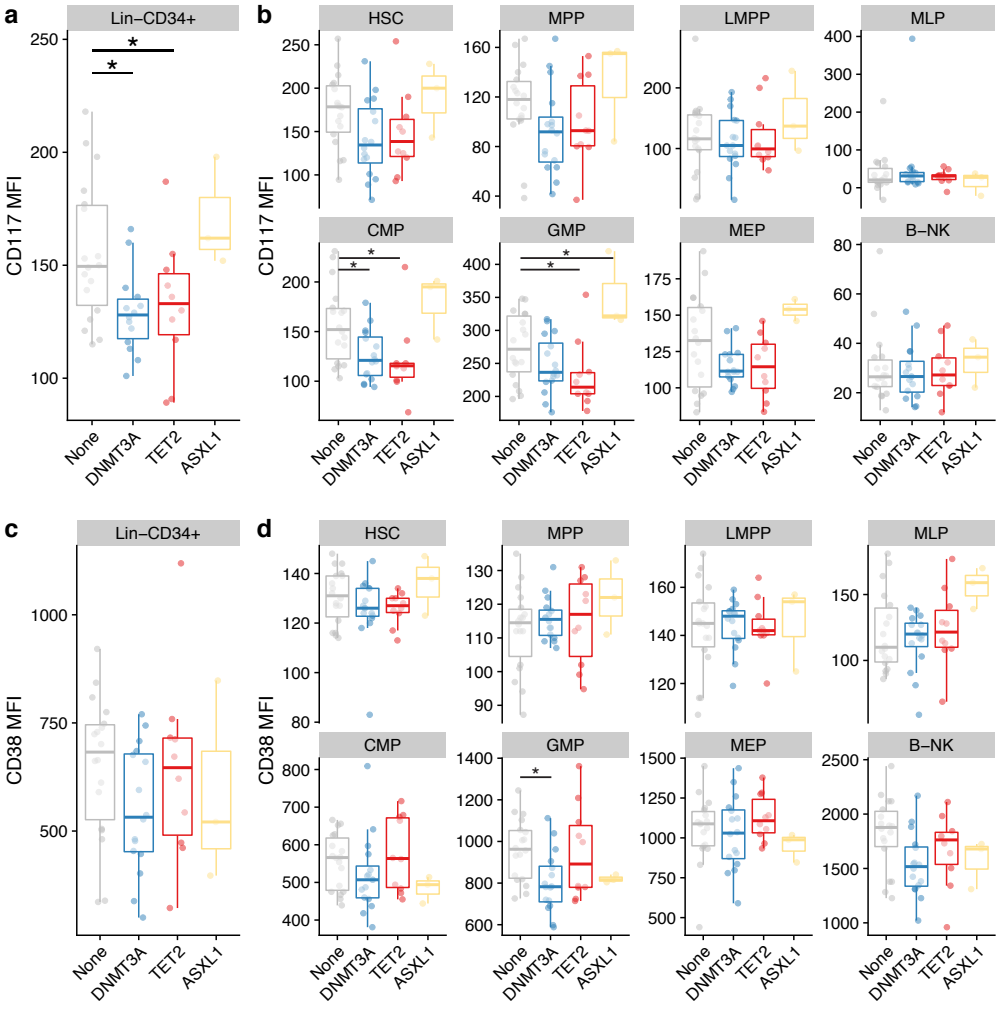


Figure 3.16: Effect of CH on CD117 and CD38 expression. (a-b) Expression of CD117 (c-Kit) in **(a)** total Lin–CD34+ HSPC and **(b)** HSPC subpopulations with respect to mutation. Tests for the effect of mutations on CD117 expression across HSPC populations were performed using a two-way ANOVA with Tukey's multiple comparison; P for *DNMT3A*-mutant vs controls = 0.018; *TET2*-mutant vs controls = 0.012; *ASXL1*-mutant vs controls = 0.11. **(c-d)** Expression of CD38 in **(c)** total Lin–CD34+ HSPC and **(d)** HSPC subpopulations with respect to mutation. Tests for the effect of mutations on CD38 expression across HSPC populations were performed using a two-way ANOVA with Tukey's multiple comparison; P for *DNMT3A*-mutant vs controls = 0.003; *TET2*-mutant vs controls = 0.98; *ASXL1*-mutant vs controls = 0.22. * $P < 0.05$ by pairwise t tests with Benjamini–Hochberg multiple testing correction.

TET2-mutant samples also had a 68% higher frequency of HSC. Interestingly, expression of the stem cell marker, CD90, was also higher within HSCs from *TET2*-mutant samples than controls (Figure 3.15d).

Haematopoietic expression of CD117 (c-Kit) and CD38 have been used as markers

of immaturity and multipotency. To examine whether the presence of somatic mutations in CH samples altered the expression profile of these markers, the MFI across HSPC populations was analysed. CD117 is normally expressed in early stem and progenitor cells, as well as late myeloid progenitors and precursors. Expression of CD117 was lower in total Lin–CD34+ cells and across HSPC populations in *DNMT3A*- and *TET2*-mutant samples than controls (Figure 3.16a-b). This was particularly pronounced in CMP and granulocyte-monocyte progenitors (GMP).

CD38 distinguishes early (low expression) and late (high expression) progenitors. There was no significant difference in expression of CD38 in total Lin–CD34+ cells or in CD38– cells between CH samples and controls. However expression was lower within CD38+ CMP, GMP and B-natural killer cell progenitors (B-NK), in samples with *DNMT3A* mutations (Figure 3.16b-d).

3.5 Conclusions: Collection and characterisation of a clonal haematopoiesis cohort

To study clonal haematopoiesis in humans, a source of haematopoietic stem and progenitor cells was required from individuals with CH. Over a two year period, 195 blood and bone marrow samples were collected from individuals undergoing total hip replacement surgery who had no history of haematological disease.

Bone marrow DNA was sequenced, identifying 85 cases of clonal haematopoiesis. The profile of mutations identified was similar to that seen in previous studies of healthy individuals, with a predominance of C→T transitions and a similar spectrum of genes affected. However, the overall prevalence of CH was significantly higher than in previous studies, and more cases had multiple somatic mutations. The most commonly affected genes were *DNMT3A*, *TET2* and *ASXL1*, all of which were characterised by loss-of-function mutations. Comparison of variant allele frequencies between bone marrow and peripheral blood showed that there are sometimes substantial differences in clone size between the two, but the mean clone size was similar.

Finally, immunophenotyping of bone marrow CD34+ stem and progenitor cells showed that frequencies of the classically-defined stem and progenitor populations are normal in most CH samples as compared to age-matched controls. Subtle gene-specific aberrations were identified, including an association between *TET2*-mutant clonal haematopoiesis and increased frequency of CD90+ HSCs.

To understand the effects of somatic mutations on haematopoietic stem and progenitor differentiation, further cellular and molecular analysis of mutant clones was needed. To do this, it would be necessary to use methods that enable comparisons of WT and mutant cells from the same sample.

Chapter 4

Development of TARGET-seq+

Contents

4.1 Introduction and rationale for the development of TARGET-seq+ 81
4.1.1 Current methods for single cell genotyping and transcriptome sequencing 81
4.1.2 Review of the TARGET-seq method 84
4.1.3 Review of Smart-seq3 chemistry 86
4.1.4 Aims in developing TARGET-seq+ 87
4.2 Modifications to the method and pilot experiments . . 87
4.3 Validation of TARGET-seq+ in cell lines and primary haematopoietic cells 91
4.3.1 Targeted genotyping with TARGET-seq+ 91
4.3.2 Transcriptome sequencing with TARGET-seq+ 93
4.4 Mutational calling in single-cell genotyping 98
4.5 Allelic dropout and accuracy of single-cell genotyping 101
4.6 Conclusions: Development of TARGET-seq+ 102

I am grateful to Bilyana Stoilova for her help with the experiments described in Chapter 4. The experimental work was done together and I performed the analysis.

4.1 Introduction and rationale for the development of TARGET-seq+

4.1.1 Current methods for single cell genotyping and transcriptome sequencing

Having established a cohort of non-malignant samples from individuals with clonal haematopoiesis, my next aim was to investigate the molecular consequences of somatic mutations in *DNMT3A* and *TET2* on haematopoietic stem and progenitor cells. Given that both enzymes regulate DNA methylation, it was felt likely that their effect on cell function is mediated through dysregulated gene expression.

A major obstacle to studying the consequences of somatic mutations in the setting

of clonal haematopoiesis is the absence of any marker to distinguish mutant cells from their wild-type (WT) counterparts, except for the mutation itself. Since mutant cells in most cases only make up a small fraction of a sample, studying bulk populations is unlikely to be an informative way to understand the consequences of mutations. Furthermore, the non-mutant cells that co-exist with a mutant clone might be phenotypically different from equivalent cells in normal individuals without CH. The cell-intrinsic effects of a mutation on mutant cells and the cell-extrinsic effects on non-mutant cells that co-exist within a tissue would be confounded by studying gene expression at the bulk level. A potentially powerful approach to overcome this problem, is to link mutation detection to gene expression analysis in single cells, enabling the comparison of wild-type and mutant cells from the same individual.

Two methods for combined single-cell genotyping and transcriptomic profiling have recently been developed. TARGET-seq combines targeted amplification of cDNA and genomic DNA, in parallel with whole-transcriptome single-cell RNA-seq (scRNA-seq) [226]. Based on the Smart-seq2 protocol for scRNA-seq [263], TARGET-seq is a plate-based method, where single cells are isolated by fluorescence-activated cell sorting (FACS) and polyA$^{(+)}$ RNA and genotyping targets are amplified together in the same well. Aliquots of the amplified material are then taken to construct separate transcriptome and genotyping libraries. By detecting mutational information in both cDNA and genomic DNA, it achieves a high sensitivity for mutational calling (>98% detection rate of 4 mutations). It also incorporates modified RT-PCR conditions that increase the sensitivity for transcript detection by about 25% over the Smart-seq2 technique. In primary human haematopoietic samples, an average of approximately 3500 genes were detected per cell. The original protocol was designed to obtain full-length transcriptome information, but a modified version was developed to increase the cell throughput by incorporating cell barcodes and pooling of RNA-seq libraries, generating 3'-biased libraries.

Genotyping of Transcriptomes (GoT) is a droplet-based method that combines the 10x Genomics platform for scRNA-seq with targeted amplification of a transcript locus of interest [264]. Single cells are isolated by mixing the sample with barcoded

gel beads and a partitioning oil to form single-cell emulsion droplets. Nam *et al.* modified this by spiking a gene-specific primer in to the RT or PCR steps to increase the amplification of transcripts containing mutations of interest. After the RT-PCR step, an aliquot of the cDNA library is taken for targeted genotyping. This technique achieved a 60–89% detection rate for a highly expressed gene such as *CALR*, but only obtained genotyping information for 7% of cells for *JAK2*, which is lowly-expressed. The sensitivity for gene expression profiling is typical of droplet based techniques, detecting approximately 2500 genes per cell in primary human haematopoietic samples. As the 10x scRNA-seq protocol is a 3' or 5'-biased method, detection of a mutation is dependent on the distance of the mutant locus from transcript ends. The authors developed a modification of their technique involving sequential rounds of circularisation and inverse PCR to remove the intervening sequence between the region of interest and the cell barcode which went some way to mitigate this, improving the detection of an *SF3B1* mutation from 9% to 24% of cells.

An ideal method for combined single-cell genotyping and RNA-sequencing would have a number of characteristics:

1. High sensitivity mutational detection with low allelic dropout to minimise mis-categorisation of mutant cells.

2. High sensitivity RNA detection to maximise phenotypic characterisation.

3. High cell throughput to enable analysis of rare populations and maximise power to detect subtle differences between distinct genotypes.

The TARGET-seq technique fulfils the first and partially fulfils the latter two of these criteria, whereas GoT fails the first two and fulfils the last. TARGET-seq achieves highly-accurate genotyping that enables mutant and WT cells to be confidently assigned, in contrast to GoT, where for some mutations a minority of cells are successfully genotyped. Genotyping in GoT is highly dependent on mutational position and the transcript expression level. For genes such as *DNMT3A* and *TET2* that are expressed at low levels and where mutations are found throughout the

length of the transcript, GoT is likely to achieve a low detection rate. This is particularly problematic when attempting to detect small clones or dissect clonal hierachies in the setting of multiple mutations. TARGET-seq also achieves a greater sensitivity than GoT for gene expression detection. GoT has an advantage in cell throughput, but the 3' modification of TARGET-seq makes profiling of medium-to-large cell numbers feasible.

Evidence to date suggests the phenotype resulting from mutations in *DNMT3A* and *TET2* is subtle and gene expression differences are few [166, 172, 183, 219]. To maximise the ability to distinguish phenotypic differences in gene expression between WT and mutant cells in CH samples, a highly sensitive method was required. For these reasons, it was decided to apply the TARGET-seq technique to the analysis of clonal haematopoiesis samples. While performing preliminary experiments, an optimised protocol for plate-based scRNA-seq was published. Developed by the authors of Smart-seq2, the new Smart-seq3 protocol provides greater sensitivity for gene expression profiling, typically detecting thousands more transcripts per cell [265]. This was achieved by iterative optimisation of the reverse transcription conditions to increase the efficiency of mRNA capture. Furthermore, the authors introduced a 5' unique molecular identifier (UMI) that enables transcripts to be assigned to their allelic origin and isoforms reconstructed *in silico*.

To determine if this would increase sensitivity, we decided to modify the TARGET-seq technique to incorporate the optimised chemistry from the Smart-seq3 protocol and, if successful, apply this to the analysis of clonal haematopoiesis samples. Principal features of these two techniques are summarised below.

4.1.2 Review of the TARGET-seq method

The principle of the TARGET-seq method is to amplify targeted genomic DNA (gDNA) and cDNA loci harbouring mutations of interest in parallel with whole transcriptome amplification (Figure 4.1) [226, 227]. The 3'-TARGET-seq method has a number of notable features that should be highlighted:

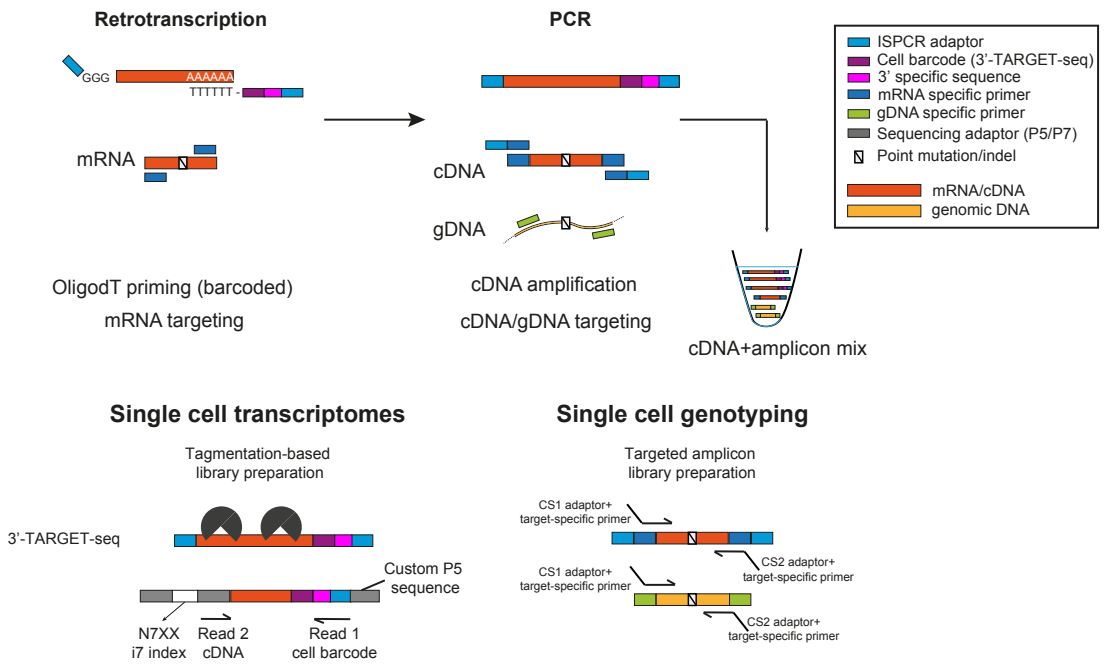


Figure 4.1: Schematic representation of the 3'-TARGET-seq method. The steps in the protocol are summarised in the main text.
Reproduced from Rodriguez-Meira *et al.* [226] (licensed under CC BY 4.0).

1. Single cells are isolated by FACS, enabling data on cell-surface proteomics to be collected (by index sorting).

2. Cells are sorted into a lysis buffer containing a protease, which assists the release of gDNA from the nucleus. This is subsequently heat inactivated to prevent inhibition of the RT and PCR steps.

3. Barcoded oligodT-ISPCR primers are used to prime polyadenylated mRNA molecules from each single cell in the RT reaction. In parallel, target-specific primers are added to prime reverse transcription of mRNA molecules harbouring the mutant locus.

4. In the preamplification PCR, ISPCR adapters are used to amplify polyA-cDNA, and target-specific cDNA and gDNA primers flanking mutations are added to amplify amplicons of interest.

5. The RT step is performed using the SMARTScribe reverse transcriptase, and the PCR uses the SeqAmp DNA polymerase, which were found to improve the sensitivity for gene expression detection, as compared to Smart-seq2.

6. An aliquot of the resulting cDNA+amplicon mix is then used for preparing the genotyping library and another aliquot for preparing the transcriptome library for scRNA-seq.

7. Genotyping libraries are prepared by two sequential PCR reactions; PCR1 adds a plate barcode and universal CS1/CS2 adapters; PCR2 adds a well-specific barcode to enable multiplexed sequencing.

8. Transcriptome libraries are prepared in a tagmentation-based library preparation, where a 3'-specific primer sequence is added to preferentially amplify fragments containing the 3'-end of the cDNA molecule. Because cDNA molecules contain a cell-specific barcode from the oligodT-ISPCR primer, cDNA from up to 384 cells can be pooled together prior to transcriptome library preparation.

Obtaining genotyping information from gDNA was found to be particularly critical; the addition of gDNA targeting improved the detection of all four mutations in a cell line from 38.9% of cells with mRNA targeting alone to 98.4% when combined. By mapping genotyping and transcriptome barcodes to the original well in which a cell was sorted, three modalities of information can be collected for every cell:

1. Immunophenotype

2. Genotype

3. Transcriptome

4.1.3 Review of Smart-seq3 chemistry

The Smart-seq3 protocol was developed by systematic evaluation of hundreds of different RT conditions in HEK293FT cells [265]. The principal modifications to the RT-PCR conditions from the original Smart-seq2 protocol are summarised below:

1. The RT reaction is performed using the Maxima H-minus reverse transcriptase (Maxima), which was found to obtain the highest sensitivity for gene expression detection.

2. Switching the salt during reverse transcription from KCl to NaCl improved sensitivity in single-cell reactions using Maxima, most likely due to the effect of Na^+ in reducing RNA secondary structures [266].

3. Adding 5% polyethylene glycol (PEG) to the RT reaction improved cDNA yields by molecular crowding.

4. Adding guanosine triphosphate (GTP) to stabilise the template-switching reaction also improved the efficiency of reverse transcription.

Several DNA polymerases were tested and KAPA HiFi Hot-Start polymerase remained the most compatible with the reaction chemistry and had the greatest sensitivity.

In addition, a UMI sequence was incorporated into the template-switching oligo (TSO), which enables 5' UMI transcript quantification, along with internal reads from either strand that are not UMI-tagged. However to make use of this, a full-length transcriptome library preparation and sequencing is required, which is costly and limits cell throughput.

4.1.4 Aims in developing TARGET-seq+

In developing a modified TARGET-seq method incorporating elements of the chemistry from the Smart-seq3 protocol, we had the following aims:

1. Improve the efficiency of reverse transcription to improve the sensitivity for gene expression detection.

2. Maintain, or closely match, the high sensitivity genotyping analysis from the TARGET-seq protocol.

3. Maintain compatibility with the 3'-biased high throughput transcriptome library preparation from 3'-TARGET-seq to enable cost- and time-efficient analysis of medium-to-large cell numbers.

4. Automate liquid handling steps as far as possible to increase cell throughput.

The modifications made to the protocol are discussed in the next section. The modified method was validated in a cell line and in primary human haematopoietic stem and progenitor cells by comparing sequencing results across methods. I refer to the new method as TARGET-seq+ (TS+), and the previous method is abbreviated as TS (TARGET-seq). Finally, analysis strategies used for calling mutant and wild-type cells from single-cell-genotyping data are discussed at the end of this chapter.

4.2 Modifications to the method and pilot experiments

To improve the efficiency of the RT reaction, we adopted the Maxima H-minus reverse transcriptase with buffer containing NaCl, 5% PEG and GTP from Smart-seq3 while adding a protease to enable the release of gDNA for genotyping. First, the compatibility of the new RT-PCR reaction conditions with the protease and heat inactivation steps were tested in a cell line (JURKAT) and in primary human HSPCs. PCR amplification with both SeqAmp and KAPA HiFi Hot-Start (KAPA) DNA polymerases was tested, and it was found that only KAPA was compatible with the new reaction chemistry (Figure 4.2a-b). Further testing showed that using the TSO with locked nucleic acid (LNA) modification from the Smart-seq2 and TARGET-seq protocols improved the cDNA yield in primary samples when compared to the TSO without LNA modification from the Smart-seq3 protocol (Figure 4.2c).

Targeted primers flanking mutations of interest and genotyping library preparation steps were largely kept as in the original TARGET-seq method, with a few modifications. In the original method, cDNA target-specific primers are attached to an ISPCR adapter sequence to increase the amplification efficiency of cDNA

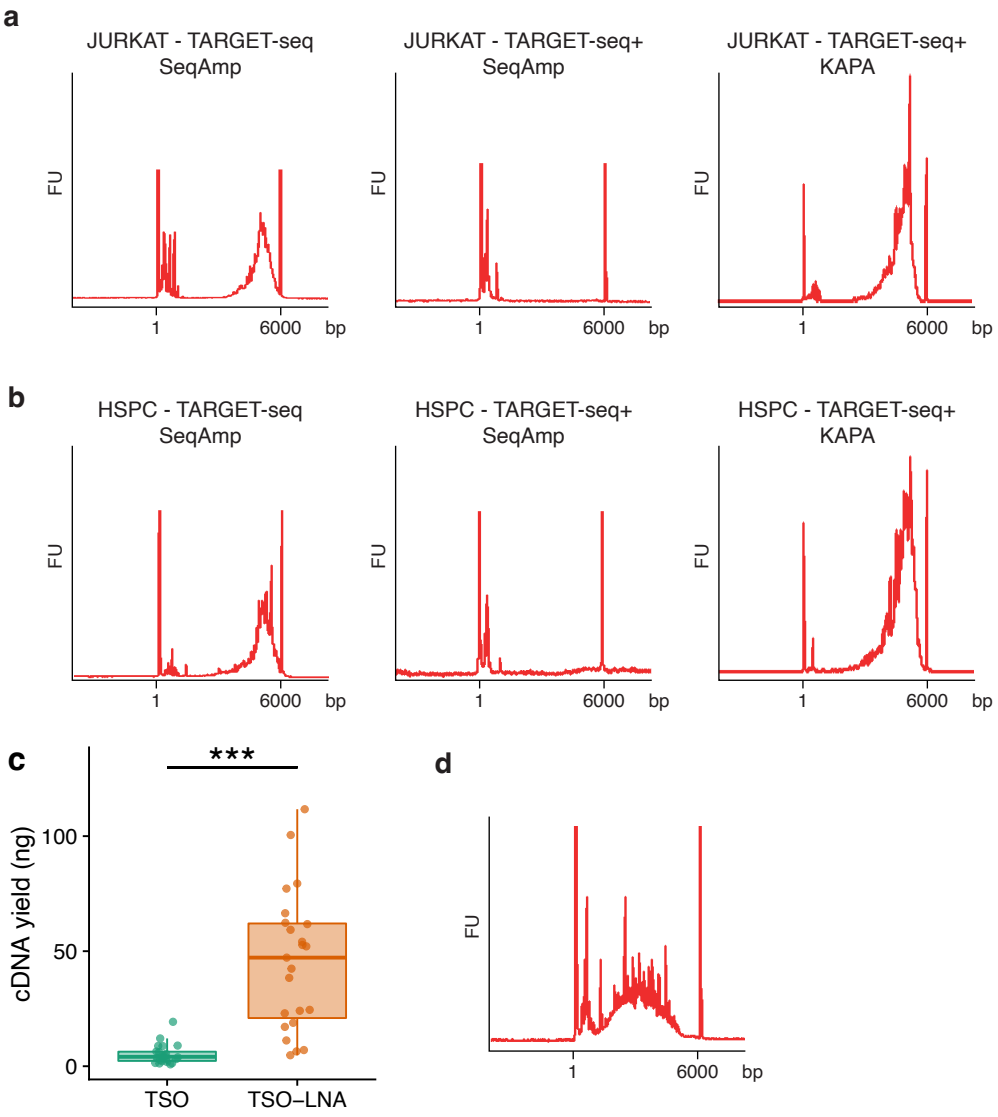


Figure 4.2: Testing RT-PCR conditions for TARGET-seq+. (a-b) Fragment analyser traces of representative pre-amplified single-cell cDNA libraries synthesised using TARGET-seq, TARGET-seq+ with SeqAmp polymerase or TARGET-seq+ with KAPA HiFi polymerase in JURKAT cells **(a)** or primary human HSPCs **(b)**. The y axis shows fluorescence units (FU), a measure of DNA concentration. **(c)** cDNA yields in primary HSPCs in an experiment comparing the use of the Smart-seq3 TSO without LNA (left) with the TSO-LNA (right) in the RT reaction. Each dot represents a single-cell library ($P < 0.001$, Wilcoxon rank-sum test). **(d)** Example of a fragment analyser trace showing concatemer formation, commonly seen when using high concentrations of targeted mRNA primers.

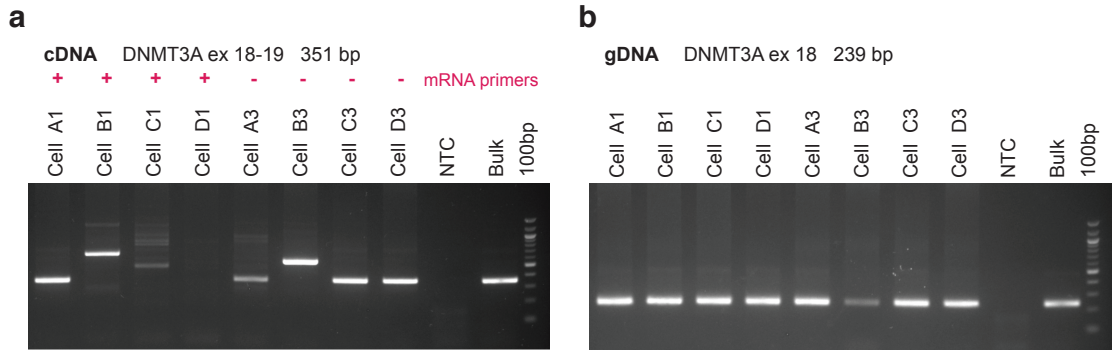


Figure 4.3: Testing targeted amplification of a genotyping locus with TARGET-seq+. Agarose gel electrophoresis showing targeted amplification of *DNMT3A* exon 18-19 cDNA **(a)** and exon 18 gDNA **(b)** from single HSPCs. Red markers indicate whether targeted mRNA primers were added at the RT step. cDNA amplification is often not successful in every cell for genes with low expression. NTC: no template control, Bulk: positive control using bulk cDNA or gDNA.

targets [267]. It was found that these increase the formation of concatemers and interfere with the generation of cDNA libraries, so the ISPCR adapter sequences were removed. Second, the concentrations of targeted primers used in the original method is high (700 nM and 800 nM for mRNA and gDNA targeting primers respectively). It was found that such high concentrations reduced the efficiency of cDNA amplification and were unnecessary for successful genotyping (perhaps due to the molecular crowding effect of PEG) and were therefore reduced (Figure 4.2d).

During testing, it was discovered that many targeted primers previously validated for the original TARGET-seq method were not specific for the intended target and had low amplification efficiency using the new PCR reaction conditions. For the majority of targets, this could be solved by redesigning and testing new primers. Note that for the validation experiment comparing the new method with the original, the primers used were those validated for the original protocol (see Figure 4.3 for an example of successful targeted amplification). In the course of testing new primers, it was also found that using an annealing temperature of 67°C for the PCR step, as is the case in TARGET-seq and Smart-seq2, achieved better results than 65°C as is used in the Smart-seq3 protocol. Due to the constraints imposed by a need for unbiased amplification of total polyA$^{(+)}$ cDNA alongside multiple

genotyping targets, customising the annealing temperature for each targeted primer pair or sample is neither possible nor desirable.

To enable high throughput RNA-sequencing, barcoded oligodT-ISPCR primers were used to incorporate a cell-specific barcode at the RT step as in the 3'-TARGET-seq protocol. The 3'-biased transcriptome libraries generated with this method prevent the use of the 5' UMI sequence introduced in Smart-seq3. While it would have been preferable to incorporate UMIs, it is important to note that the major improvements in reproducibility of transcript quantification in Smart-seq3 were down to improvements in chemistry and only a minor effect was gained by the addition of UMIs (see Figure 1e in Hagemann-Jensen *et al.* [265]). To make use of UMIs, they would have to be incorporated in the oligodT primer but this approach is hampered by oligodT-UMI primers remaining during the PCR reaction re-priming and causing a false increase in the UMIs detected (unpublished data from the Mead lab).

For transcriptome library sequencing, cells within each plate of 384 wells have a unique cell-specific barcode, and cells from different plates can be identified by unique i7 indexes added in the transcriptome library preparation. A total of 26 Nextera i7 indexes are commercially available, enabling multiplexing of up to 9,984 cells together on a sequencing run. However, with sequencing on platforms using novel patterned flow cell technology, there have been reports of index misassignment due to "index hopping" [268, 269]. To mitigate this effect, an additional i5 index was incorporated in the primers used to amplify the final transcriptome library. This introduces redundancy between the i7 and i5 barcode, such that if index hopping occurs, there is a greater chance of the read being flagged as unidentifed and filtered out rather than misassigned to the wrong cell [269]. The modified transcriptome library structure is illustrated in Figure 4.4.

In developing the new protocol, all reactions were designed to be performed in 384-well plates and reaction volumes were reduced in order to minimize the use of enzymes and cost. To increase throughput, we established most steps on automated liquid handling platforms, using the Formulatrix Mantis for mastermix dispensing steps and the Integra Viaflow for plate-to-plate transfers using some of

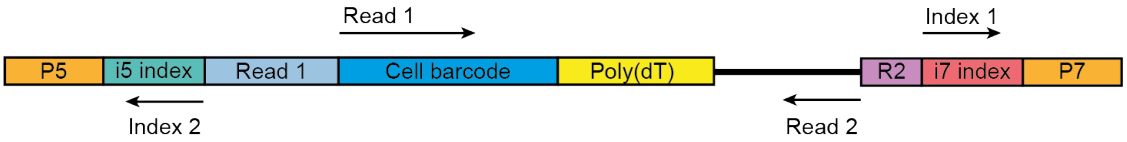


Figure 4.4: Structure of 3' transcriptome libraries for the TARGET-seq+ method. Read 1 is used to sequence the cell barcode introduced in the OligodT-ISCPR primer during reverse transcription. There are 384 unique cell barcodes, identifying the well the cell was sorted into. Read 2 sequences the cDNA insert (black line) used to assign the read to a transcript. The i7 and i5 indexes are incorporated with Illumina sequencing adapters (P7 and P5), which are introduced in the final transcriptome library amplification, and are unique for each pool of up to 384 cells pooled together from a plate. Indexes i7 and i5 are sequenced with the Index 1 and Index 2 reads respectively.

the techniques described by Rodriguez-Meira *et al.* [227]. While this does not remove the need for manual interaction at each step, it does increase the "processivity" and means more plates can be processed in a day.

4.3 Validation of TARGET-seq+ in cell lines and primary haematopoietic cells

To evaluate the performance of the TARGET-seq+ method, a validation experiment was performed in a diploid cell line (JURKAT) and primary human HSPCs.

4.3.1 Targeted genotyping with TARGET-seq+

First, the performance of single cell genotyping with TARGET-seq+ was evaluated. Four coding mutations that are heterozygous in JURKAT cells were profiled, *NOTCH1*, *PTEN*, *RUNX1* and *TP53*.

We used the same primer sequences as those used in the original TARGET-seq study to amplify both gDNA and mRNA across each mutation. As the cell line is clonal and mutations are heterozygous, both WT and mutant alleles should be detected in every cell if the reaction is 100% efficient. Using the original TARGET-seq method, the frequency of biallelic detection was >87% for *NOTCH1*, *RUNX1* and *TP53*, similar to the original study but detection of *PTEN* was poor (Figure 4.5); the reason for this is unclear. In comparison, TS+ achieved better detection rates for *PTEN*, *RUNX1* and *TP53* but only 37% biallelic detection of the *NOTCH1* mutation. For the latter, the reduced performance was due to inefficient

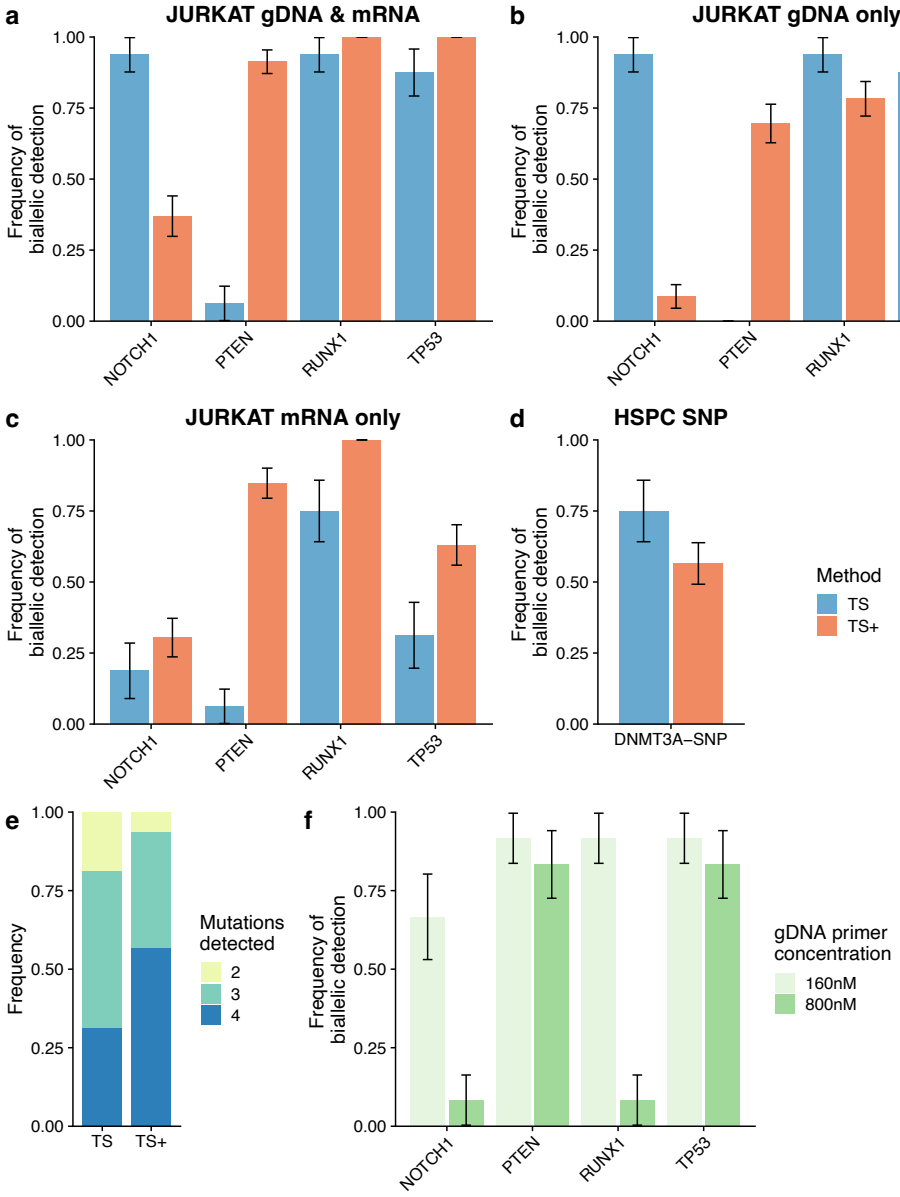


Figure 4.5: Detection of mutations with TARGET-seq+. **(a-c)** Frequency of biallelic detection of heterozygous mutations in four coding genes in the JURKAT cell line, comparing results from TARGET-seq (TS) with TARGET-seq+ (TS+). **(a)** shows the results when combining the results from gDNA and mRNA amplicons. **(b-c)** show results exclusively from gDNA amplicon sequencing **(b)** and mRNA amplicon sequencing **(c)**. **(d)** Frequency of biallelic detection of a heterozygous intronic SNP in the *DNMT3A* gene in primary HSPCs. **(e)** Frequency of detection of heterozygous mutations in JURKAT cells when four different mutations (*NOTCH1*, *PTEN*, *RUNX1* and *TP53*) were analysed in the same single cells. **(f)** Frequency of biallelic detection in gDNA amplicons, comparing PCR amplification with 160 nM and 800 nM gDNA primers. Number of cells = 16 JURKAT for TS; 46 JURKAT for TS+; 16 HSPC for TS; and 46 HSPC for TS+. Error bars show the standard error in all plots.

amplification of the gDNA amplicon (Figure 4.5a). Interestingly, the detection rate for mRNA was better using the new method for three of the four mRNA targets, consistent with increased efficiency of the RT reaction. In primary human HSPCs, an intronic heterozygous SNP was profiled (using the gDNA amplicon only) and the detection rate was slightly lower than with TS ($P = 0.24$, Fisher's exact test).

Combining the data for the four mutations analysed in JURKAT cells, we found that all four mutations were detected in approximately half of cells, and 3 mutations in 93%. In summary, mutation detection was equivalent or slightly better with TS+ than TS in a direct comparison. Nonetheless, it is clear that some amplicons, such as the *NOTCH1* gDNA amplicon in this experiment are not amplified efficiently under the new PCR conditions. In a separate experiment, it was found that reducing the concentration of targeted gDNA primers to 160 nM improved the detection rate for some of these amplicons (Figure 4.5f). In further testing of other loci, it was clear that redesigning primers improved the efficiency of amplification with the KAPA Hifi enzyme, an observation that is unsurprising given the influence of differing salt concentrations on primer annealing.

4.3.2 Transcriptome sequencing with TARGET-seq+

Next, the performance of transcriptome profiling with TARGET-seq+ was evaluated in JURKAT cells and human HSPCs. For JURKAT, transcriptome libraries were generated from 24 cells with TARGET-seq and 46 cells with TARGET-seq+, and for HSPCs, libraries generated from 482 cells with TARGET-seq were compared with 46 cells with TARGET-seq+. Libraries were processed and sequenced in the same way with 69 bp single-end sequencing and identical mapping pipelines. Sequencing quality control metrics were similar for the two methods (Figure 4.6). Using the same cutoffs, >93% of cells passed quality controls (QC) when processed with the new method compared to 80–89% with TARGET-seq.

Quantification of transcripts from endogenous genes was highly correlated between the two methods within each cell type (Figure 4.7a,c). This was also the case with ERCC spike-in controls, where read counts were highly correlated with

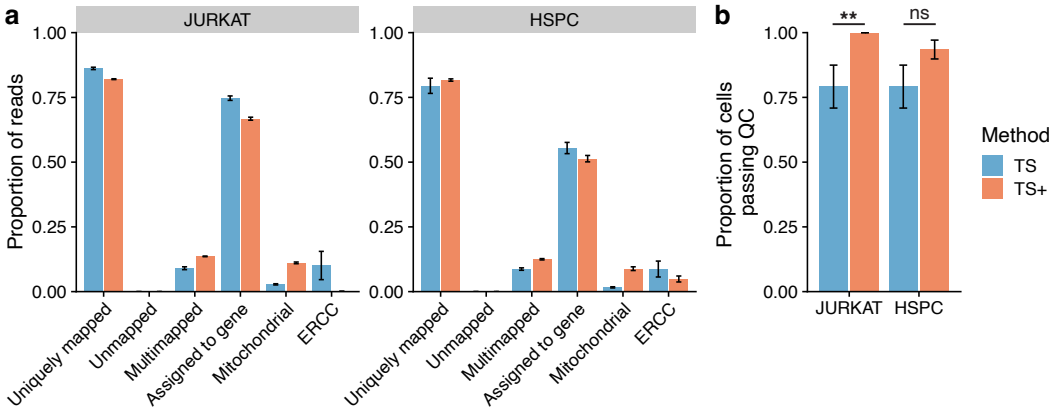


Figure 4.6: Quality metrics for TARGET-seq+ validation. **(a)** Sequencing statistics of single cell libraries from JURKAT and primary HSPCs processed with TARGET-seq (TS) and TARGET-seq+ (TS+). The bars represent the proportion of reads for each statistic and error bars represent the standard error of the mean. **(b)** Number of cells passing passing quality control (QC) for each method. N = 24 JURKAT for TS; 46 JURKAT for TS+; 24 HSPC for TS; and 46 HSPC for TS+. Error bars represent the standard error of the mean. ** $P < 0.01$, ns: non-significant, Fisher's exact test.

the concentration of each transcript (Pearson correlation coefficient > 0.9 for both cell types) and between cells processed with TS and TS+ protocols (Figure 4.7b,d).

Dimensionality reduction by principal component analysis (PCA) of the highly variable genes showed that JURKAT and HSPCs analysed by both methods were clearly separated along the first principal component (PC1) (Figure 4.8a-b). JURKAT and HSPCs also clustered separately on a UMAP. While HSPCs processed by both methods clustered together, JURKAT cells separated according the to method used (Figure 4.8c-d).

To compare the sensitivity of each method, data were downsampled to 5×10^5 reads per cell to obtain equal library sizes across methods. TARGET-seq+ significantly improved the detection of expressed polyA$^+$ genes in both JURKAT cells and HSPCs ($P < 0.001$, two-tailed t-test). For JURKAT cells, TARGET-seq+ detected a median of 969 (13.5%) more genes per cell ($P = 0.0019$, two-tailed t-test). For HSPCs, the median number of genes detected increased from 3817 to 4533, an increase of 19% ($P = 0.0030$, two-tailed t-test). In comparison, typical droplet-based experiments detect 2000–3000 genes per cell in HSPCs. TARGET-seq+ had

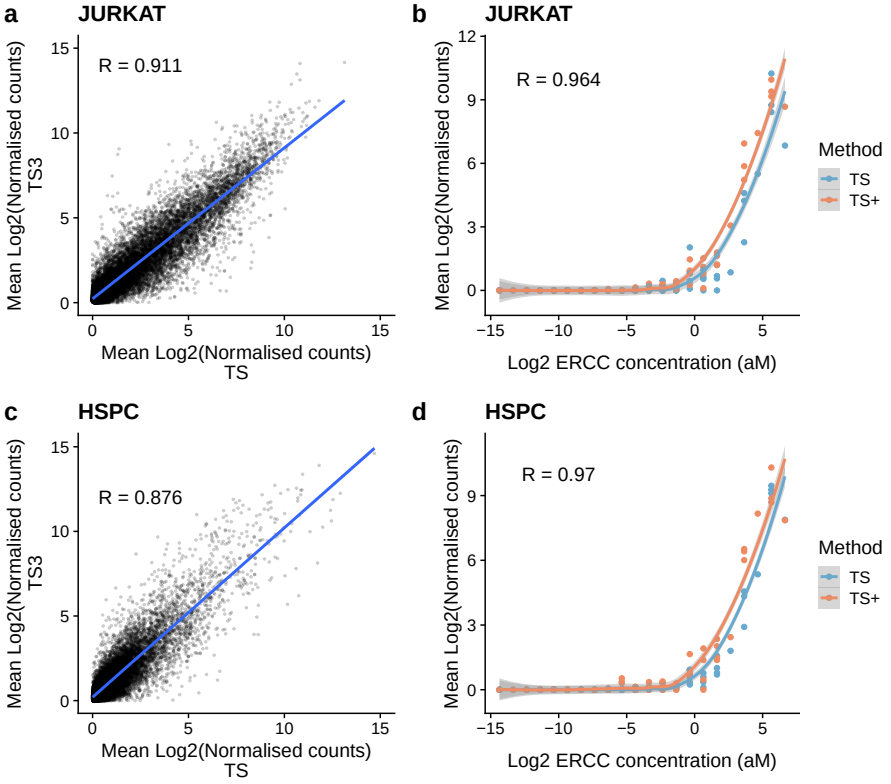


Figure 4.7: Transcript quantification with TARGET-seq and TARGET-seq+. Pearson's correlation between transcript quantification results for JURKAT **(a-b)** and HSPC **(c-d)** libraries processed with TARGET-seq and TARGET-seq+. **(a)** and **(c)** show comparisons of mean normalised counts for all expressed genes; each point represents a gene. The linear regression line is shown in blue. **(b)** and **(d)** show the mean normalised counts for ERCC spike-ins against the known concentration of each spike-in transcript in the lysis buffer; each point represents a spike-in transcript. R indicates the Pearson correlation coefficient for the comparison between TS and TS+. aM: attomolar.

a greater sensitivity across sequencing depths from 25,000 to 1 million reads per cell ($P < 0.001$, nested ANOVA). In line with previous reports, the number of genes detected approaches saturation at around 1 million reads per cell [270].

Single cell RNA-seq data are typically characterised by high dropout rates, where genes are not detected even though they are expressed due to technical limitations such as inefficient reverse transcription. To determine if the modified RT conditions in TARGET-seq+ reduces technical dropouts, the dropout rate was estimated as the fraction of cells with zero counts for a given set of genes. Note that HSPCs are a heterogeneous population, and many genes are expected to be expressed only in a subset of cells, so the fraction of cells with no detection is expected to

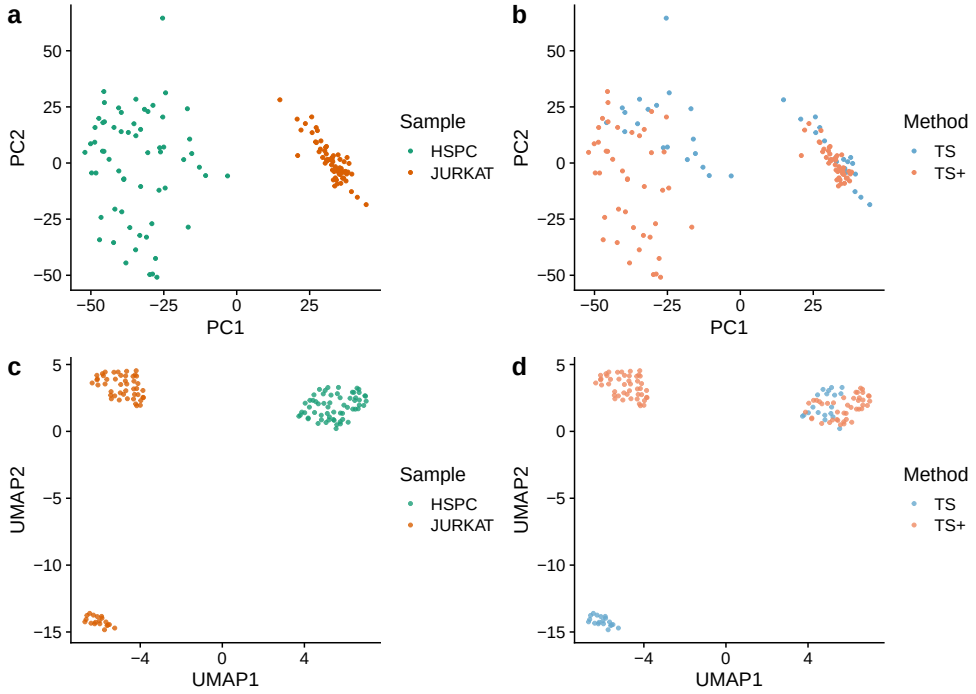


Figure 4.8: PCA and UMAP of cells analysed with TARGET-seq and TARGET-seq+. (a-b) Principal component analysis (PCA) showing JURKAT and HSPCs analysed with both methods, using 2452 highly variable genes. Cells are coloured by the cell type in **(a)** and the method used in **(b)**. **(c-d)** UMAP projection of the same cells as above using the top 20 PCs. Cells are coloured by the cell type in **(c)** and the method used in **(d)**.

be higher than for a relatively homogeneous cell line such as JURKAT. Dropout rates were significantly lower with TS+ across all expressed genes, including those expressed in over half of cells, and lowly expressed genes. This was the case for both JURKAT and HSPCs. Consistent with improved reproducibility of gene expression data, cell-to-cell correlations improved significantly for JURKAT cells with the new method. As the same culture conditions were used for cells analysed by both methods, the amount of biological variation in JURKAT expression profiles should be the same, so this shows the TS+ protocol reduces technical variation.

In summary, TARGET-seq+ was robust and clearly separated cell types by their expression profile. It had greater sensitivity than the original TARGET-seq protocol for detecting gene expression with lower dropout rates and reduced technical variation.

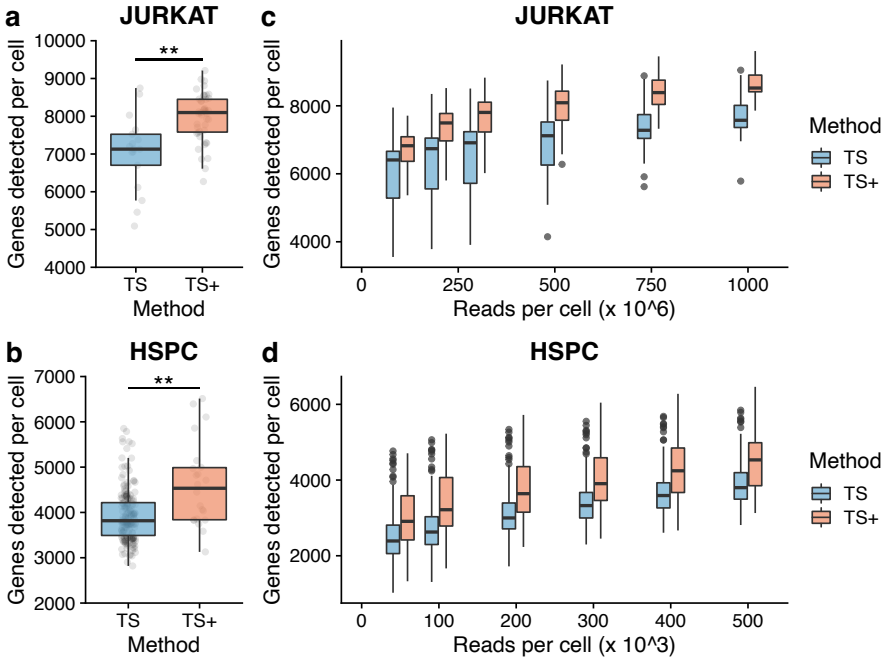


Figure 4.9: Sensitivity of TARGET-seq+ for transcript detection. **(a-b)** Number of genes detected per cell (counts ≥ 1) for JURKAT **(a)** and HSPC **(b)**. In both cases reads were downsampled to 5×10^5 reads per cell. Each dot represents a cell and each box represents the median and first and third quartiles for each method. $P < 0.01$ for both cell types, unpaired two-tailed t–test. **(c-d)** Number of genes detected per cell (counts ≥ 1) when downsampling total read counts to the indicated depths for JURKAT **(c)** and HSPC **(d)**. For JURKAT, the read depth ranges from 100,000 (n = 19 cells for TS, 46 cells for TS+) to 1,000,000 (n = 12 cells for TS, 27 cells for TS+). For HSPC, the read depth ranges from 25,000 (n = 404 cells for TS, 36 cells for TS+) to 500,000 (n = 178 cells for TS, 22 cells for TS+). Each box represents the median and first and third quartiles for each method.

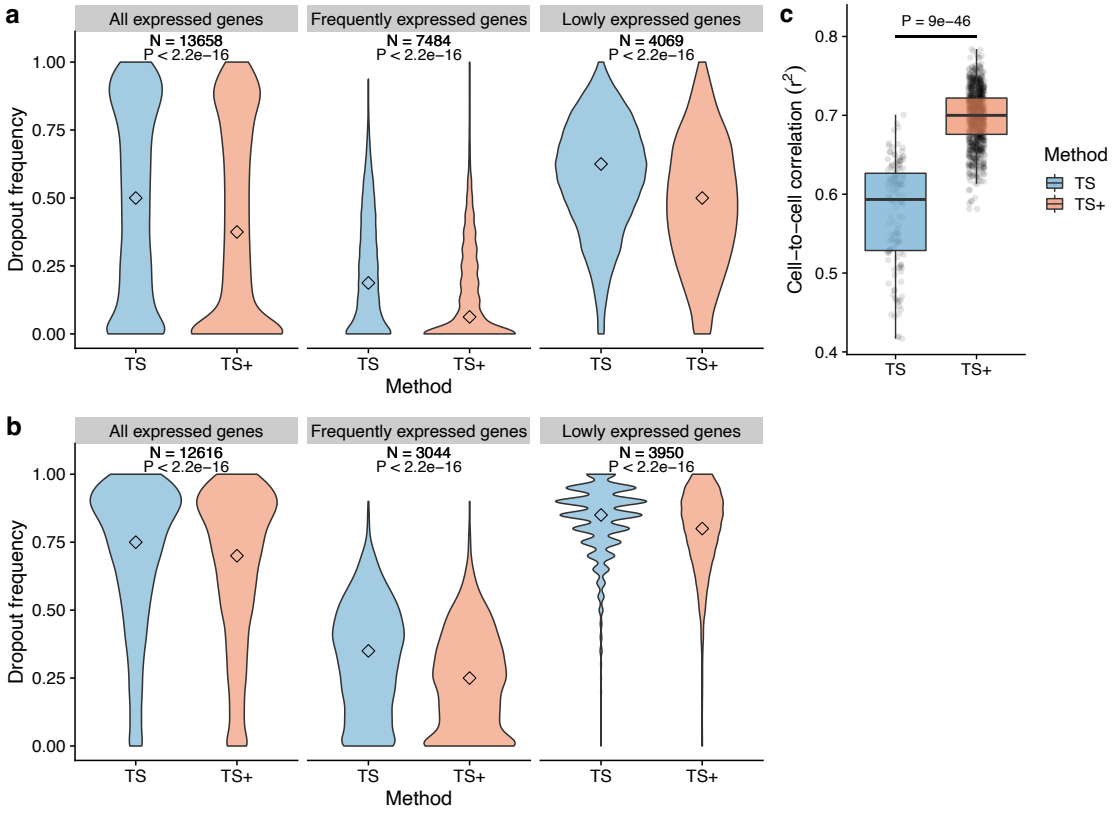


Figure 4.10: Dropout rates and reproducibility with TARGET-seq+. Comparison of dropout frequency between TARGET-seq (TS) and TARGET-seq+ (TS+) methods for JURKAT cells **(a)** and HSPCs **(b)**. A subsample of 16 cells per chemistry were analysed for JURKAT and 20 cells per chemistry for HSPC; in both cases reads were downsampled to 5×10^5 reads per cell. Dropout frequencies are shown for all genes expressed in at least 2 cells (left), for genes expressed in >50% of cells by any method (middle) and genes with low expression (mean normalised counts of 2-10 per cell; right). Diamonds show the median dropout rate for each condition. The number of genes included in each comparison and *P*-values from two-tailed Wilcoxon rank-sum test are shown on the top of the graph. **(c)** Reproducibility in gene expression quantification for JURKAT cells with TS (n = 17 cells) and TS+ (n = 43 cells). Shown are r^2 values for all pairwise cell-to-cell Pearson's correlations in libraries downsampled to 5×10^5 reads per cell.

4.4 Mutational calling in single-cell genotyping

In order to apply the TARGET-seq+ method to the analysis of samples with clonal haematopoiesis, it was first necessary to develop a strategy for calling mutations in single cells. Genotyping sequencing data were mapped and processed as described in 2.10.1. This generates tables of allelic counts per cell for every variant analysed. A strategy for calling the mutational status of each cell, taking into account technical noise, was devised and is briefly described below.

Experiments were designed such that every plate of FACS-sorted single cells included empty wells used as no template controls, and wells with single cells from a control sample. The same control sample was used across every experiment, so that it could also be used to control for batch effect in transcriptome sequencing. As control cells are known to be WT for the mutation, they can be used to determine the error rate or noise of the mutational readout.

Where it was possible to generate separate cDNA and gDNA amplicons across a mutation, this enabled independent mutational readouts from each. In these cases, allelic counts were analysed independently from each amplicon and then combined to determine the consensus genotype of each cell.

First, the total coverage across the mutant locus was calculated and compared to the no template controls to set a filtering threshold for cells with successful amplification. An example is illustrated in Figure 4.11. Coverage of gDNA amplicons was generally higher and more uniform than for cDNA amplicons.

Mutational calling was then performed on the filtered cells. The variant allele frequency (VAF) on a per-cell basis was calculated, as the frequency of variant alleles divided by the total coverage. If amplification of a heterozygous mutation was perfectly balanced, the VAF of each mutant cell would be 0.5 and the VAF of WT cells would be 0. However, in reality one allele is often amplified more than the other due to PCR bias, resulting in a broad distribution of VAFs, which for heterozygous cells is centred around 0.5. The challenge is to distinguish mutant cells from noise

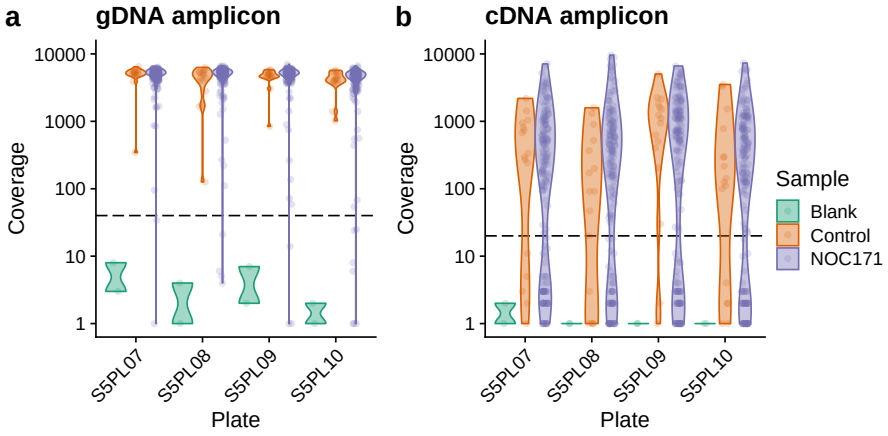


Figure 4.11: Setting the minimum coverage in single cell genotyping. Coverage of the *DNMT3A* pW581fs mutation in single cells from the NOC171 CH sample and a normal control. **(a)** Coverage in gDNA amplicon reads. **(b)** Coverage in cDNA amplicon reads. Each dot represents a single cell and are arranged by the plate. The minimum coverage thresholds (dashed lines) were set based on coverage in the blank (no cell) controls.

due to PCR error at the lower extreme. The threshold for calling mutations in each amplicon was set based on the VAF observed in control cells (Figure 4.12a-d). Genotyping information from gDNA and cDNA were then combined, and a consensus genotype assigned, such that if the mutation was identified in either amplicon, the cell was called mutant (Figure 4.12e). Allelic dropout was more common in cDNA reads than gDNA reads (visible as cells with a VAF of 0 or 1 in cDNA reads in the example in Figure 4.12e).

In four cases, it was possible to amplify a germline heterozygous SNP in the same gDNA amplicon as the mutation. Since germline variants are present in every cell, the frequency of homozygous calls for the SNP could be used to measure allelic dropout for the amplicon. Further, by correlating the VAF of the SNP with the mutation, it could be determined which allele was in phase with the mutation (Figure 4.12f-g). If the allele in phase with the mutation was not detected, the cell was assigned an undetermined genotype, as allelic dropout had occurred. In the absence of a heterozygous SNP, it is not possible to distinguish cells with allelic dropout of the mutant allele from WT cells.

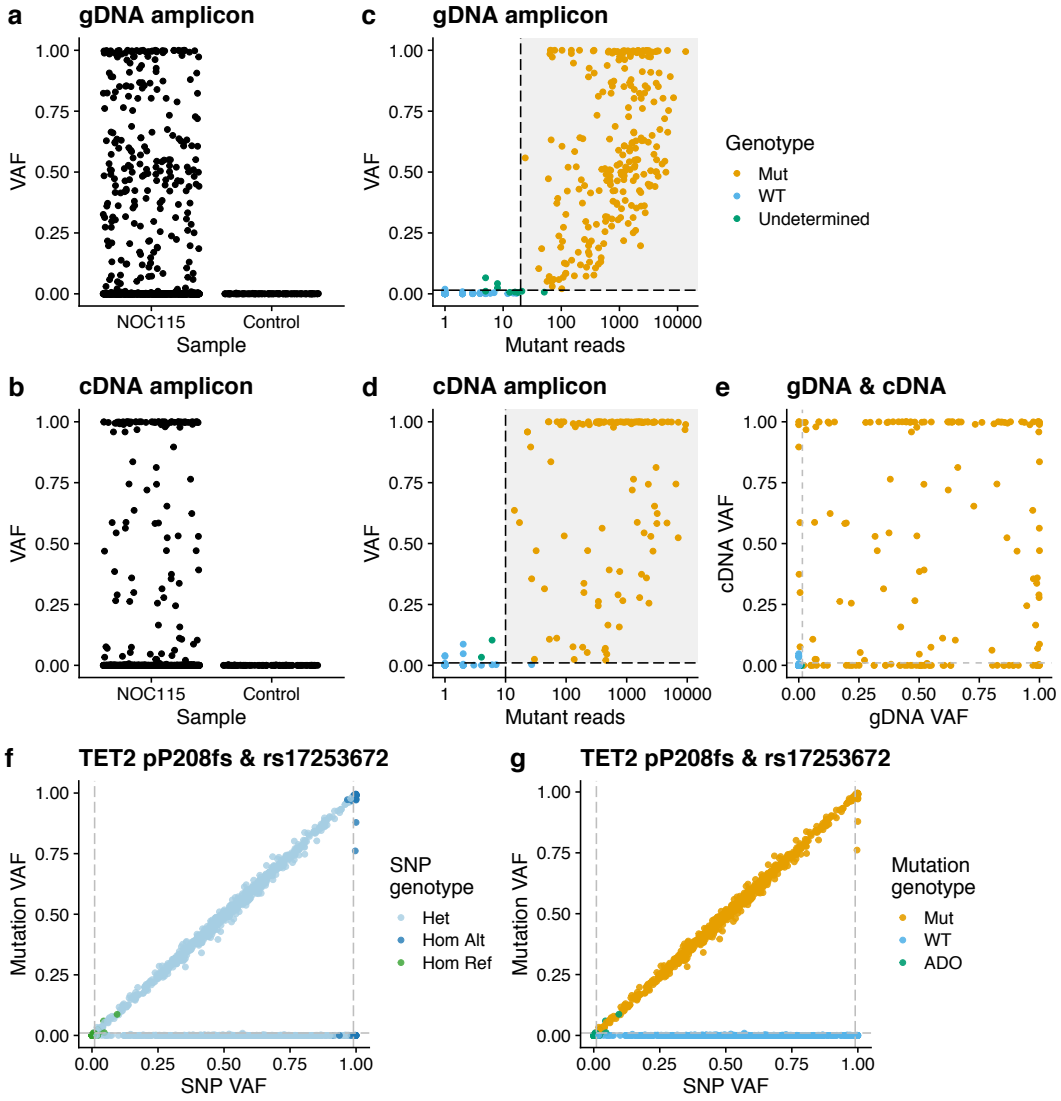


Figure 4.12: Mutation calling in single cell genotyping. **(a-b)** Distribution of VAFs in gDNA **(a)** and cDNA **(b)** amplicon reads, comparing the CH sample (NOC115 in this case) and the control. Each dot represents a cell. The VAF observed in the control cells are used to set the threshold for calling mutant cells. **(c-d)** Calling WT and mutant (Mut) cells in gDNA **(c)** and cDNA **(d)** reads. A cell is called mutant if the VAF and the number of mutant reads exceed the thresholds indicated by the dashed lines (those that fall in the grey area). Cells with borderline VAF or mutant reads are called undetermined. **(e)** Comparison of VAF in gDNA and cDNA reads from each cell. Cells are called mutant if the mutation is detected in either gDNA or cDNA. **(f-g)** Calling genotype in gDNA amplicons where there is a heterozygous germline SNP. The VAF of the mutation is plotted against the VAF of the germline SNP for each cell. In this case the SNP alternate allele is in phase with the mutation, as evident from the correlation in VAF within mutant cells (on the diagonal). **(f)** Cells are assigned as heterozygous (Het), homozygous reference (Hom Ref) or homozygous alternate (Hom Alt) for the SNP genotype to determine whether allelic dropout of either allele had occurred. **(g)** Cells are called mutant or WT as described above. When the SNP alternate allele is not detected, cells are called undetermined due to allelic dropout (ADO, green, bottom left).

4.5 Allelic dropout and accuracy of single-cell genotyping

Allelic dropout (ADO) is common in single-cell genotyping and can confound the results due to misassignment of mutant cells as WT when the mutant allele is not amplified. ADO may be estimated through two methods. First, it may be directly measured in cases where a heterozygous germline SNP is present within the genotyping amplicon, such as the four cases mentioned above (Figure 4.12f and Figure 4.13). The four mutations where heterozygous SNPs were present were:

1. *DNMT3A* pQ606X and rs2289093, in sample NOC131
2. *DNMT3A* pR771X and rs2289195, in sample NOC132
3. *TET2* pP208fs and rs17253672, in sample NOC117
4. *TET2* pP1194H and rs55838312, in sample NOC117

The frequency of biallelic detection was calculated by assigning cells as heterozygous (Het), homozygous reference (Hom Ref) or homozygous alternate (Hom Alt) for the SNP. The frequency of biallelic detection was 75–85% and dropout was balanced between the alleles. From this, the dropout of the mutant allele was found to be 7–11%. In a sample with 50% mutant cells, 3.5–5.5% of cells would therefore erroneously be classified as WT when they are actually mutant.

An alternative way of estimating ADO is by analysing the frequency with which a heterozygous variant is detected homozygously; i.e. the dropout rate of the WT allele. The frequency of homozygous detection was calculated across 16 mutations with gDNA amplicons, of which 11 also had a cDNA amplicon, and the corresponding ADO rate for the mutant allele estimated. The mean estimated ADO rate for gDNA amplicons was 8.8% (range 1.6–14%) and for cDNA amplicons it was 16.2% (range 0–29%) (Figure 4.14a).

As a result, the accuracy of genotyping from gDNA readouts was higher than for cDNA. On average, gDNA amplicons identified 95% of mutant cells, with the

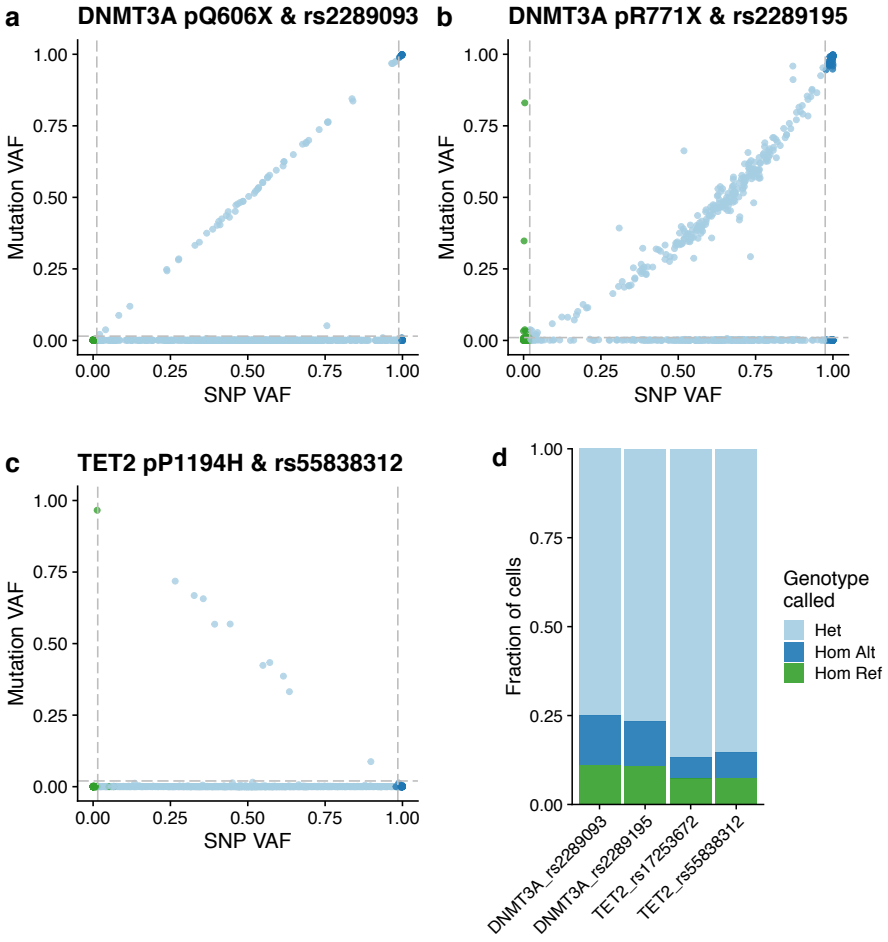


Figure 4.13: Measurement of allelic dropout in germline SNPs. (a-c) Plot of VAFs for the mutations indicated against the VAF of the SNP in each amplicon in single cells. Dashed lines show the threshold for calling the SNP homozygous (vertical) and for calling the cell mutant (horizontal). For **(a)** and **(b)** the mutation is in phase with the alternate allele, for **(c)** it is phase with the reference allele. **(d)** Frequency of biallelic detection for each of the germline SNPs. Het: heterozygous detection of both alleles, Hom Ref: homozygous detection of reference allele, Hom Alt: homozygous detection of alternate allele.

remaining 5% identified in cDNA. cDNA genotyping alone identified only 50% of mutant cells, and the accuracy was particularly low for *TET2* (Figure 4.14b). In some cases, cDNA genotyping alone would have missed 85% of mutant cells, highlighting the importance of gDNA targeting.

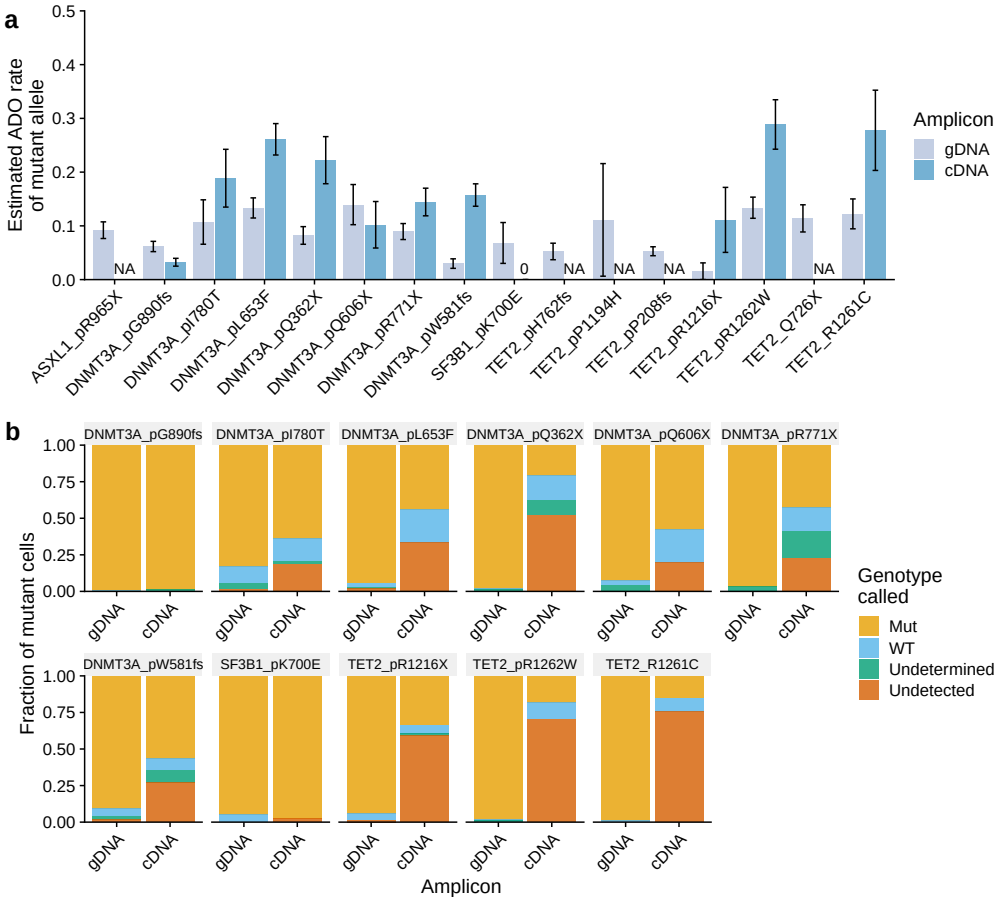


Figure 4.14: Estimation of allelic dropout in gDNA and cDNA amplicons. **(a)** Estimation of allelic dropout rate for the mutant allele across mutations from quantifying the frequency of homozygous mutant cells. NA indicates there was no cDNA amplicon for the given mutation; 0 indicates the dropout rate was zero. Error bars represent the standard error. **(b)** Fraction of mutant cells detected by gDNA and cDNA for mutations where independent readouts were possible.

4.6 Conclusions: Development of TARGET-seq+

Methods for simultaneous analysis of genotype and transcriptome from single cells enable cell phenotype to be linked to the presence of somatic mutations in mixed populations of cells from the same individual. Of the current methods for concurrent genotyping and transcriptome sequencing, TARGET-seq is uniquely suited to the study of clonal haematopoiesis due to its high sensitivity for mutation detection. To incorporate optimised reverse transcription chemistry from Smart-seq3 into the method, we developed TARGET-seq+.

The modified RT-PCR conditions were compatible with genomic and cDNA targeting, enabling high-sensitivity mutational profiling from single cells comparable to the original TARGET-seq method. TARGET-seq+ produced unbiased transcriptomic profiles that were highly correlated with results from the TARGET-seq method and clearly distinguished primary HSPCs from a cell line. Transcriptome profiling with TARGET-seq+ had a greater sensitivity and reduced dropout rates, improving reproducibility.

An analysis strategy for calling mutations in single cells was developed based on setting thresholds in non-mutant control cells. Allelic dropout rates measured in heterozygous germline SNPs and estimated from rate of homozygous mutant cells were found to be significantly lower in genomic amplicons than cDNA. Without gDNA targeting, a large proportion of mutant cells would have been misidentified as WT, particularly for *TET2*.

Chapter 5

Single-cell genomic analysis of *DNMT3A* and *TET2*-mutant clonal haematopoiesis

Contents

5.1 Approach to single-cell analysis of clonal haematopoiesis samples 105
5.1.1 Aims in single-cell analysis 105
5.1.2 Experimental design 108
5.1.3 Sample selection 108
5.2 Single-cell genotyping analysis of clonal haematopoiesis samples 110
5.2.1 Data QC and filtering 110
5.2.2 Clonal analysis of immunophenotypic populations for *DNMT3A*-mutant samples 113
5.2.3 Surface marker expression in *DNMT3A*-mutant populations 119
5.2.4 Clonal analysis of immunophenotypic populations for *TET2*-mutant samples 119
5.3 Single-cell transcriptome landscape of normal and clonal haematopoiesis samples 127
5.3.1 Data QC and filtering 127
5.3.2 Copy number variation analysis 129
5.3.3 Dimensionality reduction and clustering 129
5.3.4 Integrated analysis of transcriptomic data for normal and clonal haematopoiesis samples 135
5.3.5 Expression of *DNMT3A* and *TET2* in HSPCs 147
5.4 Transcriptomic analysis of *DNMT3A*-mutant clonal haematopoiesis 149
5.4.1 Analysis of *DNMT3A*-mutant clonal size through haematopoietic differentiation 149
5.4.2 Transcriptional consequences of *DNMT3A* mutations in progenitors 154
5.4.3 Transcriptional consequences of *DNMT3A* mutations in HSC 161
5.5 Transcriptomic analysis of *TET2*-mutant clonal haematopoiesis 166
5.5.1 Analysis of *TET2*-mutant clonal size through haematopoietic differentiation 166
5.5.2 Transcriptional consequences of *TET2* mutations in progenitors 170
5.5.3 Transcriptional consequences of *TET2* mutations in HSC 173
5.5.4 Consequences of biallelic *TET2* mutations 179

I am grateful to Sven Turkalj who helped me in performing the TARGET-seq3 experiments in Chapter 5, particularly in helping to process genotyping plates.

5.1 Approach to single-cell analysis of clonal haematopoiesis samples

5.1.1 Aims in single-cell analysis

The most commonly mutated genes in age-related clonal haematopoiesis are *DNMT3A*, a *de novo* DNA methyltransferase, and *TET2*, a methylcytosine dioxygenase. Murine models have suggested that *Dnmt3a* loss of function results in increased HSC self-renewal, leading to clonal advantage in competitive transplantation assays [166, 167, 175, 176]. *Tet2* knockout models have also shown increased self-renewal and myeloproliferation [213–215]. Given their role in regulating DNA methylation, it is thought likely that the phenotypic effects of mutations are mediated through altered epigenetic and transcriptional regulation [91, 165, 271]. However, as both enzymes act genome-wide, it is difficult to develop reductionist hypotheses for the mechanism by which mutations lead to clonal expansion. An unbiased approach to studying the effect of mutations on the phenotype of early stem and progenitor cells was therefore adopted.

Single-cell transcriptomics provide a large amount of information on cell state and have been used to distinguish phenotypically-distinct cells within heterogeneous populations. The ability to measure genotype and mRNA transcripts concurrently within the same cells with multi-omic techniques provides one approach to compare the phenotype of cells with and without mutations. For the reasons outlined in the previous chapter, I adapted the TARGET-seq method to develop TARGET-seq3, which provides highly accurate mutational detection with high sensitivity gene expression profiling to maximise the ability to detect subtle differences between phenotypically similar groups of cells. By enabling the comparison of mutant cells with their wild-type counterparts within the same individual, this also provides

an ideal control, avoiding confounding factors related to variations in germline background when comparing different individuals.

TARGET-seq+ was applied to CH cases from the MARCH cohort with *DNMT3A* and *TET2* mutations to identify the phenotypic effects of these mutations on early haematopoietic stem and progenitor cells. The aims of this analysis were:

1. To identify the cell of origin where mutations are acquired.

2. To identify the patterns of clonal expansion and effect of mutations on haematopoietic differentiation, for example whether maximal expansion occurs in stem cells or further downstream, and whether mutant cells have a lineage bias.

3. To determine differentially expressed genes in stem and progenitor cells that might indicate pathways involved in clonal advantage in the cell types where clonal expansion occurs.

4. To explore early patterns of clonal evolution in non-leukaemic samples with multiple driver mutations.

In total, 10 samples with CH were analysed, including 5 cases with *DNMT3A* mutations, 4 cases with *TET2* mutations, and 1 case with both.

5.1.2 Experimental design

In order to study the effect of mutations on haematopoietic differentiation, cells were sorted from the Lin–CD34+ HSPC compartment which includes stem, early and late progenitor cells for the major haematopoietic lineages (Figure 5.1). To enrich for the stem and early multipotent progenitor cells, which normally make up approximately 15% of CD34+ cells, approximately 30–40% of cells were specifically sorted from the CD38– compartment. These are the cell populations where early lineage commitment takes place so we were particularly interested to study the effects of the mutation on these cell types. In addition, a small proportion of cells expressing lineage markers (Lin+, including erythroid, myeloid and lymphoid

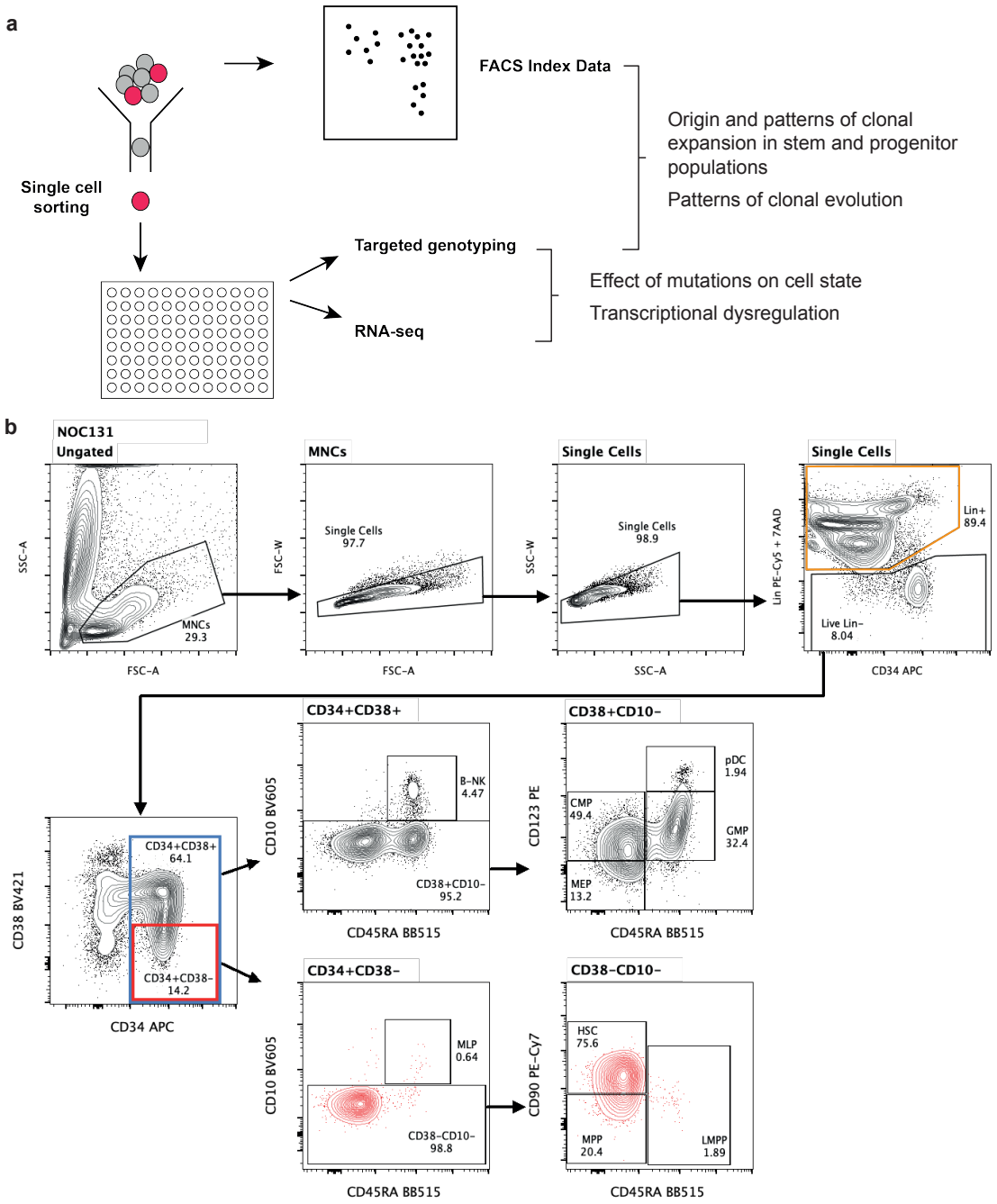


Figure 5.1: Sorting single cells for TARGET-seq+ analysis. (a) Schematic of the experimental plan. Single cells from CH samples are FACS index sorted, and TARGET-seq+ is used to perform targeted genotyping and whole-transcriptome RNA-sequencing. **(b)** Sorting strategy for isolating single cells from BM MNCs. Representative flow cytometry plots are shown for the NOC131 CH sample. Numbers represent percentage of gated cells. Cells were sorted from Lin–CD34+ (blue gate), enriching for Lin–CD34+CD38– (red gate) and a number of Lin+ cells (orange gate) for each sample. The gating strategy used to define populations on the flow cytometry index data equivalent to the gates shown. Cells that fell outside these gates were classified as CD38+CD10–CD123– or Lin–CD34–.

surface markers) were sorted to obtain cells from the major lineage trajectories and identify any lineage bias in mutant clones.

Single cells were sorted into plates and flow cytometry indexing data was collected so that the immunophenotype of each cell was known. In addition to the markers shown in Figure 5.1, index data was collected for CD117 (c-Kit) and CD49f. Three to four samples were sorted on any one day and processed identically to minimise batch effects. Cells from the NOC153 control sample were sorted onto every plate, making up approximately 10% of wells on each plate. These cells were used as a negative control in genotyping (as described in Chapter 4) and as a batch control for transcriptome analysis.

Thus from every cell there were three types of data available: flow cytometry indexing data on cell surface markers, genotyping data and transcriptome profiling data. This enabled the presence of mutations to be linked to both immunophenotypically-defined subpopulations and to cell states as defined by gene expression.

5.1.3 Sample selection

Targeted sequencing of the MARCH cohort had identified a large number of clonal haematopoiesis cases with a spectrum of driver mutations and varied clonal sizes. In order to obtain sufficient mutant cells for meaningful analysis, it was necessary to focus on samples with a sufficiently high VAF, greater than 0.04 (equivalent to 8% mutant cells). Of the 38 samples with *DNMT3A* mutations in the cohort, 14 had a VAF greater than 0.04. Furthermore, to ensure the cases chosen had mutations with an impact on protein function and to increase comparability across samples, cases with truncating mutations were chosen. There were eight of these, and the details of the mutations present in the five cases analysed are shown in Table 5.1.

Three of the cases had a single *DNMT3A* nonsense or frameshift mutation and one case had two *DNMT3A* mutations. The last case had a *DNMT3A* mutation at very high VAF with an *SF3B1* hotspot mutation and a *BCORL1* missense variant

Sample	Sex	Age	Gene	Consequence	Exon	cDNA change	Protein change	BM VAF	PB VAF	Genotyping	Germline SNP
NOC108	Female	80	DNMT3A	Nonsense	exon9	c.1084C>T	p.Q362X	0.072	0.021	gDNA & cDNA	-
NOC132	Male	72	DNMT3A	Nonsense	exon19	c.2311C>T	p.R771X	0.128	0.249	gDNA & cDNA	rs2289195
NOC171	Female	62	DNMT3A	Frameshift insertion	exon15	c.1740dupC	p.W581fs	0.230	0.233	gDNA & cDNA	-
NOC131	Female	55	DNMT3A	Nonsense	exon15	c.1816C>T	p.Q606X	0.061	0.045	gDNA & cDNA	rs2289093
			DNMT3A	Missense	exon20	c.2339T>C	p.I780T	0.031	0.033	gDNA & cDNA	-
NOC062	Male	86	DNMT3A	Frameshift insertion	exon23	c.2669dupG	p.G890fs	0.366	0.428	gDNA & cDNA	-
			SF3B1	Missense	exon15	c.2098A>G	p.K700E	0.014	0.008	gDNA & cDNA	-
			BCORL1	Missense (VUS)	exon3	c.3185T>C	p.L1062S	0.013	0.004	gDNA only	-

Table 5.1: *DNMT3A*-mutant samples analysed by TARGET-seq+. Details of *DNMT3A*-mutant samples showing the mutations identified and the VAF in BM and PB by bulk targeted DNA sequencing. The last two columns indicate whether single-cell genotyping was performed on gDNA and cDNA and the germline SNP used for allelic dropout detection if one was present. VUS: Variant of unknown significance.

of unknown significance, both at a VAF < 0.02. The median age was 72 (range 55–86), similar to the cohort overall and there were three females and two males.

For *TET2*, there were 10 samples in the cohort with a mutation of VAF greater than 0.04. Five of these were chosen, prioritising those with the highest VAF (Table 5.2. Four samples had *TET2*-truncating mutations, of which one had no additional mutations, two had a second *TET2* mutation, and one had an *ASXL1* mutation. Finally, one case had both *TET2* and *DNMT3A* missense mutations.

Sample	Sex	Age	Gene	Consequence	Exon	cDNA change	Protein change	BM VAF	PB VAF	Genotyping	Germline SNP
NOC031	Female	65	*TET2*	Frameshift insertion	exon3	c.2286dupC	p.H762fs	0.080	0.091	gDNA only	-
NOC117	Female	89	*TET2*	Frameshift deletion	exon3	c.622delC	p.P208fs	0.130	0.043	gDNA only	rs17253672
			TET2	Missense	exon5	c.3581C>A	p.P1194H	0.016	0.012	gDNA only	rs55838312
NOC002	Female	77	*TET2*	Nonsense	exon3	c.2176C>T	p.Q726X	0.162	NA	gDNA only	-
			TET2	Missense	exon6	c.3781C>T	p.R1261C	0.146	NA	gDNA & cDNA	-
			ATM	Missense	exon43	c.6218T>C	p.L2073P	0.014	NA	gDNA & cDNA	-
NOC115	Male	77	*TET2*	Missense	exon6	c.3784C>T	p.R1262W	0.101	0.175	gDNA & cDNA	-
			DNMT3A	Missense	exon17	c.1959G>C	p.L653F	0.029	0.007	gDNA & cDNA	-
NOC137	Male	73	*TET2*	Nonsense	exon6	c.3646C>T	p.R1216X	0.041	0.030	gDNA & cDNA	-
			ASXL1	Nonsense	exon13	c.2893C>T	p.R965X	0.141	0.163	gDNA only	-

Table 5.2: *TET2*-mutant samples analysed by TARGET-seq+. Details of *TET2*-mutant samples showing the mutations identified and the VAF in BM and PB by bulk targeted DNA sequencing. The last two columns indicate whether single-cell genotyping was performed on gDNA and cDNA and the germline SNP used for allelic dropout detection if one was present. NA: Sample not available.

The median age for *TET2*-mutant cases was 77 (range 65–89) and there were three females and two males.

As controls, three normal samples with no mutations identified in targeted DNA sequencing were selected to be age matched as far as possible (aged 63, 75 and 80) with the CH cases.

5.2 Single-cell genotyping analysis of clonal haematopoiesis samples

5.2.1 Data QC and filtering

In total, 17,100 cells were sorted and genotyped across the 10 CH and 3 control samples. Plates from one of the FACS sorts had a high failure rate (approximately 75% cells failed genotyping), presumably due to a problem with sorter alignment and were excluded from the analysis (Figure 5.2). After exclusion of these plates and

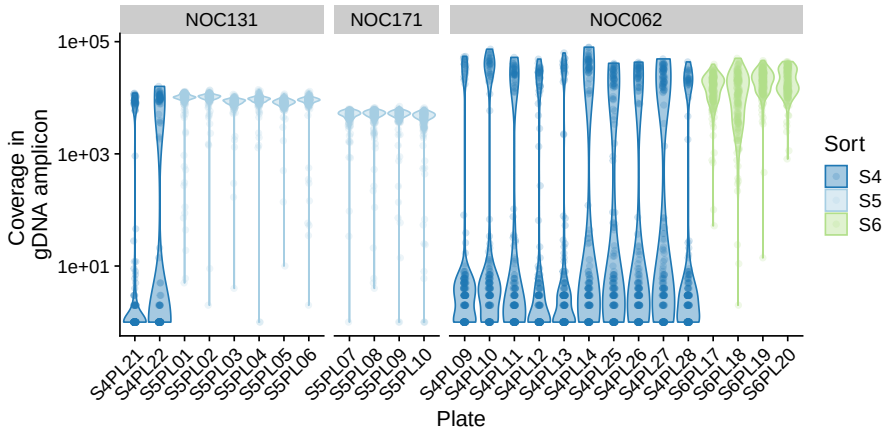


Figure 5.2: Coverage of gDNA amplicons in plates across FACS sorts. Coverage in *DNMT3A* gDNA amplicons for three samples by plate. Each dot represents a cell (n = 192 per plate). High failure rate of cells in Sort S4 (dark blue) is evident as a large proportion of cells with a low number of reads. These were excluded from analysis and cell sorting for these samples was repeated in S5 and S6.

no template controls, there was data for 14,250 single cells. Overall, a genotype was obtained for at least one mutation in 97.2% of cells analysed from the CH samples (Figure 5.3). As described in the last chapter, the success rate for gDNA amplicons was higher than for cDNA amplicons. The frequency of mutant cells identified for each mutation was consistent with that observed in bulk BM DNA sequencing. The

outliers were all explained by a difference in contribution of mutant cells to CD34+ cells relative to Lin+ cells, which make up the majority of bulk BM MNCs.

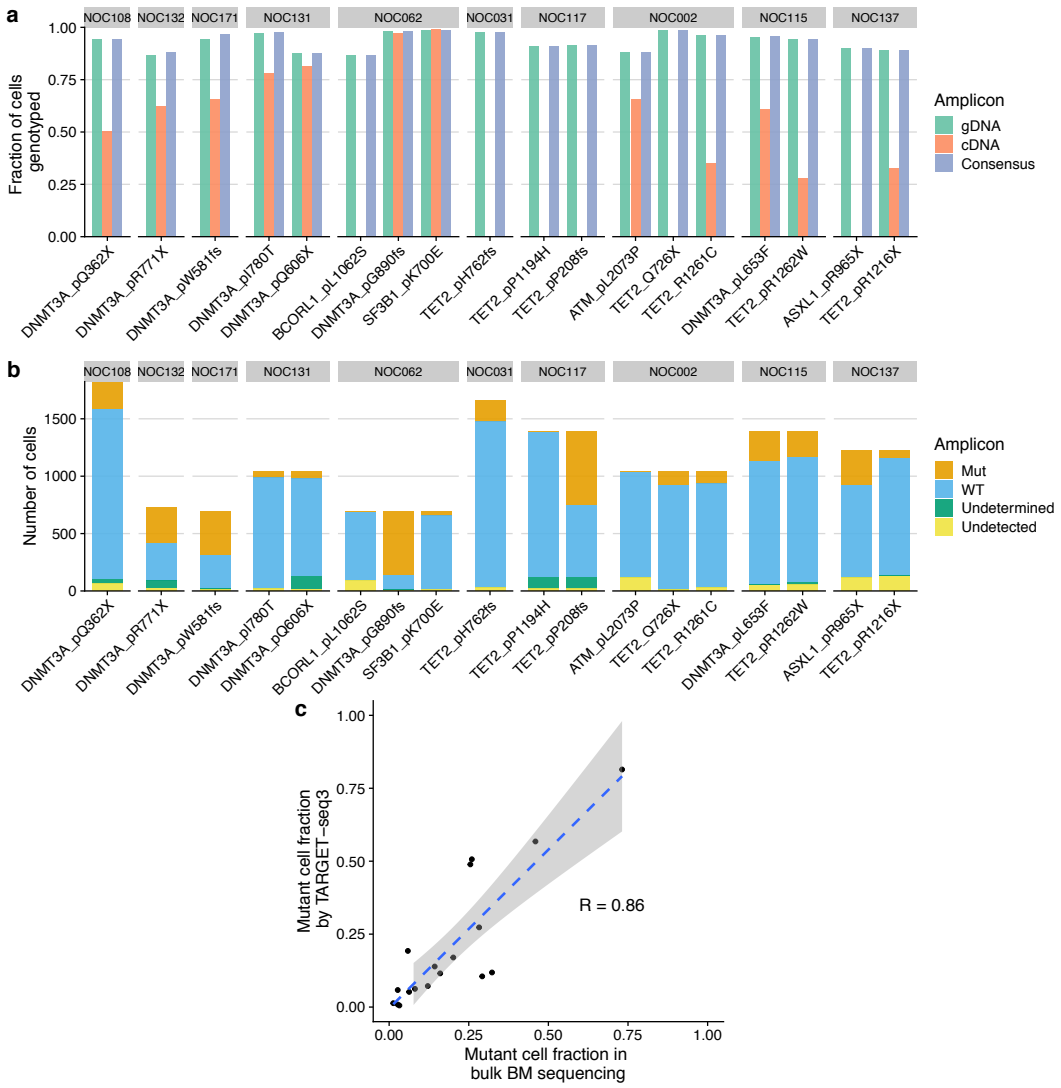


Figure 5.3: Single-cell genotyping of CH samples. **(a)** Fraction of cells with successful genotyping readout according to the mutation and sample. The number of cells analysed for each sample is shown in **(b)**. The consensus genotype is derived from combining gDNA and cDNA calls as described in Chapter 2. **(b)** Number of cells assigned a genotype for each mutation. Note that the fraction of cells with undetermined genotype is highest for *DNMT3A* pR771X, *DNMT3A* pQ606X, *TET2* pP1194H and *TET2* pP208fs gDNA amplicons because for these a germline SNP was present, enabling alleic dropout of the mutant allele to be identified. **(c)** Mutant cell fraction for each of the mutations identified in single cells by TARGET-seq+ versus the clonal cell fraction as measured in bulk BM DNA sequencing. The linear regression line is shown and R indicates the Pearson correlation coefficient.

5.2.2 Clonal analysis of immunophenotypic populations for *DNMT3A*-mutant samples

Integration of single cell genotyping with flow cytometry index data enabled the clonal contribution to immunophenotypically-defined HSPC populations to be analysed. First, the three samples with a single *DNMT3A* truncating mutation were analysed (Figures 5.4, 5.5 and 5.6). In all cases, the *DNMT3A*-mutant clone was identified within the HSC compartment, where mutant cells comprised between 10% (in NOC108) and 57% (in NOC171) of cells (panel **(a)**). This was also the case within the CD90+CD49f+ subpopulation which enriches for long-term repopulating stem cells (LT-HSCs) (panel **(c)**, far right) [272].

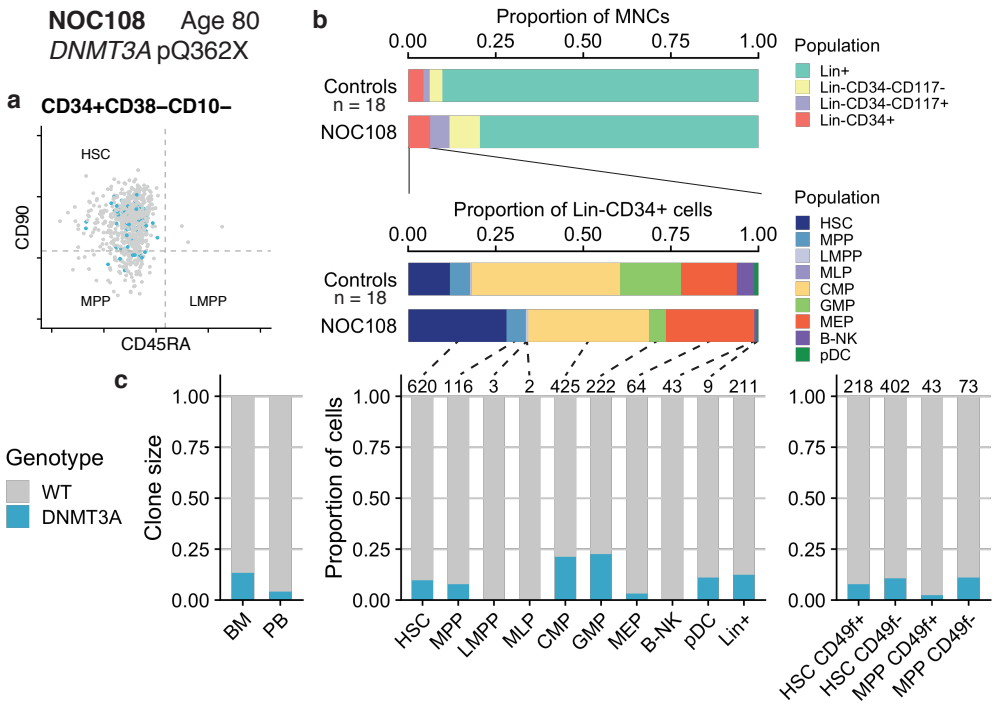


Figure 5.4: Frequency of *DNMT3A*-mutant cells in NOC108 HSPC compartments. **(a)** Flow cytometry plot of index expression data for CD45RA and CD90 in WT (grey) and *DNMT3A*-mutant (blue) cells showing presence of mutant cells within the HSC compartment. **(b)** Compartment sizes as determined by flow cytometry. The top bars show compartments as a proportion of total BM MNCs, and the bottom show HSPC compartments within Lin–CD34+ cells. Data from the sample are compared with the median population sizes from 18 age-matched control samples. **(c)** Fraction of mutant cells within each of the HSPC compartments and Lin+ cells (middle), and in HSC/MPP populations split by CD49f expression (right). The number of cells analysed within each compartment are shown above. For comparison, the clone size from bulk BM and PB sequencing is shown (left).

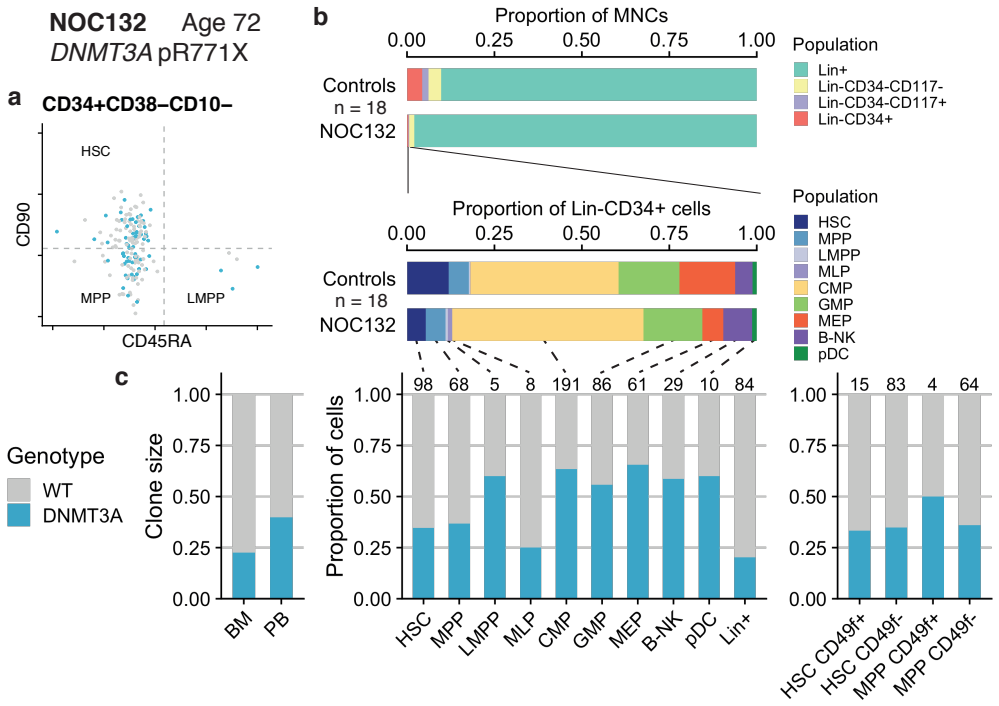


Figure 5.5: Frequency of *DNMT3A*-mutant cells in NOC132 HSPC compartments. **(a)** Flow cytometry plot of index expression data for CD45RA and CD90 in WT (grey) and *DNMT3A*-mutant (blue) cells showing presence of mutant cells within the HSC compartment. **(b)** Compartment sizes as determined by flow cytometry. The top bars show compartments as a proportion of total BM MNCs, and the bottom show HSPC compartments within Lin–CD34+ cells. Data from the sample are compared with the median population sizes from 18 age-matched control samples. **(c)** Fraction of mutant cells within each of the HSPC compartments and Lin+ cells (middle), and in HSC/MPP populations split by CD49f expression (right). The number of cells analysed within each compartment are shown above. For comparison, the clone size from bulk BM and PB sequencing is shown (left).

By comparing the frequency of mutant cells with the overall compartment size, the patterns of clonal expansion could be determined (Figures 5.4, 5.5 and 5.6b-c). Note that while smaller subpopulations were enriched during FACS sorting, the frequency of each population in unenriched BM samples was known from previous flow cytometry analysis. The size of HSPC compartments was compared with data from the 18 age-matched controls presented in Chapter 3. There was substantial heterogeneity between the samples studied. In NOC108, the oldest of the three individuals, there was a relative expansion of the CD34+CD38– compartment, particularly of HSC, yet mutant cells made up only 10% of this population. The clone size was larger in CMP and GMP myeloid progenitors, where it constituted

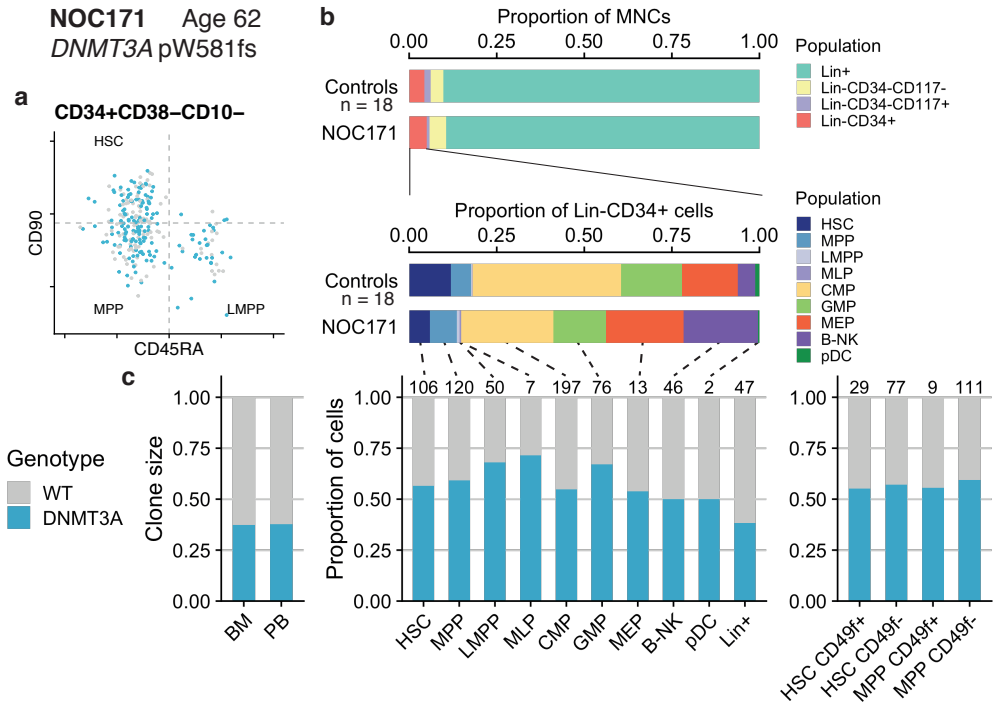


Figure 5.6: Frequency of *DNMT3A*-mutant cells in NOC171 HSPC compartments. **(a)** Flow cytometry plot of index expression data for CD45RA and CD90 in WT (grey) and *DNMT3A*-mutant (blue) cells showing presence of mutant cells within the HSC compartment. **(b)** Compartment sizes as determined by flow cytometry. The top bars show compartments as a proportion of total BM MNCs, and the bottom show HSPC compartments within Lin–CD34+ cells. Data from the sample are compared with the median population sizes from 18 age-matched control samples. **(c)** Fraction of mutant cells within each of the HSPC compartments and Lin+ cells (middle), and in HSC/MPP populations split by CD49f expression (right). The number of cells analysed within each compartment are shown above. For comparison, the clone size from bulk BM and PB sequencing is shown (left).

about 20% of cells and was almost absent from MEP and lymphoid progenitors. In contrast, in NOC132 and NOC171 the CD34+CD38– compartment was smaller than in controls; particularly in NOC132 the Lin–CD34+ frequency was only 10% of the average size in control samples. In both these samples, the clone size was large, contributing to over half of myeloid progenitor cells. Notably, the clone size was larger in stem and progenitor cells than in differentiated Lin+ cells. This explained why more mutant cells were identified for NOC132 than had been predicted from the BM VAF.

Next, data from the two samples with multiple mutations were analysed. By

analysing the co-occurrence of mutations in single cells, the clonal structure could be identified. NOC131 had two separate *DNMT3A* mutations; one nonsense mutation and one missense mutation located in the methyltransferase domain of the protein, at VAFs of 0.061 and 0.031 respectively in bulk BM sequencing. Notably, the I780T missense mutation was recently shown to cause *DNMT3A* loss of function through reduced protein stability [162]. Single-cell genotyping showed that these two mutations were never present in the same cell, and had therefore occurred independently creating two separate clones (Figure 5.7a,b). Similar to NOC108, this sample had a relatively large HSC compartment where the two clones constituted 7% and 4% of cells respectively. Similarly to NOC108, both clones expand to about twice that size in CMP. Notably, the two clones followed similar patterns with no obvious clone-specific effects in terms of contribution to immature or mature cells.

Sample NOC062 had a very large *DNMT3A*-mutant clone, with a VAF of 0.366 in BM and 0.428 in PB, as well as *SF3B1* and *BCORl1* mutations at low VAF. While the VAF of the *DNMT3A* mutation might be consistent with a germline variant, the mutation was not detected in sequencing of hair follicle DNA. Overall, the *DNMT3A* mutation was detected in 81% of cells. Whilst some mutant cells will be missed due to allelic dropout, the dropout rate for this mutation was estimated to be <6% (Figure 4.14), suggesting a fraction of BM cells in this sample were truly WT. *SF3B1* and *BCORl1*-mutant cells constituted two separate subclones within the larger *DNMT3A*-mutant clone (Figure 5.8a,b). The Lin–CD34+ compartment for NOC062 was half the median size in controls, but HSC were relatively expanded within this. The *DNMT3A*-mutant clone size was 72–90% in all the compartments analysed. The *SF3B1*-mutant clone was also identified in the HSC compartment and downstream progenitors. *BCORL1*-mutant cells were not identified in HSC but this is likely due to the small number of cells present with this genotype.

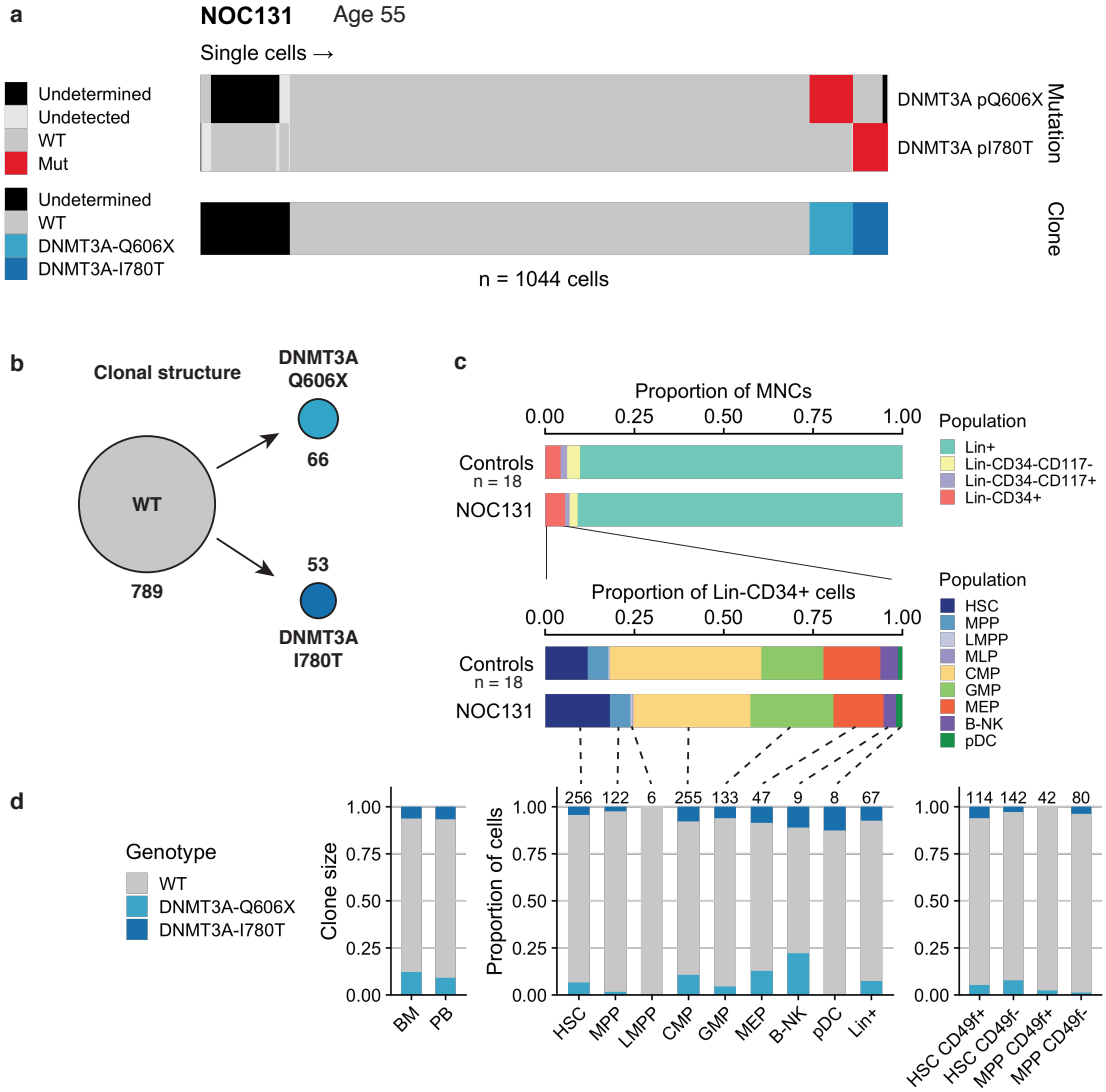


Figure 5.7: Clonal hierarchy and frequency of mutant cells in NOC131 HSPC compartments. (a) Top: Matrix of mutational sequencing results in NOC131 single cells. Each mutation is in a row and each column is a single cell (n = 1044 single cells analysed). Colours indicate whether a cell was called mutant, WT, undetected or undetermined. Bottom: Clone assignment based on the occurrence of the two mutations. Cells with either mutation were assigned as mutant; those where both loci were detected and called WT were assigned as WT; cells where either mutation could not be called were assigned as undetermined. **(b)** Clonal structure based on the results in **(a)**. The number of cells in each clone are shown. **(c)** Compartment sizes as determined by flow cytometry. The top bars show compartments as a proportion of total BM MNCs, and the bottom show HSPC compartments within Lin–CD34+ cells. Data from the sample are compared with the median population sizes from 18 age-matched control samples. **(d)** Fraction of mutant cells within each of the HSPC compartments and Lin+ cells (middle), and in HSC/MPP populations split by CD49f expression (right). The number of cells analysed within each compartment are shown above. For comparison, the clone size from bulk BM and PB sequencing is shown (left).

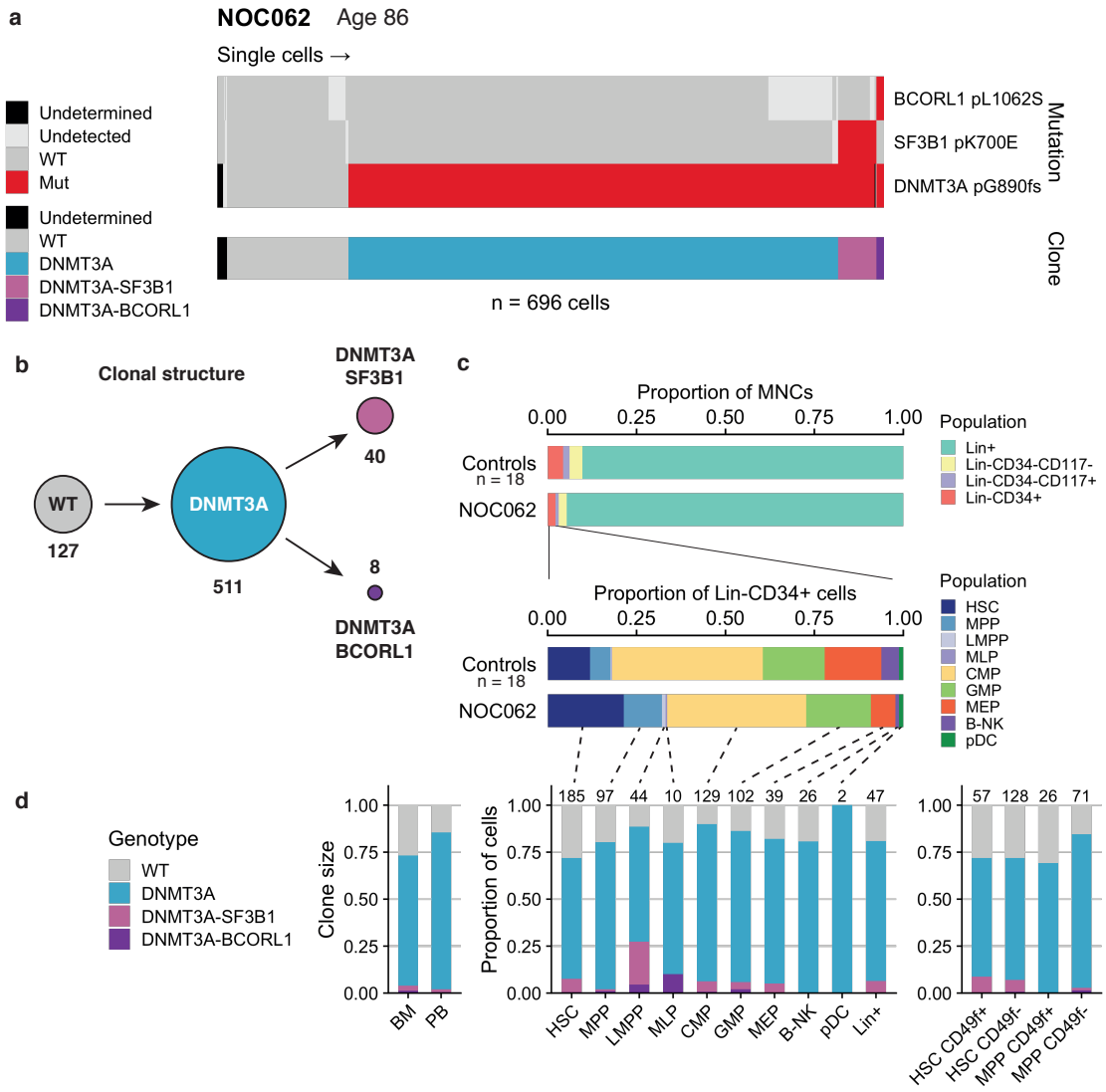


Figure 5.8: Clonal hierarchy and frequency of mutant cells in NOC062 HSPC compartments. **(a)** Top: Matrix of mutational sequencing results in NOC062 single cells. Each mutation is in a row and each column is a single cell (n = 696 single cells analysed). Colours indicate whether a cell was called mutant, WT, undetected or undetermined. Bottom: Clone assignment based on the occurrence of mutations. Cells with either mutation were assigned as mutant; those where *DNMT3A* was called WT were assigned as WT; cells where *DNMT3A* could not be called were assigned as undetermined. **(b)** Clonal structure based on the results in **(a)**. **(c-d)** Compartment sizes and fraction of mutant cells within each compartment as described in Figure 5.7.

5.2.3 Surface marker expression in *DNMT3A*-mutant populations

Linking flow cytometry index data to the genotype of single cells enabled differential expression of cell surface markers to be determined. No consistent immunophenotype distinguishing mutant cells from WT cells was identified amongst the markers used in the panel. The most consistent change was an increase in CD34 expression in various HSPC compartments, which was observed in four of the five samples (Figure 5.9). However, changes were modest and not consistently observed in any one compartment across samples.

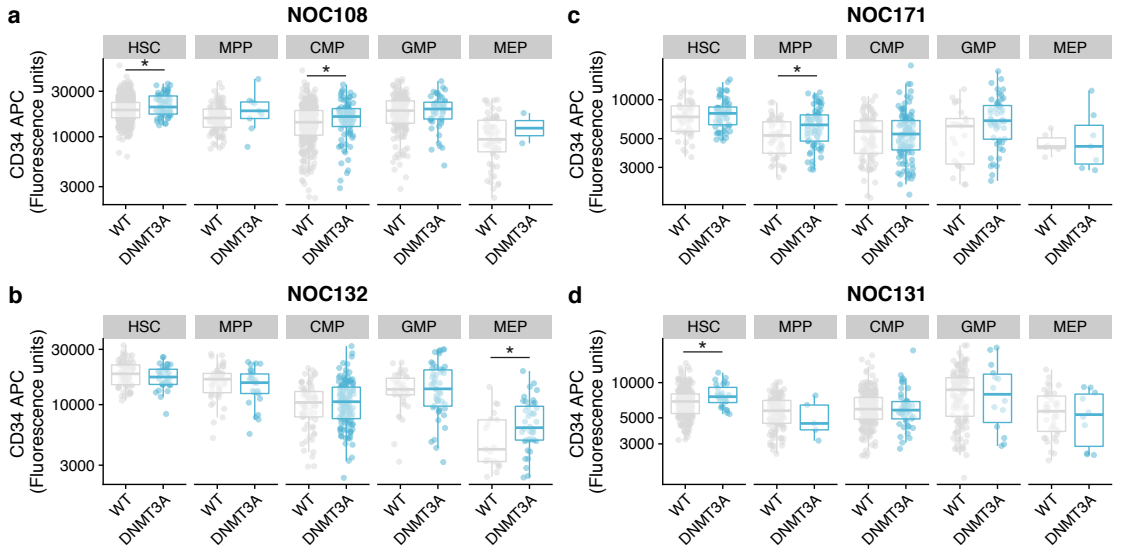


Figure 5.9: CD34 expression in *DNMT3A*-mutant versus WT cells. Expression of CD34 measured by flow cytometry indexing in WT and *DNMT3A*-mutant cells across selected HSPC populations. Each dot represents a single cell with the genotype shown. * $P < 0.05$ by Wilcoxon rank-sum test.

5.2.4 Clonal analysis of immunophenotypic populations for *TET2*-mutant samples

Next, the clonal contribution across HSPC populations was analysed in the *TET2*-mutant samples. Sample NOC031 had a single *TET2* frameshift mutation in exon 3 with a VAF of 0.08 in bulk BM and 0.09 in PB. This clone was identified in 11% of the immunophenotypic HSC compartment, which was slightly expanded (Figure 5.10). The clone expanded further within GMP and pDC where mutant cells made up 20% and of 38% of cells respectively.

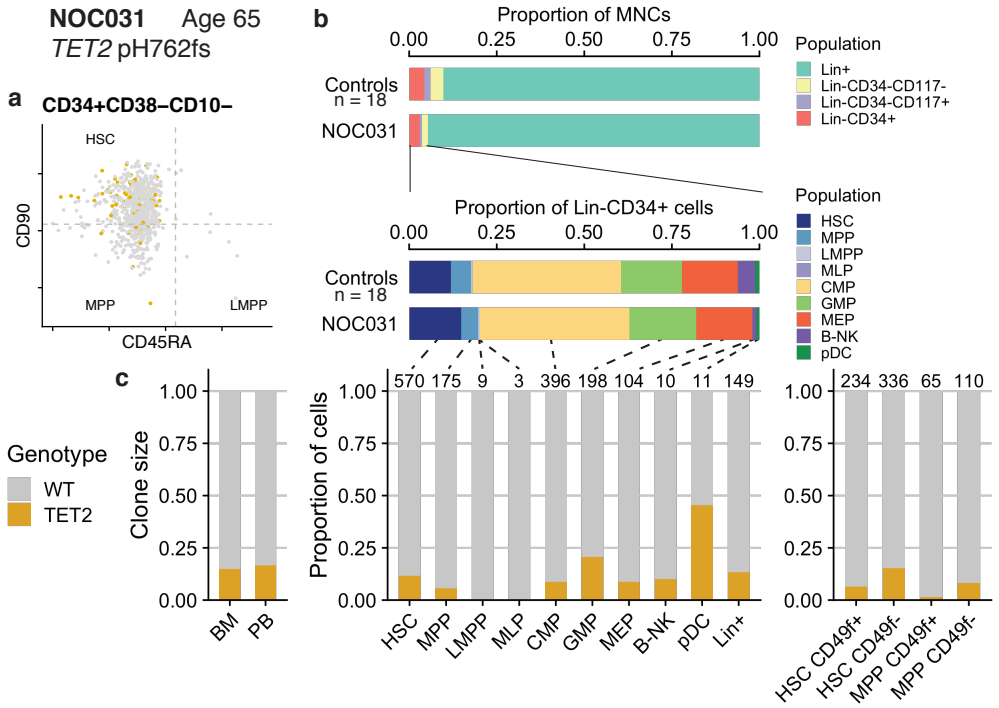


Figure 5.10: Frequency of *TET2*-mutant cells in NOC031 HSPC compartments. **(a)** Flow cytometry plot of index expression data for CD45RA and CD90 in WT (grey) and *TET2*-mutant (blue) cells showing presence of mutant cells within the HSC compartment. **(b)** Compartment sizes as determined by flow cytometry. The top bars show compartments as a proportion of total BM MNCs, and the bottom show HSPC compartments within Lin–CD34+ cells. Data from the sample are compared with the median population sizes from 18 age-matched control samples. **(c)** Fraction of mutant cells within each of the HSPC compartments and Lin+ cells (middle), and in HSC/MPP populations split by CD49f expression (right). The number of cells analysed within each compartment are shown above. For comparison, the clone size from bulk BM and PB sequencing is shown (left).

Sample NOC117 had two *TET2* mutations, one was a frameshift mutation in exon 3 with a VAF of 0.13 in BM and the other was missense mutation with a VAF of 0.016. Single cell genotyping showed that these mutations occurred in separate clones (Figure 5.11). The *TET2*-P208fs clone was found in 33% of HSC and expanded to a clone size of 84% in GMP. The *TET2*-P1194H mutation was detected in only 7 (0.5%) of the 1392 cells genotyped. While this was lower than expected from the bulk BM sequencing, the frequency within Lin+ cells was 1.4%, which is consistent with the bulk sequencing data. The mutation was detected in a single HSC (corresponding to a clone size of 0.2%, n = 474 HSC genotyped),

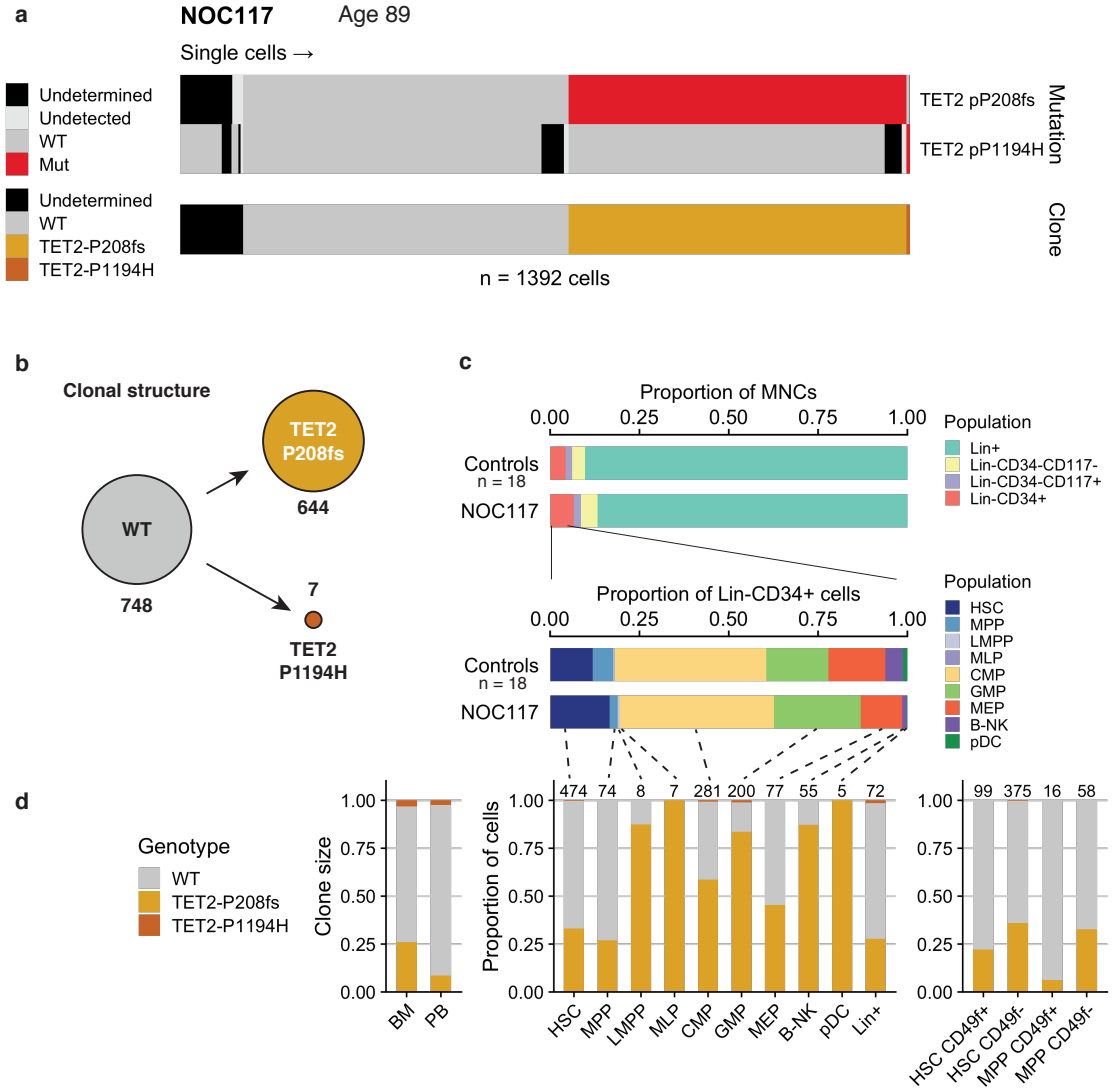


Figure 5.11: Clonal hierarchy and frequency of mutant cells in NOC117 HSPC compartments. (a) Top: Matrix of mutational sequencing results in NOC117 single cells. Each mutation is in a row and each column is a single cell (n = 1392 single cells analysed). Colours indicate whether a cell was called mutant, WT, undetected or undetermined. Bottom: Clone assignment based on the occurrence of the two mutations. Cells with either mutation were assigned as mutant; those where both loci were detected and called WT were assigned as WT; cells where *TET2* pP1194 could not be called, but *TET2* pP208 was WT (with no allelic dropout) were assigned as WT. **(b)** Clonal structure based on the results in **(a)**. The number of cells in each clone are shown. **(c)** Compartment sizes as determined by flow cytometry. The top bars show compartments as a proportion of total BM MNCs, and the bottom show HSPC compartments within Lin–CD34+ cells. Data from the sample are compared with the median population sizes from 18 age-matched control samples. **(d)** Fraction of mutant cells within each of the HSPC compartments and Lin+ cells (middle), and in HSC/MPP populations split by CD49f expression (right). The number of cells analysed within each compartment are shown above. For comparison, the clone size from bulk BM and PB sequencing is shown (left).

suggesting the clone size was larger in mature cells than the stem cell compartment.

Sample NOC002 also had two *TET2* mutations identified; one was an exon 3 nonsense mutation with a VAF of 0.162 and the other a missense mutation in the Cys-rich domain with a VAF of 0.146 in bulk BM. The R1261C mutation is notable in that this amino acid residue forms part of the TET2 catalytic core where it stabilises 2-oxoglutarate, and mutation of this residue has been shown to reduce enzymatic activity [186]. Additionally, a missense mutation in the cell cycle checkpoint kinase *ATM* was identified with a VAF of 0.013. Single-cell genotyping showed co-occurrence of all three mutations within the same cells (Figure 5.12a), suggesting a linear clonal structure. Most *TET2*-mutant cells had both mutations, and a subset had one or the other. Within bulk BM sequencing, *TET2*-PQ726X had the higher VAF, which suggested this mutation occurred first. This was consistent with single-cell sequencing where more cells were detected with *TET2*-PQ726X than *TET2*-R1261C. Of 1044 cells genotyped, 29 cells (2.8%) had just *TET2*-PQ726X detected, and 13 (1.2%) had just *TET2*-R1261C. The latter group, where *TET2*-R1261C was detected but *TET2*-PQ726X was not, made up 9.6% of all *TET2*-mutant cells, which was in line with the estimated allelic dropout rate for *TET2*-PQ726X of 11%, so it was inferred that these cells were in fact double-mutant. Cells were therefore annotated with a linear clonal structure, where *TET2*-PQ726X was the founder mutation, *TET2*-R1261C was then acquired creating a double-mutant subclone, and finally *ATM*-L2073P was acquired within that subclone (Figure 5.12b). Flow cytometry analysis showed the HSC compartment was approximately twice the size of age-matched controls, there were very few CD10+ cells and a substantial expansion of the $CD123^{hi}$ pDC compartment. Within HSC, 3% of cells were mutant, of which 5 cells had a single *TET2* mutation and 3 were *TET2*-double mutant. However, there was a striking expansion of the double-mutant clone progressing from CMP to GMP where the size of this clone was 6.7% and 40% respectively. This suggests the acquisition of the second *TET2*-mutation conferred a significant clonal advantage within the myeloid lineage. *ATM*-mutant cells were detected in CMP, GMP and

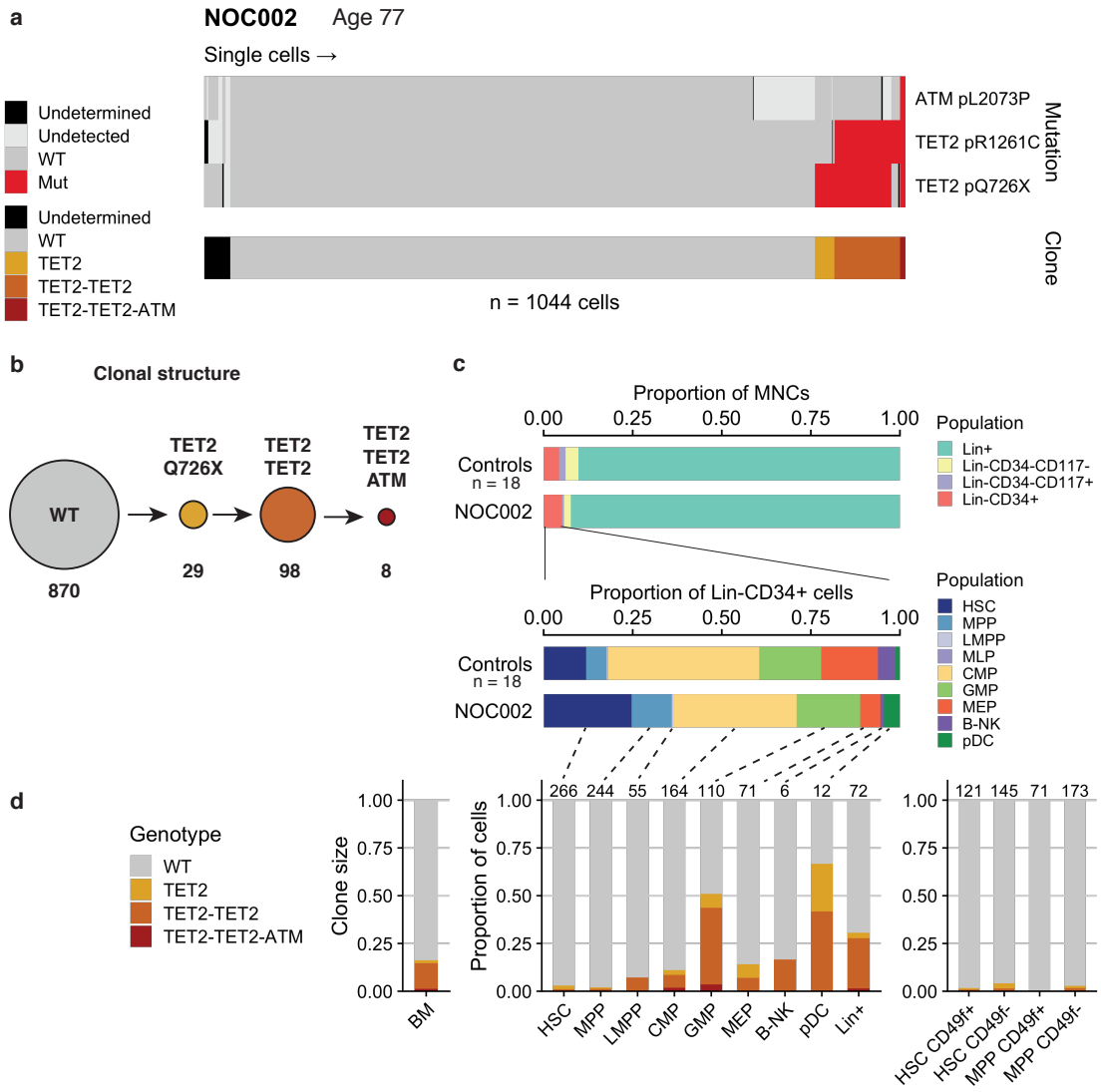


Figure 5.12: Clonal hierarchy and frequency of mutant cells in NOC002 HSPC compartments. **(a)** Top: Matrix of mutational sequencing results in NOC002 single cells. Each mutation is in a row and each column is a single cell (n = 1392 single cells analysed). Colours indicate whether a cell was called mutant, WT, undetected or undetermined. Bottom: Clone assignment based on the occurrence of the two mutations. Cells with either mutation were assigned as mutant; those where both loci were detected and called WT were assigned as WT; cells where either mutation could not be called were assigned as undetermined. **(b)** Clonal structure based on the results in **(a)**. **(c-d)** Compartment sizes and fraction of mutant cells within each compartment as described in Figure 5.11.

Lin+ compartments and were not detected in CD38– cells, but the occurrence of this mutation in HSCs cannot be ruled out due to the small number of *TET2*-mutant cells analysed in this compartment.

Sample NOC115 had a *TET2* missense mutation in the Cys-rich domain of

the protein with a VAF of 0.101 in BM and a *DNMT3A* missense mutation in the methyltransferase domain with a VAF of 0.029. The *DNMT3A* L653F missense mutation occurs in the same residue as one which was recently shown to cause protein stability (L653W) [162]. Single-cell genotyping showed that the two mutations occurred in independent clones (Figure 5.13). Surprisingly, more *DNMT3A*-mutant cells were detected than *TET2*-mutant cells. This was explained by a large *DNMT3A*-mutant clonal fraction within the CD38– compartment, where the clone constituted 34% of HSC, compared to just 5.2% in Lin+ cells. The *TET2*-mutant clone showed an inverse pattern of expansion, contributing to 13% of HSC, but 31–36% of GMP and MEP, and 11% of Lin+ cells. Thus, while the NOC115 *TET2*-mutant clone showed a pattern of expansion that was similar to the other samples analysed, the pattern of expansion for the *DNMT3A*-mutant contrasted with that seen in other samples, in that the clone size decreased between HSC and downstream progenitors.

Finally, the NOC137 sample had an *ASXL1* nonsense mutation with a VAF of 0.141 in BM and a *TET2* nonsense mutation with a VAF of 0.041. Here, single-cell genotyping showed the *TET2*-mutant clone was a subclone within the *ASXL1*-mutant clone. Both clones were again identified within the HSC compartment, where in total they contributed 33% of cells. The *ASXL1* single-mutant and *ASXL1-TET2* double-mutant clones followed similar patterns of expansion, and the maximal clone size was found in the MEP compartment, where 54% of cells were mutant.

Overall, the data from *TET2*-mutant samples showed a consistent pattern of clonal expansion occurring within the HSC compartment, but reaching maximal clone size in GMP and/or pDC. Furthermore, within CD38– cells, the *TET2* clone was consistently larger within CD49f– than CD49f+ cells.

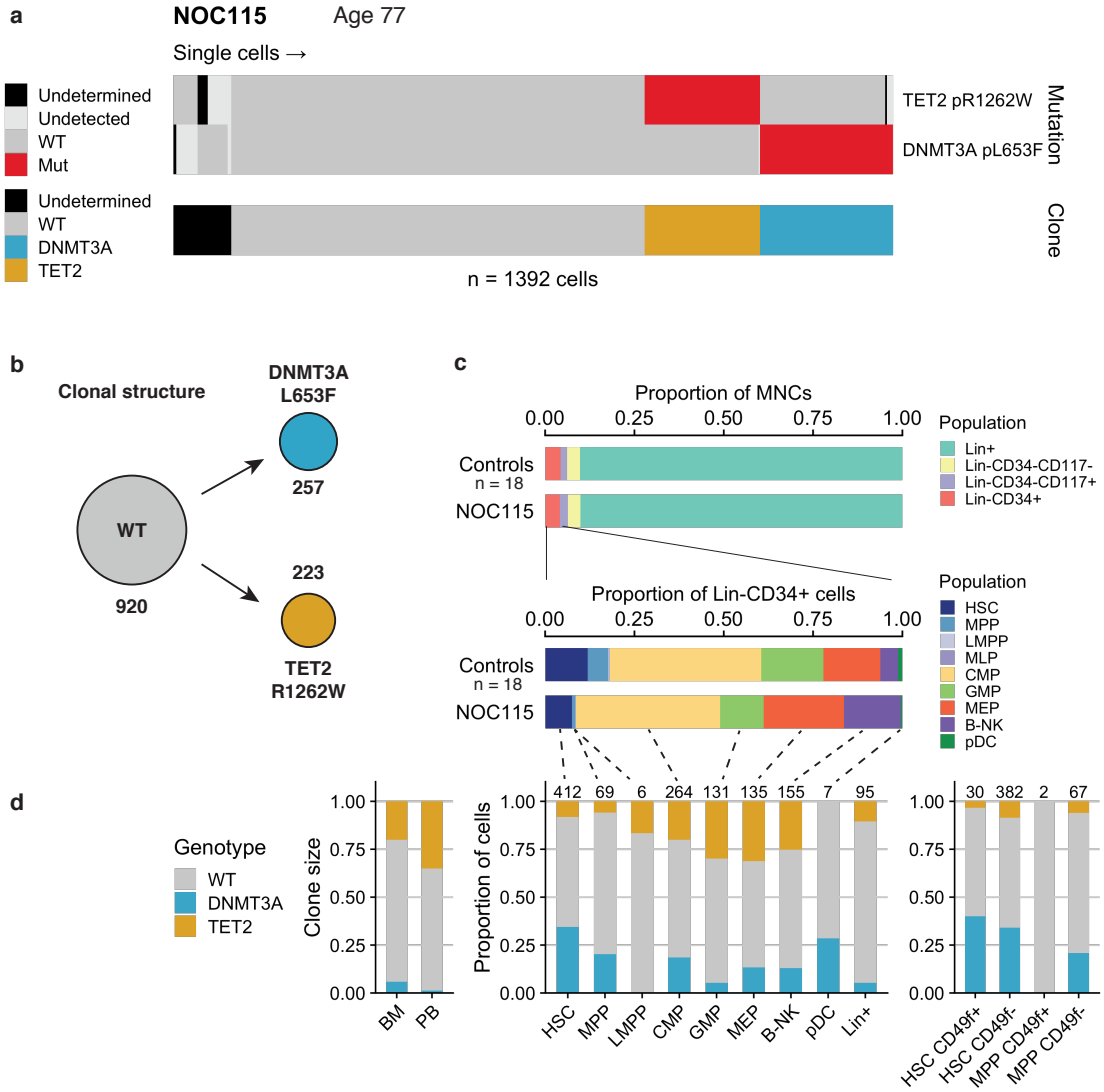


Figure 5.13: Clonal hierarchy and frequency of mutant cells in NOC115 HSPC compartments. **(a)** Top: Matrix of mutational sequencing results in NOC115 single cells. Each mutation is in a row and each column is a single cell (n = 1392 single cells analysed). Colours indicate whether a cell was called mutant, WT, undetected or undetermined. Bottom: Clone assignment based on the occurrence of the two mutations. Cells with either mutation were assigned as mutant; those where both loci were detected and called WT were assigned as WT; cells where either mutation could not be called were assigned as undetermined. **(b)** Clonal structure based on the results in **(a)**. **(c-d)** Compartment sizes and fraction of mutant cells within each compartment as described in Figure 5.13.

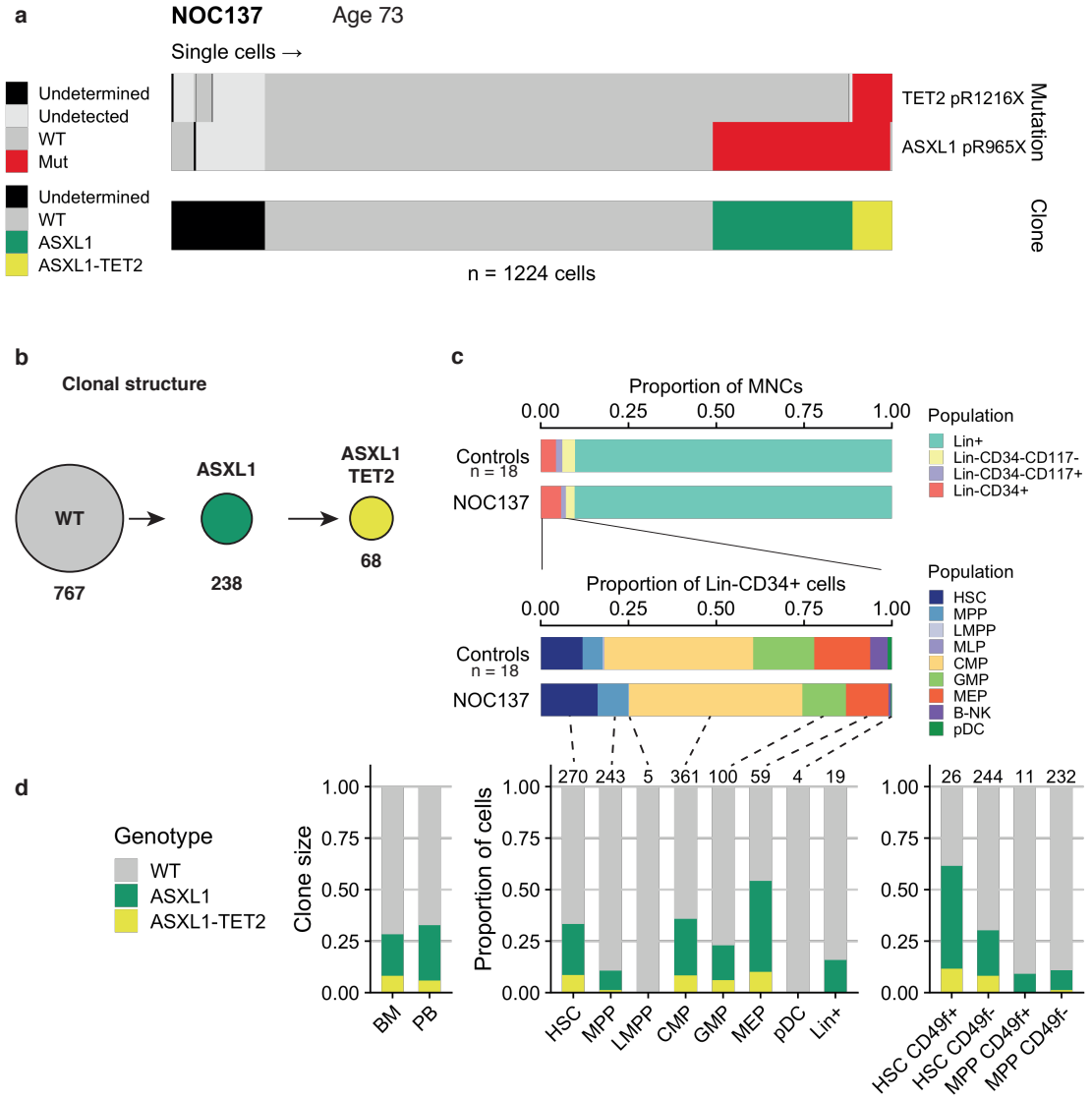


Figure 5.14: Clonal hierarchy and frequency of mutant cells in NOC137 HSPC compartments. **(a)** Top: Matrix of mutational sequencing results in NOC137 single cells. Each mutation is in a row and each column is a single cell (n = 1392 single cells analysed). Colours indicate whether a cell was called mutant, WT, undetected or undetermined. Bottom: Clone assignment based on the occurrence of the two mutations. Cells with either mutation were assigned as mutant; those where both loci were detected and called WT were assigned as WT; cells where either mutation could not be called were assigned as undetermined. **(b)** Clonal structure based on the results in **(a)**. **(c-d)** Compartment sizes and fraction of mutant cells within each compartment as described in Figure 5.14.

5.3 Single-cell transcriptome landscape of normal and clonal haematopoiesis samples

5.3.1 Data QC and filtering

Transcriptome libraries were sequenced, obtaining an average of 9.3×10^5 raw reads per cell. The plates with a high genotyping failure rate (described above) were excluded. Cells were filtered based on a minimum library size, the number of genes detected, and the percentage of reads mapping to mitochondrial and ERCC genes (Figure 5.15a). Based on this, 85–95% of cells per sample passed filtering

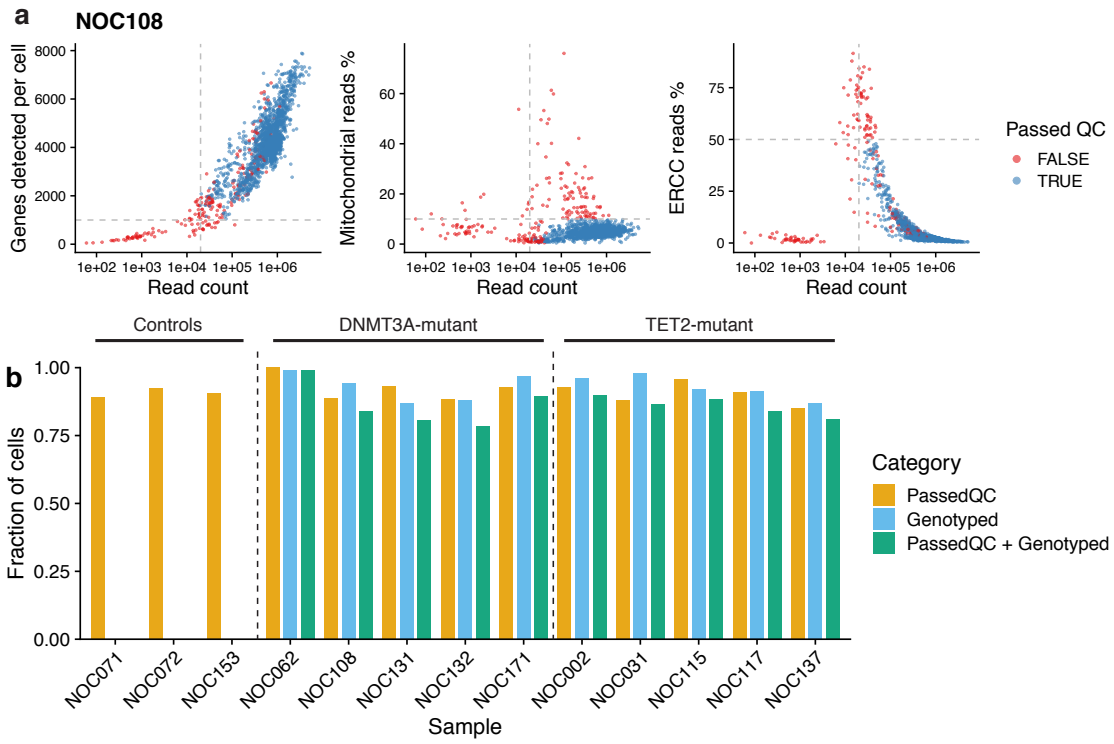


Figure 5.15: Quality control of RNA sequencing. (a) Example of cell filtering based on RNA-seq quality metrics. Cells are filtered with a minimum library size of 20,000 reads, minimum 1500 detected genes, maximum 10–14% mitochondrial reads (threshold individually set per sample) and maximum 50% ERCC reads. Dashed lines show the filtering thresholds. **(b)** Fraction of cells passing quality filtering by sample. For CH samples, the fraction of cells with a genotype called is also shown.

(Figure 5.15b). Of these, 87–99% had a genotype called. Cells with undetermined genotype were kept in the analysis, but excluded in all comparisons between WT

and mutant cells. Following filtering, there were a total of 12,847 cells across the 13 samples. Within the 10,562 filtered cells from the 10 CH samples, 6,854 were WT and 3,082 were mutant (107–590 per sample).

After filtering, an average of 4,750 genes were detected per cell (range 4,276–5,826 per sample) (Figure 5.16a,b). Despite pooling equal volumes of cDNA library from every cell, library size differed amongst populations likely due to differences in mRNA content. MEP progenitors consistently had larger library size and more genes detected per cell than other cell types (Figure 5.16c,d).

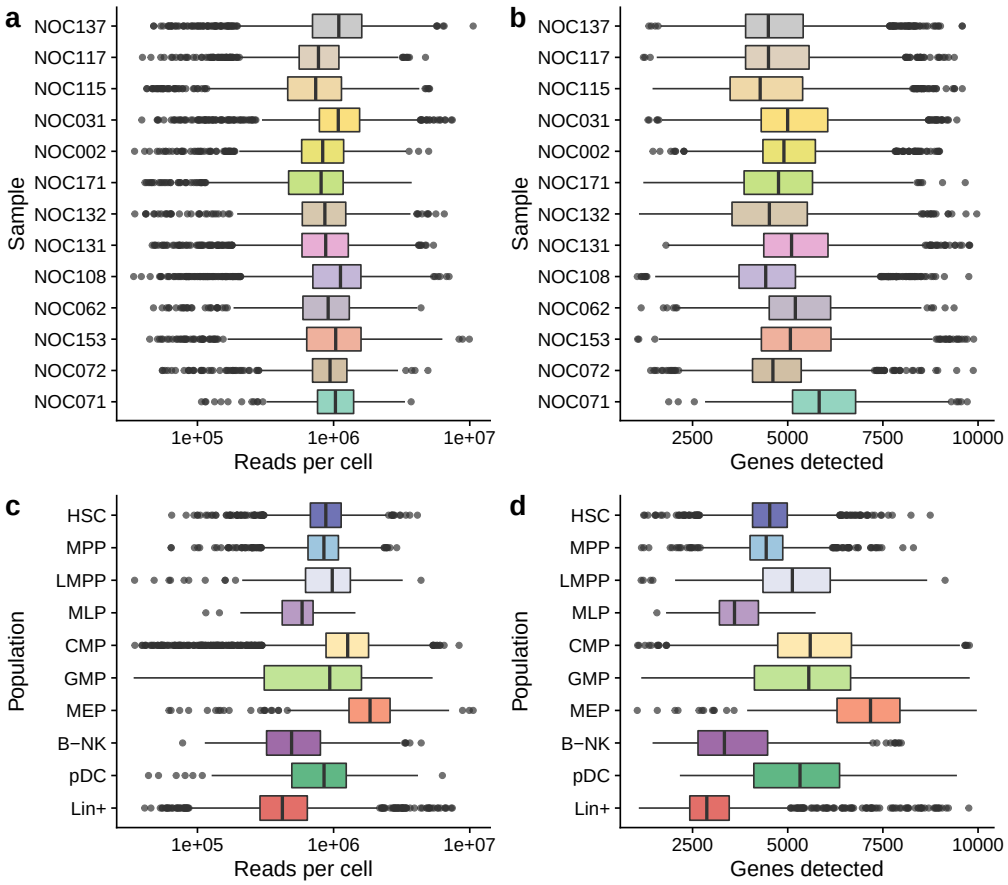


Figure 5.16: Library size and genes detected per cell. (a) Number of raw reads per cell by sample. **(b)** Number of genes detected per cell by sample. **(c)** Number of raw reads per cell by cell type, as defined by flow cytometry immunophenotype. **(b)** Number of genes detected per cell by cell type.

5.3.2 Copy number variation analysis

In addition to single nucleotide variants, clonal haematopoiesis may be associated with mosaic chromosomal alterations such as deletions, duplications, or copy number neutral loss of heterozygosity [65–67]. To explore whether additional somatic genetic variation was present in HSPCs, copy number analysis was performed on the scRNA-seq data using the InferCNV package [273]. No significant copy number variants were identified across any of the CH samples (Figure 5.17).

5.3.3 Dimensionality reduction and clustering

Transcriptome data were initially analysed on a sample-by-sample basis to visualise the global transcriptional landscape in an unbiased fashion, without influence from mutational status. Gene expression data were normalised, corrected for the plate batch effect and highly variable genes (HVGs) identified for each sample based on modelling technical variation of the ERCC spike-ins. 3963–8970 highly variable genes were identified per sample. Dimensionality reduction using principal components analysis and UMAP visualisation were then performed.

Labelling cells with their immunophenotype and mutational status showed cells clustered by stage of haematopoietic differentiation and lineage across all the *DNMT3A*-mutant clonal haematopoiesis samples (Figure 5.18). Consistent with the findings from the clonal analysis of immunophenotypic populations (earlier in this chapter), *DNMT3A*-mutant cells were found throughout all HSPC populations interspersed with WT cells, and did not create any novel cell clusters. This was the case even in samples with a very large clone size such as NOC171 and NOC062.

Similarly, in *TET2*-mutant samples, cells were broadly separated by their immunophenotype with no novel clusters created by mutant cells (Figure 5.19). There was a high density of mutant cells visible in cell clusters overlapping with immunophenotypic GMP, consistent with the findings from integrating genotyping with flow cytometry data. This was particularly clear in the NOC117 and NOC002 samples. Additionally, in all samples there were Lin+ clusters that either had few or no *TET2*-mutant cells.

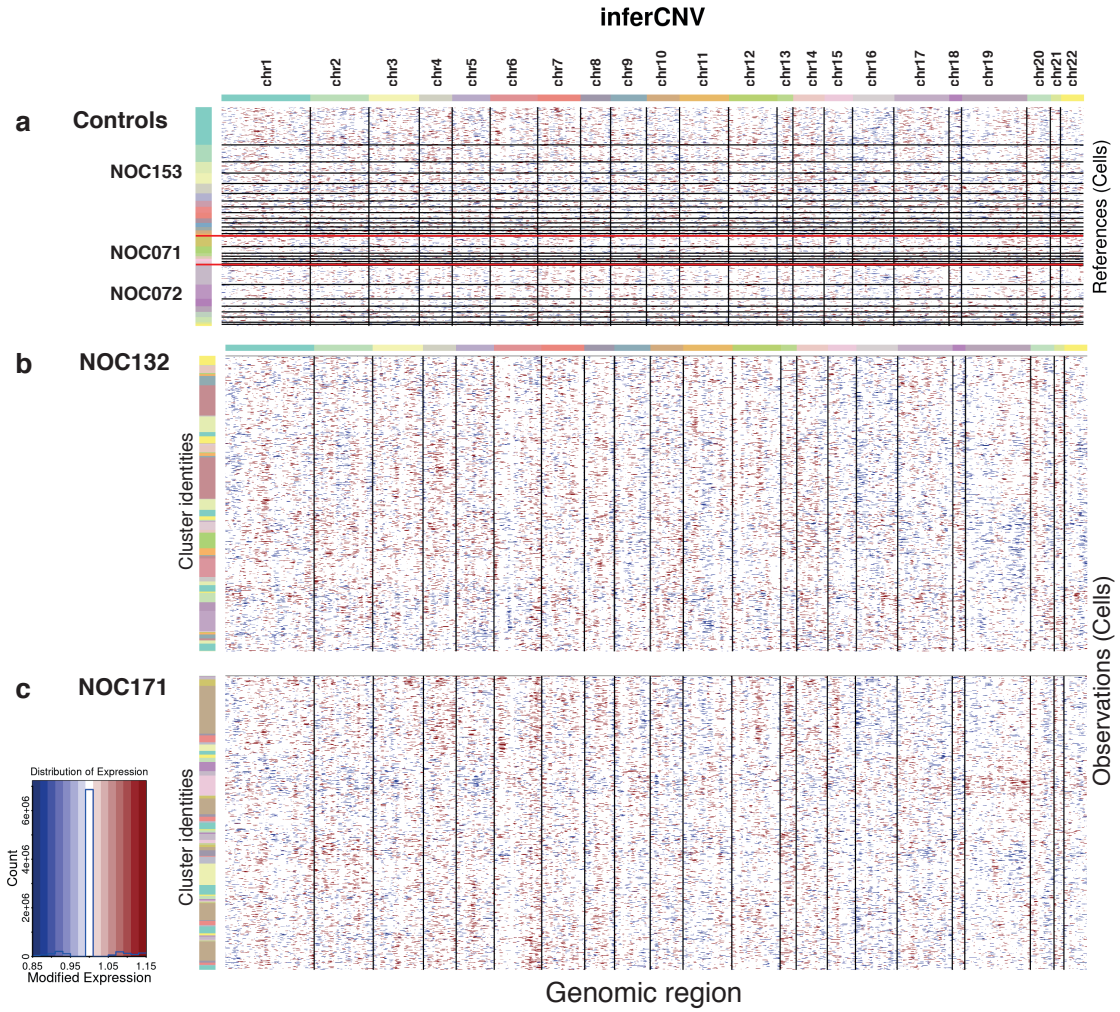
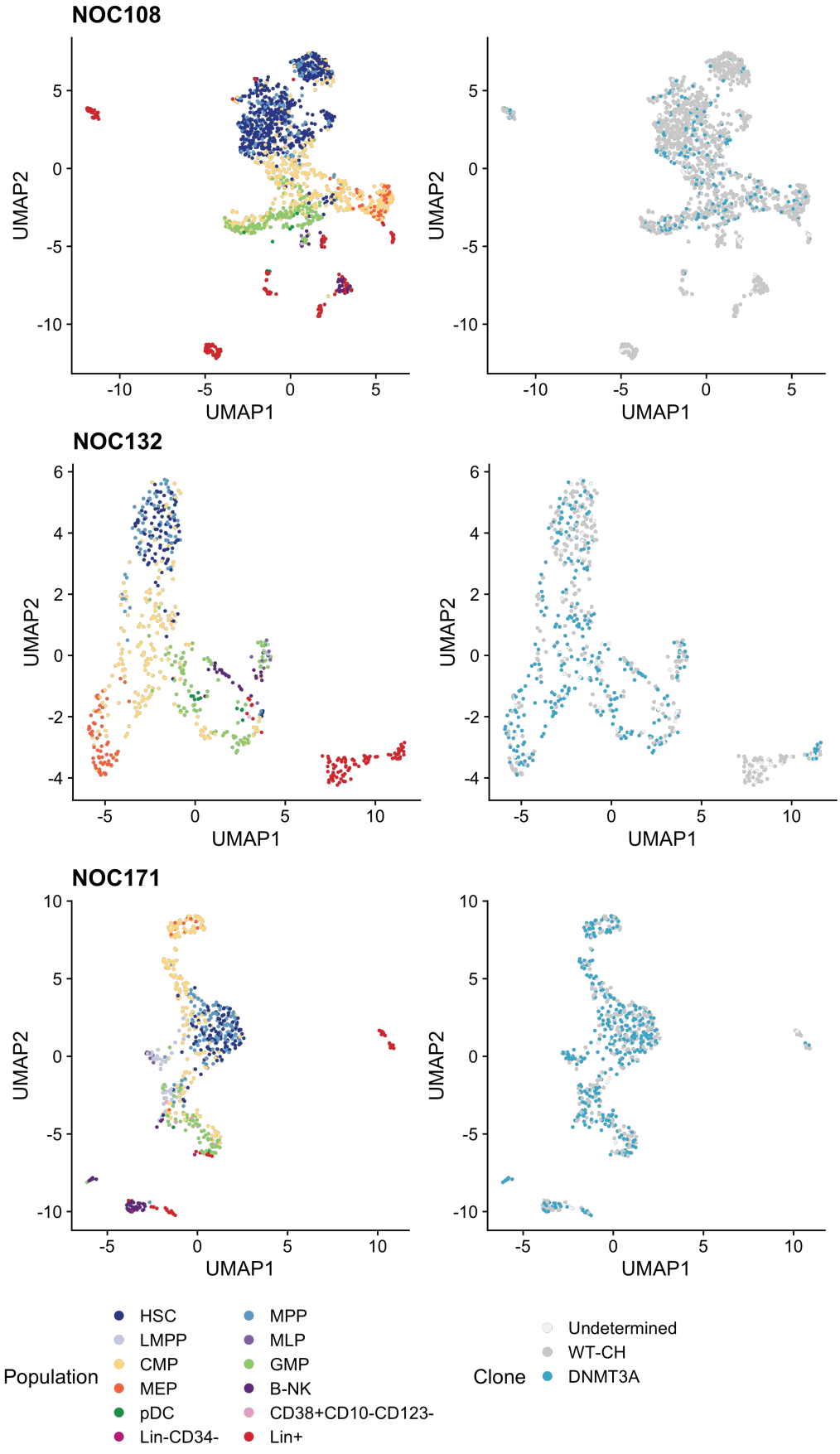


Figure 5.17: Copy number analysis of cells from control and clonal haematopoiesis samples. Heatmaps of relative expression of genes ordered by chromosome and chromosomal position following copy number variation analysis using the InferCNV package. Each row represents a single cell. **(a)** Control samples, which are used as a reference. Cells are stratified by the sample and cell type as determined by gene expression clustering. **(b-c)** Example of copy number analysis on the NOC132 **(b)** and NOC171 **(c)** samples. Cells are stratified by hierarchical clustering of the relative expression analysis, and annotated with the cell type as determined by gene expression clustering (left bar).


NOC108
NOC132
NOC171
UMAP1
UMAP2
Population
HSC
LMPP
CMP
MEP
pDC
Lin-CD34-
MPP
MLP
GMP
B-NK
CD38+CD10-CD123-
Lin+
Clone
Undetermined
WT-CH
DNMT3A

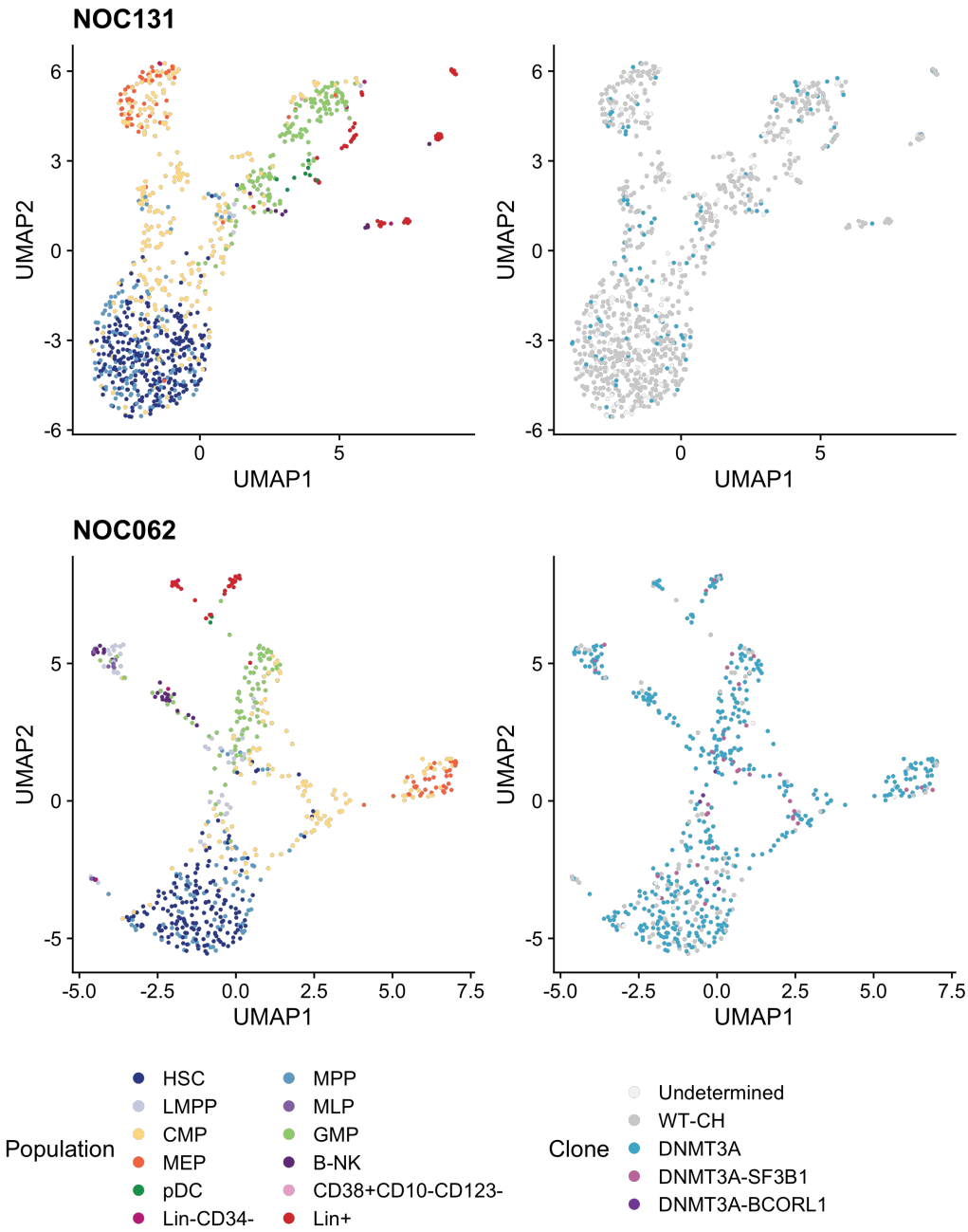
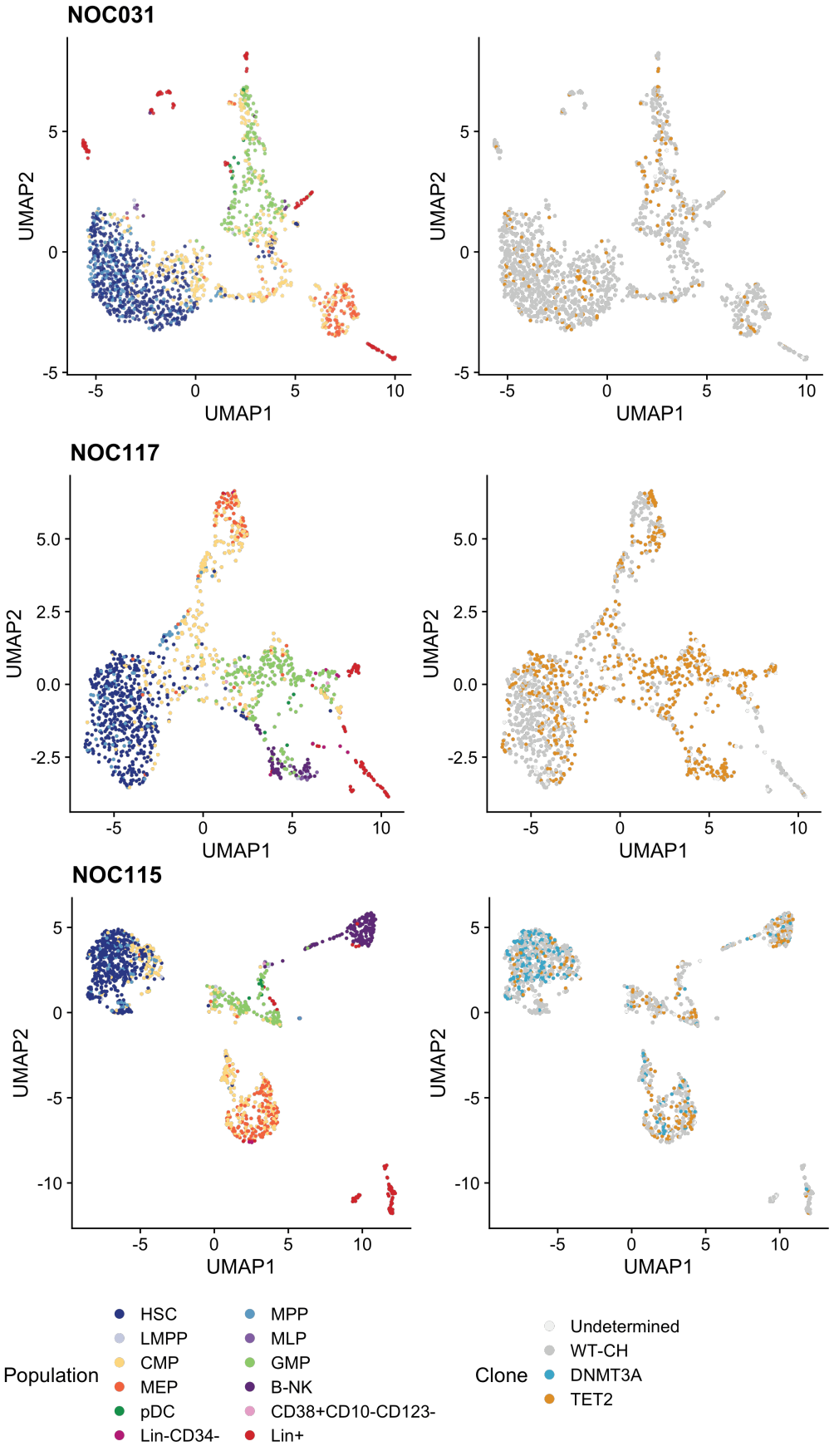


Figure 5.18: UMAP of individual *DNMT3A*-mutant samples. *(This page and previous page).* Uniform manifold approximation and projection (UMAP) of single cells from *DNMT3A*-mutant samples after batch effect correction. Cells are coloured by their immunophenotype (left) and mutational status (right).


NOC031
NOC117
NOC115
UMAP1
UMAP2
Population
HSC
LMPP
CMP
MEP
pDC
Lin-CD34-
MPP
MLP
GMP
B-NK
CD38+CD10-CD123-
Lin+
Clone
Undetermined
WT-CH
DNMT3A
TET2

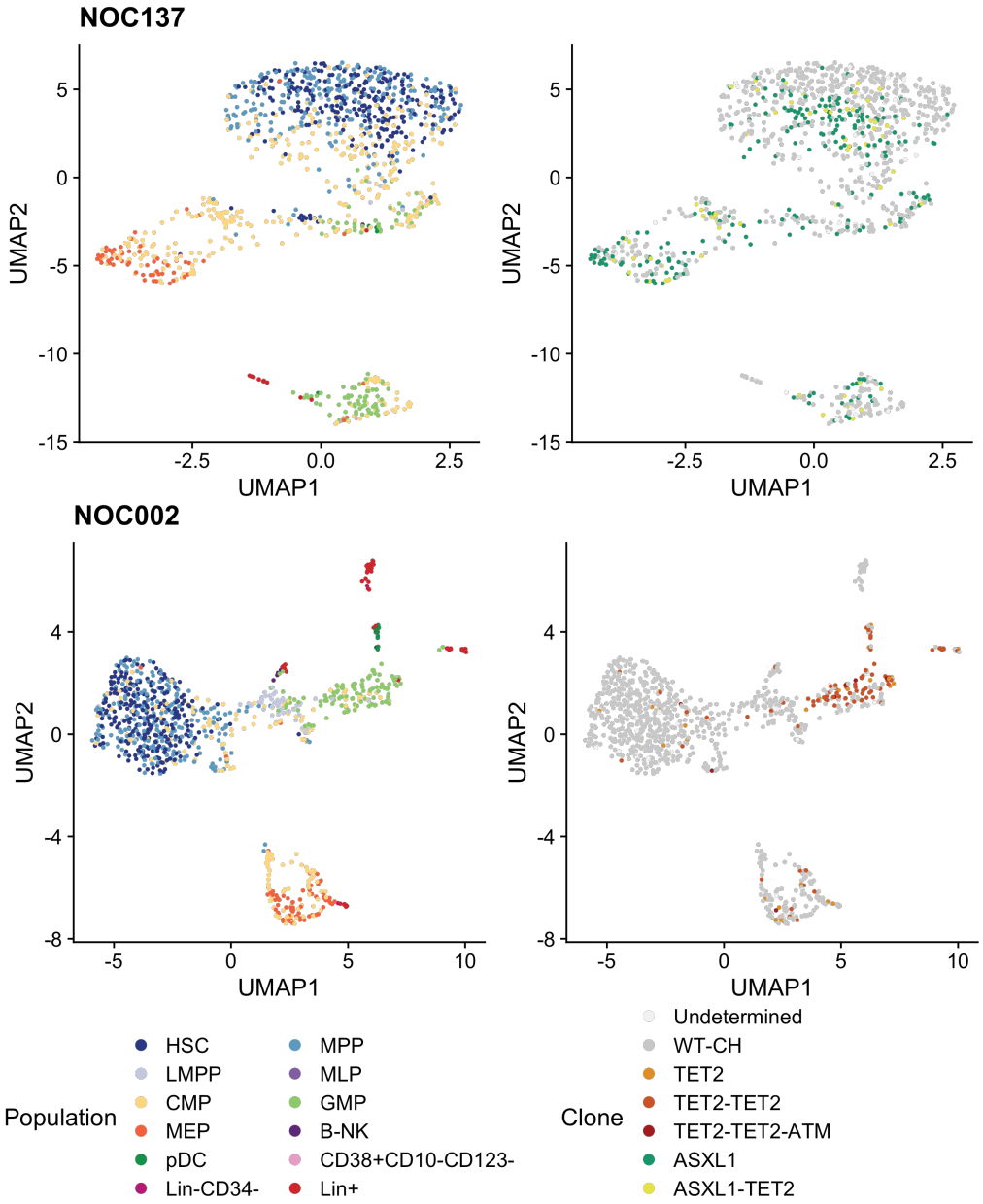


Figure 5.19: UMAP of individual *TET2*-mutant samples. *(This page and previous page).* UMAP of single cells from *TET2*-mutant samples after batch effect correction. Cells are coloured by their immunophenotype (left) and mutational status (right).

5.3.4 Integrated analysis of transcriptomic data for normal and clonal haematopoiesis samples

To achieve greater resolution in the transcriptional landscape, the data from all 10 clonal haematopoiesis and 3 control samples were combined into a single dataset of 12,847 cells. UMAP visualisation of the combined dataset showed good overlap between samples for late progenitors and mature cells, but there was a substantial donor effect amongst early HSPCs (HSC, MPP and LMPP) (Figure 5.20a,b). While there may be relevant biological differences between samples, such as between CH samples and controls, it appeared that other factors including the day of sample processing and sex were also contributing (Figure 5.20c.d). To obtain a unified landscape of haematopoietic differentiation, the data were integrated using Harmony [235]. This effectively removed the donor effect while preserving biological variation, which was visible by labelling cells with their immunophenotype (Figure 5.20e,f).

Unsupervised clustering was performed using the Louvain community-detection method, which identified 25 cell clusters (Figure 5.22a). Superimposing the expression of lineage signature gene sets on the UMAP identified the main megakaryocytic, erythroid, lymphoid and myeloid lineages. Within lineages, cell cycle genes also separated cells between cycling and non-cycling cells (Figure 5.22b,c). Expression of lineage markers was consistent with the immunophenotypic data (Figure 5.22d,e). Note that HSC and MPP are over-represented in frequency due to the enrichment for CD38– cells in FACS sorting. There was marked heterogeneity within immunophenotypically-defined HSPC populations. CMP were particularly heterogeneous, being split across 11 clusters.

To annotate the clusters, gene set enrichment analysis (GSEA) was performed for genes differentially expressed genes between clusters. Curated gene signatures specific to haematopoietic stem, progenitor and mature cells from 9 publications were used for enrichment analysis [236, 238–245] (Figure 5.22). These results, in combination with the immunophenotype and examination of selected marker genes, were used to manually annotate cell types.

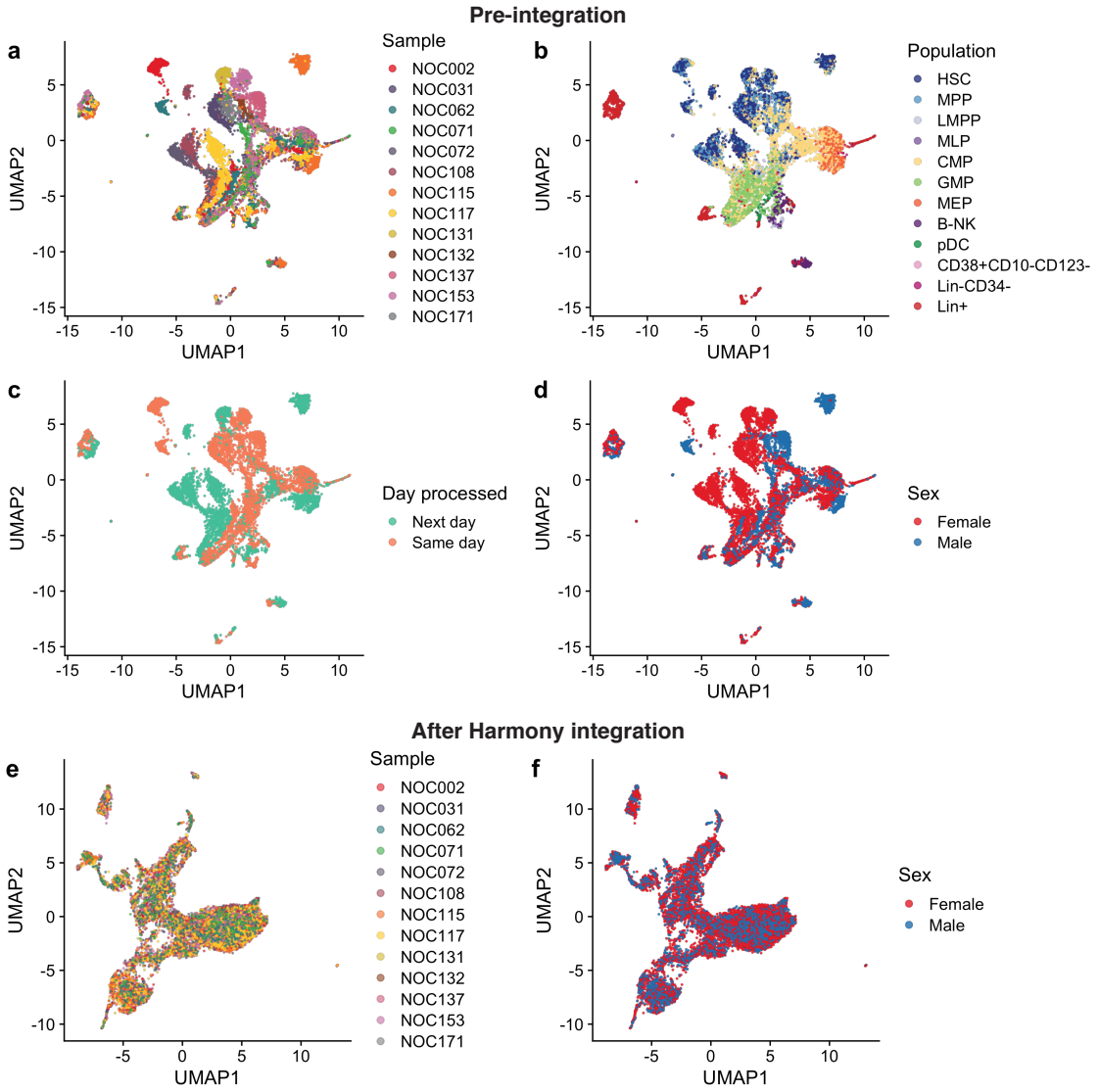


Figure 5.20: Harmony integration of transcriptomic data from control and clonal haematopoiesis samples. UMAP projections of 12,847 single cells from all 3 control and 10 clonal haematopoiesis samples analysed. 11935 variable genes were identified and the top 100 principal components used for building UMAPs and Harmony integration. **(a-d)** Combined dataset prior to data integration using the raw principal components (PCs) to construct the UMAP. Cells are labelled by the donor sample, immunophenotype, day of sample processing and sex. **(e-f)** Combined dataset after Harmony integration using donor sample as the covariate. Cells are coloured by the donor sample and sex.

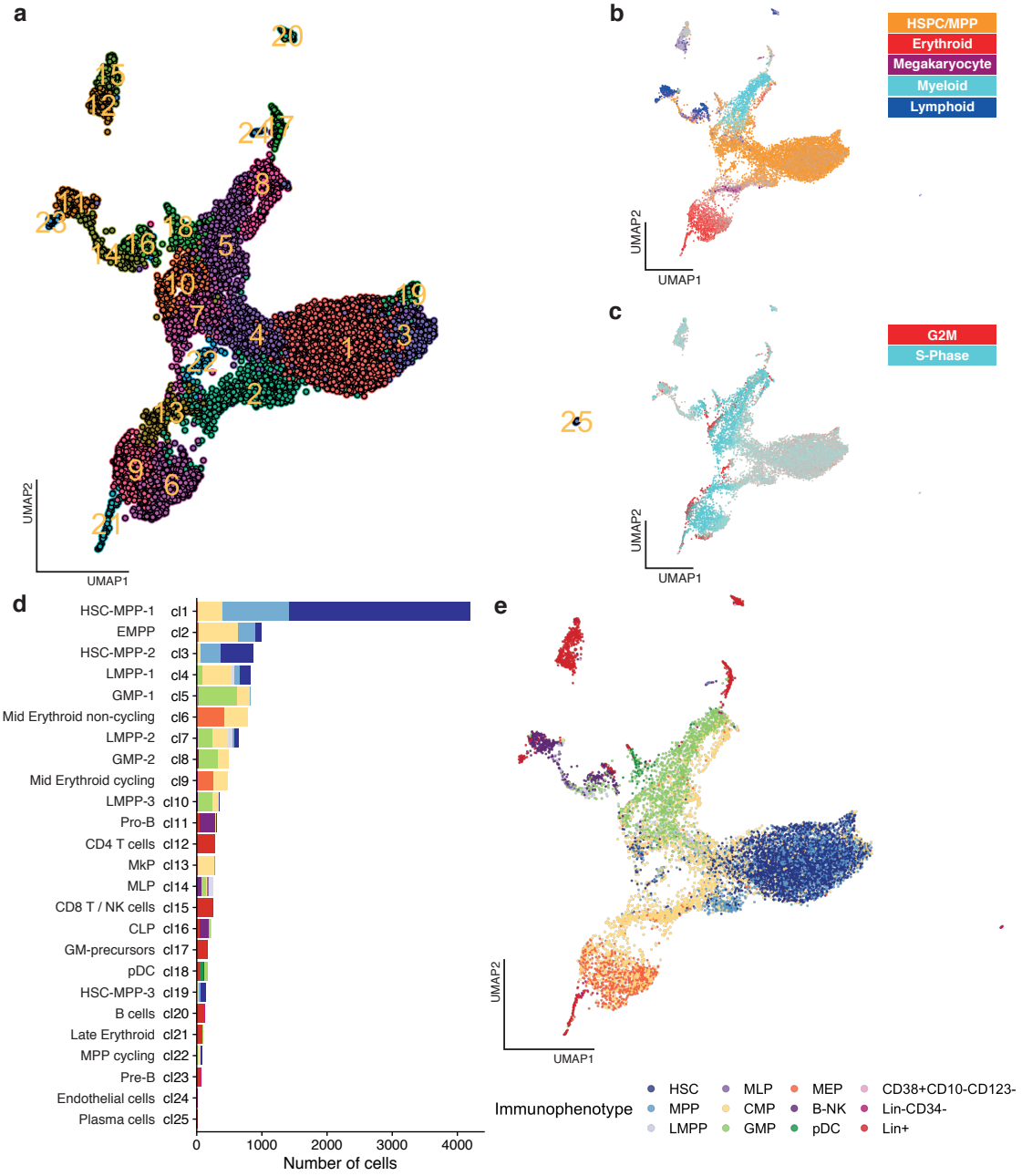


Figure 5.21: Identification of clusters in single cell transcriptome data. **(a)** UMAP projection labelled with the clusters identified using Louvain community-detection. **(b)** UMAP superimposed with the expression score for lineage signature gene sets. Grey indicates uncommitted or expression of >1 lineage gene set. **(c)** UMAP superimposed with the expression score for cell cycle modules. **(d)** Number of cells in each cluster coloured by their immunophenotype and labelled with the descriptive name given to the cluster. See main text and following figures for the basis for the labels given. **(e)** UMAP coloured by the immunophenotype of single cells obtained from flow cytometry index sorting data.

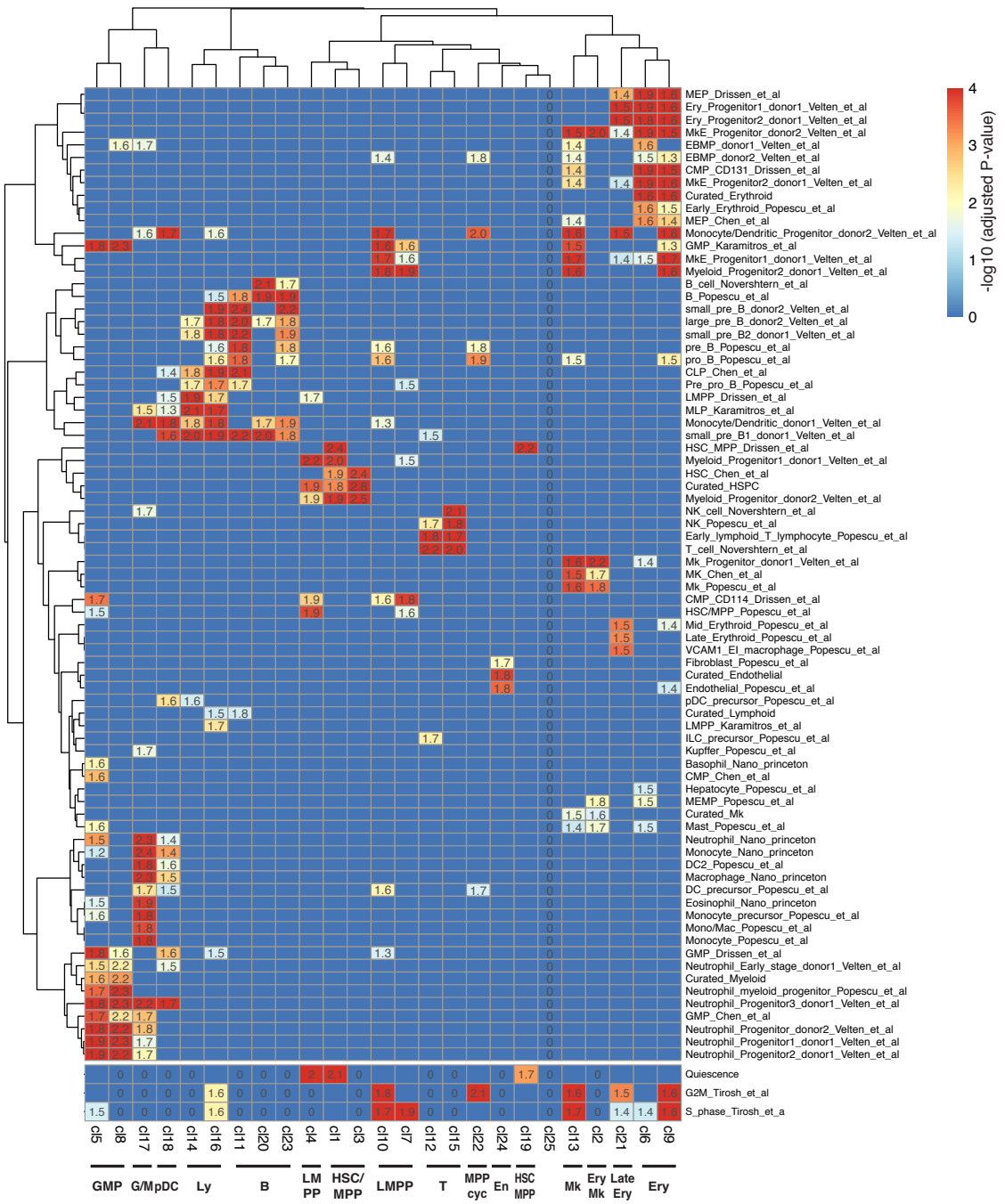


Figure 5.22: Gene set enrichment analysis of cluster marker genes. Heatmap of the gene set enrichment (GSEA) results using published gene signatures for HSPCs. Clusters are numbered as in Figure 5.21. Heatmap colours represent the –log10 of the adjusted *P*-value, and the normalised enrichment score is shown for each comparison with an adjusted *P*-value < 0.05. Hierarchical clustering was performed on the *P*-values. GSEA results for cell cycle and quiescence gene sets are shown below; these were not used for clustering. G/M: Granulocytic/monocytic, Ly: Lymphoid, B: B cell, LMPP: lymphoid-primed multipotent progenitor, T: T cell, En: Endothelial, Mk: megakaryocytic, Ery: Erythroid, Cyc: cycling.

Clusters 1 and 3 were enriched in HSC/MPP genesets and represent the majority of CD38– cells. They were composed of 91% and 94% immunophenotypic HSC and MPP respectively, the majority of which were HSC. Additionally, cluster 19 was a small HSC/MPP cluster of 138 cells. Clusters 2, 4 and 22 made up the earliest branches and were mainly composed of immunophenotypic CMP with some CD38– cells and GMP (cluster 4). GSEA and the UMAP showed that cluster 2 was enriched in megakaryocytic/erythroid genes, while cluster 4 was enriched in both lympho/myeloid and HSC/MPP genesets. Beyond this, clusters 7, 10 and 5 were lymphoid/myeloid progenitors with CMP/GMP immunophenotype.

To distinguish the transition from HSC/MPP to megakaryocytic/erythroid and lymphoid/myeloid progenitors in more detail, differential gene expression analysis was performed amongst these 6 clusters (Figure 5.23a). Cluster 1 and 3 had very similar profiles and were distinguished from other clusters by expression of genes including *HLF*, *DLK1*, *EMCN* (endomucin) and the long non-coding RNA *MEG3*, all of which have recently been identified as HSC markers [257, 272, 274–276]. Cluster 19 was marked by upregulation of *DUSP6*, *PTGS2* and *EMP1*, which have been associated with quiescence, as well as genes that positively regulate transcription, such as *KLF4* and *NFKB2*. It also had a higher percentage of mitochondrial reads than the other HSC/MPP clusters, so may represent a subset of a subset of HSC with activated stress response. These clusters were labelled HSC-MPP-1 to 3.

Amongst the early progenitors, cluster 4 expressed lymphoid and myeloid genes such as *IGHM*, *MZB1*, *TRBC2* (T Cell Receptor Beta Constant 2) and *CSF3R* (Figure 5.23a,b). Importantly, *CSF3R* (CD114) has been identified as marker that distinguishes CMP with neutrophil/monocyte and lymphoid potential from those with megakaryocytic/erythroid and mast cell/basophil/eosinophil potential [241]. Cluster 2 expresses key megakaryocytic/erythroid transcription factors (*GATA1*, *KLF1*) and basophil genes (*FCER1A*) (Figure 5.23a,b). This therefore represents the EMPP (erythroid/megakaryocyte-primed multipotent progenitor, see also below). Cluster 22 was half CD38– cells, and shared genes with clusters 1

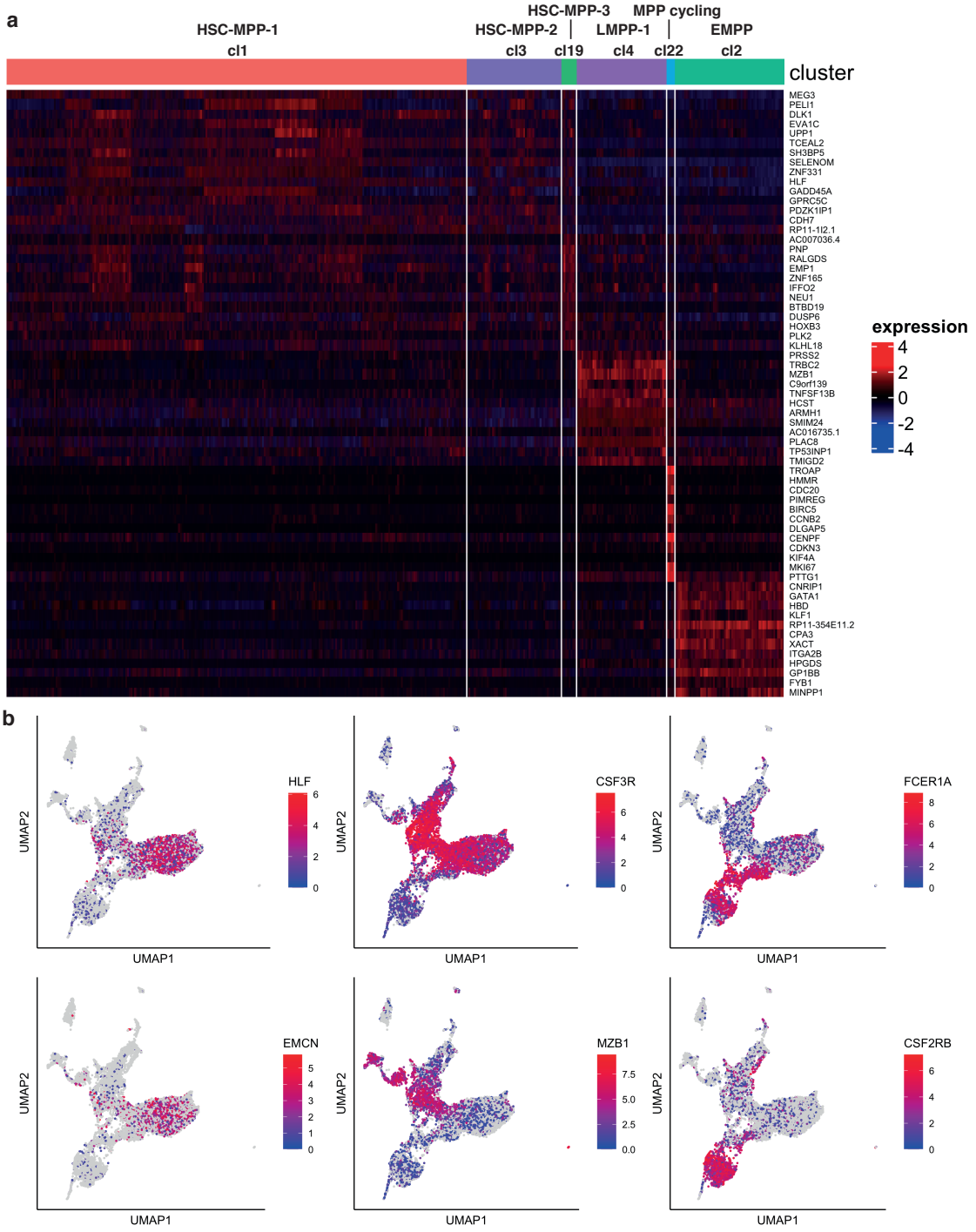


Figure 5.23: Differentially gene expression amongst HSC and MPP clusters. **(a)** Heatmap showing the top 15 differentially expressed genes between each of the 6 HSC and MPP clusters. Genes were ranked based on the fold change. Rows represent genes and columns represent single cells. Clusters are labelled as in Figure 5.21. **(b)** UMAPs coloured by the expression of the genes shown. The colour scale shows the log of the raw transcript counts for each cell.

and 3, but additionally expressed cell cycle genes, so this may represent a small population of cycling HSC/MPP.

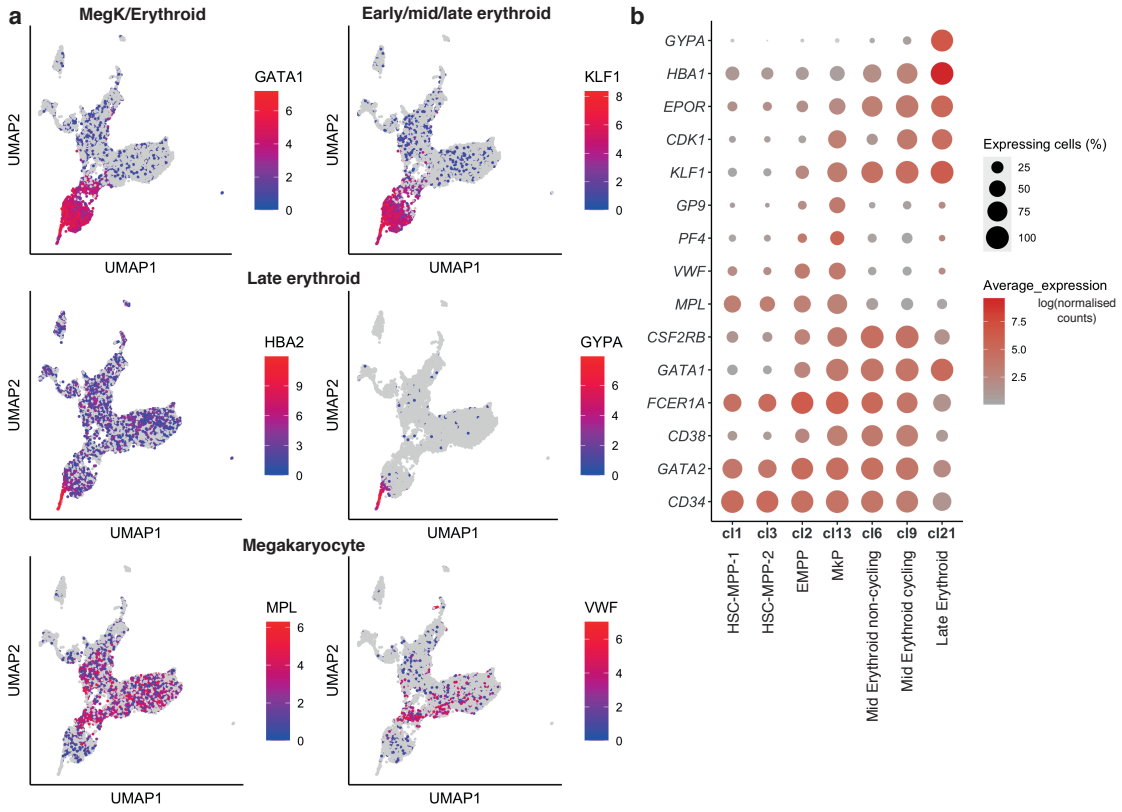


Figure 5.24: Identification of megakaryocytic and erythroid progenitors. (a) UMAPs coloured by the expression of megakaryocytic and erythroid genes. The colour scale shows the log of the raw transcript counts for each cell. **(b)** Dot plot showing mean normalised expression values within megakaryocytic/erythroid clusters for the genes shown. The percentage of cells expressing each gene is indicated by the spot size and the log of the mean level of expression by the colour scale.

Clusters 2, 13, 6, 9 and 21 represented progenitor cells of the erythroid and megakaryocytic lineage. Examination of canonical marker genes within this lineage enabled them to be separated into stages of maturity (Figure 5.24). Cluster 2 is the most immature, composed of immunophenotypic CD38– MPP and CD38+ CMP. It is the earliest compartment in which *GATA1* is expressed, along with *VWF*, *KLF1* and *CSF2RB* (CD131) which has been used to define progenitors capable of generating megakaryocytes, erythroid cells, basophils/mast cells and eosinophils. It was therefore labelled as EMPP (erythroid/megakaryocyte-primed MPP). Cluster 13 expressed *VWF* and megakarytocytic markers *PF4* and *GP9* so was marked as

megakaryocyte progenitors (MkP). MkP were all immunophenotypic CMP. Clusters 6 and 9 represent intermediate erythroid progenitors marked by *KLF1*, *EPOR* (EPO receptor) and low expression of *HBA1/2*, and are separated by greater expression of cell cycle genes in cluster 9. Cluster 21 cells are late erythroid progenitors with high expression of *HBA1/2* and Glycophorin A (*GYPA*, CD235a).

Within the lympho-myeloid progenitors, immunophenotypic CMP, GMP and LMPP were split across eight clusters (clusters 4, 7, 10, 5, 8, 18, 16 and 14). As mentioned above, cluster 4 was the earliest cluster expressing both stem cell genes such as *HLF* and *MLLT3* as well as lympho-myeloid genes (Figure 5.25). It was therefore labelled lymphoid-primed multipotent progenitor 1 (LMPP-1). Clusters 7 and 10 shared genes with cluster 4 and were marked by expression of cell cycle genes (S-phase and G2M respectively) along with low expression of lymphoid (*DNTT*, *MZB1*, *VPREB1*, *IGHM*) and myeloid (*MPO*, *CEBPA*, *SPI1* (PU.1), *IRF8*) genes (Figure 5.26 and 5.26). They were therefore named LMPP-2 and LMPP-3. Clusters 5 and 8 were mainly immunophenotypic GMP and were marked by key myeloid genes such as *MPO*, *CEBPA*, *SPI1*/PU.1 and *RNASE3* while cluster 8 also contained a small population of immunophenotypic CMP that expressed mast cell/basophil genes such as *CLC* and *HDC*. They were labelled as GMP-1 and GMP-2 respectively. Cluster 18 were CD123 (IL3RA)-high expressing cells marked by expression of *IRF8*, *IRF4* and *LILRA4* and were labelled as plasmacytoid dendritic cells (pDC). Cluster 17 (not shown on the heatmap) contained Lin+ cells that emerged from the GMP and expressed neutrophil and monocyte markers including *CD14*, *ELANE* and *LYZ* (lysozyme) (Figure 5.26). These were therefore labelled granulocyte-monocyte (GM)-precursors.

Clusters 14 and 16 were predominantly immunophenotypic LMPP/MLP and B-NK cells respectively, and expressed lymphoid genes including *PRSS2*, *PAX5* and *DNTT* (Figure 5.25). Cluster 14 appeared more immature, and most closely represented multi-lymphoid progenitors (MLP), as it was CD38–CD10+ while cluster 16 was CD38+CD10+ and expressed *IL7R* (CD127) which are markers of the common lymphoid progenitor (CLP). Notably cluster 16 also expressed

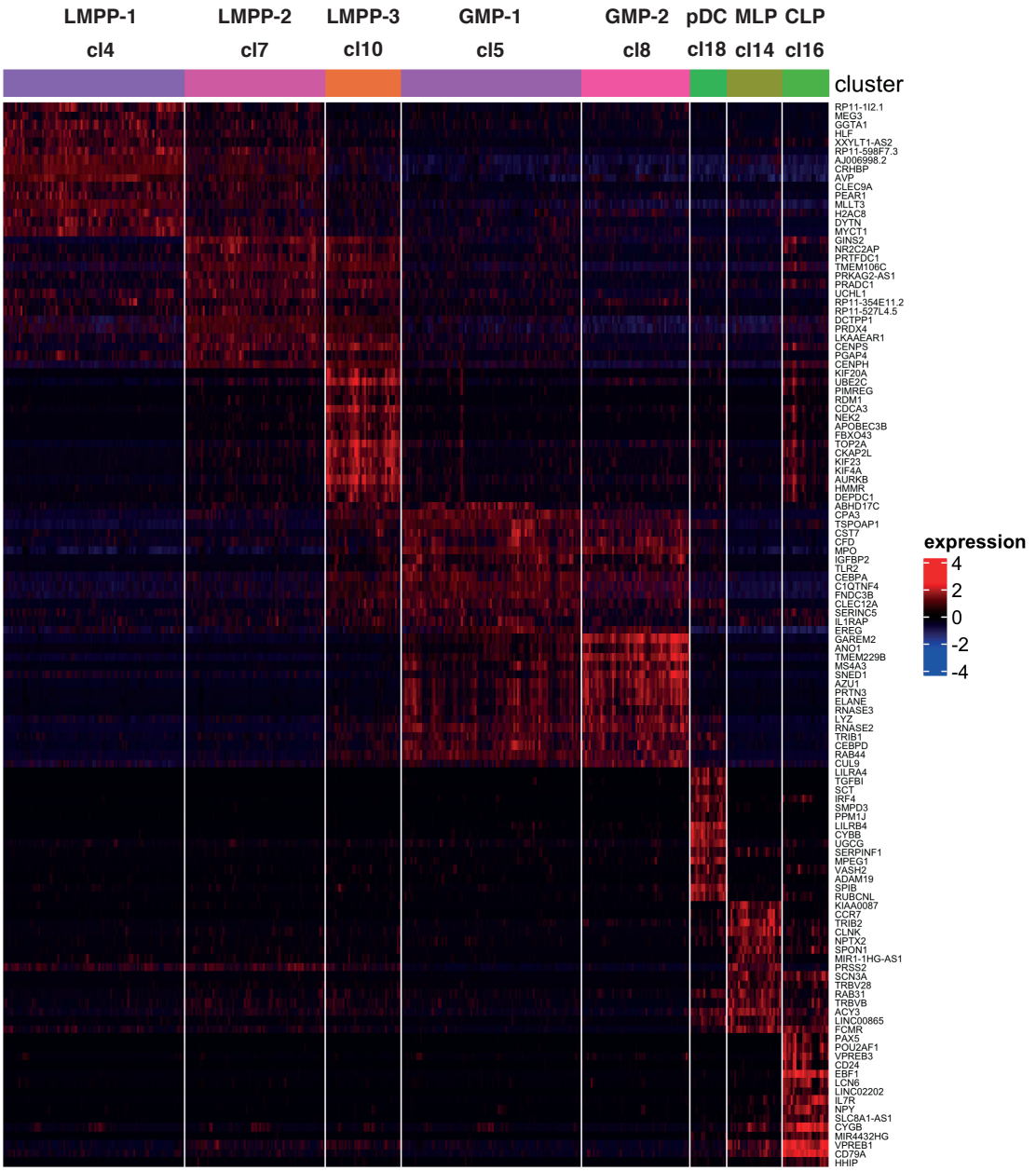


Figure 5.25: Differential gene expression amongst lympho-myeloid progenitors. Heatmap showing the top 15 differentially expressed genes between each of the 8 lympho-myeloid progenitor clusters. Genes were ranked based on the fold change. Rows represent genes and columns represent single cells.

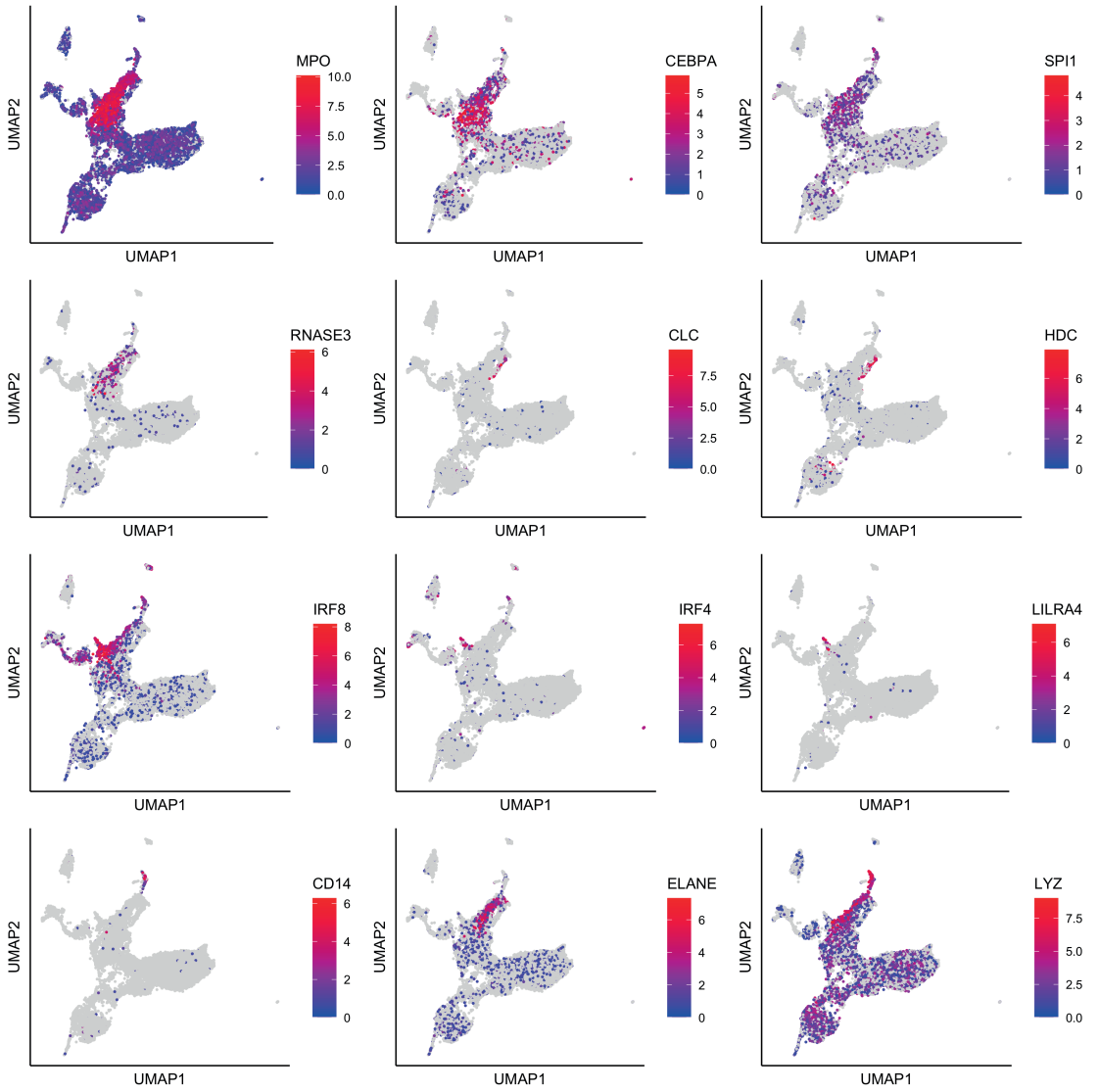


Figure 5.26: Identification of myeloid progenitors and precursors. UMAPs coloured by the expression of myeloid genes. The colour scale shows the log of the raw transcript counts for each cell.

NKG7, a natural killer (NK) cell gene. Other clusters in the lymphoid lineage were 11, 23 and 20, which expressed B cell markers as well as clusters 12 and 15 which expressed *CD3D*. By examining genes expressed in the B cell lineage, it was possible to delineate the key stages of B cell development progressing from cluster 11 through cluster 23 to cluster 20 (Figure 5.27). Cells in cluster 11 gain *CD24* while retaining CD34, so were labelled B cell progenitors (Pro-B). Cluster 23 cells lost CD34 expression and gained *IL4R* and *MS4A1* (CD20) so were labelled B cell precursors (Pre-B), and finally cluster 20 cells expressed high levels of *IL4R* and

MS4A1 and lost CD10 expression, so were labelled B cells.

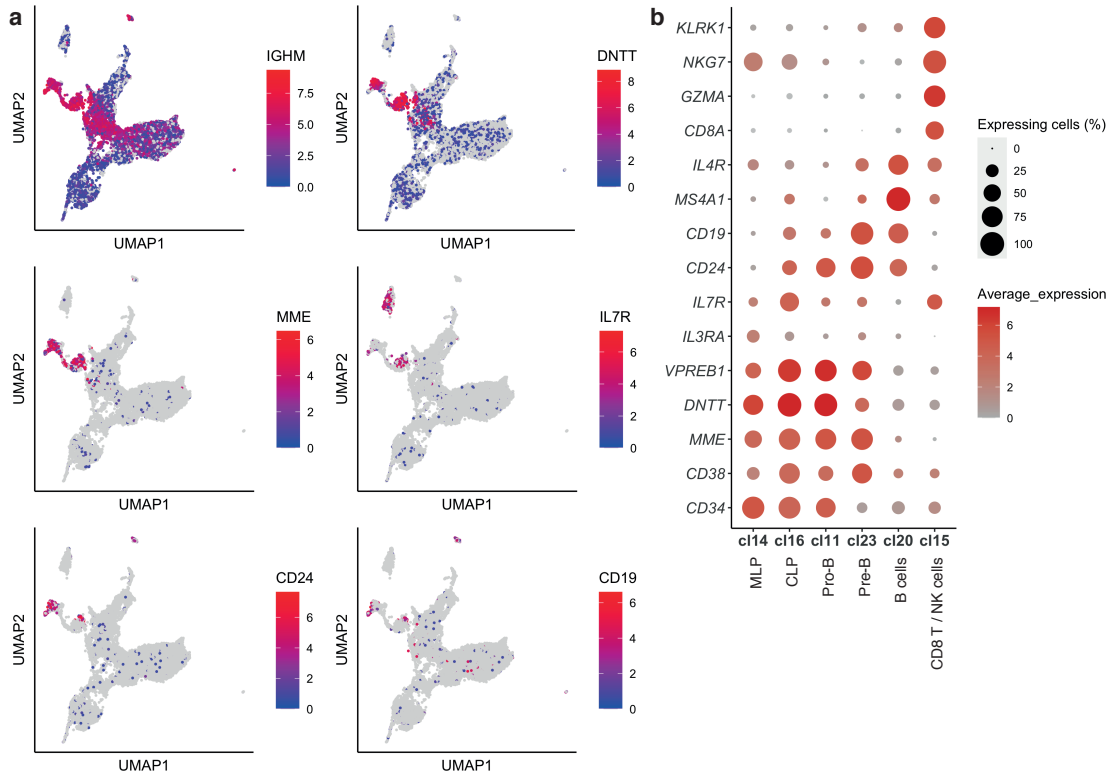


Figure 5.27: Identification of lymphoid progenitors and B cell precursors. **(a)** UMAPs coloured by the expression of lymphoid genes. **(b)** Dot plot showing mean normalised expression values within lymphoid cell clusters for the genes shown. The percentage of cells expressing each gene is indicated by the spot size and the log of the mean level of expression by the colour scale.

The remaining clusters were mature cell populations. Clusters 12 were CD4+ T cells and clusters 15 was composed of a CD8+ T cells and NK cells (Figure 5.27 and 5.28). Cluster 25 was a small population of plasma B cells, identified by the expressiong of *SDC1* (CD138) and *CD27*. Finally, cluster 24 was a population of endothelial which expressed *KDR*, *ESAM*, *NRFR2* and *TCIM*.

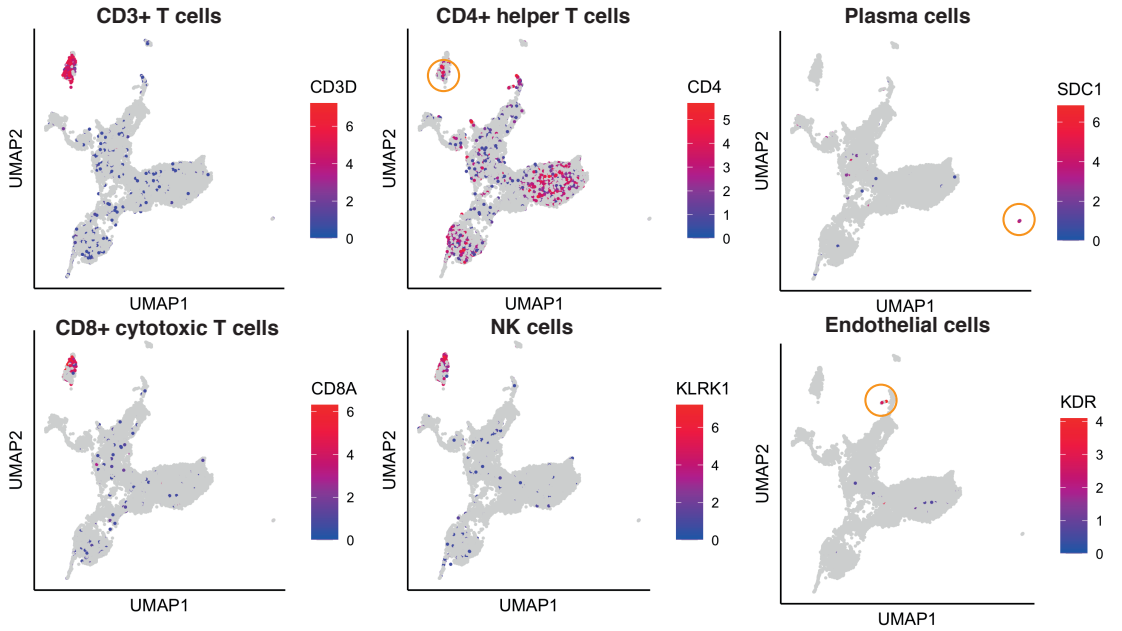


Figure 5.28: Identification of T, NK, plasma and endothelial cells. UMAPs coloured by the expression of myeloid genes. The CD4+ helper T cell, plasma cell and endothelial cell clusters are highlighted in the relevant plots.

The frequency of cells assigned to each cluster were compared between controls, *DNMT3A*-mutant and *TET2*-mutant samples. Consistent with the findings from flow cytometry analysis, the representation of cells across the transcriptionally-defined cell types was broadly similar between controls and CH samples (Figure 5.29). There were no significant differences in cluster assignment amongst CD38– cells. Within CD38+ cells, there tended to be fewer LMPP in both *DNMT3A*- and *TET2*-mutant samples while amongst Lin+ cells, *DNMT3A*-mutant samples had fewer late erythroid cells, and *TET2*-mutant samples had more plasma cells. None of these differences were statistically significant (paired Wilcoxon test).

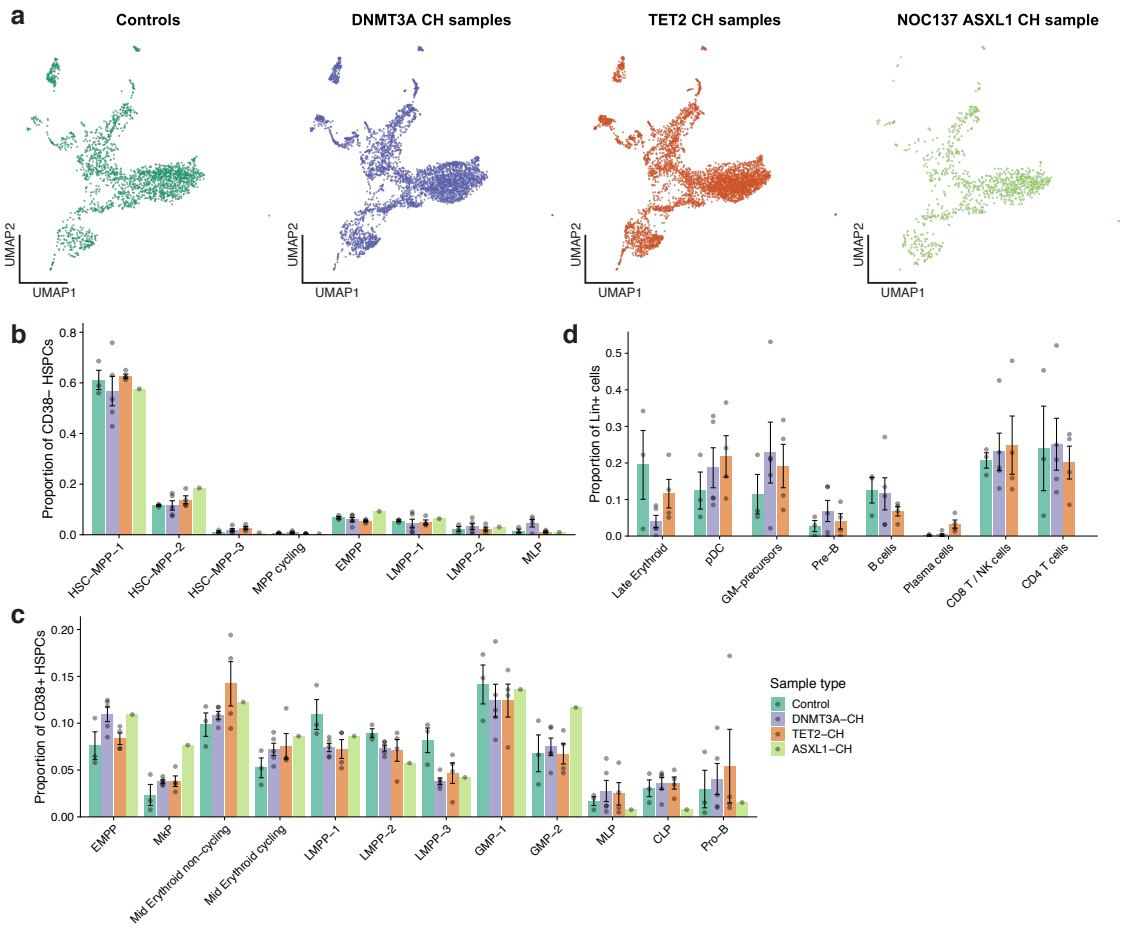


Figure 5.29: Frequency of cell types in control and CH samples. **(a)** UMAP plots showing all cells from the 3 control samples, 5 *DNMT3A*-mutant samples, 4 *TET2*-mutant samples, and NOC137 which has an *ASXL1*-mutant dominant clone. Sample NOC115 was included with the *TET2*-mutant samples. **(b-d)** Frequency of cells in each cluster by sample type for CD38– cells **(b)**, CD38+ cells **(c)**, and Lin+ cells **(d)**. Lin+ cells are not shown for NOC137 as there were too few to calculate frequencies.

5.3.5 Expression of *DNMT3A* and *TET2* in HSPCs

An important determinant of mutational phenotype is the pattern of expression of the mutated gene through differentiation. The expression of *DNMT3A* and *TET2* were therefore examined across the transcriptional cell clusters defined above.

DNMT3A was expressed in 47% of HSC-MPP cells. Its expression increased in early and late progenitors and then decreased in mature cells across all the lineages (Figure 5.30). For example, in lympho-myeloid progenitors, expression was highest in LMPP-2, LMPP-3 and GMP-1 clusters where expression was detected in 67%, 87% and 72% of cells respectively. The pattern was similar in megakaryo-

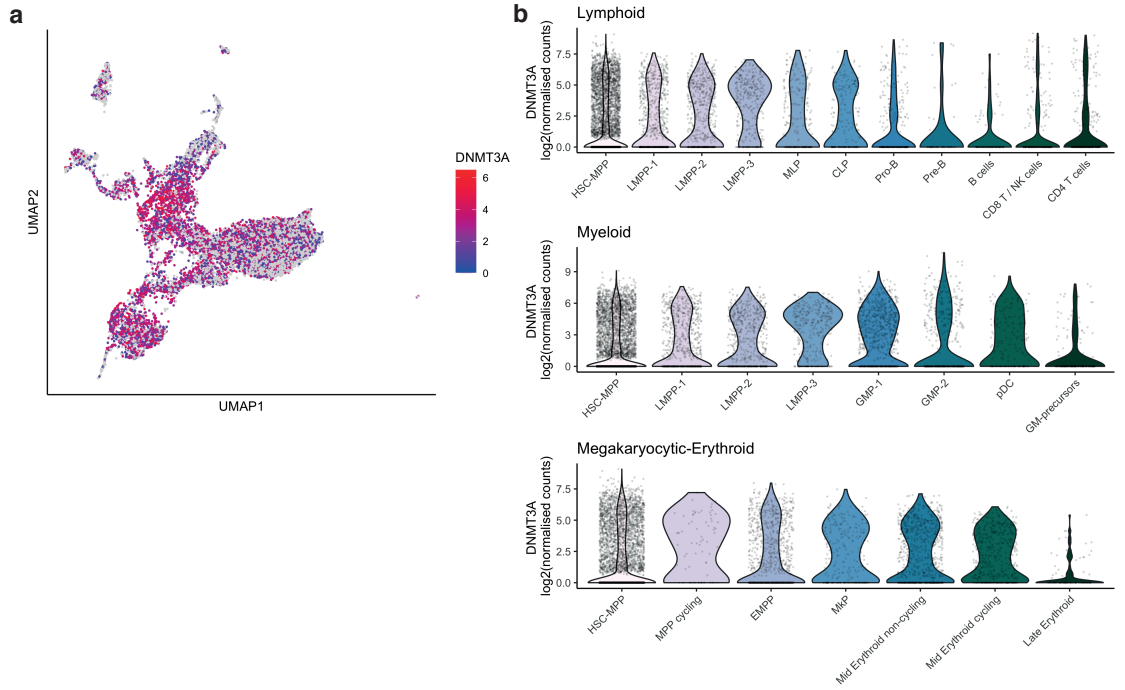


Figure 5.30: *DNMT3A* expression through differentiation. (a) UMAP showing the expression of *DNMT3A*. **(b)** Violin plots of normalised, log2 transformed *DNMT3A* expression values for each cell cluster in lymphopoiesis (top), myelopoiesis (middle) and megakaryo-erythropoiesis (bottom). Clusters 1, 3, and 19 (HSC-MPP-1, HSC-MPP-2, HSC-MPP-3 were merged as they showed similar expression levels.

erythropoiesis where expression was highest in MkP and mid-erythroid progenitors. Thus the pattern of expression is similar to CD38 which increases during lineage commitment and is then downregulated in most mature cells.

TET2 was expressed in 48% of HSC-MPP cells and was also upregulated in progenitors (Figure 5.31). In the lympho-myeloid lineages, expression was highest in LMPP-3, GMP-1 and pDC where it expressed in 60–67% of cells. In contrast to *DNMT3A*, *TET2* was expressed in very few cells of the lymphoid lineage. Similarly, in the megakaryocyte-erythroid lineage, *TET2* was uprgulated in progenitors, and downregulated in late erythroid cells.

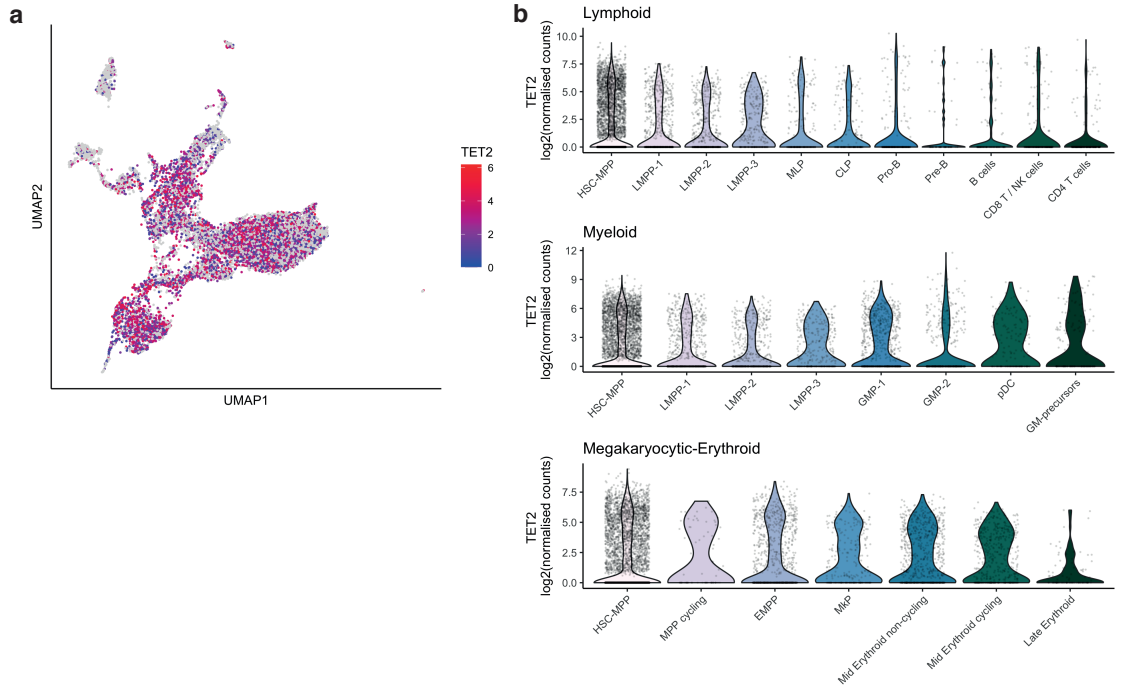


Figure 5.31: *TET2* expression through differentiation. (a) UMAP showing the expression of *TET2*. **(b)** Violin plots of normalised, log2 transformed *TET2* expression values for each cell cluster in lymphopoiesis (top), myelopoiesis (middle) and megakaryo-erythropoiesis (bottom).

5.4 Transcriptomic analysis of *DNMT3A*-mutant clonal haematopoiesis

5.4.1 Analysis of *DNMT3A*-mutant clonal size through haematopoietic differentiation

To explore how the acquisition of *DNMT3A* mutations perturbs normal haematopoiesis, the frequency of mutant cells was analysed across the haematopoietic differentiation landscape. As was seen in the analysis of individual samples, *DNMT3A*-mutant cells intermingled with WT cells, suggesting they follow a common differentiation trajectory and do not show a differentiation arrest (Figure 5.32a). Nonetheless, mutations might lead to a differentiation bias towards one lineage or another, or a relative differentiation delay, which would be visible as differences in mutant cell frequency across the haematopoietic landscape. Differential abundance of *DNMT3A*-mutant cells (with no additional mutations) versus WT cells was tested

using the MiloR package [246]. This method assigns cells to partially overlapping neighbourhoods on a k-nearest neighbour graph enabling the composition of cell states to be compared between genotypes while accounting for variation between the 5 individuals.

DNMT3A-mutant cells were relatively depleted in some HSC-MPP neighbourhoods and showed increased abundance in early EMPP and LMPP progenitors (Figure 5.32b,c). Within the megakaryocyte-erythroid lineage, mutant cells were specifically enriched in non-cycling mid-erythroid progenitors. In the myeloid lineage, mutant cells were increased in GMP-1 and GM precursors, while in lymphoid cells, there was an increase in the CLP and a relative reduced frequency in B cell progenitors. The most significant difference was in T cells, where there was a very low frequency of mutant cells, in particular in CD4+ T cells. The reduced mutant frequency in T cells is consistent with previous studies of peripheral blood [85, 277, 278].

To corroborate these findings with an alternative method, the frequency of WT and mutant cells was compared across the cell types identified by Louvain clustering (Figure 5.32d). In order to enable comparison across samples, the frequency of mutant cells was normalised to the mutant cell frequency in the HSC-MPP clusters for each sample. This confirmed that mutant cells were enriched in early EMPP and LMPP progenitors, with a mean frequency 50–75% higher than in HSC-MPP. The increased frequency in non-cycling erythroid progenitors, GMP and GM precursors and depletion in T cells was also consistent with findings from MiloR. Thus while *DNMT3A*-mutant clones expand in the HSC-MPP compartment, they tended to reach the largest clone size in early progenitor cells. Furthermore, in this analysis there was no clear evidence of lineage bias in *DNMT3A*-mutant cells, aside from a depletion in T cells.

To determine whether *DNMT3A* mutations alter the dynamics of haematopoietic differentiation, pseudotime trajectory analysis was performed using Slingshot [247]. Slingshot performs lineage inference in two stages: First, lineages are identified by constructing a minimum spanning tree across the clusters from an origin cluster

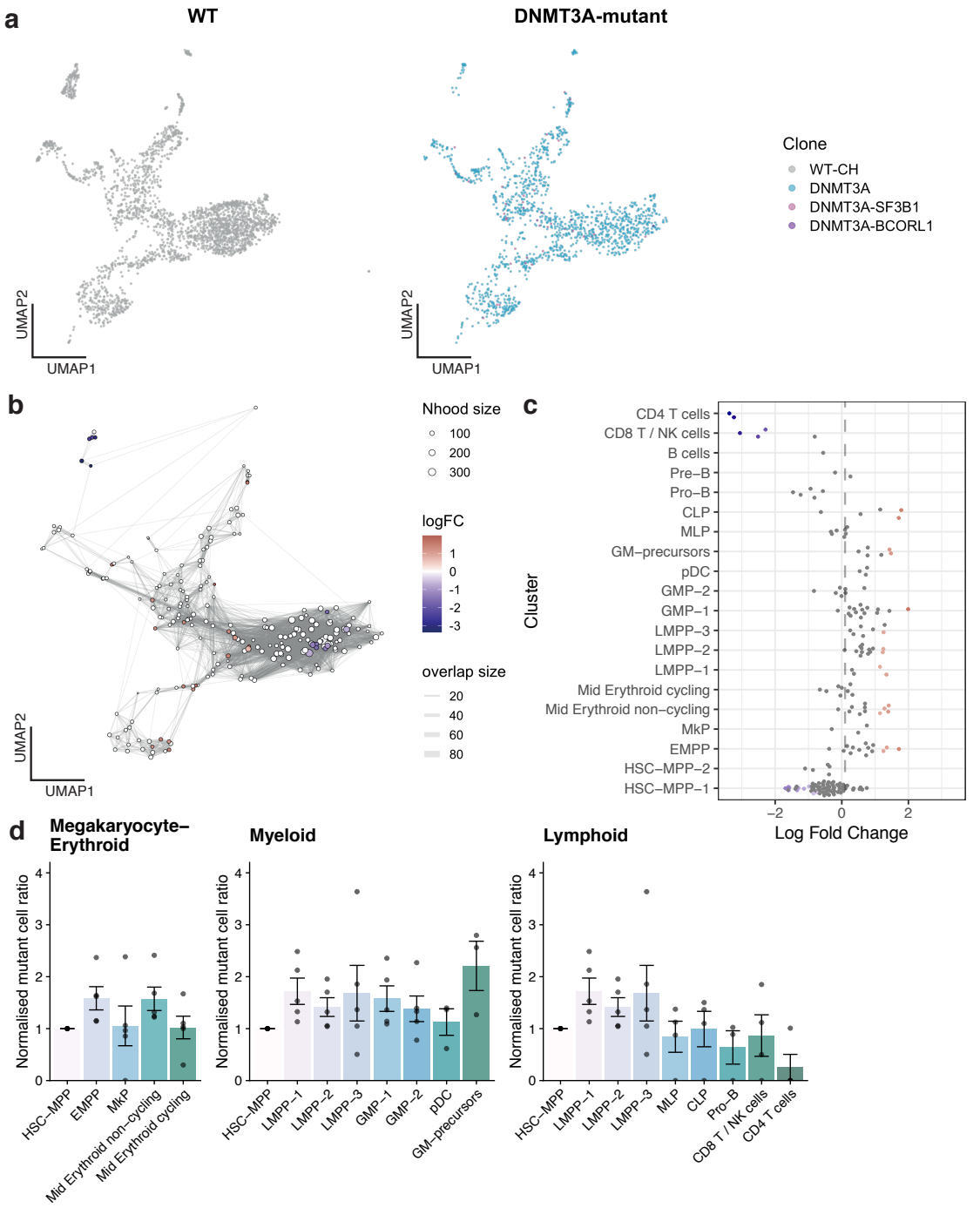


Figure 5.32: Abundance of *DNMT3A*-mutant cells across the haematopoietic hierarchy. (a) UMAP plots showing WT and mutant cells from the 5 samples with a dominant *DNMT3A*-mutant clone. Cells are coloured by genotype, where WT-CH indicates WT cells (grey), and mutant cells are coloured by the genotypes listed in the legend. **(b)** Neighbourhood graph of the results from MiloR differential abundance testing. Nodes represent neighbourhoods, the size of each node corresponds to the number of cells in the neighbourhood and graph edges depict the number of cells shared between neighbourhoods. The layout of nodes is determined by the position of an index cell in the UMAP. *(continued next page)*

Figure 5.32 ***(previous page)*****:** Neighbourhoods with a significant difference in abundance between WT and *DNMT3A*-mutant cells (with FDR < 0.2) are coloured by their log fold change; red indicates a relative increase in abundance of *DNMT3A*-mutant cells and blue indicates a relative decrease. **(c)** Beeswarm plot of the distribution of log fold change between WT and *DNMT3A*-mutant cells. Neighbourhoods are labelled with the most abundant cell type cluster (with at least 60% purity). Neighbourhoods with a significant difference in abundance between WT and *DNMT3A*-mutant cells are coloured by their log fold change as in **(b)**. **(d)** Frequency of *DNMT3A*-mutant cells in each Louvain cluster. For each sample, the frequency of mutant cells was normalised to the mutant cell frequency in the HSC-MPP clusters, in order to enable comparison across samples. Bars show the mean and error bars represent the standard error. Points show the values for individual samples. For each cluster, samples were excluded if there were fewer than 8 cells assigned to that cluster. Nhood: neighbourhood, logFC: log fold change.

to terminal clusters. Second, for each lineage, cells are ordered in pseudotime, representing each cell's transcriptional progression toward the terminal state. To construct lineage trajectories, the merged HSC-MPP clusters were assigned as the origin, and late erythroid, MkP, GM precursors, pDC and B cell precursors were assigned as the terminal clusters. Analysis was performed on the principal components to capture the full biological variance in the data, rather than rely on the two-dimensional UMAP as some methods do.

Slingshot identified trajectories across the five lineages that were consistent with the cell type annotation of clusters (Figure 5.33a). To determine whether there was a difference in progression from HSC to mature cells, the density of WT and mutant cells was compared along each lineage trajectory (Figure 5.33b). *DNMT3A*-mutant cells were significantly shifted towards more differentiated cell states than WT cells in all five lineages ($P < 0.05$ in all lineages, generalised linear mixed model accounting for sample effects). Notably, this shift towards more differentiated states was consistent across all the *DNMT3A*-mutant samples with the exception of NOC171 and NOC115, which has a co-existing *TET2*-mutant clone (Figure 5.34). In NOC115, *DNMT3A*-mutant cells were shifted towards a more primitive state, suggesting impaired differentiation, whereas *TET2*-mutant cells were on average more differentiated.

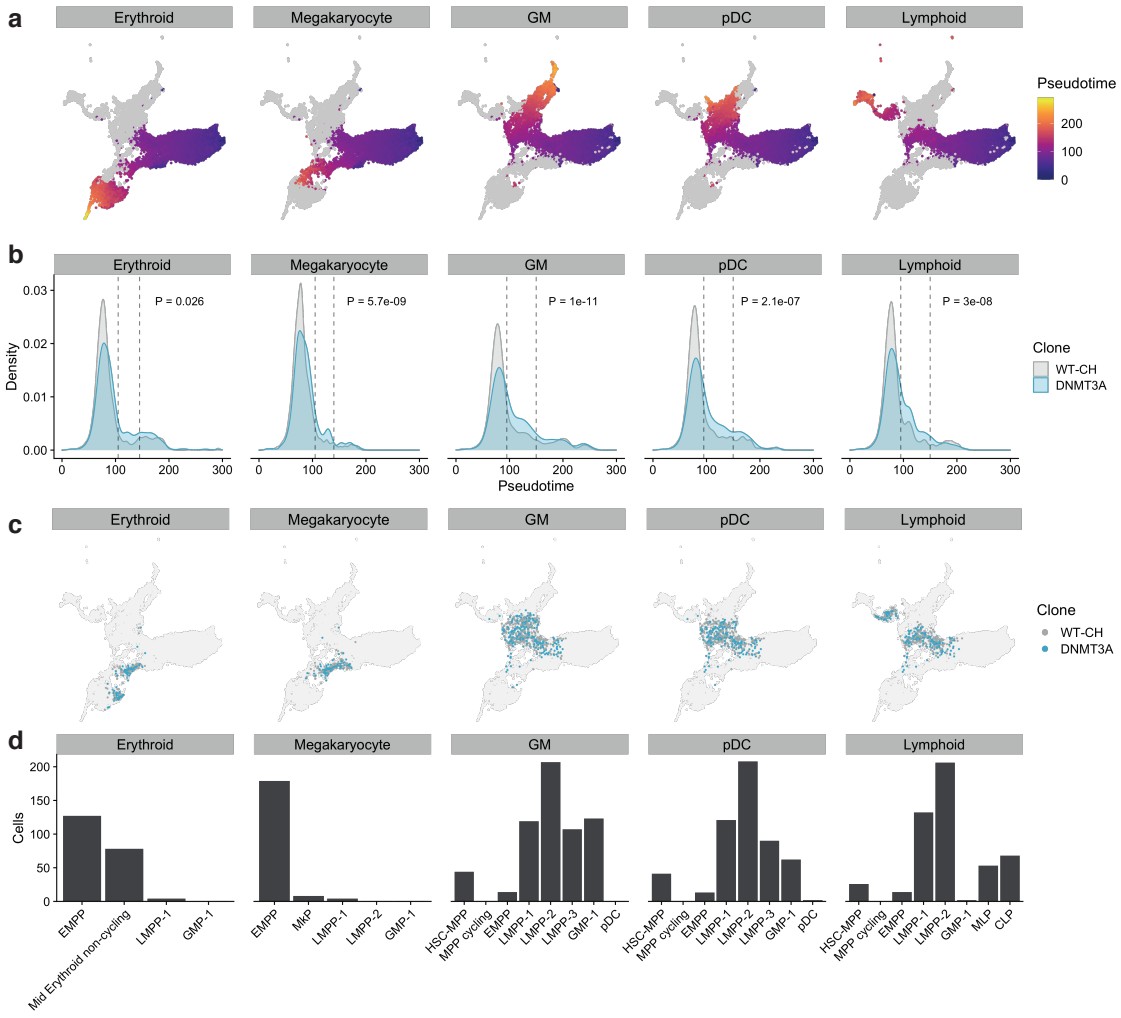


Figure 5.33: Pseudotime analysis of *DNMT3A*-mutant cells. (a) UMAP plots showing the lineage trajectories identified by Slingshot analysis. Cells in each lineage are coloured by their pseudotime score, originating in the HSC-MPP clusters (pseudotime of 0) and progressing towards mature cells (greater pseudotime values). **(b)** Density plots showing the distribution of WT and *DNMT3A*-mutant cells through pseudotime in each lineage. *DNMT3A*-mutant cell density is shown in blue, overlayed on WT cell density (grey). Data include the samples with a dominant *DNMT3A*-mutant clone, excluding NOC115. Differential progression was tested using a generalised linear mixed model accounting for individual sample effects. Regions of each lineage with greatest expansion of mutant cells are indicated by the dashed lines and the identity of these cells are shown in **(c)** and **(d)**. **(c)** UMAP plots showing the location of cells within the regions bounded by the dashed lines in **(b)** for each lineage. Cells are coloured by their genotype. **(d)** Cluster identity of the cells in the regions bounded by the dashed lines in **(b)** for each lineage.

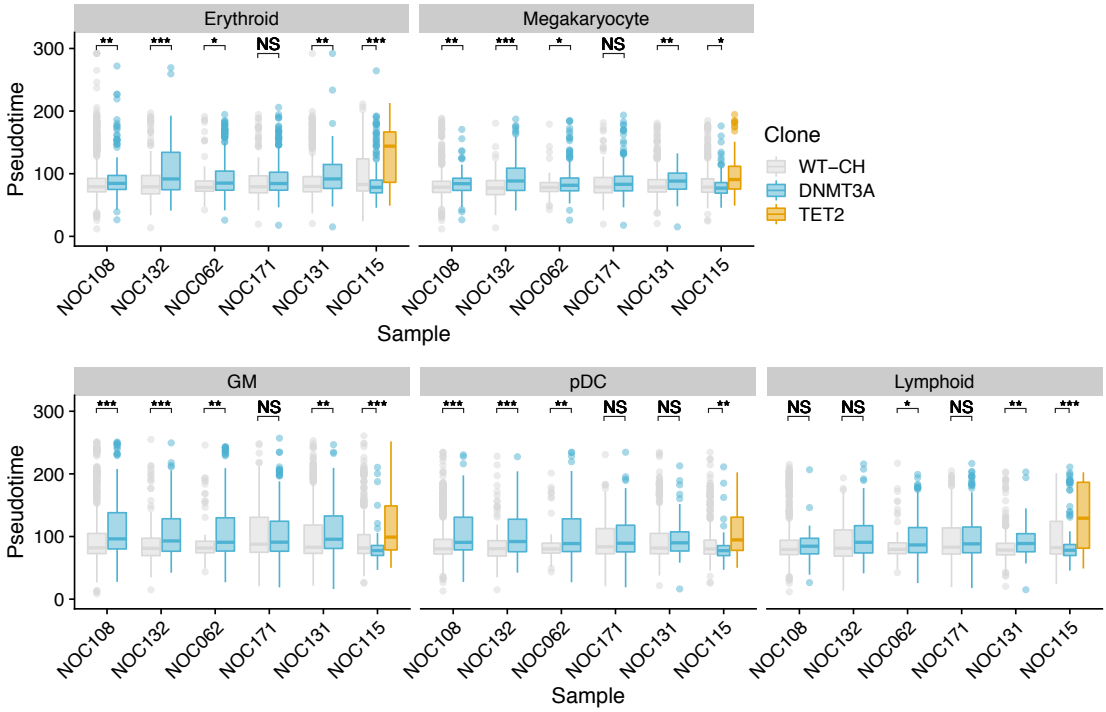


Figure 5.34: Pseudotime for individual *DNMT3A*-mutant samples. Comparison of the pseudotime scores between genotypes for individual *DNMT3A*-mutant samples. *TET2*-mutant cells are also shown for sample NOC115 for comparison. Boxplots show the median and interquartile range of pseudotime scores for single cells within each lineage. Difference in median pseudotime scores tested using Wilcoxon rank-sum test with Benjamini–Hochberg multiple testing correction. NS = non-significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Consistent with the findings from differential abundance analysis, the greatest expansion of mutant cells was in early progenitors, corresponding to EMPP and LMPP (Figure 5.33c-d). Beyond this stage, the density of *DNMT3A*-mutant cells decreased again, suggesting that mutant cells specifically have a differentiation delay in the early progenitor stage.

5.4.2 Transcriptional consequences of *DNMT3A* mutations in progenitors

To understand the phenotype observed in early progenitors, differential gene expression analysis was performed between WT and *DNMT3A*-mutant cells. Paired analysis was performed within CH samples, so that mutant cells were directly compared to WT cells from the same patient, and sample-to-sample differences were accounted for.

First, cells from lympho-myeloid progenitor clusters (LMPP 1–3) were analysed. Gene expression differences were subtle, with only 3 genes reaching significance with a FDR < 0.05 (Figure 5.35a). The upregulated genes were *MPIG6B* (megakaryocyte and platelet inhibitory receptor G6B) and *FUOM* (fucose mutarotase). Notably, *CD9*, which is involved in megakaryocyte differentiation [280] was also upregulated in LMPPs. The most significantly downregulated gene was the interferon-induced gene *IFITM1*, which is normally highly expressed in HSCs and lympho-myeloid progenitors and is associated with cell quiescence [255, 257]. Gene set enrichment analysis (GSEA) was performed against gene signatures collated from a number of studies of human and murine haematopoiesis [50, 236, 238–245, 251–258]. Strikingly, this showed that gene signatures linked to megakaryocyte and erythroid differentiation were enriched in mutant cells, whereas myeloid and lymphoid progenitor gene signatures were negatively enriched (Figure 5.35b-c). To explore this further, AUCell analysis was performed using the same gene signatures. This confirmed that the positively enriched gene signatures were specific to the megakaryocyte-erythroid lineage and negatively enriched gene signatures were specific to the lympho-myeloid lineage (Figure 5.35d).

Next, differential gene expression analysis was performed within the EMPP and mid-erythroid clusters where *DNMT3A* clonal expansion was also seen. No individual gene reached statistical significance but *IFITM1* was again the most significantly downregulated gene (FDR = 0.07). GSEA identified enrichment of megakaryocyte, lymphoid and myeloid gene signatures in mutant cells, as well as antigen presentation signatures (Figure 5.36b). Upregulated genes included class II HLA genes (*HLA-DR* and *DQ*) which are normally expressed in HSC and the lymphoid/DC lineages, as well as *IGHM*, the constant region of the IgM heavy chain, which is an early marker of the lympho-myeloid lineage (Figure 5.27). Concurrently, there was negative enrichment of erythroid and cell cycle signatures. The latter was consistent with the finding that mutant cells were specifically enriched in the non-cycling erythroid cells in the MiloR analysis (Figure 5.32). Note that

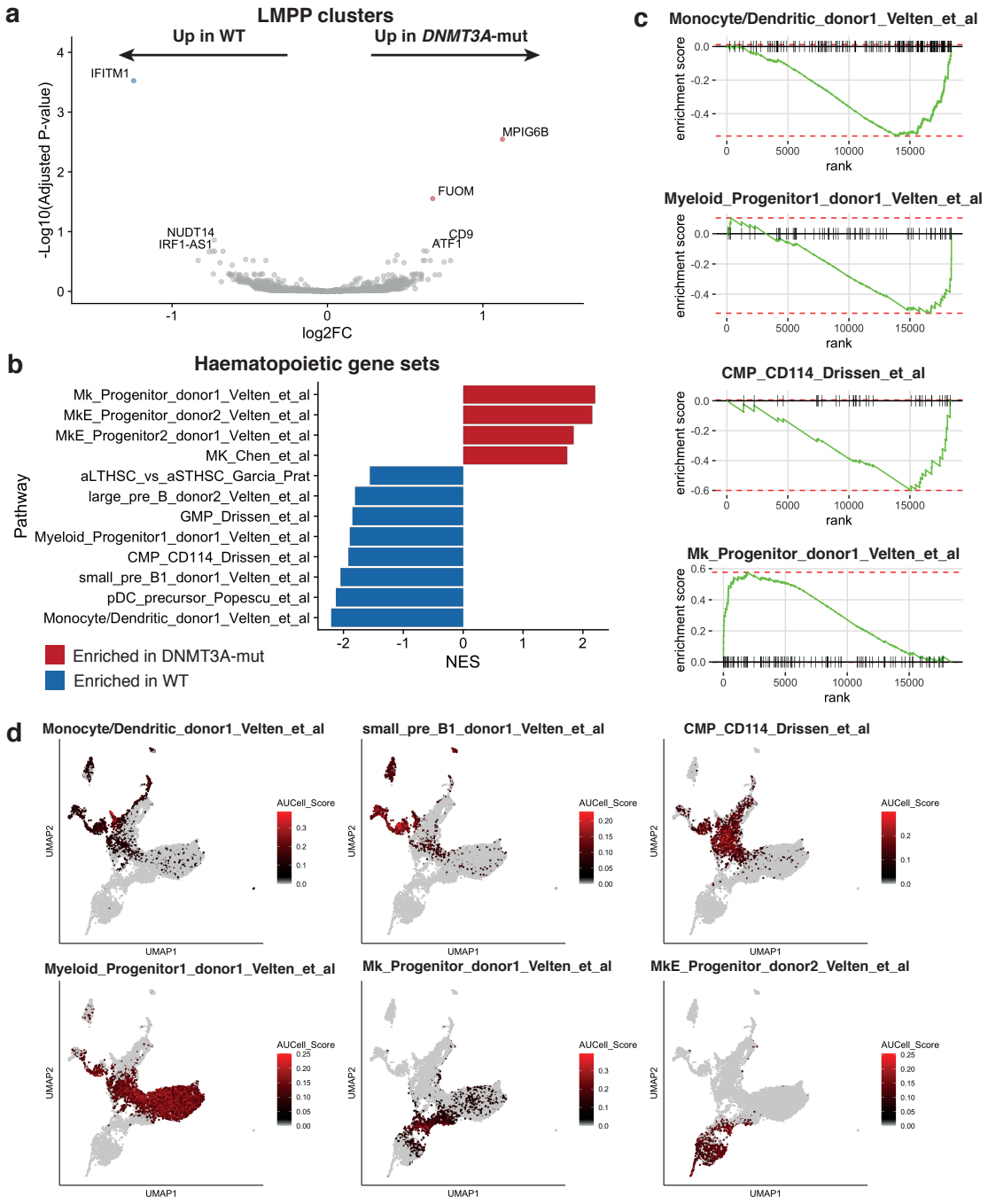


Figure 5.35: Differential gene expression analysis in *DNMT3A*-mutant LMPP. **(a)** Differentially expressed (DE) genes between *DNMT3A*-mutant and WT cells within the LMPP-1, LMPP-2 and LMPP-3 clusters determined by limma (n = 360 WT cells, 275 mutant cells). Genes highlighted in red are significantly upregulated in *DNMT3A*-mutant cells and genes highlighted in blue are downregulated (FDR < 0.05). **(b)** Gene set enrichment analysis (GSEA) against gene signatures from studies of human and murine haematopoiesis [50, 236, 238–245, 251–258, 279]. Gene sets significantly enriched (with FDR < 0.01) in mutant cells are coloured red, and gene sets enriched in WT cells are coloured blue. NES: Normalised enrichment score. **(c)** Enrichment plots for selected gene signatures from **(b)**. **(d)** UMAP plots showing the AUCell expression scores for selected gene signatures from **(b)**.

MYC/MAX targets were negatively enriched, in contrast to the findings from another recent study by Nam *et al.* [184].

The negative enrichment of lineage-specific genes in *DNMT3A*-mutant early progenitors was reminiscent of the impaired differentiation observed in *Dnmt3a*$^{-/-}$ murine HSPCs [166]. To explore this phenotype further, marker genes were obtained for EMPP, LMPP and four lineage-committed progenitor clusters, mid-erythroid progenitors, megakaryocyte progenitors (MkP), GMP and common lymphoid progenitors (CLP). The mean expression of the markers genes for each progenitor type was compared between WT and *DNMT3A*-mutant cells (Figure 5.37). Little difference was seen within EMPP and LMPP but in mid-erythroid progenitors, GMP and CLP, 60–65% of marker genes were downregulated, showing that mutant cells fail to upregulate lineage-specific genes appropriately. Interestingly, the opposite pattern was seen in MkP, where more genes were upregulated than downregulated.

Murine models have also shown that *Dnmt3a*$^{-/-}$ HSPCs fail to repress stem-cell-specific genes during differentiation [166, 172]. To determine if a similar phenotype is observed in human *DNMT3A*-mutant clonal haematopoiesis, HSC marker genes were identified by differential gene expression analysis between HSC and all other cells. To verify that the selected marker genes were specific to HSCs, the mean expression of each gene was compared between WT HSC and WT MPP, EMPP and LMPP. Of the 373 marker genes, 67.3% were downregulated in MPP, 96.0% in EMPP and 99.7% in LMPP, as compared to HSC (Figure 5.38a). The expression of these HSC marker genes was then compared between WT and *DNMT3A*-mutant stem and early progenitor cells. In contrast to the findings from murine experiments, no generalised upregulation of stem-cell-specific genes was observed within mutant HSC, and the majority of these genes were further downregulated in mutant MPP and EMPP. Marginal downregulation was seen in mutant LMPP. Thus, retained expression of stem-cell-specific genes is not a feature of *DNMT3A*-mutant early progenitors.

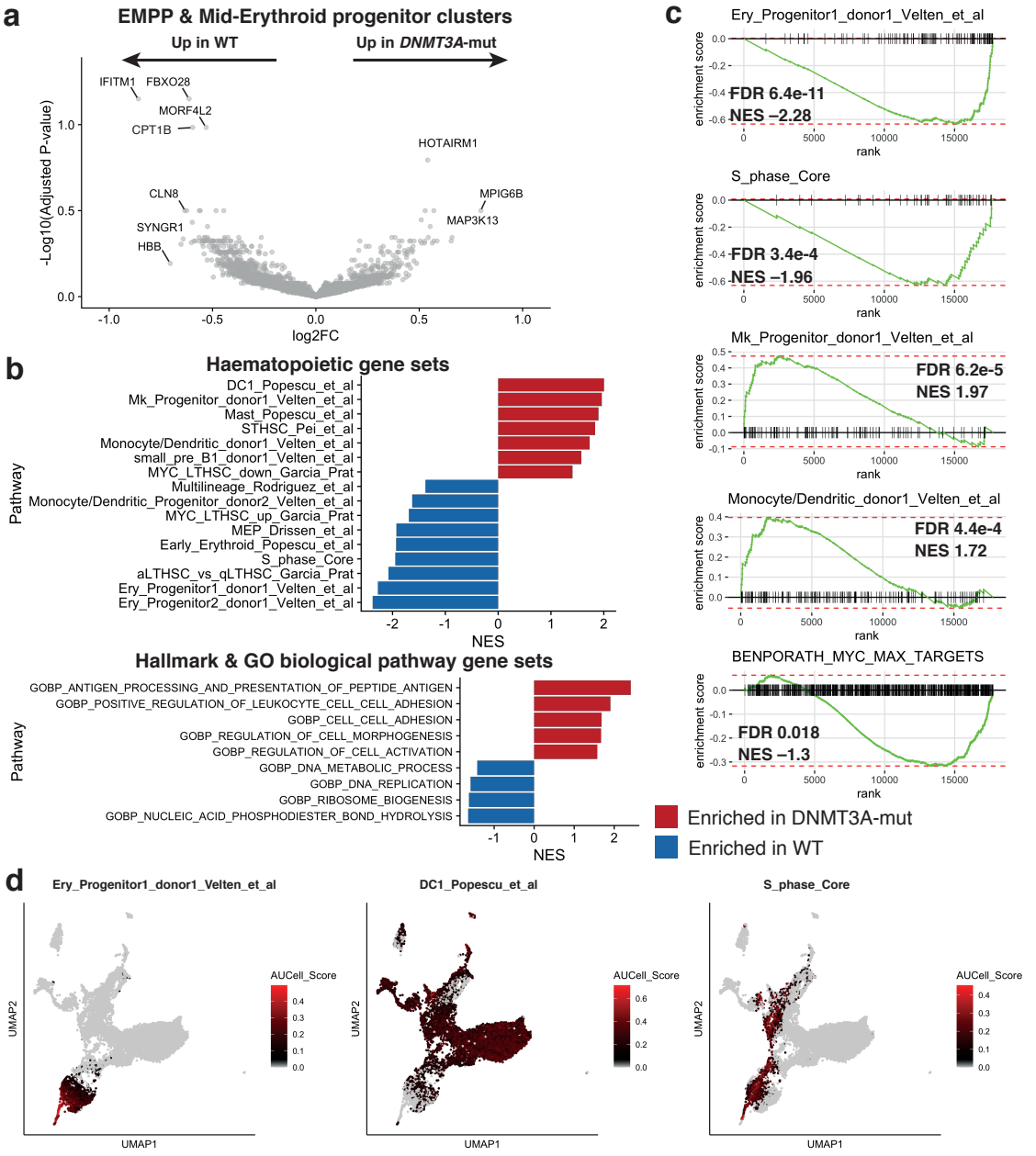


Figure 5.36: Differential gene expression analysis in *DNMT3A*-mutant erythroid/megakaryocyte progenitors. **(a)** Differentially expressed (DE) genes between *DNMT3A*-mutant and WT cells within the EMPP, Mid-erythroid cycling and Mid-erythroid non-cycling clusters determined by limma (n = 604 WT cells, 346 mutant cells). **(b)** Gene set enrichment analysis (GSEA) against haematopoietic gene signatures (top) and Hallmark and Gene Ontology Biological Process gene sets (bottom). Gene sets significantly enriched (with FDR < 0.01) in mutant cells are coloured red, and gene sets enriched in WT cells are coloured blue. NES: Normalised enrichment score. **(c)** Enrichment plots for selected gene signatures from **(b)** and Benporath MYC MAX targets from the MSigDB CGP dataset. **(d)** UMAP plots showing the AUCell expression scores for selected gene signatures from **(b)**.

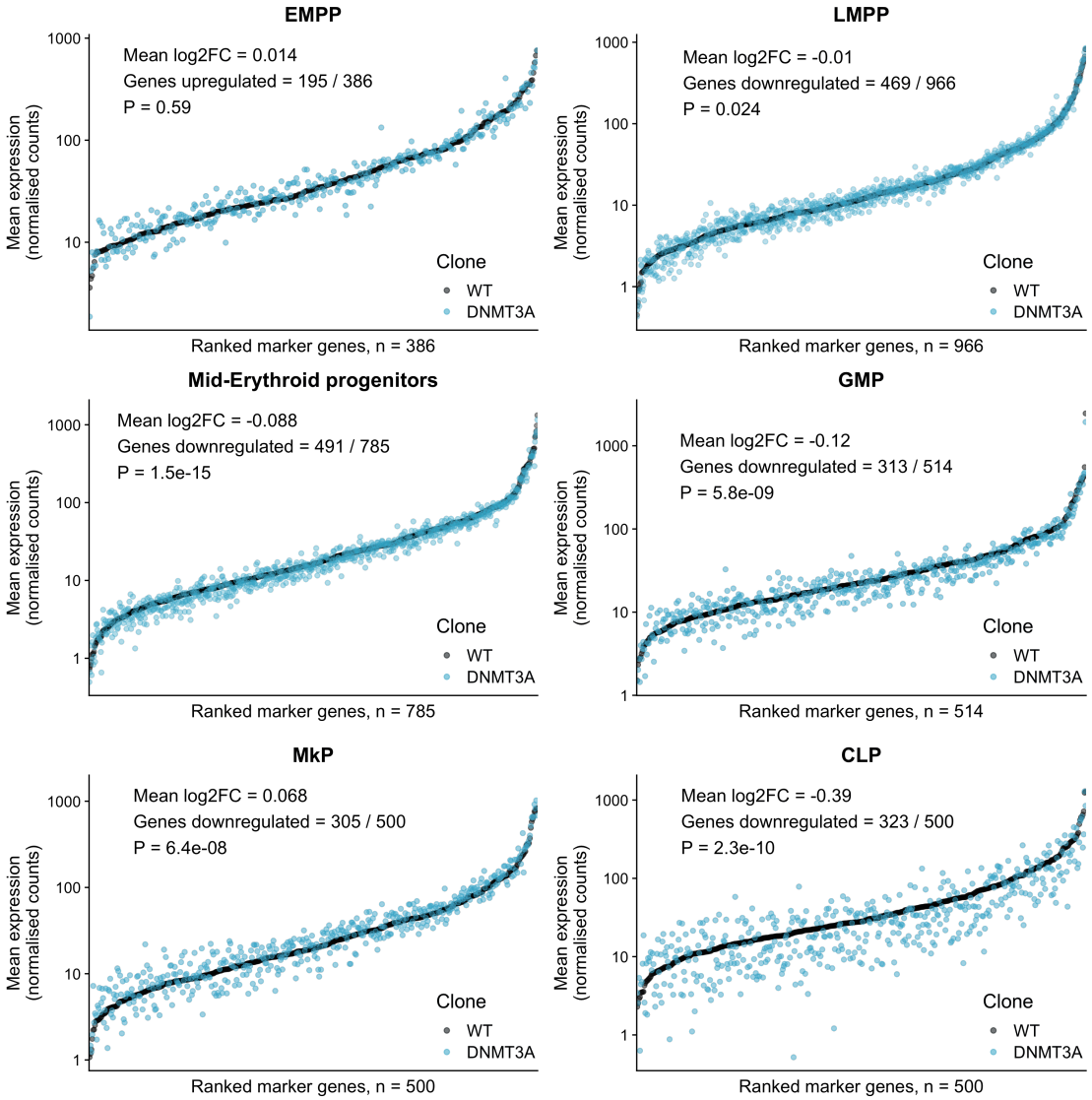


Figure 5.37: Expression of lineage marker genes in *DNMT3A*-mutant progenitors. Comparison of marker gene expression in WT and *DNMT3A*-mutant lineage-restricted progenitors. Marker genes for each progenitor type were identified as described in Section 2.13.6 and the mean expression of each gene was calculated in WT and *DNMT3A*-mutant cells. Genes are ranked along the x-axis according to the mean expression in WT cells (black), and the mean expression in *DNMT3A*-mutant cells (blue) is shown. The mean log2FC of all marker genes is shown for each cell type (positive values = upregulation in mutant). LMPP (LMPP-1, LMPP-2 & LMPP-3 clusters): n = 360 WT, 275 *DNMT3A*-mutant cells. EMPP: n = 254 WT, 163 *DNMT3A*-mutant cells. Mid-erythroid progenitors (cycling & non-cycling): n = 350 WT, 183 *DNMT3A*-mutant cells. Mk progenitors: n = 73 WT, 33 *DNMT3A*-mutant cells. GMP (GMP-1 & GMP-2): n = 323 WT, 177 *DNMT3A*-mutant cells. CLP: n = 53 WT, 41 *DNMT3A*-mutant cells. The number of marker genes for each cell type is shown in the plots. Difference in mean expression tested using paired Wilcoxon signed-rank test.

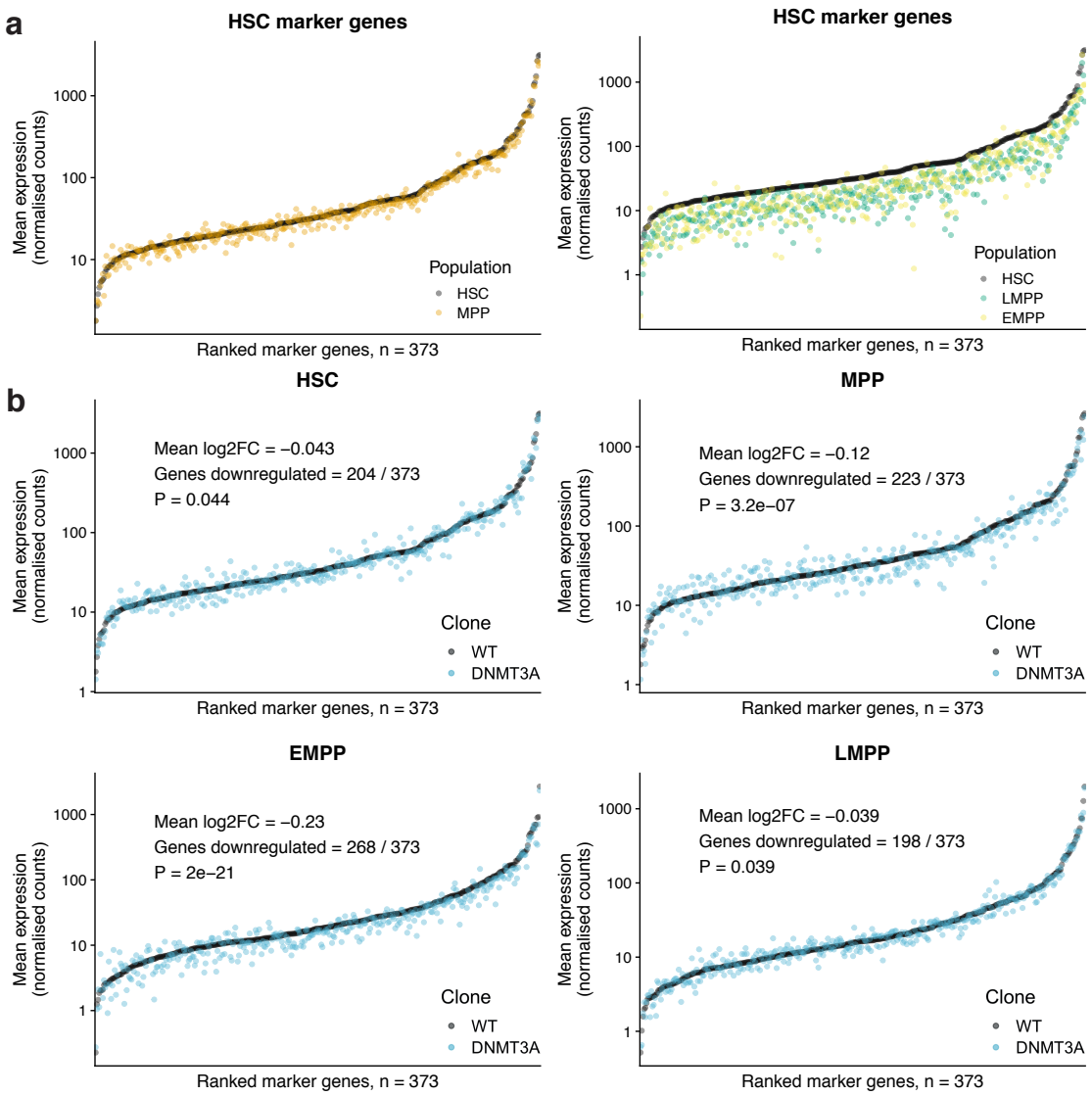


Figure 5.38: Expression of HSC marker genes in *DNMT3A*-mutant HSC and early progenitors. 373 marker genes for immunophenotypic HSC within the HSC-MPP cluster were identified as described in Section 2.13.6, using a minimum log2FC of 0.75. **(a)** Comparison of the mean expression of each gene in WT HSC (CD38–CD45RA–CD90+) and MPP (CD38–CD45RA–CD90–) (left panel), and between WT HSC, EMPP and LMPP (LMPP-1, LMPP-2 & LMPP-3 clusters) (right panel). Genes are ranked along the x-axis according to the mean expression in WT HSC (black). **(b)** Comparison of the mean expression of the 373 HSC marker genes in WT and *DNMT3A*-mutant cells for each of the cell types shown. Genes are ranked along the x-axis according to the mean expression in WT cells (black), and the mean expression in *DNMT3A*-mutant cells (blue) is shown. The mean log2FC of all marker genes is shown for each cell type (positive values = upregulation in mutant). HSC (immunophenotypic CD38–CD45RA–CD90+): n = 1022 WT, 375 *DNMT3A*-mutant cells. MPP (immunophenotypic CD38–CD45RA–CD90–): n = 276 WT, 124 *DNMT3A*-mutant cells. EMPP: n = 254 WT, 163 *DNMT3A*-mutant cells. LMPP (LMPP-1, LMPP-2 & LMPP-3 clusters): n = 360 WT, 275 *DNMT3A*-mutant cells. Difference in mean expression tested using paired Wilcoxon signed-rank test.

5.4.3 Transcriptional consequences of *DNMT3A* mutations in HSC

While *DNMT3A*-mutant clones reach maximal expansion in early progenitors, the origin of clonal expansion was clearly identified as being in the HSC-MPP compartment.

Mutant cells in the HSC-MPP cluster were slightly shifted towards more differentiated states than WT cells, but there was no significant difference within immunophenotypic HSC (Figure 5.39a). To interrogate the phenotype of mutant HSC/MPP further, AUCell analysis was performed using published gene signatures from studies of human haematopoiesis. This enables the expression of gene sets to be quantified in single cells and compared between WT and mutant populations. Signatures associated with cell cycle, quiescence, long-term CD90+CD49f+ HSC (LT-HSC) and downstream progenitors from the myelo-lymphoid and megarkarocyte-erythroid lineages were chosen [236, 272, 281]. First, to determine if the high frequency of mutant cells in the LMPP and EMPP populations could be explained by increased proliferation or exit from the HSC compartment, the expression of cell cycle and quiescence gene signatures was analysed within the HSC-MPP cluster. While there was no difference within MPP, HSC had a lower expression of cell cycle genes suggesting greater quiescence (Figure 5.39b). Further supporting this, *DNMT3A*-mutant HSC had increased expression of genes differentially expressed between quiescent and activated LT-HSCs, but showed no difference in expression of genes that distinguish LT-HSCs from ST-HSC (MPPs) (Figure 5.39c,d). Furthermore, there was no significant difference in the frequency of mutant cells between CD49f+ and CD49f– HSC and MPP subpopulations or in cell size, which increases on differentiation (Figure 5.39d,e). Consistent with the absence of a lineage bias, *DNMT3A*-mutant HSC and MPP did not show differences in expression of signatures derived from lympho-myeloid or megakaryocyte-erythroid progenitors (Figure 5.39f).

To explore the phenotype of mutant HSC-MPP in more detail, differential gene expression analysis was performed. A total of 1298 WT and 499 *DNMT3A*-mutant cells were included in the analysis, restricted to immunophenotypic HSC and MPP within the HSC-MPP cluster. This analysis identified 25 differentially expressed genes with a FDR < 0.05, of which 19 were upregulated and 6 were

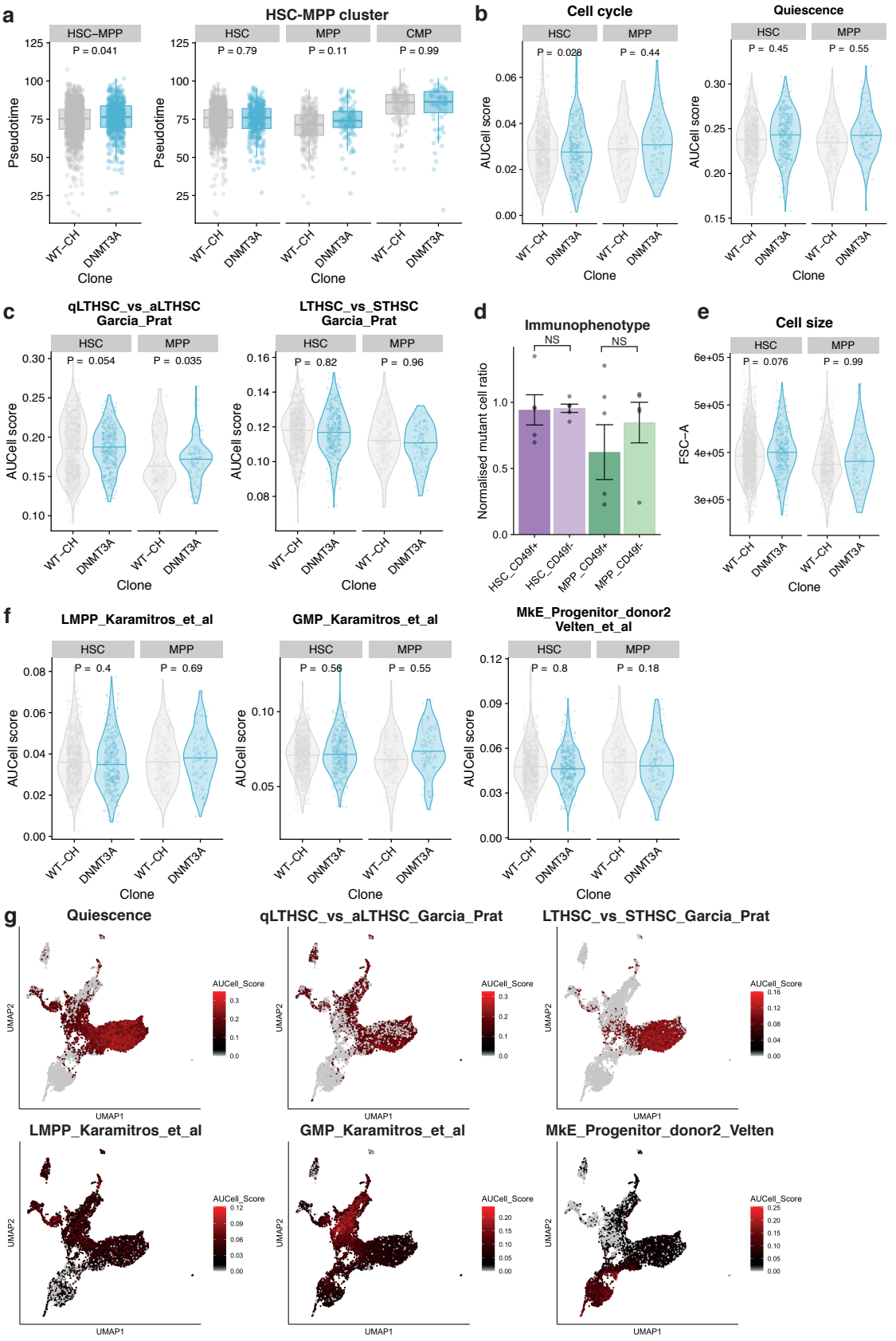


Figure 5.39: *(legend on next page)*

Figure 5.39 *(previous page)*: Quiescence and lineage bias in *DNMT3A*-mutant HSC/MPP. (a) Comparison of pseudotime between WT and *DNMT3A*-mutant cells within the HSC-MPP cluster. All cells in the cluster are shown on the left, and broken down by immunophenotype on the right. Differential testing of pseudotime scores was done using a generalised linear mixed model accounting for individual sample effects. **(b-c)** AUCell expression scores in immunophenotypic HSC and MPP within the HSC-MPP cluster. **(b)** AUCell scores for cell cycle genes (combined S-phase and G2M-phase signatures [236]) and a quiescence gene signature [230]. **c)** AUCell scores for genes upregulated in quiescent LT-HSC relative to activated LT-HSC (left), and genes upregulated in LT-HSC relative to ST-HSC (right) [254]. **(d)** Frequency of *TET2*-mutant cells in immunophenotypic HSC/MPP split by CD49f expression. For each sample, the frequency of mutant cells was normalised to the mutant cell frequency of the whole HSC-MPP cluster. Bars show the mean and error bars represent the standard error. Points show the values for individual samples. Paired t-test with Bonferroni multiple testing correction; ** $P < 0.01$. **(e)** Cell size of immunophenotypic HSC/MPP populations split by CD49f expression, measured by flow cytometry forward scatter (FSC-A). **(f)** AUCell scores for lineage signatures. **(g)** UMAP plots showing the AUCell expression scores for the gene signatures used in **(b**,**c** and **f)**. Difference in AUCell scores **(b**,**c** and **f)** and cell size **(e)** was tested using a two-way ANOVA with sample interaction.

downregulated. The most upregulated genes were *IGHM*, *ARMH1* (armadillo like helical domain containing 1) a gene of unknown function which is widely expressed in haematopoiesis, and, in the male samples, *RPS4Y1* (Ribosomal Protein S4 Y-Linked 1), a component of the 40S ribosomal subunit. Other upregulated genes included two serine proteases, *PRSS21* and *PRSS57*, *LST1* (Leukocyte Specific Transcript 1), a transmembrane adaptor protein expressed in leukocytes of myeloid lineage [282], *GATAD2B*, which is part of the methyl-CpG-binding protein-1 complex (MeCP1/NuRD complex) that represses gene expression by deacetylating methylated nucleosomes, and *BCL3*, a transcriptional co-activator that associates with NF-kappa B homodimers. The top downregulated gene was *FCER1A*, the alpha subunit of the immunoglobulin epsilon receptor, which is a marker for the megakaryocyte/erythroid/eosinophil/basophil/mast cell lineage (Figure 5.23). Other downregulated genes included *PMAIP1*, a pro-apoptotic BH3-only protein that inhibits BCL2, and *CDK2AP2* (cyclin dependent kinase 2 associated protein 2), which inhibits cell cycle G1/S phase transition by repressing CDK2 expression and activation [283]. Gene set enrichment analysis showed enrichment of B cell-associated signatures, with genes in the leading edge including

IGHM, *IGHD* and *CD52*, as well as genes upregulated in quiescent LT-HSCs as compared to cultured (or activated) HSCs [254]. In addition, gene signatures linked to biosynthetic processes such as ribosome biogenesis and protein folding were downregulated. These are amongst the most upregulated pathways in HSCs in the transition from quiescence to cell cycle entry [254].

To focus on the effect of mutations in stem cells, differential expression was performed within immunophenotypic CD90+ HSC. 25 DE genes were identified, many of which were shared with those identified in HSC/MPP. The gene signature associated with quiescent LT-HSCs was again significantly enriched (FDR < 0.001), with genes in the leading edge including *JUP* (junction plakoglobin), *NEDD4L* (NEDD4 Like E3 Ubiquitin Protein Ligase) and *BCL3*.

Finally, to determine if these expression changes were present in the most primitive stem cell compartment, the analysis was repeated for CD90+CD49f+ LT-HSCs (Figure 5.40e-f). Sample NOC132 had only 11 CD90+CD49f+ cells, so was excluded from this analysis; for all other samples there were at least 12 cells of either genotype. The only differentially expressed gene that reached significance was *IGHM* which was upregulated in mutant cells. Other upregulated genes included *ATG9B*, a phospholipid scramblase that participates in autophagosome formation, and *TNIP1*, a negative regulator of nuclear factor κB (NF-κB) [284]. The most significantly downregulated gene was *EIF4G3*, which is involved in the recognition of the mRNA cap and recruitment of mRNA to the ribosome. Indeed GSEA also identified negative enrichment of genes related to translation, mainly ribosomal protein genes (Figure 5.40f). Additionally, there was positive enrichment of the quiescent LT-HSC signature, supporting increased quiescence of *DNMT3A*-mutant stem cells.

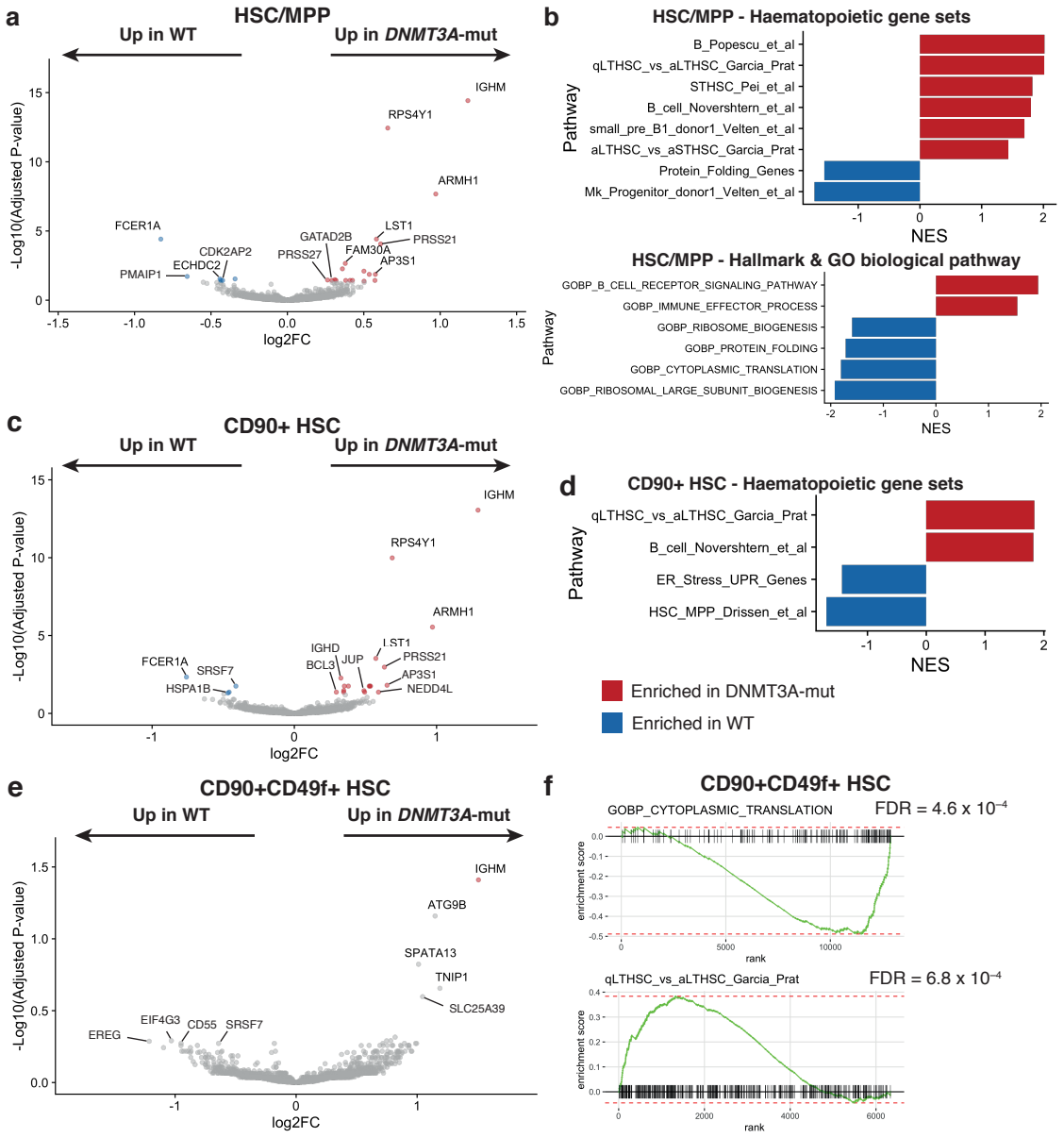


Figure 5.40: Differential gene expression analysis in *DNMT3A*-mutant HSC/MPP. **(a)** Differentially expressed (DE) genes between *DNMT3A*-mutant and WT cells within CD38–CD45RA– cells of the HSC-MPP cluster determined by limma (n = 1298 WT cells, 499 mutant cells across 6 samples). **(b)** GSEA against haematopoietic gene signatures (top) and Hallmark and Gene Ontology Biological Process gene sets (bottom) for DE genes in HSC/MPP. Gene sets significantly enriched (with FDR < 0.05) in mutant cells are coloured red, and gene sets enriched in WT cells are coloured blue. **(c)** DE genes between *DNMT3A*-mutant and WT cells within CD38–CD45RA–CD90+ HSC within the HSC-MPP cluster determined by limma (n = 1022 WT cells, 375 mutant cells across 6 samples). **(d)** GSEA against haematopoietic gene signatures for DE genes in HSC, showing gene sets significantly enriched (with FDR < 0.05). **(e)** DE genes between *DNMT3A*-mutant and WT cells within CD38–CD45RA–CD90+CD49f+ LT-HSC in the HSC-MPP cluster (n = 315 WT cells, 85 mutant cells across 4 samples, excluding NOC132). **(f)** Enrichment plots for enriched gene signatures from haematopoietic and Gene Ontology gene sets.

5.5 Transcriptomic analysis of *TET2*-mutant clonal haematopoiesis

5.5.1 Analysis of *TET2*-mutant clonal size through haematopoietic differentiation

To understand how the acquisition of *TET2* mutations perturb normal haematopoiesis, the frequency of mutant cells was analysed across the haematopoietic landscape. As for *DNMT3A*, *TET2*-mutant cells were found intermingled with WT cells and did not form any novel cell clusters (Figure 5.41a).

Differential abundance analysis with MiloR showed relative depletion of *TET2*-mutant cells in HSC-MPP and EMPP neighbourhoods and relatively increased abundance in myeloid progenitors (LMPP-3, GMP-1, GMP-2 and pDC) (Figure 5.41b-c). This was corroborated by analysis of normalised mutant cell frequencies across the cell types identified by Louvain clustering, which showed that the average frequency of *TET2*-mutant cells was 5-fold higher in GMP clusters than in the HSC-MPP cluster (Figure 5.41d). While *TET2*-mutant clonal expansion was most prominent in myeloid progenitors, there was no clear deficiency in differentiation towards the lymphoid and megakaryocyte-erythroid lineages. Indeed, compared to HSC-MPP, mutant cell frequencies were higher in B cell progenitors in two out of four samples, and in erythroid progenitors in three out of four, though this increase was not statistically significant with MiloR testing. However, *TET2*-mutant cells were significantly depleted in T cells, particularly in the CD4+ T cell cluster, similar to the findings for *DNMT3A*-mutant cells. Thus, while *TET2*-mutant clones were identified in HSC-MPP, further clonal expansion was observed in late progenitor cells of all lineages, with a distinct increase in the myeloid lineage where clones reached their maximal size in the GMP and pDC.

Pseudotime trajectory analysis was used to quantify how *TET2* mutations alter the frequency of cell states through haematopoietic differentiation. The Slingshot trajectories used for analysis of *DNMT3A*-mutant clones were applied to *TET2*-mutant clones (Figure 5.42a). In all lineages, the median pseudotime score for

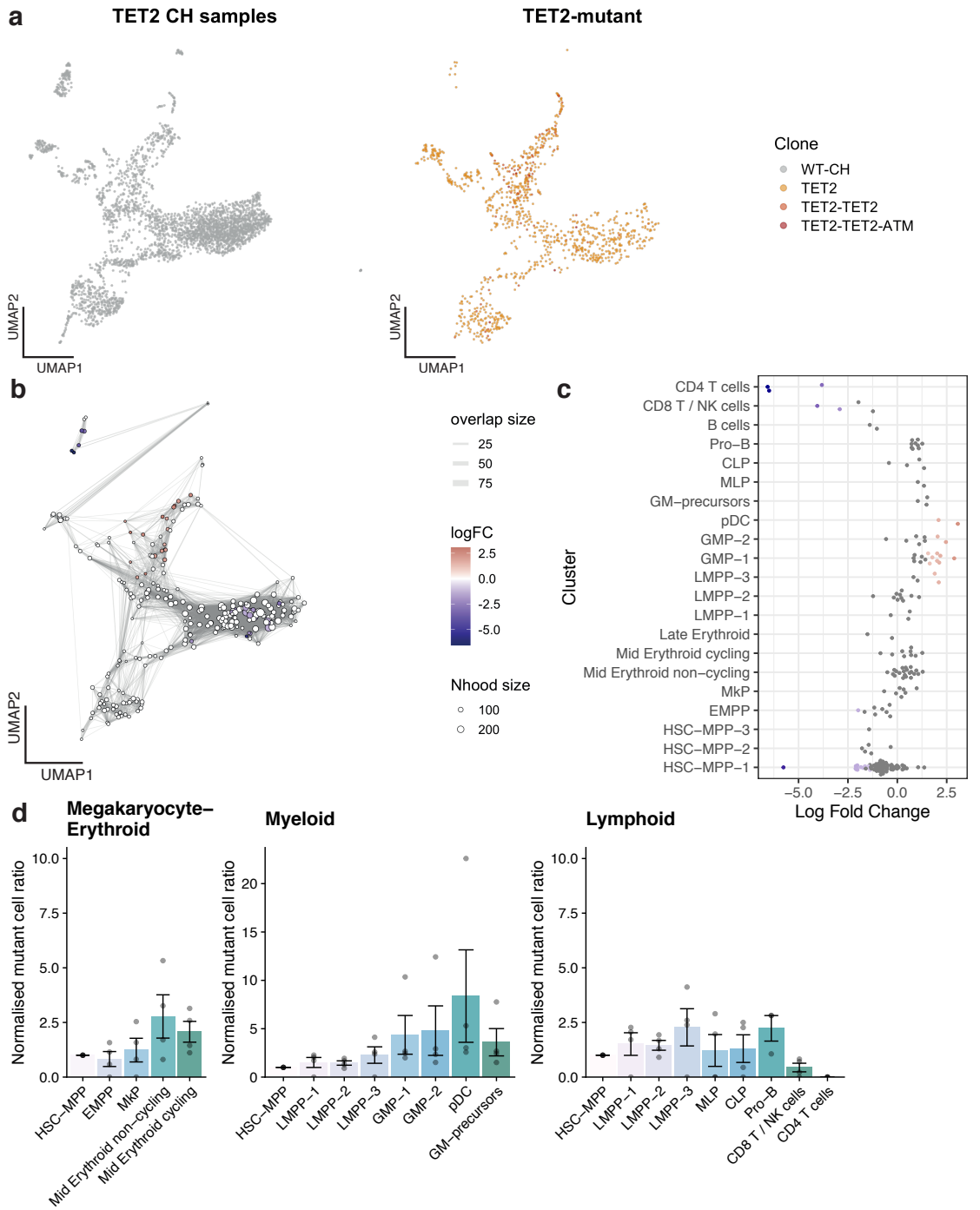


Figure 5.41: Abundance of *TET2*-mutant cells across the haematopoietic hierarchy. (a) UMAP plots showing WT and mutant cells from the 4 samples with a dominant *TET2*-mutant clone. Cells are coloured by genotype, where WT-CH indicates WT cells (grey), and mutant cells are coloured by the genotypes listed in the legend. **(b)** Neighbourhood graph of the results from MiloR differential abundance testing. Nodes represent neighbourhoods, the size of each node corresponds to the number of cells in the neighbourhood and graph edges depict the number of cells shared between neighbourhoods.
(continued next page)

Figure 5.41 ***(previous page)*****:** The layout of nodes is determined by the position of an index cell in the UMAP. Neighbourhoods with a significant difference in abundance between WT and *TET2*-mutant cells (with FDR < 0.2) are coloured by their log fold change; red indicates a relative increase in abundance of *TET2*-mutant cells and blue indicates a relative decrease. **(c)** Beeswarm plot of the distribution of log fold change between WT and *TET2*-mutant cells. Neighbourhoods are labelled with the most abundant cell type cluster (with at least 60% purity). Neighbourhoods with a significant difference in abundance between WT and *TET2*-mutant cells are coloured by their log fold change as in **(b)**. **(d)** Frequency of *TET2*-mutant cells in each Louvain cluster. For each sample, the frequency of mutant cells was normalised to the mutant cell frequency in the HSC-MPP clusters, in order to enable comparison across samples. Bars show the mean and error bars represent the standard error. Points show the values for individual samples. For each cluster, samples were excluded if there were fewer than 8 cells assigned to that cluster. Nhood: neighbourhood, logFC: log fold change.

TET2-mutant cells was significantly higher than for WT cells indicating a greater frequency of mutant cells in more differentiated cell states (generalised linear mixed model accounting for individual sample effects, Figure 5.42b). In the GM and pDC myeloid lineages, the shift of *TET2*-mutant cells towards more differentiated states was seen in every sample, whereas for the erythroid, megakaryocyte and lymphoid lineages, it was only seen in a subset of the samples (Figure 5.43). Comparison of the density of WT and mutant cells along each lineage trajectory, confirmed that the frequency of *TET2*-mutant cells was greatest in parts of the trajectory that correspond to CD38+ late progenitors, and this increase was most prominent in the GM and pDC myeloid lineages (Figure 5.42b-d). This was a noticeable difference from *DNMT3A*-mutant clones, which reach maximal expansion in early progenitors (Figure 5.33).

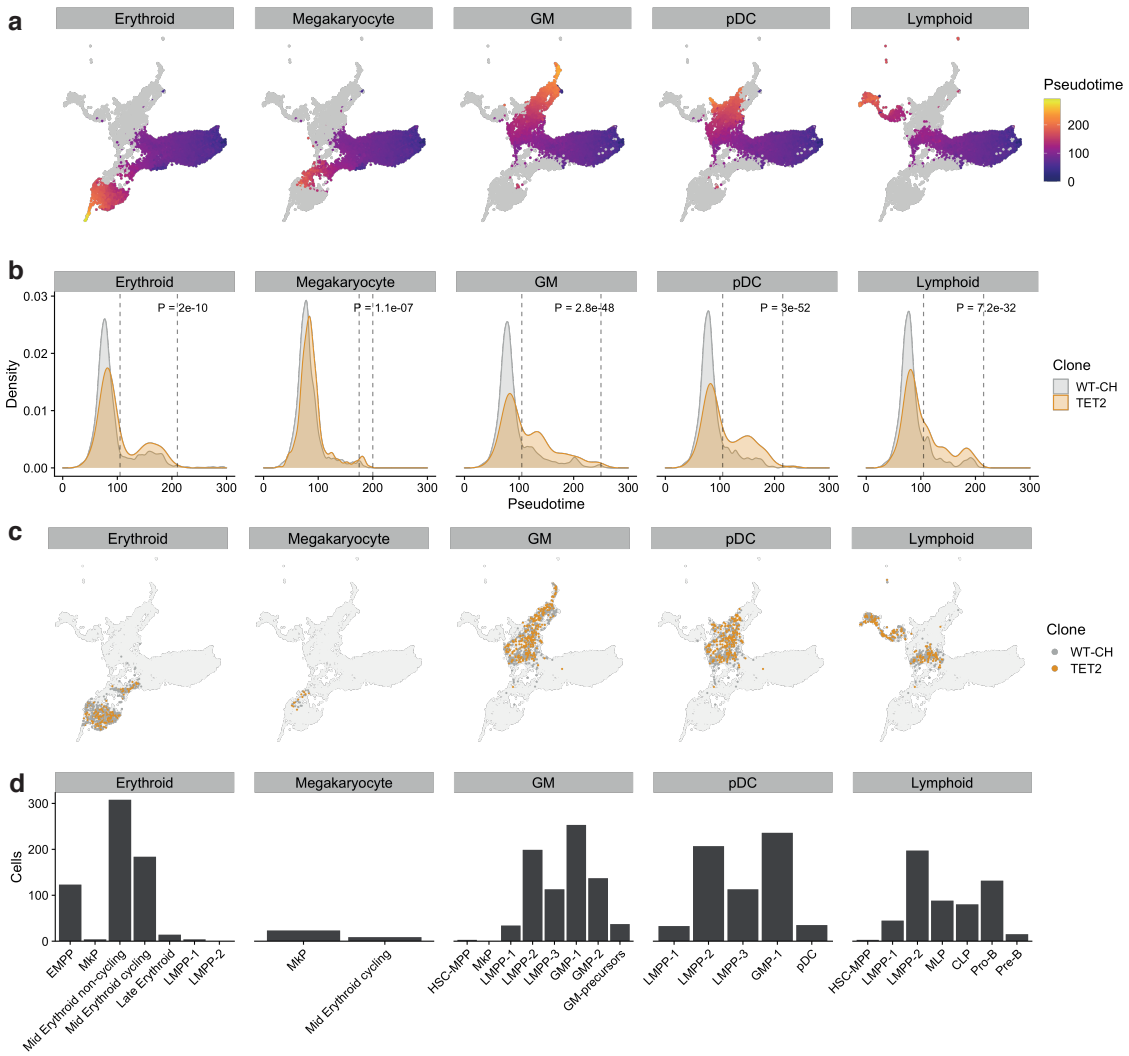


Figure 5.42: Pseudotime analysis of *TET2*-mutant cells. **(a)** UMAP plots showing the lineage trajectories identified by Slingshot analysis. Cells in each lineage are coloured by their pseudotime score, originating in the HSC-MPP clusters (pseudotime of 0) and progressing towards mature cells (greater pseudotime values). **(b)** Density plots showing the distribution of WT and *TET2*-mutant cells through pseudotime in each lineage. *TET2*-mutant cell density is shown in orange, overlayed on WT cell density (grey). Data include the 4 samples with a dominant *TET2*-mutant clone. Differential progression was tested using a generalised linear mixed model accounting for individual sample effects. Regions of each lineage with greatest expansion of mutant cells are indicated by the dashed lines and the identity of these cells are shown in **(c)** and **(d)**. **(c)** UMAP plots showing the location of cells within the regions bounded by the dashed lines in **(b)** for each lineage. Cells are coloured by their genotype. **(d)** Cluster identity of the cells in the regions bounded by the dashed lines in **(b)** for each lineage.

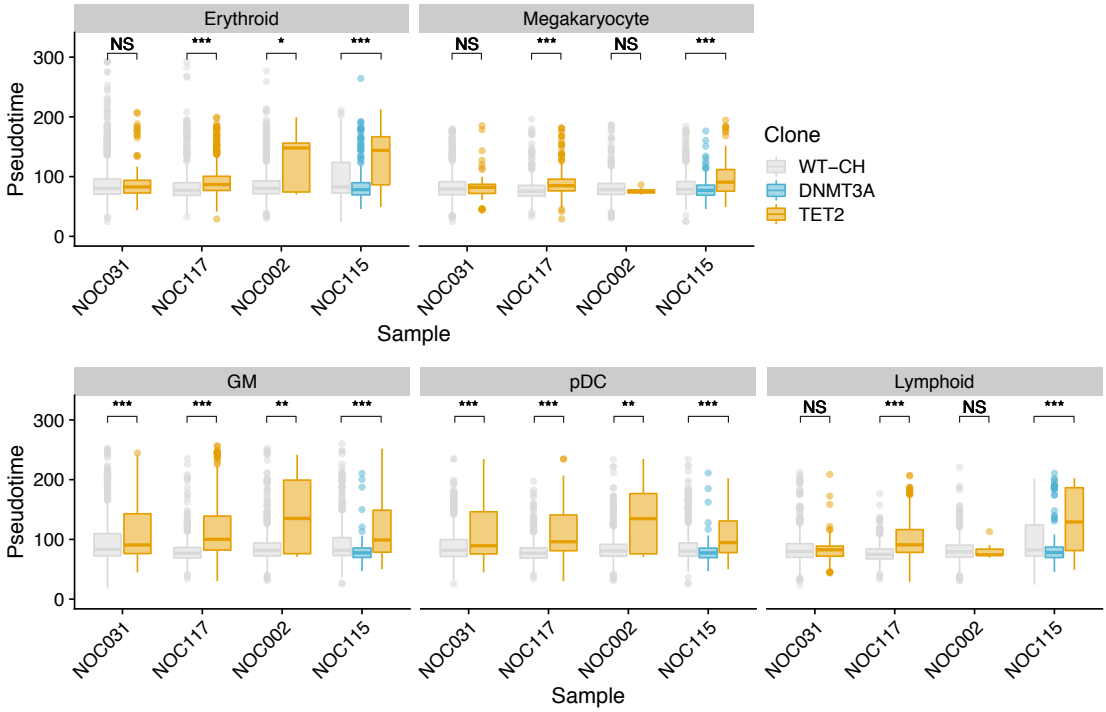


Figure 5.43: Pseudotime for individual *TET2*-mutant samples. Comparison of the pseudotime scores between WT and *TET2*-mutant cells for individual samples. *DNMT3A*-mutant cells are also shown for sample NOC115 for comparison. Boxplots show the median and interquartile range of pseudotime scores for single cells within each lineage. Difference in median pseudotime scores tested using Wilcoxon rank-sum test with Benjamini–Hochberg multiple testing correction. NS = non-significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

5.5.2 Transcriptional consequences of *TET2* mutations in progenitors

To identify transcriptional dysregulation in *TET2*-mutant progenitor cells, differential gene expression analysis was performed between WT and *TET2* single-mutant cells.

First, GMP progenitors were analysed, as that is where maximal *TET2*-clonal expansion was observed. Differential expression analysis in the GMP-1 and GMP-2 clusters identified 10 genes that reached statistical significance (limma, FDR $<$ 0.05). Notably, 3 of the upregulated genes were transcription factors involved in myeloid differentiation, *CEBPD*, *KLF4* and *IRF8*. *CEBPD* (CCAAT/enhancer-binding protein delta) is a transcriptional activator that is involved in granulocyte differentiation and regulation of inflammatory responses, including as a regulator

of IL-6 transcription [285–287]. *KLF4* promotes monocytic differentiation [288], and *IRF8* is required for both monocytic and dendritic cell differentiation [289]. Other upregulated genes included *MPO* (myeloperoxidase), *GPR183*, a chemotactic receptor found on dendritic cells and the angiogenin receptor *PLXNB2*. The only downregulated gene was *GATA2*, which is normally expressed in early HSPCs and becomes downregulated on differentiation. Gene set enrichment analysis showed an enrichment of neutrophil, monocyte and dendritic cell gene signatures in *TET2*-mutant GMP and negative enrichment of signatures linked to HSC/MPP and the megakaryocyte/erythroid lineage. Interestingly, G2M checkpoint and E2F target genes were also downregulated in mutant cells, suggesting reduced cell cycle activity, while genes involved in oxidative phosphorylation were upregulated.

To determine if the myeloid differentiation bias could be detected earlier in differentiation, differential gene expression analysis was performed within the LMPP-1 and LMPP-2 clusters, where the myeloid and lymphoid lineages diverge. Eight genes were significantly differentially expressed (limma, FDR < 0.05). Of the seven upregulated genes, four were mitochondrially-encoded components of enzymes in the mitochondrial electron transport, *MT-ND4* (subunit of NADH dehydrogenase, complex I), MT-CO2 and MT-CO3 (subunits of succinate dehydrogenase, complex II) and *MT-ATP6* (subunit of ATP synthase, complex V). Other upregulated genes included *CLEC11A* (osteolectin), *ARHGAP32*, a GTPase-activating protein, and *MEF2C-AS1*, a long non-coding RNA (lncRNA). The serine protease *PRSS21* was downregulated. Gene set enrichment analysis showed enrichment of neutrophil progenitor gene signatures, indicating that *TET2*-mutant LMPP have a transcriptional bias towards the myeloid lineage. Interestingly, a number of stem cell-associated signatures were also enriched, including two signatures of HSC dormancy, one derived from a study of label-retaining HSCs in mice (Cabezas_sc_dHSC [255]) and the other derived from a study of retinoic acid dim-expressing HSCs which are slow cycling (RACFP_DIM_Lauridsen_et_al, [290]). The leading edge for these gene sets included *FTL* (Ferritin light chain), *LST1*, and ribosomal protein genes. Indeed, gene ontology pathways linked to ribosomal biogenesis and translation as well as

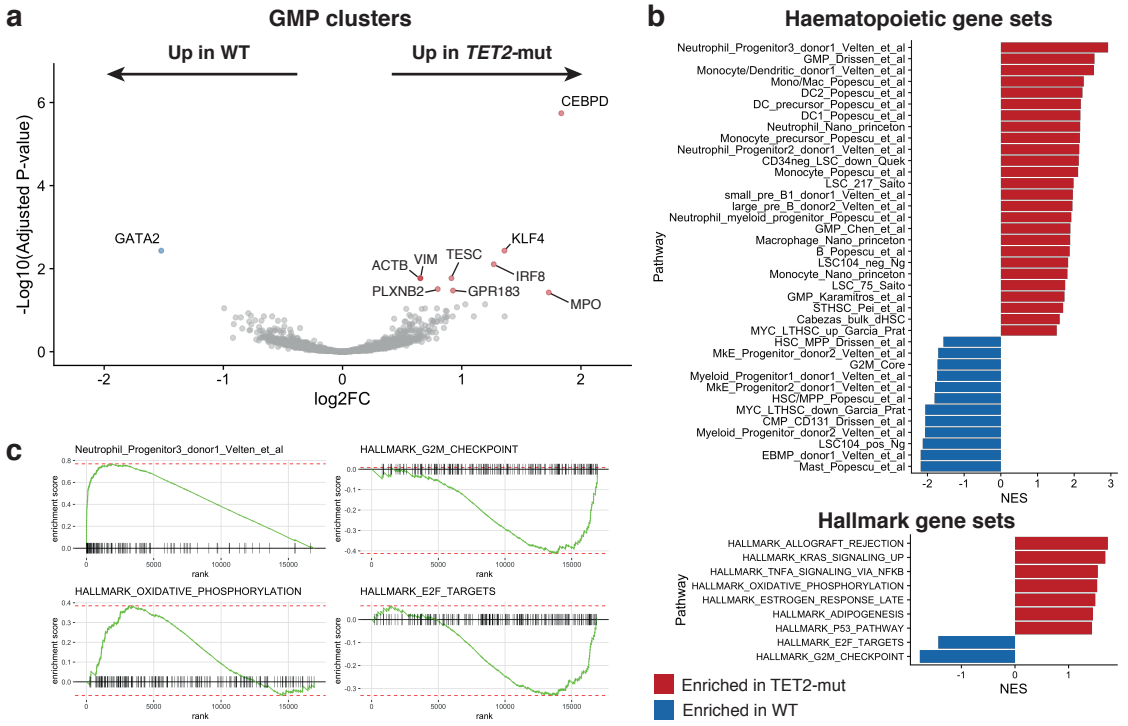


Figure 5.44: Differential gene expression analysis in *TET2*-mutant GMP. (a) Differentially expressed (DE) genes between *TET2*-mutant and WT cells within the GMP-1 and GMP-2 clusters determined by limma (n = 237 WT cells, 157 mutant cells). Genes highlighted in red are significantly upregulated in *TET2*-mutant cells and genes highlighted in blue are downregulated (FDR < 0.05). **(b)** Gene set enrichment analysis (GSEA) against haematopoietic gene signatures (top) and Hallmark and Gene Ontology Biological Process gene sets (bottom). Gene sets significantly enriched (with FDR < 0.01) in mutant cells are coloured red, and gene sets enriched in WT cells are coloured blue. NES: Normalised enrichment score. **(c)** Enrichment plots for selected gene signatures from **(b)**.

the mitochondrial electron transport chain were enriched in mutant cells, whereas signatures linked to cell cycle and DNA replication were negatively enriched. One of the most downregulated genes was *CDC5L*, a DNA-binding protein that forms part of the spliceosome and promotes G2 progression and mitotic entry [291, 292]. Overall, this shows that some of the transcriptional differences observed in *TET2*-mutant GMP are already established at the LMPP stage, including upregulation of myeloid genes and reduced expression of genes linked to cellular proliferation.

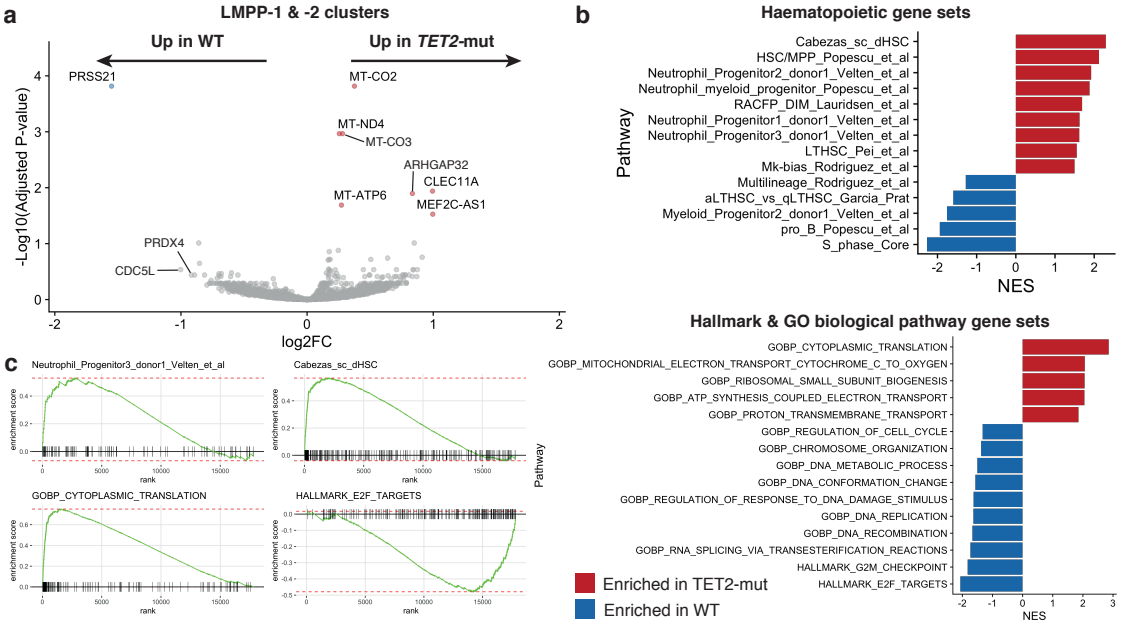


Figure 5.45: Differential gene expression analysis in *TET2*-mutant LMPP. **(a)** Differentially expressed (DE) genes between *TET2*-mutant and WT cells within the LMPP-1 and LMPP-2 clusters determined by limma (n = 360 WT cells, 134 mutant cells). Genes highlighted in red are significantly upregulated in *TET2*-mutant cells and genes highlighted in blue are downregulated (FDR < 0.05). **(b)** Gene set enrichment analysis (GSEA) against haematopoietic gene signatures (top) and Hallmark and Gene Ontology Biological Process gene sets (bottom). Gene sets significantly enriched (with FDR < 0.05) in mutant cells are coloured red, and gene sets enriched in WT cells are coloured blue. NES: Normalised enrichment score. **(c)** Enrichment plots for selected gene signatures from **(b)**.

5.5.3 Transcriptional consequences of *TET2* mutations in HSC

TET2-mutant cells were identified in the HSC-MPP cluster in all four samples analysed, where they constituted 1.6–33% of cells in each sample. Comparison of pseudotime scores showed *TET2*-mutant cells were significantly closer to differentiating states than WT cells. This was case for all cells within the HSC-MPP cluster and within immunophenotypic HSC (Figure 5.46a). AUCell analysis using the cell cycle, quiescence and HSC/MPP gene signatures described earlier was performed. There was no significant difference in expression of cell cycle or quiescence gene signatures between WT and *TET2*-mutant HSCs and MPPs. However, mutant HSCs had significantly lower expression of the long-term HSC signature suggesting that they are transcriptionally closer to MPPs or short-term HSCs (ST-HSCs),

which have been defined as CD34+CD38–CD90–CD49f– immunophenotypic cells (Figure 5.46b-c) [272, 281]. In support of this finding, the frequency of *TET2*-mutant cells was significantly higher within CD49f– HSCs than CD49f+ HSCs in all of the four samples analysed (Figure 5.46d). Exit from quiescence is also associated with an increase in cell size as HSCs increase biosynthetic processes [254, 281], and *TET2*-mutant HSC were found to be significantly larger than WT HSCs from the same samples as measured by flow cytometry indexing (Figure 5.46e).

To determine if there was evidence of a lineage bias in the HSC compartment, expression of lineage signatures was quantified. Indeed, *TET2*-mutant HSCs had significantly higher expression of signatures derived from LMPP and GMP progenitors and decreased expression of the megakaryocyte-erythroid progenitor signature (Figure 5.46f), showing that mutant HSC are transcriptionally biased towards the myeloid lineage.

Differential gene expression analysis was performed to identify transcriptional dysregulation underlying the phenotype of *TET2*-mutant stem cells. Analysis of all immunophenotypic HSC and MPP in the HSC-MPP cluster identified 65 genes that were significantly upregulated in *TET2*-mutant HSC-MPP and 22 that were downregulated (Figure 5.47a). The upregulated genes included the pseudogene *WASH6P*, a small Rho GTPase, *RHOB*, mitochondrial electron transport chain components *MT-CO1* and *MT-CO2*, and the cell surface receptors *CD9*, *ADGRE5* (CD97) and *SELL* (Selectin L, CD62L). They also included the transcription factors *NFIB*, *ANXA1*, *KLF6* and *KLF2*. Downregulated genes included *PRSS21*, *MTLN* (mitoregulin), which regulates mitochondrial complex assembly and the transcription factor *EGR1*, which regulates HSC proliferation [293]. GSEA showed enrichment of progenitor signatures and a ST-HSC signature, whereas LT-HSC signatures were negatively enriched, confirming the findings from AUCell analysis (Figure 5.47b). Although GSEA against Hallmark and Gene Ontology genesets was performed, no pathways were significantly enriched. Differential expression analysis within CD90+ HSC yielded similar results both in DE genes and enriched pathways (Figure 5.47c-d).

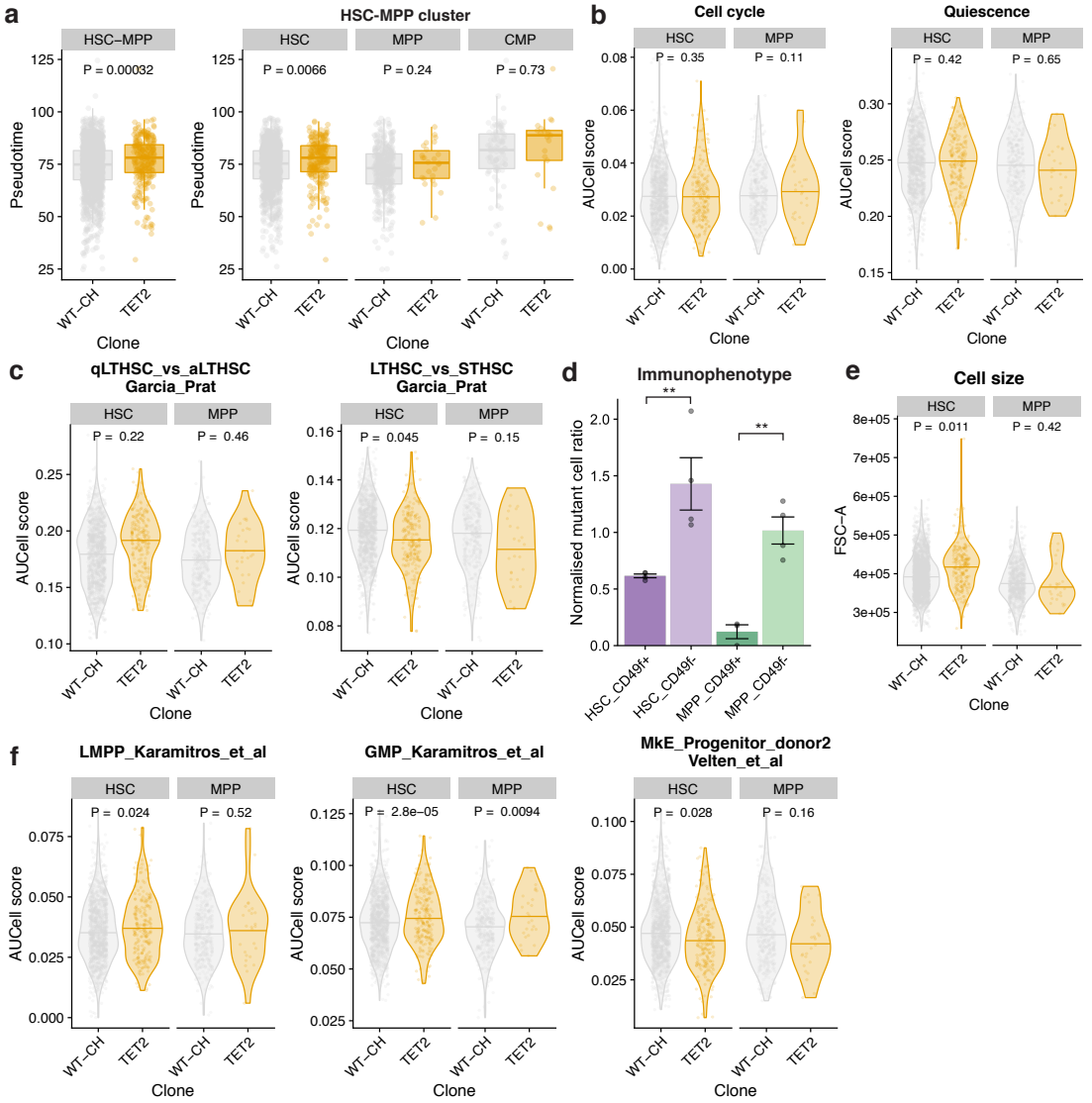


Figure 5.46: Quiescence and lineage bias in *TET2*-mutant HSC/MPP. (a) Comparison of pseudotime between WT and *TET2*-mutant cells within the HSC-MPP cluster. All cells in the cluster are shown on the left, and broken down by immunophenotype on the right. Differential testing of pseudotime scores was done using a generalised linear mixed model accounting for individual sample effects. **(b-c)** AUCell expression scores in immunophenotypic HSC and MPP within the HSC-MPP cluster. **(b)** AUCell scores for cell cycle genes (combined S-phase and G2M-phase signatures [236]) and a quiescence gene signature [230]. **c)** AUCell scores for genes upregulated in quiescent LT-HSC relative to activated LT-HSC (left), and genes upregulated in LT-HSC relative to ST-HSC (right) [254]. **(d)** Frequency of *TET2*-mutant cells in immunophenotypic HSC/MPP split by CD49f expression. For each sample, the frequency of mutant cells was normalised to the mutant cell frequency of the whole HSC-MPP cluster. Bars show the mean and error bars represent the standard error. Points show the values for individual samples. Paired t-test with Bonferroni multiple testing correction; ** $P < 0.01$. **(e)** Cell size of immunophenotypic HSC/MPP populations split by CD49f expression, measured by flow cytometry forward scatter (FSC-A). **(f)** AUCell scores for lineage signatures. Difference in AUCell scores **(b**,**c** and **f)** and cell size **(e)** was tested using a two-way ANOVA with sample interaction.

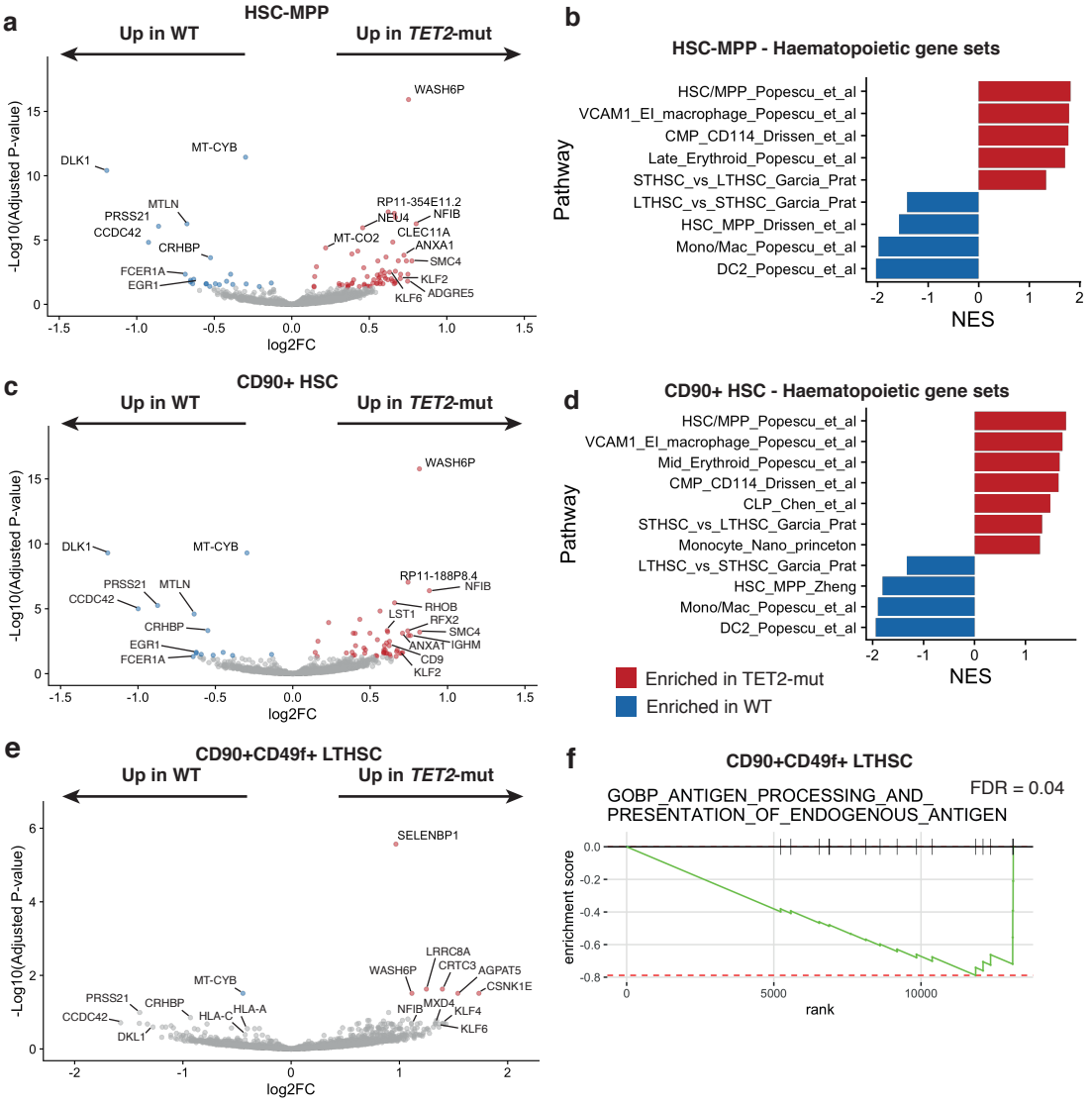


Figure 5.47: Differential gene expression analysis in *TET2*-mutant HSC/MPP. **(a)** Differentially expressed (DE) genes between *TET2*-mutant and WT cells within CD38–CD45RA– cells of the HSC-MPP cluster determined by limma (n = 1484 WT cells, 236 mutant cells across 4 samples). **(b)** GSEA against haematopoietic gene signatures for DE genes in HSC/MPP. Gene sets significantly enriched (with FDR < 0.1) in mutant cells are coloured red, and gene sets enriched in WT cells are coloured blue. NES: Normalised enrichment score. **(c)** DE genes between *TET2*-mutant and WT cells within CD38–CD45RA–CD90+ HSC in the HSC-MPP cluster determined by limma (n = 1118 WT cells, 210 mutant cells across 4 samples). **(d)** GSEA against haematopoietic gene signatures for DE genes in HSC, showing gene sets significantly enriched (with FDR < 0.05). **(e)** DE genes between *TET2*-mutant and WT cells within CD38–CD45RA–CD90+CD49f+ LT-HSC in the HSC-MPP cluster determined by limma (n = 261 WT cells, 32 mutant cells across 2 samples, NOC031 and NOC117). **(f)** Enrichment plot for the GO signature which was negatively enriched in mutant cells.

Finally, the analysis was repeated for cells in CD90+CD49f+ LT-HSCs (Figure 5.47e-f). For samples NOC002 and NOC115 there was only a single mutant cell identified in this population, so these samples were excluded from analysis. Seven genes were differentially expressed with FDR < 0.05. Upregulated genes included a selenium binding protein, *SELENBP1*, a protein kinase, *CSNK1E*, and the transcriptional coactivator, *CRTC3*. The only significantly downregulated gene was mitochondrially encoded cytochrome B *MT-CYB*. Notably several of the DE genes identified in HSC, were also differentially expressed in LT-HSC but did not reach significance, including *KLF4*, *KLF6*, *DLK1*, *PRSS21* and *CCDC42*. GSEA identified negative enrichment of the endogenous antigen presentation pathway in mutant LT-HSCs, which was driven by *HLA-A*, *HLA-B*, *HLA-C* and *B2M* which were amongst the most significantly downregulated genes. No haematopoietic gene sets were significantly enriched, suggesting the enrichment of ST-HSC and progenitor gene signatures observed in mutant HSC/MPP was due to increased frequency of CD49f– ST-HSCs and not due to altered gene expression in LT-HSCs.

It was noted that a number of genes encoding components of the mitochondrial electron transport chain and ATP synthase were differentially expressed in HSC/MPP and LMPP. Visualising expression of these genes through the myeloid lineage, showed that some were already upregulated in *TET2*-mutant cells in the HSC-MPP cluster (*MT-ND4* and *MT-CO2*) (Figure 5.48). The most prominent upregulation occurred in the LMPP and GMP-1 clusters, where the greatest expansion of *TET2*-mutant clones was observed.

Differential gene expression analysis had also identified transcription factors that are known to have key roles in regulating haematopoietic differentiation. *GATA2* was identified as downregulated in *TET2*-mutant GMP, but visualising its expression showed that it is already downregulated in mutant HSC-MPP and LMPP (Figure 5.49). *KLF2* and *KLF6* were upregulated in HSC-MPP and for *KLF2*, this upregulation is maintained into the LMPP and GMP compartments. In contrast, the upregulation of *KLF4* appears to be confined to the GMP compartment, so is less likely to be important to the mutant phenotype in stem cells. *IRF8* and

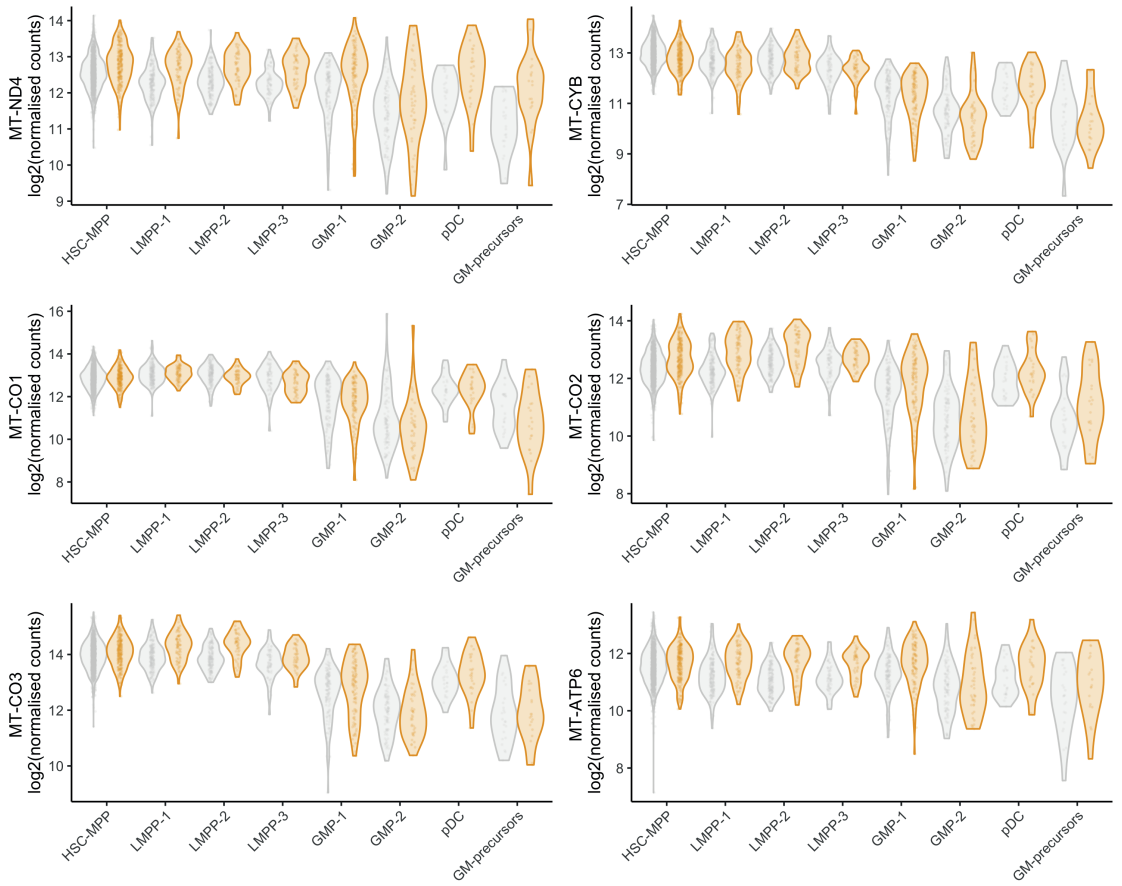


Figure 5.48: Expression of mitochondrial genes in the myeloid lineage in *TET2*-mutant samples. Violin plots of normalised, log2 transformed expression values for mitochondrial-encoded genes that were differentially expressed between WT and *TET2*-mutant cells. The expression in each cluster of the myeloid lineage shown.

CEBPD are first expressed later in myeloid differentiation, in the GMP, and this upregulation is more pronounced in *TET2*-mutant cells.

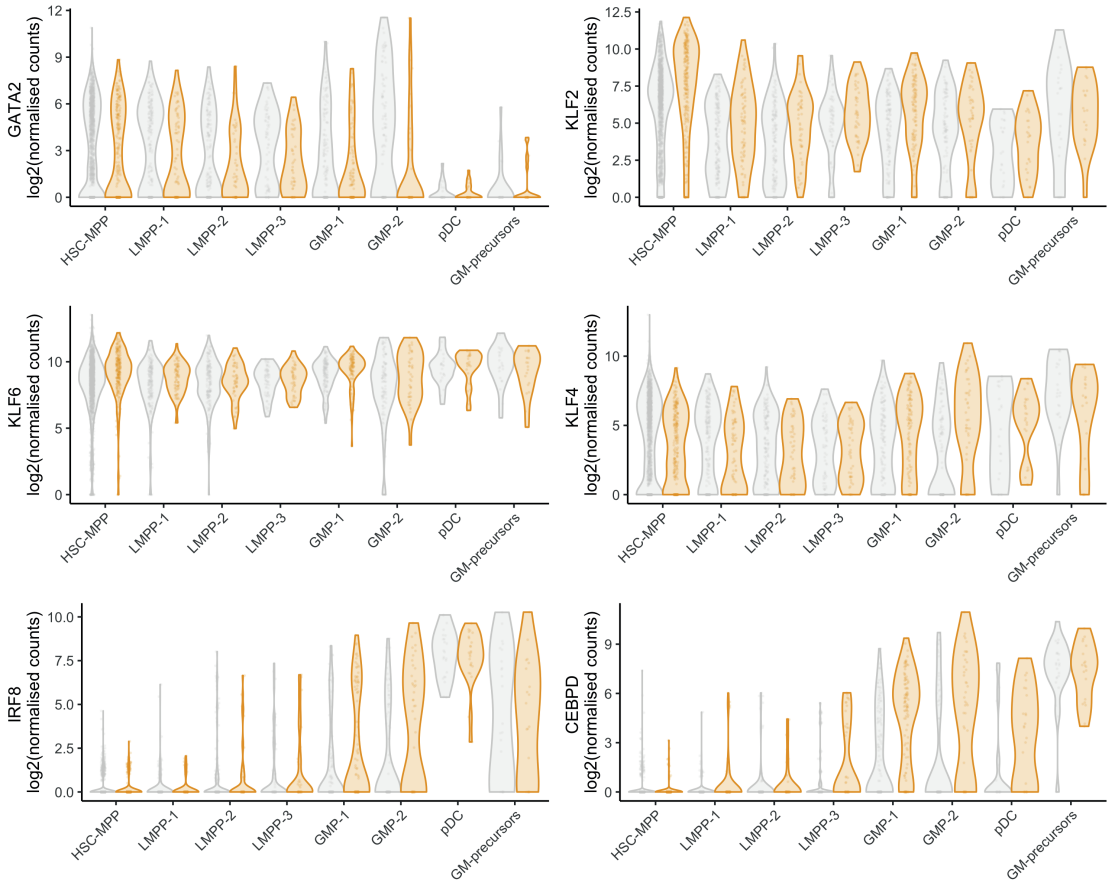


Figure 5.49: Differentially expressed transcription factors in *TET2*-mutant samples. Violin plots of normalised, log2 transformed expression values for transcription factors that were differentially expressed between WT and *TET2*-mutant cells. The expression in each cluster of the myeloid lineage shown.

5.5.4 Consequences of biallelic *TET2* mutations

Biallelic *TET2* mutations are often observed in myeloid malignancy. Analysis of single and double *TET2*-mutant clones in sample NOC002 enabled the effects of a second mutation to be explored.

Within the HSC-MPP the single and double-mutant clones were approximately of equal size — 13 out of 441 cells (2.9%) were mutant, of which 7 had a single *TET2* mutation and 6 were *TET2*-double mutant (Figure 5.50a). Each *TET2*-mutant clone was identified in only 1 out of 109 (0.9%) CD49f+ LT-HSCs. The single mutant clone showed a similar pattern of expansion as *TET2*-mutant cells in the other CH samples, reaching maximal clonal size in mid-erythroid and myeloid progenitors. However, the double-mutant clone gained a considerable advantage in

the myeloid lineage, where it contributed to more than half of cells in GMP, pDC and GM-precursor clusters. Interestingly, while this was associated with an absolute expansion of immunophenotypic pDC, the GMP population was of normal size, whereas CD38– HSC and MPP were expanded (Figure 5.12). There were very few CD10+ lymphoid progenitors in this sample, and they were almost exclusively WT.

Differential expression analysis between subclones was limited by low cell numbers, but comparison between the double-mutant clone and WT cells within the GMP clusters showed negative enrichment for neutrophil signatures. Furthermore, cell cycle genesets were negatively enriched, suggesting increased quiescence in double-mutant cells. Notably the top downregulated gene was *MS4A3* a modulator of the G1-S cell cycle transition. Analysis of single cells by AUCell enrichment confirmed these trends.

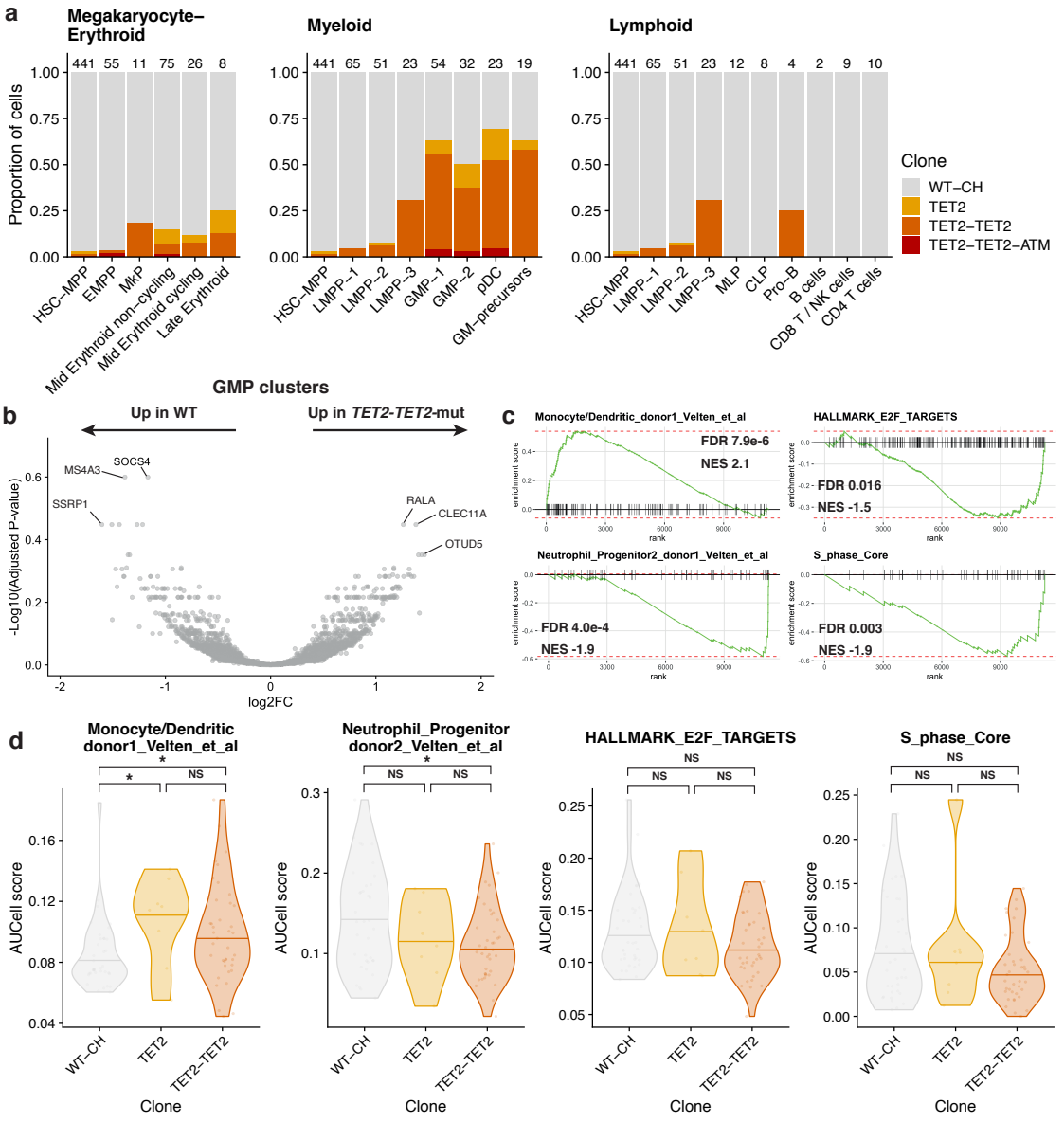


Figure 5.50: Consequences of biallelic *TET2*-mutations in sample NOC002. **(a)** Clonal composition within each cluster for sample NOC002. The number of cells analysed in each cluster are shown above. **(b)** Differentially expressed genes between *TET2*-double-mutant and WT cells within the GMP-1 and GMP-2 clusters determined by limma (n = 36 WT cells, 39 mutant cells). **(c)** Enrichment plots for GSEA against haematopoietic and Hallmark gene signatures for genes differentially expressed between WT and *TET2*-double-mutant GMP, showing gene sets significantly enriched (with FDR < 0.05). **(d)** AUCell expression scores for myeloid and S-phase gene sets in the GMP clusters. * $P < 0.05$, Wilcoxon rank-sum test.

Chapter 6

Discussion

Contents

6.1 Establishing a human clonal haematopoiesis cohort . . 183
6.2 Developing the TARGET-seq+ method 185
6.3 Transcriptional consequences of *DNMT3A* and *TET2* mutations in haematopoietic stem and progenitor cells 187
6.4 Limitations of this study and future work 193

The discovery that clonal haematopoiesis harbouring cancer-associated mutations is common in healthy older people and confers an increased risk of blood cancer has generated enormous interest and opened a new field of research into somatic clonal selection. Expanded mutant clones have been identified in multiple adult tissues and are likely to affect any replicating tissue maintained by long-lived stem cells [1, 11, 12, 294]. However, clones in the haematopoietic system are arguably of unique interest, since their consequences may not only affect the blood, but any tissue of the body that interacts with immune cells which are the progeny of a mutant clone.

While the clinical implications of clonal haematopoiesis are increasingly clear, many questions remain. The mechanisms by which CH-associated mutations lead to clonal expansion are unclear, and it is not known why some clones progress to malignancy and others do not. Because clone size is associated with the risk of leukaemia and other adverse outcomes, it critical to understand the mechanisms that drive clonal expansion and the consequences of mutations on the haematopoietic system.

One of the most surprising and puzzling observations from the population studies of clonal haematopoiesis is that most cases are associated with mutations in just two genes, *DNMT3A* and *TET2*. The fact that loss-of-function mutations in enzymes that regulate global DNA methylation patterns should confer a selective advantage seems counter-intuitive. The mechanisms by which mutations in these

genes lead to clonal expansion have been the subject of much speculation and study, but remain unexplained.

Mouse models using conditional knockout or mutation of these genes have led to important insights into their function in haematopoiesis. They have confirmed that homozygous or heterozygous loss of either *DNMT3A* or *TET2* confers a selective advantage, and perturbs haematopoietic stem and progenitor cell function [166, 213, 214]. *DNMT3A* or *TET2* loss is associated with focal changes in DNA methylation and subtly altered gene expression. However, no studies have yet shown the importance of any of these changes in driving clonal expansion or indeed that similar changes are observed in humans with clonal haematopoiesis. All of these studies are limited by the use of competitive transplantation or genetic alteration of the entire haematopoietic system, neither of which faithfully represent the human condition. Furthermore, there are likely to be important differences between mice and humans in haematopoietic stem and progenitor gene regulation, so findings from mice may not necessarily translate to humans. Finally, the lifespan of mice makes it a difficult species in which to study a condition that appears to take many years to develop. No mouse has yet been found to have naturally-occurring clonal haematopoiesis.

In this project, I aimed to characterise the cellular and molecular consequences of somatic *DNMT3A* and *TET2* mutations in healthy human individuals with clonal haematopoiesis. Specifically, my questions were: Where in the haematopoietic hierarchy do mutant clones expand, how do they alter haematopoietic stem and progenitor cell differentiation, and how is gene expression dysregulated in mutant cells? By identifying the consequences of mutations on gene expression in the stem or progenitor cell type that gains a selective advantage, it was hoped that insights might be gained into the mechanisms of clonal expansion.

6.1 Establishing a human clonal haematopoiesis cohort

Studying clonal haematopoiesis in humans is associated with significant challenges. The first of these is the limited availability of samples from individuals with CH. To study the effects of mutations on haematopoietic differentiation, a source

of haematopoietic stem and progenitor cells was required. Adult CD34+ HSPCs can be obtained from bone marrow or from peripheral blood following treatment with a mobilising agent such as G-CSF (for example in haematopoietic stem cell donors). Of these, bone marrow is preferable as it avoids any confounding effects of G-CSF treatment, which has been shown to induce HSC proliferation, promote granulopoiesis, and induce gene expression changes in HSPCs [295–298]. However, since the majority of individuals with CH have normal blood counts and don't meet clinical indications for bone marrow biopsy, samples are extremely scarce. One approach would be to use samples from patients in remission from acute myeloid leukaemia, who often have residual *DNMT3A* or *TET2*-mutant clones. However, the effects of leukaemic disease and treatment on cell phenotype are likely to be considerable, and the profile of haematopoietic cells in such patients often does not return to normal [19, 299]. Similarly, samples might be obtained from patients with a non-leukaemic condition that receive bone marrow sampling, for example multiple myeloma patients undergoing autologous stem cell transplantation. This approach was used in a recent study that analysed the consequences of *DNMT3A* R882 mutations in 4 such patients and 1 healthy individual with CH [184], but again, the effects of a co-existing haematological malignancy and prior chemotherapy are unknown. I therefore decided to obtain samples from individuals with unperturbed haematopoiesis, and established a study to collect paired bone marrow and blood from patients undergoing total hip replacement (THR) surgery.

The cohort described here included 195 individuals, of which 44% were found to have CH with a pathogenic mutation at VAF ≥ 0.01. Overall, the profile of mutations across the cohort was similar to that observed in other studies of non-cancer populations, with a predominance of small clones and loss-of-function mutations in *DNMT3A* and *TET2*. The age-specific prevalence of CH was significantly higher than in a comparable study by Abelson *et al.* [59] and I identified more cases with multiple mutations. These findings are consistent with another recent study that identified a high prevalence of CH in THR patients when sequencing BM DNA [94]. There are a number of possible explanations for this. Individuals

in the Abelson study were selected from the general population, whereas those undergoing THR may be more likely to have CH due to selective pressures conferred by inflammation and co-existing conditions. Indeed, individuals with osteoarthritis are more likely to have other chronic conditions, including stroke and metabolic syndrome, which have been associated with CH [24, 60, 300]. Another potential reason is increased detection of clones in BM DNA as compared to PB DNA. Although I did not systematically compare BM and PB sequencing across the entire cohort, re-sequencing of selected cases showed that 75% of variants present in BM were also detected in PB with a VAF ≥ 0.01. While this may account for some of the difference, mean clone sizes were not larger in BM and it may also be true that variants in PB were not detected in BM.

Myeloid malignancies are characterised by abnormally expanded progenitor or precursor populations that have impaired differentiation. In MDS, there may be a relative expansion of the immunophenotypic HSC and CMP populations and a depletion of GMP and MEP [48, 301, 302]. In AML, there is a dramatic expansion of leukaemic progenitors, often with LMPP, GMP, MPP or CMP phenotype and an almost complete absence of HSC [49]. Immunophenotyping of 27 samples with *DNMT3A*, *TET2* and *ASXL1* mutations showed that while there is substantial heterogeneity between samples, stem and progenitor frequencies are largely normal. This was expected, considering that most of these individuals had normal blood counts and a small clone size.

6.2 Developing the TARGET-seq+ method

The second major challenge to studying the effect of somatic mutations in human CH is that clone sizes are often small, and there is no way to prospectively separate mutant from WT cells. To overcome this problem, I adopted a single-cell technique that combines targeted genotyping with whole transcriptome sequencing, enabling comparison of gene expression between wild-type and mutant cells from the same individual. Two techniques have recently been developed, and the TARGET-seq method was felt to be most appropriate as it is more sensitive for both genotype

and transcript detection than genotyping of transcriptomes (GoT) [226, 264]. Furthermore, there were reasons to believe that the gene expression differences in *DNMT3A* and *TET2*-mutant HSPCs would be subtle. Humans with these mutations have normal blood counts and the mutations have been identified in mature blood cells, indicating that progenitors differentiate normally. Furthermore, murine studies have shown subtle gene expression differences, even in HSCs that have been serially transplanted following complete gene knockout [166, 172, 183, 219]. To maximise the ability to distinguish phenotypic differences in gene expression, I therefore decided to modify the TARGET-seq method, incorporating optimised reverse transcription conditions from the Smart-seq3 method [265].

The modified TARGET-seq+ method was more sensitive, detecting 19% more genes expressed per cell, with lower dropout rates and reduced technical variation, while maintaining the high-sensitivity mutational detection of the TARGET-seq method. High transcript dropout rates are a considerable problem in single-cell RNA-seq, which poses challenges for data analysis particularly in reducing the power to detect differentially expressed genes [248, 270, 303]. The improved sensitivity and reduced technical variation of the TARGET-seq+ method is therefore an important step towards distinguishing cells with subtly different expression profiles. Incorporation of unique molecular identifiers (UMIs) to reduce noise from PCR amplification bias would further reduce technical variation, but would require a modification to prevent re-priming by residual UMI-containing primers during the PCR reaction [304, 305].

Highly-accurate genotyping is important for similar reasons. The importance of gDNA genotyping was highlighted by the observation that cDNA genotyping alone would have identified only 50% of mutant cells, and in some samples only 15%. This detection rate is likely to be even lower if the droplet-based GoT method were used. In the recent study by Nam *et al.* [184] that analysed four *DNMT3A*^{R882}-mutant samples, the authors detected the R882 locus in 23.5% of cells using GoT, compared to 50–90% detection of *DNMT3A* targets using cDNA alone with TARGET-seq+ (Figure 5.3). Not only does a low detection rate mean more cells need to be profiled

to identify an equal number of mutant cells, it also means WT cells cannot be identified with the same level of confidence because the allelic dropout rate is higher.

The main limitation of the plate-based methods is their throughput. Both TARGET-seq and TARGET-seq+ can be set up with automated liquid handling in most steps, and this greatly increases the number of plates that can be processed in a day [227]. However, they remain more labour-intensive than droplet-based methods and most suited to applications requiring high-depth analysis.

6.3 Transcriptional consequences of *DNMT3A* and *TET2* mutations in haematopoietic stem and progenitor cells

To investigate the transcriptional consequences of *DNMT3A* and *TET2* mutations on haematopoietic stem and progenitor cells, I applied the TARGET-seq+ method to 12,847 single cells from 10 clonal haematopoiesis samples and 3 controls. Integration of transcriptomic data across these samples generated a high-resolution landscape of haematopoietic differentiation. By combining this with single-cell genotyping and surface marker expression, the patterns of clonal expansion could be precisely determined across stem and progenitor populations and along each differentiation trajectory.

A number of broad observations emerge from this analysis. First, neither *DNMT3A* nor *TET2* mutant cells occupy distinct transcriptional states — instead, they intermingle with wild-type stem and progenitor cells. This suggests they follow a normal differentiation trajectory but have subtly altered differentiation dynamics, leading to differences in mutant cell frequency at different stages of differentiation. Second, the differences in gene expression observed between WT and mutant cells are subtle, suggesting that haploinsufficiency of these enzymes does not have a major impact on gene regulation. Finally, although *DNMT3A* and *TET2* mutations both lead to clonal expansion, the phenotype observed is distinct between the two.

DNMT3A-mutant clones were expanded in the immunophenotypic LT-HSC population and reached their maximal size in early progenitors, principally in the LMPP and EMPP. Pseudotime analysis demonstrated a distinct accumulation of

mutant cells in these early multipotent progenitors, suggesting the mutation may lead to a delayed exit from this compartment. There was no clear evidence of lineage bias in *DNMT3A*-mutant clones — clone sizes were broadly similar across lineage-committed progenitors, and no clear transcriptional bias was identified in mutant HSC/MPP. Instead, I observed subtle downregulation of marker genes in lineage-committed progenitors, particularly in lymphoid, myeloid and erythroid progenitors. These expression changes were small and shared across many genes. It is notable that this is exactly where *DNMT3A* expression is highest — from early multipotent progenitors into lineage-committed progenitors — and may indicate that the enzyme plays a role in the transition from multipotency to lineage commitment. However, in all lineages, mutant cells progress through to fully differentiated cells, showing that the normal gene regulatory programmes are still established in the setting of *DNMT3A* haploinsufficiency.

Interestingly, the downregulation of lineage markers was not observed in the MkP. It has been proposed that megakaryocytic differentiation can occur directly from stem cells, and megakaryocyte-biased HSCs are associated with increased quiescence and self-renewal [256, 306]. One might hypothesise that megakaryocytic differentiation programmes are less dependent on *de novo* DNA methylation. In support of this, an analysis of DNA methylation changes in human haematopoiesis showed only mild changes in DNA methylation at regulatory regions in maturing megakaryocytes, with the majority of sites losing methylation [307].

In the study of individuals with *DNMT3A*$^{\mathrm{R882}}$ mutant CH, Nam *et al.* [184] observed expansion of *DNMT3A*$^{\mathrm{R882}}$-mutant myeloid-biased LMPPs and immature progenitors that were biased towards the megakaryocyte-erythroid lineage. The latter upregulated cell cycle genes and were enriched for MYC/MAX targets, suggestive of increased proliferation. These "immature progenitors" may correspond to LMPP/EMPP progenitors identified in the present study. In contrast, I found negative enrichment of cell cycle, DNA replication gene sets and the MYC/MAX targets signature used in that study. Furthermore the differential abundance analysis showed that in the erythroid lineage, mutant cells are specifically enriched in the

non-cycling erythroid progenitor cluster. These differences may be explained by a difference in phenotype between R882 mutations and *DNMT3A* loss-of-function mutations. R882 mutations alter the flanking sequence preference of DNMT3A, disfavouring CpGs with a T at the N+1 position [160, 161]. It is possible that this accounts for the increase in MYC/MAX target gene activity and the increased predisposition to myeloid leukaemia conferred by this mutation as compared to loss-of-function mutations.

While *DNMT3A*-mutant clones reached their maximal size in early progenitors, the increase in size from the HSC/MPP compartment to EMPP/LMPP was only a 1.5–1.75 fold change. To put this in context, the expansion from a single originating stem cell to a clone size of 10% is a 5000–fold increase (assuming humans have 50,000 HSCs) [101]. Thus, it is clear that the majority of clonal expansion occurs in HSCs. In support of this, the frequency of mutant cells was equal between CD90+CD49f+ LT-HSCs and MPPs. As in progenitors, gene expression differences in HSC/MPPs were subtle.

In murine studies *Dnmt3a* knockout is associated with increased expression of stem cell-associated genes both in HSCs and differentiated B cells, suggesting incomplete repression of stem-cell-specific genes in differentiated progeny [172]. However, this was not seen in *DNMT3A*-mutant HSPCs in the current analysis. Instead, there was subtle downregulation of HSC marker genes in MPP and EMPP. HSCs showed increased expression of a quiescent LT-HSC signature and downregulated cell cycle and ribosomal genes. Interestingly, cell cycle genes were also downregulated in *Dnmt3a*$^{-/-}$ murine HSCs [172]. While increased quiescence may seem at odds with clonal expansion, it has previously been shown that mutations or environmental stressors that promote HSC proliferation lead to exhaustion [178–180, 296, 308]. Perhaps increased quiescence protects *DNMT3A*-mutant HSCs from environmental challenges that normally promote proliferation and differentiation, and, over many years, results in increased frequency of the mutant clone in the bone marrow. This is consistent with studies showing resistance of *Dnmt3a*$^{-/-}$ HSCs to IFN-γ signalling [181].

What might cause *DNMT3A*-mutant HSCs to be more quiescent? The current correlative analysis is limited in its ability to identify a mechanism, though the negative enrichment of genes involved in cytoplasmic translation is of interest. Protein synthesis is tightly controlled in HSCs, and becomes strongly activated during the exit from quiescence [239, 254, 309]. Downregulation of ribosomal proteins may simply be a correlate of increased quiescence, but it is known that stem cells are sensitive to the levels of ribosomes. Diamond–Blackfan anaemia, which is caused by heterozygous mutations in one of many possible ribosomal proteins, is associated with impaired differentiation of HSCs to the erythroid lineage [310]. Furthermore, knockdown of individual ribosomal proteins decreases total ribosome levels in HSCs and recapitulates the differentiation defects towards the erythroid lineage [311]. Given the number of ribosomal proteins required to assemble a functional ribosome and the tight control of their expression, this might provide a plausible explanation for how subtle, global differences in DNA methylation could lead to a phenotype.

TET2-mutant clones were expanded in the immunophenotypic LT-HSC population but showed much greater expansion in downstream progenitors than *DNMT3A*-mutant clones. In all samples, *TET2*-mutant clones had a clear myeloid bias, reaching their maximal size in GMP and pDC progenitors, where the frequency of mutant cells was on average 5-fold higher than in HSC/MPP. This was associated with increased expression of neutrophil, monocyte and dendritic cell gene signatures in LMPP and GMP, suggesting these progenitors are more likely to differentiate towards the myeloid than the lymphoid lineage. This is entirely consistent with findings from mouse models, where *Tet2*-deficient mice show expansion of granulocyte-monocyte progenitors [213–215].

However, expansion downstream of HSC was not exclusively observed in the myeloid lineage of *TET2*-mutant CH samples. In three out of four samples, there was additional expansion in erythroid progenitors and in two there was expansion in the lymphoid lineage. Aberrant expansion of erythroid and lymphoid progenitors was also observed in one of the *Tet2*-inactivation mouse models [213]. It is not clear why *TET2*-mutant clones are associated with erythroid and lymphoid expansion

in some individuals and not others, but this roughly reflects the distribution of malignancies seen with *TET2* mutations.

In all lineages, *TET2*-mutant clones expanded somewhat later than *DNMT3A*-mutant clones, corresponding to CD38+ late progenitors. This reflects where *TET2*-expression is highest. Perhaps surprisingly, *TET2*-mutant progenitors were more quiescent, with reduced expression of cell cycle genes, DNA replication and E2F target genes. However, genes related to protein synthesis and ATP production, including key members of the mitochondrial electron transport chain were upregulated in mutant LMPP and to a lesser extent in GMP. These are hallmarks of high metabolic and biosynthetic activity, processes which often are upregulated prior to cell division. Intriguingly, G2M checkpoint genes were downregulated, including a key regulator of G2 progression, *CDC5L* [291]. This may suggest that *TET2*-mutant progenitors have delayed entry into mitosis and spend more time in S phase. Cell cycle phase transitions are tightly linked to terminal differentiation in erythropoiesis, where developing erythroid progenitors spend an increasing fraction of their time in S phase, and their transition to terminal differentiation occurs after a single, short S phase [312]. One might hypothesise that similar cell-cycle-dependent mechanisms regulate myeloid differentiation and that *TET2*-mutant progenitors accumulate in the LMPP/GMP compartment due to delayed mitotic progression.

TET2-mutant HSCs were also characterised by increased expression of lympho-myeloid gene signatures and were transcriptionally closer to differentiated cells than their WT counterparts. This shift towards a progenitor state was validated by flow cytometry index data, which showed that *TET2*-mutant HSCs were less likely to express the LT-HSC marker CD49f and were significantly larger in size. It is also consistent with the downregulation of *GATA2* in *TET2*-mutant early progenitors. *GATA2* is essential for the maintenance of HSCs and loss of *GATA2* promotes myeloid differentiation [313–315].

This raises a key question: If *TET2*-mutant HSPCs are pushed towards differentiation, why does the clone persist and not exhaust? Other oncogenic

mutations that promote proliferation, such as *FLT3* internal tandem duplications (*FLT3* ITDs) and *KRAS* activating mutations lead to HSC depletion, which probably explains why they are always late events and are not observed in the absence of pre-leukaemic mutations [308, 316, 317]. It is possible that the *TET2*-mutant clone is maintained by LT-HSCs and the differentiation-promoting effects of the mutation do not become apparent until cells divide and enter the early progenitor state. Although *TET2*-mutant HSCs on average had reduced expression of the LT-HSC signature, a fraction of mutant cells retained high expression (Figure 5.46), and enrichment of ST-HSC signatures was not observed within the CD49f+ LT-HSC population (Figure 5.47).

The differential expression of several *KLF* transcription factors is of interest. KLF2 and KLF4 both promote embryonic stem cell self-renewal and KLF4 promotes monocyte/macrophage differentiation but neither has been studied in HSCs in detail [288, 318, 319]. *KLF6* is among the top downregulated genes in HSCs during human ageing, and loss of *KLF6* in human CD34+ cells results in increased colony-forming potential and impaired differentiation *in vitro* [320]. These may therefore be candidate mediators of the *TET2*-mutant phenotype in HSCs. The most upregulated gene within the CD49f+ LT-HSC population was *CSKN1E*, a casein kinase that inhibits G-CSF–induced haematopoietic cell differentiation by stabilising SOCS3, a negative regulator of STAT3 signalling [321]. Interestingly, another of the upregulated genes, *CRTC3*, regulates SOCS3 expression in BM stromal cells [322]. Upregulation of these genes may play a role in giving *TET2*-mutant clones a selective advantage in the setting of inflammation.

While the convergent phenotype of *DNMT3A* and *TET2* mutations has often been emphasised [91, 165, 271, 323, 324], it is clear from this analysis that the two mutations have quite distinct effects on haematopoietic stem and progenitor cell differentiation. It has been proposed that both mutations increase HSC self-renewal, and thereby lead to clonal expansion over time. While HSC-based clonal expansion is consistent with the findings in *DNMT3A*-mutant CH samples, the effect of *TET2* mutations appears to be much more pronounced in progenitor cells.

Most studies investigating *Tet2* loss of function have used serial-replating colony assays as a measure of self-renewal [214, 219, 220], and the only study to perform serial transplantation experiments showed that *Tet2*$^{-/-}$ HSCs exhaust at the same rate as wild-type HSCs [216]. Overall, this suggests that the major effect of *TET2* mutations is to increase the self-renewal and proliferative capacity of progenitors, while stem cell function is relatively unperturbed.

The progenitor phenotype was particularly pronounced in the one sample where biallelic *TET2* mutations were present. Here, the double mutant clone was barely detectable in LT-HSCs ($< 1\%$ of cells), but was present in more than half of GMPs. This accumulation of mutant cells in the GMP population is reminiscent of GMP-like leukaemia stem cell expansion in AML, but unlike in AML, it was not associated with loss of normal stem cells and progenitor compartment sizes were relatively preserved. Although limited conclusions can be drawn from a single sample, these observations reinforce the importance of *TET2* function in differentiating progenitors and suggests biallelic *TET2* loss perturbs haematopoietic differentiation in a manner that is one step closer to myeloid malignancy.

6.4 Limitations of this study and future work

Single-cell transcriptome analysis can identify changes in cell state and changes in expression at the level of individual genes. However, the findings described in this study are necessarily correlative and causative links between somatic *DNMT3A* and *TET2* mutations and cell phenotype cannot be proven without further experiments. A number of confounding factors may be at play. First, in order to identify samples with somatic mutations by DNA sequencing, it was necessary to cryopreserve viable cells for later transcriptomic analysis. The process of sample collection, cryopreservation and thawing is know to have substantial impact on the transcriptome profile of primary cells. Mitigating against this, all analyses were performed on samples that had been processed as identically as possible, and the direct comparison of WT and mutant cells within the same individual means the effects of sample processing are controlled for.

While the targeted DNA sequencing used to identify somatic mutations covered all the most commonly mutated genes in CH and myeloid malignancies, the presence of further genetic alterations in these samples cannot be ruled out. InferCNV analysis did not identify any clear chromosomal copy number changes but this method is not able to detect small copy number alterations [273]. Furthermore, there may be point mutations in genes not covered by our panel or in regulatory regions that influence haematopoietic stem and progenitor cell function. Identification of such changes would require deep whole genome sequencing, which is being done for selected samples.

As previously discussed, single-cell RNA-seq is known to have limited sensitivity with high dropout rates and high technical variation, which may limit the differences observed between WT and mutant cells. Profiling a greater number of samples and cells might increase the resolution of the data and detect further clone-specific differences. Added to this, it is known that gene expression profiles do not always distinguish cells with different functional properties. It is thought that cell potential (something that is particularly important when considering stem cell properties) is better reflected by epigenetic profiles, such as patterns of chromatin accessibility and histone modifications. Analysis of chromatic accessibility might also identify regulatory elements including transcription factor binding sites that differ between WT and mutant cells. As proposed by Izzo *et al.* [183], differential methylation or accessibility of transcription factor motifs may give additional insights into how subtle changes in DNA methylation lead to altered haematopoietic differentiation. This would require a method for concurrent chromatin accessibility analysis (such as ATAC-seq) with genotyping analysis.

It is likely that mutations in *DNMT3A* and *TET2* have effects on many genes and pathways, and identifying which of these contribute to clonal expansion and the predisposition to malignancy will be important. In this regard, *in vitro* or *in vivo* perturbation experiments will be needed to identify the effects of dysregulated pathways on clonal growth. For *DNMT3A*, the downregulation of ribosomal proteins and the mechanisms controlling HSC quiescence might be explored. For *TET2*,

there are multiple potential avenues of investigation, including the role of *KLF* transcription factors, *GATA2* downregulation and the role of cell cycle phase transitions in mediating clonal expansion in myeloid progenitors.

Appendices

Chapter A

Appendices

Contents

A.1 Clinical characteristics of the MARCH cohort 196
A.2 Targeted DNA sequencing 200
A.3 Flow cytometry antibodies 204
A.4 Reagents and conditions used for cDNA synthesis in TARGET-seq 205
A.5 Reagents and conditions used for cDNA synthesis in TARGET-seq+ 206
A.6 Oligonucleotides used in TARGET-seq+ 207

A.1 Clinical characteristics of the MARCH cohort

Table A.1: Clinical characteristics of the MARCH cohort.
BM: Bone marrow, PB: Peripheral blood, BMI: Body mass index, TIA: Transient ischaemic attack. COPD: Chronic obstructive pulmonary disease, VTE: Venous thromboembolism, NSAID: Non-steroidal anti-inflammatory drugs, Hb: Haemoglobin, MCV: Mean corpuscular volume, WCC: White cell count.

Sample	Gender	Age	Ethnicity	BMI	Ischaemic heart disease	Cardiac arrhythmia	Valvular heart disease	Stroke/ TIA	Hyper-tension	Diabetes	Hyper-cholesterol-aemia	Asthma	COPD	Previous VTE	Chronic kidney disease	Hypo-thyroidism	Inflammatory bowel disease	Other autoimmune disease	Osteo-arthritis	Previous cancer	Previous chemo-therapy	Previous radio-therapy	Smoking	NSAID	Immuno-suppressants	Hb (g/L)	MCV (fL)	WCC (x10^9/L)	Platelets (x10^9/L)	Neutrophils (x10^9/L)	Lymphocytes (x10^9/L)
NOC001	Male	53.5	White British	35.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	146	87.4	4.84	250	2.39	1.68
NOC002	Female	77.6	White British	27.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	125	87.9	3.75	232	2.31	1.11
NOC003	Female	61	White British	37.6	None	None	None	None	Yes	Type II	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	135	89.2	12.98	353	8.34	3.41
NOC004	Female	53.7	White British	29.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	138	84.6	10.64	339	6.97	2.73
NOC006	Male	72.4	White British	21.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	143	87.7	4.51	234	2.70	1.19
NOC007	Female	79.5	White British	28.4	None	None	None	None	None	None	None	None	None	Yes	Yes	None	None	None	Yes	None	No	No	Never	Yes	None	133	92.0	5.83	215	2.30	2.46
NOC008	Female	81.6	White British	29.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	139	93.0	9.67	282	7.13	1.92
NOC009	Female	74.9	White British	22.5	None	None	None	None	None	None	None	Yes	None	None	None	Yes	None	None	Yes	None	No	No	Never	Yes	None	132	92.9	4.51	267	2.60	1.49
NOC010	Male	85.7	Other White	26.3	None	None	None	Yes	Yes	None	None	Yes	None	None	None	Yes	None	None	Yes	None	No	No	Never	None	None	140	90.6	7.30	286	4.82	1.47
NOC011	Female	73.2	White British	29.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	125	89.6	5.59	291	3.07	1.96
NOC012	Female	82.4	White British	25.4	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	125	91.5	6.08	333	4.05	1.44
NOC013	Female	70.4	White British	25.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	141	90.3	6.48	247	4.80	1.21
NOC014	Female	78.3	White British	24.8	None	None	None	None	Yes	Type II	None	None	None	None	None	None	None	None	Yes	None	No	No	Unknown	None	None	159	88.9	8.02	239	5.86	1.49
NOC015	Male	63.4	White British	22.7	None	None	None	None	Yes	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	148	94.2	5.89	287	3.40	1.39
NOC016	Female	75.8	White British	28.2	None	None	None	None	Yes	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	134	91.0	6.08	261	4.38	1.22
NOC017	Female	78.4	White British	27.5	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	119	87.7	6.08	233	3.62	1.77
NOC018	Male	73.5	White British	33.3	None	None	None	Yes	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	144	92.8	8.51	250	5.84	1.56
NOC019	Female	71	White British	25.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	148	89.5	9.74	226	6.58	2.28
NOC020	Female	79.7	White British	30.8	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	155	85.7	8.09	215	4.29	2.73
NOC021	Male	78.3	White British	22.3	None	None	None	None	None	Type II	Yes	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Never	None	None	150	92.3	9.04	225	6.51	1.45
NOC022	Male	75.6	White British	32.7	Yes	None	None	None	None	Type II	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	138	86.6	6.24	210	4.66	0.95
NOC023	Male	54.3	White British	29.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	151	91.2	5.35	262	2.52	2.03
NOC024	Male	43.5	White British	28.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	135	87.6	6.65	225	4.24	1.86
NOC025	Male	70.3	White British	31.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	145	90.1	8.15	275	5.99	1.45
NOC026	Female	74	White British	28.9	None	None	None	None	None	None	Yes	Yes	Yes	None	None	Yes	None	None	None	Yes	Yes	No	No	Ex-smoker	None	None	124	92.9	5.49	291	3.92	0.87
NOC027	Female	65.7	White British	32.2	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	126	84.4	5.24	343	2.56	1.88
NOC028	Male	64.4	White British	27.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	155	87.7	6.78	245	2.90	2.99
NOC029	Female	49.8	White British	36.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	143	91.3	8.36	215	5.53	2.04
NOC030	Female	80.5	White British	23.8	None	None	None	None	Yes	None	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	123	80.5	6.88	410	5.09	0.97
NOC031	Female	65.1	White British	37.3	None	None	None	None	None	Type II	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	123	90.2	8.63	254	5.33	2.36
NOC032	Female	62.6	White British	29.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	161	92.6	5.40	274	2.89	1.88
NOC033	Female	61.3	White British	38.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	150	87.9	7.55	217	4.39	2.39
NOC034	Female	65	White British	36.3	None	None	None	None	None	None	None	None	None	None	None	Yes	None	None	None	Yes	None	No	No	Never	None	None	132	84.3	5.31	251	3.94	0.91
NOC035	Female	65.2	White British	35.4	None	None	None	None	Yes	None	None	None	None	None	None	Yes	None	None	None	Yes	None	No	No	Smoker	None	None	125	87.7	6.04	297	3.69	1.80
NOC036	Female	51.2	Asian	32.1	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	144	83.2	3.27	243	1.94	1.03
NOC037	Male	77.9	White British	35.6	None	None	None	None	Yes	Type II	None	None	None	None	None	Yes	None	None	None	Yes	None	No	No	Never	None	None	128	91.2	7.57	256	5.04	1.56
NOC038	Female	65.4	White British	26.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	135	95.3	7.53	273	5.30	1.53
NOC039	Female	77.9	White British	35.6	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	None	None	Yes	None	No	No	Never	None	None	100	95.1	6.12	181	3.37	1.75
NOC040	Male	63.2	White British	25.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	134	93.8	12.09	362	9.13	2.00
NOC041	Female	55	White British	27.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	129	85.0	7.04	271	3.68	2.54
NOC042	Female	66.2	White British	26.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	126	93.9	3.80	291	2.13	1.12
NOC043	Male	69.5	White British	30.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	150	94.9	4.69	213	2.99	0.95
NOC044	Male	70.1	White British	30.9	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	156	86.4	5.31	251	2.78	1.93
NOC045	Female	78.5	White British	34.8	None	None	None	None	Yes	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	143	86.2	6.20	308	3.78	1.51
NOC046	Male	54	White British	28.6	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	141	85.9	6.66	244	3.67	1.86
NOC047	Male	59.8	White British	26.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	148	96.3	7.54	281	4.87	1.52
NOC048	Female	70.1	White British	26.9	None	Yes	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	None	None	114	91.1	5.93	328	3.20	1.85
NOC049	Female	50.5	White British	34.8	None	None	None	None	None	None	None	None	None	None	None	Yes	None	None	None	Yes	None	No	No	Never	None	None	142	87.9	6.45	313	4.31	1.57
NOC050	Male	72.7	White British	32.6	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	142	94.1	7.08	330	4.40	1.81
NOC051	Male	71	White British	30.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	126	81.8	7.99	254	4.03	2.95
NOC052	Male	71.6	White British	36.9	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	122	94.4	16.85	220	13.57	1.49
NOC053	Female	66.3	Other White	34.1	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	None	None	114	87.4	9.99	234	7.63	1.59
NOC054	Female	53.5	White British	46.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	143	88.1	8.03	399	5.53	1.91
NOC055	Male	67.8	White British	30.9	None	None	Yes	None	None	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	Yes	145	92.1	7.74	240	5.87	1.05
NOC056	Female	69.3	White British	22.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	Yes	Never	None	None	131	84.6	6.40	300	3.78	2.02

(Continued on next page)

Sample	Gender	Age	Ethnicity	BMI	Ischaemic heart disease	Cardiac arrhythmia	Valvular heart disease	Stroke/ TIA	Hyper-tension	Diabetes	Hyper-cholesterol-aemia	Asthma	COPD	Previous VTE	Chronic kidney disease	Hypo-thyroidism	Inflammatory bowel disease	Other autoimmune disease	Osteo-arthritis	Previous cancer	Previous chemo-therapy	Previous radio-therapy	Smoking	NSAID	Immuno-suppressants	Hb (g/L)	MCV (fL)	WCC (x10^9/L)	Platelets (x10^9/L)	Neutrophils (x10^9/L)	Lymphocytes (x10^9/L)
NOC057	Female	69.4	White British	24.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	123	96.1	3.53	198	1.30	1.70
NOC058	Male	56.8	White British	35.2	None	None	None	None	None	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Never	None	None	117	92.9	12.72	201	9.88	1.36
NOC059	Female	62.6	White British	28.2	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	155	86.9	6.73	245	4.32	1.73
NOC060	Female	83.8	White British	25.2	None	None	None	None	Yes	None	Yes	Yes	None	None	None	Yes	None	None	Yes	None	No	No	Never	None	None	120	91.0	7.66	344	5.88	1.01
NOC061	Male	85	White British	32.5	None	None	None	None	Yes	Type II	None	None	None	None	None	None	None	None	Yes	Yes	Yes	Yes	Ex-smoker	Yes	None	132	92.0	7.10	192	4.76	1.79
NOC062	Male	86.2	White British	24.6	None	None	None	None	Yes	None	None	None	None	Yes	None	Yes	None	None	Yes	None	No	No	Ex-smoker	None	None	99	94.2	6.92	286	4.80	1.26
NOC063	Female	71.9	White British	21.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	135	87.4	7.50	359	5.25	1.72
NOC064	Male	82.9	White British	31.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	153	91.2	6.37	154	3.48	2.28
NOC065	Female	77.1	White British	27.9	None	Yes	Yes	None	Yes	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Never	None	None	144	90.7	6.23	236	4.09	1.41
NOC066	Female	76.9	Other White	36.8	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	Yes	Yes	No	Yes	Ex-smoker	None	None	139	96.3	5.27	231	2.67	1.61
NOC067	Female	68.5	White British	23.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	Yes	None	141	90.5	5.39	250	2.98	1.68
NOC068	Male	59	White British	24.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	150	85.3	4.92	220	2.88	1.48
NOC069	Female	53.3	White British	25.4	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	137	94.0	5.20	257	2.35	2.14
NOC070	Male	65.9	White British	25.6	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	127	86.3	8.01	239	6.01	1.28
NOC071	Male	80.5	White British	29.0	Yes	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	Yes	No	Yes	Smoker	None	None	152	97.6	7.07	202	4.40	1.74
NOC072	Female	75.9	White British	30.4	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	None	Yes	None	No	No	Never	None	None	131	94.8	6.83	200	3.27	2.29
NOC073	Male	54.4	White British	21.8	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	118	97.6	7.32	254	4.60	2.02
NOC074	Female	54	White British	20.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	138	92.9	5.48	345	3.51	1.59
NOC075	Female	79.5	White British	24.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	119	97.9	4.69	285	3.00	1.18
NOC076	Male	68.4	White British	33.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	158	86.6	5.67	263	3.38	1.45
NOC077	Male	30.9	White British	25.2	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	159	84.3	5.15	223	3.01	1.63
NOC078	Female	83.9	White British	26.8	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	140	89.7	5.94	207	4.09	1.35
NOC079	Male	72.8	White British	29.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	143	91.3	4.78	161	2.48	1.78
NOC080	Male	71.5	White British	24.6	Yes	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	None	None	138	96.5	4.22	209	2.25	1.37
NOC081	Male	56.4	White British	21.5	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	144	93.3	5.45	233	3.11	1.58
NOC082	Female	65.5	White British	28.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	133	87.6	7.04	262	4.49	1.83
NOC083	Female	57.8	White British	32.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	Yes	None	121	91.8	7.50	359	4.91	1.78
NOC084	Male	64.5	White British	30.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	148	90.8	5.10	242	2.89	1.62
NOC085	Female	63.6	White British	37.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	130	91.4	7.63	314	4.83	2.26
NOC086	Male	51.1	White British	23.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	157	86.3	5.22	230	2.48	1.63
NOC087	Male	66.4	White British	29.4	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	150	93.1	5.07	228	2.83	1.38
NOC088	Female	70.1	White British	29.2	None	Yes	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	149	93.8	5.91	201	3.68	1.60
NOC089	Female	57.1	White British	32.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	143	83.1	7.70	195	4.39	2.56
NOC090	Female	86.2	White British	25.2	None	None	None	None	None	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	136	85.9	10.55	287	7.48	2.15
NOC091	Male	76.6	White British	28.4	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	146	90.7	7.36	184	4.46	2.07
NOC092	Female	79.4	White British	30.5	None	None	None	Yes	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	136	82.3	7.12	349	5.07	1.34
NOC093	Male	66.3	White British	28.6	None	Yes	Yes	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	153	91.9	5.74	207	3.48	0.91
NOC094	Female	75.5	White British	22.2	None	None	None	None	None	None	None	None	None	None	None	Yes	None	None	Yes	None	No	No	Ex-smoker	None	None	127	93.3	9.49	197	5.65	2.93
NOC095	Male	75.3	White British	38.4	None	None	None	None	Yes	Type II	None	None	None	None	None	None	Yes	None	Yes	None	No	No	Ex-smoker	None	Yes	156	98.1	6.92	206	4.42	1.73
NOC096	Female	78.4	White British	21.4	None	None	None	Yes	Yes	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	None	None	123	88.3	12.30	165	9.44	1.47
NOC097	Female	80.4	White British	29.6	None	None	None	None	Yes	None	None	None	Yes	None	None	None	None	None	Yes	None	No	No	Never	None	None	131	89.9	7.75	233	5.43	1.18
NOC098	Male	72.6	White British	31.0	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	150	86.2	8.04	299	4.52	2.73
NOC099	Male	39.6	White British	32.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	158	92.4	7.22	303	4.18	1.81
NOC100	Male	62.5	White British	36.8	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	151	91.9	5.68	211	3.53	1.14
NOC101	Female	72.0	White British	24.9	None	None	None	None	Yes	Type II	None	None	None	None	None	None	None	Yes	None	Yes	No	No	Never	None	Yes	135	84.4	6.33	256	4.17	1.31
NOC102	Female	68.5	White British	26.2	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	Yes	Ex-smoker	None	None	140	88.3	5.38	263	3.33	1.62
NOC103	Male	76.6	White British	26.9	None	Yes	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	152	93.7	7.99	250	4.92	2.18
NOC104	Female	74.1	White British	24.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	130	88.1	5.00	241	3.02	1.39
NOC105	Male	64.8	White British	33.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	159	84.3	7.49	233	4.61	2.12
NOC106	Female	73.9	White British	36.4	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	Yes	Yes	No	Ex-smoker	None	None	158	91.1	8.01	261	4.37	2.97
NOC107	Female	72.6	White British	20.8	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	128	87.4	5.51	278	3.95	0.99
NOC108	Female	80.5	White British	23.1	Yes	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	124	94.1	7.72	282	4.39	2.58
NOC109	Female	59.9	White British	23.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	126	94.9	4.14	226	2.42	1.07
NOC110	Female	60.7	White British	30.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	136	90.7	5.11	268	3.40	1.36
NOC111	Female	60.3	white British	44.5	None	None	None	None	Yes	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	133	88.5	8.67	378	5.24	2.60
NOC112	Male	69.2	White British	42.4	None	Yes	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	121	93.2	8.50	212	5.73	1.89
NOC113	Male	70.4	White British	33.8	None	None	None	None	None	None	None	None	Yes	None	None	Yes	None	None	None	Yes	None	No	No	Smoker	Yes	None	151	90.7	7.70	298	5.22	1.50
NOC114	Male	80.4	White British	26.3	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	Yes	Yes	No	Ex-smoker	None	None	140	102.7	8.14	187	4.53	2.97
NOC115	Male	77.5	White British	26.2	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	154	86.5	7.46	218	5.15	1.71
NOC116	Female	72.8	White British	34.6	Yes	None	None	None	None	None	None	Yes	None	None	None	Yes	None	None	None	Yes	Yes	No	Yes	Ex-smoker	None	None	130	98.5	7.09	245	4.35	1.89
NOC117	Female	89.9	White British	17.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	139	103.2	3.18	194	2.27	0.65
NOC118	Male	84.6	White British	33.2	Yes	None	None	None	None	Type II	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	128	99.7	6.64	188	4.78	1.26
NOC119	Male	55.2	White British	22.2	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	None	None	147	93.8	4.23	242	2.35	1.40
NOC120	Female	61.1	White British	29.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	139	84.1	5.52	178	2.88	1.84
NOC121	Female	76.2	White British	15.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	132	89.8	5.87	305	3.77	1.44
NOC122	Female	60.4	White British	41.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	139	91.1	5.37	276	2.93	1.82
NOC123	Male	76.6	White British	25.7	Yes	Yes	None	None	None	Pre-Diabetic	None	None	None	None	Yes	None	None	None	None	Yes	Yes	No	No	Ex-smoker	None	None	133	94.7	5.83	191	2.91	2.24
NOC124	Male	76.3	White British	28.4	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	154	95.8	6.33	180	3.98	1.62
NOC125	Female	80.4	White British	32.9	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	141	90.7	9.25	297	6.36	2.15
NOC126	Male	65.4	White British	21.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	None	None	149	95.5	6.85	287	4.93	1.35

(Continued on next page)

Sample	Gender	Age	Ethnicity	BMI	Ischaemic heart disease	Cardiac arrhythmia	Valvular heart disease	Stroke/ TIA	Hyper-tension	Diabetes	Hyper-cholesterol-aemia	Asthma	COPD	Previous VTE	Chronic kidney disease	Hypo-thyroidism	Inflammatory bowel disease	Other autoimmune disease	Osteo-arthritis	Previous cancer	Previous chemo-therapy	Previous radio-therapy	Smoking	NSAID	Immuno-suppressants	Hb (g/L)	MCV (fL)	WCC (x10^9/L)	Platelets (x10^9/L)	Neutrophils (x10^9/L)	Lymphocytes (x10^9/L)
NOC127	Male	74.4	White British	34.0	Yes	None	None	Yes	Yes	Type II	None	None	None	None	Yes	None	None	None	Yes	Yes	No	No	Ex-smoker	None	None	123	91.2	9.66	344	4.90	3.74
NOC128	Female	79.9	White British	36.7	None	None	None	Yes	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	113	87.7	4.98	232	3.41	1.00
NOC129	Female	57.8	White British	35.4	None	None	None	None	Yes	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Ex-smoker	None	Yes	140	94.1	5.89	388	3.39	1.94
NOC130	Female	92	White British	15.6	None	None	Yes	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	148	92.7	6.76	187	5.33	0.88
NOC131	Female	55.1	White British	25.1	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	Yes	None	128	89.1	5.59	286	2.92	2.04
NOC132	Male	72.4	White British	34.3	None	None	None	None	Yes	Type II	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	151	89.3	6.33	238	4.16	1.47
NOC133	Male	51.3	Other White	24.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	173	87.1	6.39	224	3.89	1.92
NOC134	Male	73.8	White British	28.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	143	89.2	5.76	318	3.11	2.02
NOC135	Male	75.4	White British	25.8	None	Yes	Yes	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	140	91.8	7.78	241	4.76	2.07
NOC136	Male	59.9	Other White	25.8	Yes	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	164	90.2	8.51	196	6.58	1.17
NOC137	Male	73.6	White British	37.7	Yes	None	None	None	None	Type II	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	131	91.4	6.38	203	4.51	1.13
NOC138	Female	65.9	White British	24.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	150	89.8	8.42	159	5.29	2.52
NOC139	Female	62.8	White British	27.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	137	98.6	8.54	331	4.74	2.98
NOC140	Female	83.9	White British	23.0	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	115	90.5	6.66	261	4.40	1.53
NOC141	Male	62.2	White British	42.1	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	144	90.2	4.58	255	2.59	1.37
NOC142	Female	87.5	White British	21.1	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	110	98.0	8.70	205	7.19	0.74
NOC143	Male	62.7	White British	21.1	None	None	None	None	None	Type I	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Ex-smoker	None	None	127	93.2	8.87	189	5.90	1.92
NOC144	Female	69.3	White British	33.0	None	Yes	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	None	None	135	96.0	5.46	222	3.78	1.00
NOC145	Female	67.3	White British	21.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	145	93.4	6.51	230	4.23	1.57
NOC146	Male	56.8	White British	35.3	None	None	None	None	None	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Never	None	None	155	86.8	4.95	271	2.85	1.58
NOC147	Male	57.9	White British	36.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	152	91.2	6.86	254	4.32	1.82
NOC148	Male	72.5	White British	25.5	None	None	None	None	Yes	None	Yes	None	Yes	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	154	85.8	9.69	200	6.16	2.18
NOC149	Female	67.9	White British	36.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Smoker	None	None	134	91.2	7.93	292	4.18	2.70
NOC150	Male	86.6	White British	27.4	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	Yes	?	Yes	Never	None	None	128	92.7	6.81	262	4.38	1.61
NOC151	Female	74.9	White British	24.8	None	None	None	None	Yes	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	117	102.0	7.21	280	4.45	2.11
NOC152	Female	81.2	White British	32.7	None	None	None	None	Yes	None	None	Yes	None	None	Yes	None	None	None	Yes	None	No	No	Ex-smoker	None	None	125	87.5	8.32	237	5.25	1.63
NOC153	Female	63.1	White British	36.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	140	86.1	4.54	173	2.74	0.99
NOC154	Female	47.2	White British	25.0	None	None	None	None	None	None	None	Yes	None	None	None	Yes	None	None	Yes	None	No	No	Never	None	None	151	96.2	6.49	360	3.22	2.62
NOC155	Male	84.9	White British	30.1	None	None	None	Yes	Yes	Type II	None	None	Yes	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	148	78.1	9.34	357	6.41	2.10
NOC156	Female	63.1	White British	34.3	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	111	94.3	4.73	279	3.04	1.06
NOC157	Female	48.1	White British	21.0	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	110	74.6	5.65	175	4.40	0.78
NOC158	Male	72.7	White British	35.2	None	None	None	None	None	None	Yes	None	None	None	None	Yes	None	None	Yes	Yes	No	Yes	Ex-smoker	None	None	171	93.7	7.95	275	5.72	1.18
NOC159	Male	68	White British	28.7	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	155	88.6	11.56	190	8.18	2.26
NOC160	Male	57.1	White British	26.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	144	85.2	4.22	237	2.58	1.21
NOC161	Male	81.4	White British	27.3	None	None	None	Yes	Yes	None	None	None	Yes	None	None	None	None	None	Yes	None	No	No	Never	None	None	148	90.0	7.89	215	5.85	1.36
NOC162	Female	63.5	White British	33.3	None	None	None	None	None	Pre-Diabetic	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	None	None	139	86.3	9.58	387	6.91	1.59
NOC163	Male	89.5	White British	24.6	None	None	None	None	Yes	Pre-Diabetic	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	132	92.9	7.85	273	5.03	1.83
NOC164	Male	54.4	White British	28.0	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	156	94.1	6.63	331	4.14	1.67
NOC165	Female	52.2	White British	26.5	None	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	138	94.4	6.13	259	3.50	1.32
NOC166	Female	72.7	White British	24.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	127	89.8	4.31	261	2.09	1.56
NOC167	Male	55.6	White British	31.8	None	None	Yes	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	160	88.9	6.93	193	4.08	2.00
NOC168	Female	69.5	White British	22.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	Yes	None	121	91.1	4.59	184	3.29	0.92
NOC169	Female	44.4	White British	33.0	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	144	95.0	4.20	214	2.10	1.63
NOC170	Male	80.6	White British	65.0	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	Yes	No	No	Ex-smoker	None	None	148	93.9	7.69	183	4.49	2.14
NOC171	Female	62.3	White British	24.2	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	136	92.4	8.60	346	4.80	3.05
NOC172	Male	65.6	White British	28.9	None	None	None	None	Yes	None	None	None	Yes	None	None	None	None	None	None	Yes	None	No	No	Smoker	None	None	154	97.6	9.75	290	5.18	3.38
NOC173	Male	63.7	White British	25.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	137	90.0	4.87	240	3.17	1.25
NOC174	Male	72.8	White British	31.4	None	None	None	None	Yes	None	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Smoker	None	None	142	92.6	10.27	195	7.11	2.22
NOC175	Male	74.9	White British	28.4	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	136	87.9	6.79	231	4.51	1.66
NOC176	Male	62.8	White British	43.0	None	None	None	Yes	Yes	Type II	None	None	None	None	Yes	None	None	None	None	Yes	None	No	No	Never	None	Yes	123	99.7	5.43	282	4.31	0.51
NOC177	Male	77.1	White British	33.8	None	Yes	None	None	None	None	None	None	None	None	None	Yes	None	None	Yes	Yes	None	No	No	Ex-smoker	None	None	135	92.2	6.18	191	3.77	1.53
NOC179	Female	71	White British	35.2	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	None	Yes	Yes	Yes	No	Ex-smoker	None	None	139	86.4	4.17	256	2.36	1.18
NOC180	Female	76.5	White British	25.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	134	89.0	5.12	214	2.92	1.78
NOC181	Male	70.8	White British	28.8	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Smoker	None	None	150	87.2	7.39	353	4.54	2.07
NOC182	Female	75.4	White British	28.2	None	None	None	Yes	Yes	None	Yes	None	None	None	None	Yes	None	None	None	Yes	None	No	No	Ex-smoker	None	None	148	88.3	5.96	378	3.06	2.06
NOC183	Male	60.1	White British	27.6	None	None	None	None	None	None	None	None	None	Yes	None	None	None	None	None	Yes	None	No	No	Never	None	None	148	95.6	7.64	252	3.63	2.96
NOC184	Female	71.1	White British	36.8	None	None	None	Yes	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	141	92.8	5.95	339	3.95	1.20
NOC185	Female	66	White British	25.9	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	139	97.1	5.69	337	2.58	2.21
NOC186	Female	80.9	White British	30.9	None	None	None	None	Yes	Type II	None	None	None	None	None	None	None	None	None	Yes	Yes	No	No	Never	None	None	115	92.8	10.34	444	8.26	1.34
NOC187	Female	47.5	White British	30.1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	141	94.8	4.49	267	1.82	2.25
NOC188	Female	79.3	White British	18.2	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	120	97.8	5.02	328	3.52	1.13
NOC189	Male	77.1	White British	36.5	None	None	None	None	Yes	None	Yes	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	None	None	149	91.9	8.26	243	5.23	2.14
NOC190	Female	51.3	White British	24.7	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	110	99.7	7.66	427	3.97	1.86
NOC191	Female	52.1	White British	26.6	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	Yes	None	No	No	Ex-smoker	Yes	None	139	98.8	4.97	230	2.33	1.96
NOC192	Male	72.9	White British	28.6	None	None	None	None	None	None	Yes	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	147	91.8	4.75	264	2.78	1.41
NOC194	Female	91	White British	31.9	None	None	None	None	Yes	Type II	Yes	None	None	None	Yes	Yes	None	None	Yes	Yes	No	Yes	Ex-smoker	None	None	117	85.6	9.60	297	6.58	1.69
NOC195	Female	71	White British	33.5	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	Yes	None	136	94.7	5.92	232	3.06	1.96
NOC196	Male	60.5	White British	32.2	None	Yes	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	Yes	None	148	93.0	4.62	245	2.59	1.40
NOC197	Female	79.5	White British	29.7	None	None	None	None	Yes	None	None	None	None	None	None	None	None	None	None	Yes	None	No	No	Never	None	None	138	94.2	6.12	251	3.92	1.40
NOC198	Female	69.2	White British	25.4	Yes	None	None	None	Yes	None	None	Yes	None	None	None	None	None	None	None	Yes	None	No	No	Ex-smoker	None	None	138	87.1	7.20	273	3.90	1.81

A.2 Targeted DNA sequencing

Table A.2: Coverage of targeted sequencing panel.
The RefSeq transcript IDs used for annotation are shown. Where hotspots only were covered, the hg38 coordinates are shown.

Gene	Ensembl Gene ID	RefSeq Transcript ID	Exonic Coverage
ASXL1	ENSG00000171456	NM_015338	Full
ASXL2	ENSG00000143970	NM_018263	Full
ATM	ENSG00000149311	NM_000051	Full
ATRX	ENSG00000085224	NM_000489	Full
BCOR	ENSG00000183337	NM_001123385	Full
BCORL1	ENSG00000085185	NM_001184772	Full
BRAF	ENSG00000157764	NM_004333	Full
BRCC3	ENSG00000185515	NM_024332	chrX:155078578-155078743;155089249-155089378;155119916-155120208
BRINP3	ENSG00000162670	NM_199051	Full
CALR	ENSG00000179218	NM_004343	chr19:12943524-12943653;12943697-12943923
CBL	ENSG00000110395	NM_005188	Full
CBLB	ENSG00000114423	NM_170662	Full
CDKN2A	ENSG00000147889	NM_000077	Full
CEBPA	ENSG00000245848	NM_004364	Full
CREBBP	ENSG00000005339	NM_004380	Full
CSF1R	ENSG00000182578	NM_005211	Full
CSF3R	ENSG00000119535	NM_000760	Full
CTCF	ENSG00000102974	NM_006565	Full
CUX1	ENSG00000257923	NM_181552;NM_001913	Full
DDX3X	ENSG00000215301	NM_001356	chrX:41337389-41337486;41345342-41345587;41346407-41346796
DHX15	ENSG00000109606	NM_001358	Full
DNMT3A	ENSG00000119772	NM_022552	Full
EP300	ENSG00000100393	NM_001429	Full
ETV6	ENSG00000139083	NM_001987	Full
EZH2	ENSG00000106462	NM_004456	Full
FBXW7	ENSG00000109670	NM_001349798	Full
FLT3	ENSG00000122025	NM_004119	Full
GATA1	ENSG00000102145	NM_002049	Full
GATA2	ENSG00000179348	NM_032638	Full
GNAS	ENSG00000087460	NM_080425	Full
GNB1	ENSG00000078369	NM_002074	chr1:1815749-1815870
HNRNPK	ENSG00000165119	NM_031263	Full
HRAS	ENSG00000174775	NM_005343	chr11:534205-534327
IDH1	ENSG00000138413	NM_005896	Full
IDH2	ENSG00000182054	NM_002168	Full
IKZF1	ENSG00000185811	NM_006060	Full
JAK2	ENSG00000096968	NM_004972	Full
JAK3	ENSG00000105639	NM_000215	chr19:17835049-17835248;17837112-17837227;17838231-17838422
KANSL1	ENSG00000120071	NM_015443	Full
KDM5A	ENSG00000073614	NM_001042603	Full
KDM6A	ENSG00000147050	NM_001291415	Full
KIT	ENSG00000157404	NM_000222	Full
KMT2A	ENSG00000118058	NM_001197104	Full
KMT2C	ENSG00000055609	NM_170606	Full
KMT2D	ENSG00000167548	NM_003482	Full
KMT2E	ENSG00000005483	NM_182931	Full
KRAS	ENSG00000133703	NM_033360	Full
LUC7L2	ENSG00000146963	NM_016019	Full
MAU2	ENSG00000129933	NM_015329	Full
MED12	ENSG00000184634	NM_005120	Full

(Continued on next page)

Gene	Ensembl Gene ID	RefSeq Transcript ID	Exonic Coverage
MPL	ENSG00000117400	NM_005373	Full
MTA2	ENSG00000149480	NM_004739	Full
MYC	ENSG00000136997	NM_002467	Full
MYD88	ENSG00000172936	NM_002468	chr3:38141120-38141293
NF1	ENSG00000196712	NM_001042492	Full
NIPBL	ENSG00000164190	NM_133433	Full
NOTCH1	ENSG00000148400	NM_017617	Full
NPM1	ENSG00000181163	NM_002520	Full
NRAS	ENSG00000213281	NM_002524	Full
PDGFRA	ENSG00000134853	NM_006206	Full
PDS5B	ENSG00000083642	NM_015032	Full
PHF6	ENSG00000156531	NM_001015877	Full
PIGA	ENSG00000165195	NM_002641	Full
PPM1D	ENSG00000170836	NM_003620	Full
PRDM1	ENSG00000057657	NM_001198	chr6:106104757-106105953
PRPF40B	ENSG00000110844	NM_001031698	Full
PRPF8	ENSG00000174231	NM_006445	Full
PTEN	ENSG00000171862	NM_000314	Full
PTPN11	ENSG00000179295	NM_002834	Full
PTPRF	ENSG00000142949	NM_002840	Full
RAD21	ENSG00000164754	NM_006265	Full
RB1	ENSG00000139687	NM_000321	Full
RIT1	ENSG00000143622	NM_006912	Full
RUNX1	ENSG00000159216	NM_001754	Full
SETBP1	ENSG00000152217	NM_015559	Full
SETD2	ENSG00000181555	NM_014159	Full
SETDB1	ENSG00000143379	NM_001145415	chr1:150949117-150949285;150960514-150961253
SF1	ENSG00000168066	NM_004630	Full
SF3A1	ENSG00000099995	NM_005877	Full
SF3B1	ENSG00000115524	NM_012433	Full
SH2B3	ENSG00000111252	NM_005475	Full
SMC1A	ENSG00000072501	NM_006306	Full
SMC3	ENSG00000108055	NM_005445	Full
SMG1	ENSG00000157106	NM_015092	Full
SRSF2	ENSG00000161547	NM_003016	Full
STAG1	ENSG00000118007	NM_005862	Full
STAG2	ENSG00000101972	NM_006603	Full
STAT3	ENSG00000168610	NM_139276	chr17:42322259-42322511;42322994-42323150
STAT5B	ENSG00000173757	NM_012448	Full
SUZ12	ENSG00000178691	NM_015355	Full
TET2	ENSG00000168769	NM_001127208	Full
TP53	ENSG00000141510	NM_000546	Full
U2AF1	ENSG00000160201	NM_006758	Full
U2AF2	ENSG00000063244	NM_007279	Full
WAPL	ENSG00000062650	NM_015045	Full
WT1	ENSG00000184937	NM_024426	Full
ZRSR2	ENSG00000169249	NM_005089	Full

Table A.3: Variants detected by targeted BM DNA sequencing. List of variants detected by targeted panel sequencing of BM MNCs in the MARCH cohort (hg38 coordinates).
SNV: single nucleotide variant, INDEL: insertion-deletion, MNV: multiple nucleotide variant, VUS: variant of unknown significance.

Sample	Gene	Chrom-osome	Start Position	Type	Reference Allele	Variant Allele	Total Depth	Variant Reads	VAF	Variant Classification	Accession	Exon	cDNA change	Protein change	Significance
NOC002	TET2	chr4	105236118	SNV	C	T	1399	226	0.1615	stopgain	NM_001127208	exon3	c.2176C>T	p.Q726X	Pathogenic
NOC002	TET2	chr4	105243756	SNV	C	T	883	129	0.1461	nonsynonymous_SNV	NM_001127208	exon6	c.3781C>T	p.R1261C	Pathogenic
NOC002	KMT2D	chr12	49032794	SNV	G	C	763	23	0.0301	nonsynonymous_SNV	NM_003482	exon39	c.11911C>G	p.Q3971E	VUS
NOC002	ATM	chr11	108317392	SNV	T	C	591	8	0.0135	nonsynonymous_SNV	NM_000051	exon43	c.6218T>C	p.L2073P	Pathogenic
NOC007	ETV6	chr12	11885969	SNV	G	A	1198	22	0.0184	nonsynonymous_SNV	NM_001987	exon7	c.1196G>A	p.R399H	Pathogenic
NOC007	DNMT3A	chr2	25240699	SNV	A	G	600	9	0.015	nonsynonymous_SNV	NM_022552	exon18	c.2114T>C	p.I705T	Pathogenic
NOC008	PTPN11	chr12	112450496	SNV	G	A	267	19	0.0712	nonsynonymous_SNV	NM_002834	exon3	c.316G>A	p.D106N	Pathogenic
NOC008	GNB1	chr1	1815790	SNV	T	C	746	10	0.0134	nonsynonymous_SNV	NM_002074	exon5	c.169A>G	p.K57E	Pathogenic
NOC008	CBL	chr11	119278532	SNV	C	G	469	5	0.0107	nonsynonymous_SNV	NM_005188	exon9	c.1250C>G	p.P417R	Pathogenic
NOC008	TP53	chr17	7674872	SNV	T	C	944	15	0.0159	nonsynonymous_SNV	NM_000546	exon6	c.659A>G	p.Y220C	Pathogenic
NOC008	ASXL1	chr20	32434638	INDEL	A	AG	1105	184	0.1665	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs	Pathogenic
NOC009	DNMT3A	chr2	25241581	SNV	C	T	478	9	0.0188	nonsynonymous_SNV	NM_022552	exon17	c.2063G>A	p.R688H	Pathogenic
NOC010	DNMT3A	chr2	25248036	SNV	C	G	510	6	0.0118	splicing	NM_022552	exon7	c.855+1G>C	p.E285X	Pathogenic
NOC010	ASXL1	chr20	32434638	INDEL	A	AG	827	30	0.0363	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs	Pathogenic
NOC011	SH2B3	chr12	111447675	INDEL	CAG	C	994	15	0.0151	frameshift_deletion	NM_005475	exon7	c.1257_1258del	p.T419fs	Pathogenic
NOC012	TET2	chr4	105275146	SNV	C	T	1054	37	0.0351	stopgain	NM_001127208	exon11	c.4636C>T	p.Q1546X	Pathogenic
NOC012	TET2	chr4	105276050	SNV	G	A	1107	13	0.0117	stopgain	NM_001127208	exon11	c.5540G>A	p.W1847X	Pathogenic
NOC013	GNB1	chr1	1815790	SNV	T	C	656	119	0.1814	nonsynonymous_SNV	NM_002074	exon5	c.169A>G	p.K57E	Pathogenic
NOC013	PTPRF	chr1	43603697	SNV	G	T	1010	12	0.0119	stopgain	NM_002840	exon16	c.2545G>T	p.E849X	Pathogenic
NOC015	TET2	chr4	105272591	SNV	C	T	585	11	0.0188	stopgain	NM_001127208	exon10	c.4210C>T	p.R1404X	Pathogenic
NOC016	ASXL1	chr20	32434638	INDEL	A	AG	1223	19	0.0155	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs	Pathogenic
NOC017	ASXL1	chr20	32433373	SNV	C	A	802	13	0.0162	stopgain	NM_015338	exon11	c.1175C>A	p.S392X	Pathogenic
NOC018	KRAS	chr12	25245350	SNV	C	T	1948	20	0.0103	nonsynonymous_SNV	NM_033360	exon2	c.35G>A	p.G12D	Pathogenic
NOC020	DNMT3A	chr2	25240420	SNV	T	C	1290	69	0.0535	nonsynonymous_SNV	NM_022552	exon19	c.2204A>G	p.Y735C	Pathogenic
NOC021	DNMT3A	chr2	25241564	INDEL	G	GCTT	731	34	0.0465	nonframeshift_insertion	NM_022552	exon17	c.2079_2080insAAG	p.H694delinsKH	Pathogenic
NOC022	KRAS	chr12	25245351	SNV	C	T	1883	68	0.0361	nonsynonymous_SNV	NM_033360	exon2	c.34G>A	p.G12S	Pathogenic
NOC022	STAT5B	chr17	42207641	SNV	T	A	810	19	0.0235	nonsynonymous_SNV	NM_012448	exon16	c.1994A>T	p.Y665F	Pathogenic
NOC025	KMT2D	chr12	49030647	SNV	C	T	504	7	0.0139	nonsynonymous_SNV	NM_003482	exon41	c.13793G>A	p.G4598E	VUS
NOC026	TET2	chr4	105243729	INDEL	CT	C	886	14	0.0158	frameshift_deletion	NM_001127208	exon6	c.3755delT	p.L1252fs	Pathogenic
NOC027	KMT2C	chr7	152138805	SNV	T	C	1037	70	0.0675	synonymous_SNV	NM_170606	exon58	c.14634A>G	p.K4878K	Silent
NOC028	PTPRF	chr1	43603741	SNV	G	A	1621	17	0.0105	synonymous_SNV	NM_002840	exon16	c.2589G>A	p.A863A	Silent
NOC030	NF1	chr17	31232164	SNV	G	T	534	11	0.0206	stopgain	NM_001042492	exon25	c.3289G>T	p.E1097X	Pathogenic
NOC030	ASXL1	chr20	32433315	SNV	C	T	861	9	0.0105	stopgain	NM_015338	exon11	c.1117C>T	p.Q373X	Pathogenic
NOC031	TET2	chr4	105236227	INDEL	A	AC	898	72	0.0802	frameshift_insertion	NM_001127208	exon3	c.2286dupC	p.H762fs	Pathogenic
NOC034	DNMT3A	chr2	25246269	SNV	C	T	1115	22	0.0197	stopgain	NM_022552	exon11	c.1320G>A	p.W440X	Pathogenic
NOC034	DNMT3A	chr2	25247676	SNV	A	G	1446	70	0.0484	nonsynonymous_SNV	NM_022552	exon8	c.929T>C	p.I310T	Pathogenic
NOC034	SMG1	chr16	18817405	SNV	T	A	975	19	0.0195	nonsynonymous_SNV	NM_015092	exon57	c.9960A>T	p.L3320F	VUS
NOC035	KMT2E	chr7	105107213	SNV	T	C	466	7	0.015	synonymous_SNV	NM_182931	exon21	c.2895T>C	p.Y965Y	Silent
NOC037	DNMT3A	chr2	25247071	SNV	C	G	1110	88	0.0793	nonsynonymous_SNV	NM_022552	exon9	c.1102G>C	p.A368P	Pathogenic
NOC039	CBL	chr11	119278170	SNV	A	G	872	17	0.0195	nonsynonymous_SNV	NM_005188	exon8	c.1100A>G	p.Q367R	Pathogenic
NOC040	IDH2	chr15	90085308	SNV	C	G	1036	21	0.0203	synonymous_SNV	NM_002168	exon8	c.1047G>C	p.G349G	Silent
NOC042	BRCC3	chrX	155120026	SNV	T	A	1380	77	0.0558	nonsynonymous_SNV	NM_024332	exon11	c.827T>A	p.M276K	VUS
NOC045	PRPF8	chr17	1680944	SNV	T	C	1847	28	0.0152	nonsynonymous_SNV	NM_006445	exon7	c.977A>G	p.H326R	VUS
NOC048	DNMT3A	chr2	25244567	SNV	A	T	768	50	0.0651	nonsynonymous_SNV	NM_022552	exon14	c.1640T>A	p.L547H	Pathogenic
NOC048	IDH2	chr15	90087473	SNV	G	A	411	5	0.0122	nonsynonymous_SNV	NM_002168	exon6	c.781C>T	p.R261C	Pathogenic
NOC050	ASXL1	chr20	32434638	INDEL	A	AG	2097	48	0.0229	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs	Pathogenic
NOC051	TET2	chr4	105276203	SNV	C	A	1740	25	0.0144	nonsynonymous_SNV	NM_001127208	exon11	c.5693C>A	p.S1898Y	Pathogenic
NOC051	DNMT3A	chr2	25247107	SNV	G	A	1422	18	0.0127	stopgain	NM_022552	exon9	c.1066C>T	p.Q356X	Pathogenic
NOC051	SMG1	chr16	18896856	SNV	C	T	1007	19	0.0189	nonsynonymous_SNV	NM_015092	exon2	c.193G>A	p.V65M	VUS
NOC052	SETD2	chr3	47122241	SNV	C	T	1938	111	0.0573	nonsynonymous_SNV	NM_014159	exon3	c.2395G>A	p.D799N	VUS
NOC056	DNMT3A	chr2	25240705	SNV	A	G	923	117	0.1268	nonsynonymous_SNV	NM_022552	exon18	c.2108T>C	p.L703P	Pathogenic
NOC059	CEBPA	chr19	33301847	MNV	AGG	GGT	1036	18	0.0174	nonframeshift_substitution	NM_004364	exon1	c.566_568ACC	NA	VUS
NOC060	ATRX	chrX	77663455	SNV	G	A	1126	38	0.0337	synonymous_SNV	NM_000489	exon12	c.4047C>T	p.D1349D	Silent
NOC060	NOTCH1	chr9	136509948	SNV	G	A	793	13	0.0164	synonymous_SNV	NM_017617	exon18	c.2754C>T	p.N918N	Silent
NOC061	GATA2	chr3	128481275	SNV	C	T	1150	12	0.0104	nonsynonymous_SNV	NM_032638	exon6	c.1187G>A	p.R396Q	Pathogenic
NOC062	DNMT3A	chr2	25234348	INDEL	G	GC	782	286	0.3657	frameshift_insertion	NM_022552	exon23	c.2669dupG	p.G890fs	Pathogenic
NOC062	SF3B1	chr2	197402110	SNV	T	C	961	13	0.0135	nonsynonymous_SNV	NM_012433	exon15	c.2098A>G	p.K700E	Pathogenic
NOC062	BCORL1	chrX	130015957	SNV	T	C	845	11	0.013	nonsynonymous_SNV	NM_001184772	exon3	c.3185T>C	p.L1062S	VUS
NOC064	DNMT3A	chr2	25247714	SNV	C	A	1205	18	0.0149	nonsynonymous_SNV	NM_022552	exon8	c.891G>T	p.W297C	Pathogenic
NOC064	KMT2E	chr7	105110540	SNV	C	A	732	36	0.0492	stopgain	NM_182931	exon25	c.3908C>A	p.S1303X	Pathogenic
NOC064	ASXL1	chr20	32433739	INDEL	CTG	C	1720	23	0.012	frameshift_deletion	NM_015338	exon11	c.1542_1543del	p.V515Gfs	Pathogenic
NOC064	ASXL1	chr20	32434825	SNV	G	T	1359	44	0.0324	stopgain	NM_015338	exon12	c.2113G>T	p.E705X	Pathogenic
NOC064	TET2	chr4	105235611	INDEL	CAGCACTA	C	1858	45	0.0242	frameshift_deletion	NM_001127208	exon3	c.1670_1676del	p.Q557fs	Pathogenic
NOC065	DNMT3A	chr2	25239199	SNV	A	G	630	20	0.0317	nonsynonymous_SNV	NM_022552	exon20	c.2339T>C	p.I780T	Pathogenic
NOC066	TET2	chr4	105235434	SNV	G	A	1254	21	0.0167	nonsynonymous_SNV	NM_001127208	exon3	c.1492G>A	p.V498I	VUS
NOC067	BRCC3	chrX	155089338	SNV	A	G	447	8	0.0179	nonsynonymous_SNV	NM_024332	exon6	c.479A>G	p.D160G	Pathogenic
NOC068	DNMT3A	chr2	25240312	SNV	C	T	568	10	0.0176	nonsynonymous_SNV	NM_022552	exon19	c.2312G>A	p.R771Q	Pathogenic
NOC075	TP53	chr17	7674947	SNV	A	G	730	25	0.0342	nonsynonymous_SNV	NM_000546	exon6	c.584T>C	p.I195T	Pathogenic
NOC075	ATRX	chrX	77654128	SNV	C	T	597	63	0.1055	synonymous_SNV	NM_000489	exon14	c.4287G>A	p.K1429K	Silent
NOC075	PTPRF	chr1	43619078	SNV	C	T	596	6	0.0101	stopgain	NM_002840	exon27	c.4522C>T	p.Q1508X	Pathogenic
NOC078	TET2	chr4	105243636	SNV	T	A	700	39	0.0557	nonsynonymous_SNV	NM_001127208	exon6	c.3661T>A	p.C1221S	Pathogenic
NOC079	TET2	chr4	105235775	SNV	G	A	1499	28	0.018	nonsynonymous_SNV	NM_001127208	exon3	c.1833G>A	p.M611I	VUS
NOC079	SF3B1	chr2	197402635	SNV	C	G	879	10	0.0114	nonsynonymous_SNV	NM_012433	exon14	c.1998G>C	p.K666N	Pathogenic
NOC082	ASXL1	chr20	32434638	INDEL	A	AG	883	71	0.0804	frameshift_insertion	NM_015338	exon12	c.1927dupG	p.G642fs	Pathogenic
NOC083	DNMT3A	chr2	25234323	SNV	G	A	616	20	0.0325	nonsynonymous_SNV	NM_022552	exon23	c.2695C>T	p.R899C	Pathogenic
NOC083	PTPRF	chr1	43591895	SNV	C	T	780	8	0.0103	nonsynonymous_SNV	NM_002840	exon10	c.1615C>T	p.R539W	Pathogenic
NOC084	DNMT3A	chr2	25247061	MNV	TCG	ACT	511	32	0.0626	stopgain	NM_022552	exon9	c.1110_1112AGT	p.Y370X	Pathogenic

(Continued on next page)

Sample	Gene	Chrom-osome	Start Position	Type	Reference Allele	Variant Allele	Total Depth	Variant Reads	VAF	Variant Classification	Accession	Exon	cDNA change	Protein change	Significance
NOC090	DNMT3A	chr2	25244596	SNV	G	T	814	10	0.0123	stopgain	NM_022552	exon14	c.1611C>A	p.C537X	Pathogenic
NOC090	DNMT3A	chr2	25244539	SNV	C	A	541	7	0.0129	splicing	NM_022552	exon14	c.1667+1G>T	p.R556X	Pathogenic
NOC090	SF3B1	chr2	197402757	SNV	T	C	527	13	0.0247	nonsynonymous_SNV	NM_012433	exon14	c.1876A>G	p.N626D	Pathogenic
NOC091	STAG2	chrX	124049075	SNV	A	G	124	6	0.0484	nonsynonymous_SNV	NM_006603	exon9	c.890A>G	p.Y297C	Pathogenic
NOC091	DNMT3A	chr2	25244203	SNV	C	T	922	18	0.0195	stopgain	NM_022552	exon15	c.1803G>A	p.W601X	Pathogenic
NOC091	TP53	chr17	7675148	SNV	G	A	1394	16	0.0115	nonsynonymous_SNV	NM_000546	exon5	c.464C>T	p.T155I	Pathogenic
NOC092	ASXL1	chr20	32434773	SNV	T	A	1038	35	0.0337	stopgain	NM_015338	exon12	c.2061T>A	p.C687X	Pathogenic
NOC092	DNMT3A	chr2	25235749	SNV	A	C	542	12	0.0221	nonsynonymous_SNV	NM_022552	exon22	c.2555T>G	p.M852R	Pathogenic
NOC096	TET2	chr4	105275146	SNV	C	T	640	10	0.0156	stopgain	NM_001127208	exon11	c.4636C>T	p.Q1546X	Pathogenic
NOC096	ASXL1	chr20	32434636	INDEL	GGAGGGGG GGGT	G	669	13	0.0194	frameshift_deletion	NM_015338	exon12	c.1925_1935del	p.G642fs	Pathogenic
NOC096	SMC1A	chrX	53414817	SNV	T	A	523	11	0.021	nonsynonymous_SNV	NM_006306	exon3	c.352A>T	p.S118C	Pathogenic
NOC097	TET2	chr4	105236160	SNV	C	T	937	40	0.0427	stopgain	NM_001127208	exon3	c.2218C>T	p.Q740X	Pathogenic
NOC098	TET2	chr4	105259618	SNV	G	A	452	36	0.0796	splicing	NM_001127208	exon7	c.3804-1G>A	p.E1268X	Pathogenic
NOC108	DNMT3A	chr2	25247089	SNV	G	A	1021	73	0.0715	stopgain	NM_022552	exon9	c.1084C>T	p.Q362X	Pathogenic
NOC113	SMC1A	chrX	53381072	SNV	G	A	317	7	0.023	synonymous_SNV	NM_006306	exon23	c.3453C>T	p.P1151P	Silent
NOC113	DNMT3A	chr2	25241575	SNV	A	G	748	9	0.012	nonsynonymous_SNV	NM_022552	exon17	c.2069T>C	p.V690A	Pathogenic
NOC113	DNMT3A	chr2	25245256	SNV	G	T	676	13	0.0192	stopgain	NM_022552	exon13	c.1551C>A	p.C517X	Pathogenic
NOC114	SF1	chr11	64769556	SNV	C	T	1114	20	0.018	nonsynonymous_SNV	NM_004630	exon6	c.533G>A	p.R178Q	Pathogenic
NOC115	DNMT3A	chr2	25241685	SNV	C	G	1122	33	0.0294	nonsynonymous_SNV	NM_022552	exon17	c.1959G>C	p.L653F	Pathogenic
NOC115	TET2	chr4	105243759	SNV	C	T	804	81	0.1007	nonsynonymous_SNV	NM_001127208	exon6	c.3784C>T	p.R1262W	Pathogenic
NOC117	TET2	chr4	105234563	INDEL	GC	G	1032	134	0.1298	frameshift_deletion	NM_001127208	exon3	c.622delC	p.P208fs	Pathogenic
NOC117	TET2	chr4	105242914	SNV	C	A	559	9	0.0161	nonsynonymous_SNV	NM_001127208	exon5	c.3581C>A	p.P1194H	Pathogenic
NOC118	PPM1D	chr17	60663143	INDEL	ATC	A	1006	22	0.0219	frameshift_deletion	NM_003620	exon6	c.1410_1411del	p.D470fs	Pathogenic
NOC123	NOTCH1	chr9	136515676	SNV	G	A	668	29	0.0434	synonymous_SNV	NM_017617	exon11	c.1710C>T	p.C570C	Silent
NOC125	DNMT3A	chr2	25244291	SNV	G	T	677	27	0.0399	nonsynonymous_SNV	NM_022552	exon15	c.1715C>A	p.A572D	Pathogenic
NOC125	KMT2D	chr12	49034660	SNV	T	C	443	24	0.0542	synonymous_SNV	NM_003482	exon36	c.10362A>G	p.Q3454Q	Silent
NOC128	BCORL1	chrX	130056005	SNV	G	A	840	36	0.0429	nonsynonymous_SNV	NM_001184772	exon13	c.5227G>A	p.A1743T	Pathogenic
NOC129	PPM1D	chr17	60663012	INDEL	AT	A	642	18	0.024	frameshift_deletion	NM_003620	exon6	c.1279delT	p.W427Gfs	Pathogenic
NOC129	PPM1D	chr17	60663388	SNV	C	T	629	76	0.1208	stopgain	NM_003620	exon6	c.1654C>T	p.R552X	Pathogenic
NOC129	PPM1D	chr17	60663122	SNV	G	C	665	19	0.0286	nonsynonymous_SNV	NM_003620	exon6	c.1388G>C	p.G463A	VUS
NOC130	NOTCH1	chr9	136518683	SNV	A	G	902	11	0.0122	nonsynonymous_SNV	NM_017617	exon6	c.1007T>C	p.I336T	Pathogenic
NOC131	DNMT3A	chr2	25244190	SNV	G	A	626	38	0.0607	stopgain	NM_022552	exon15	c.1816C>T	p.Q606X	Pathogenic
NOC131	DNMT3A	chr2	25239199	SNV	A	G	477	15	0.0314	nonsynonymous_SNV	NM_022552	exon20	c.2339T>C	p.I780T	Pathogenic
NOC132	DNMT3A	chr2	25240313	SNV	G	A	690	88	0.1275	stopgain	NM_022552	exon19	c.2311C>T	p.R771X	Pathogenic
NOC134	SMC3	chr10	110584380	SNV	A	C	776	9	0.0116	nonsynonymous_SNV	NM_005445	exon13	c.1289A>C	p.N430T	Pathogenic
NOC137	TET2	chr4	105243621	SNV	C	T	815	33	0.0405	stopgain	NM_001127208	exon6	c.3646C>T	p.R1216X	Pathogenic
NOC137	ASXL1	chr20	32435605	SNV	C	T	1295	183	0.1413	stopgain	NM_015338	exon12	c.2893C>T	p.R965X	Pathogenic
NOC138	DNMT3A	chr2	25240661	INDEL	G	GCA	737	34	0.0461	frameshift_insertion	NM_022552	exon18	c.2151_2152insTG	p.P718fs	Pathogenic
NOC138	DNMT3A	chr2	25247683	SNV	C	T	951	23	0.0242	nonsynonymous_SNV	NM_022552	exon8	c.922G>A	p.G308S	Pathogenic
NOC139	DNMT3A	chr2	25240432	SNV	A	T	964	17	0.0176	nonsynonymous_SNV	NM_022552	exon19	c.2192T>A	p.F731Y	Pathogenic
NOC139	PTPN11	chr12	112486532	SNV	G	T	1258	17	0.0135	nonsynonymous_SNV	NM_002834	exon11	c.1282G>T	p.V428L	Pathogenic
NOC140	PPM1D	chr17	60663475	SNV	C	T	1054	12	0.0114	stopgain	NM_003620	exon6	c.1741C>T	p.R581X	Pathogenic
NOC142	TET2	chr4	105243595	SNV	A	G	776	16	0.0206	nonsynonymous_SNV	NM_001127208	exon6	c.3620A>G	p.E1207G	Pathogenic
NOC144	BRAF	chr7	140924611	SNV	G	A	119	6	0.057	synonymous_SNV	NM_004333	exon1	c.93C>T	p.A31A	Silent
NOC144	NOTCH1	chr9	136523878	SNV	C	G	1317	75	0.0569	nonsynonymous_SNV	NM_017617	exon3	c.242G>C	p.G81A	Pathogenic
NOC145	KMT2D	chr12	49033949	SNV	T	G	1094	23	0.021	nonsynonymous_SNV	NM_003482	exon39	c.10756A>C	p.K3586Q	VUS
NOC146	TET2	chr4	105259626	INDEL	T	TG	849	27	0.0318	frameshift_insertion	NM_001127208	exon7	c.3812dupG	p.C1271fs	Pathogenic
NOC146	DNMT3A	chr2	25239130	SNV	C	G	605	15	0.0248	nonsynonymous_SNV	NM_022552	exon20	c.2408G>C	p.R803T	Pathogenic
NOC146	DNMT3A	chr2	25240710	SNV	G	T	863	25	0.029	nonsynonymous_SNV	NM_022552	exon18	c.2103C>A	p.F701L	Pathogenic
NOC148	DNMT3A	chr2	25240439	SNV	G	A	834	16	0.0192	nonsynonymous_SNV	NM_022552	exon19	c.2185C>T	p.R729W	Pathogenic
NOC148	JAK2	chr9	5073770	SNV	G	T	923	10	0.0108	nonsynonymous_SNV	NM_004972	exon14	c.1849G>T	p.V617F	Pathogenic
NOC150	DNMT3A	chr2	25239157	INDEL	AAGT	A	992	17	0.018	nonframeshift_deletion	NM_022552	exon20	c.2378_2380del	p.Y793del	Pathogenic
NOC150	TET2	chr4	105236479	INDEL	A	AACAG ACT	1126	20	0.0178	frameshift_insertion	NM_001127208	exon3	c.2537_2538insACA GACT	p.E846fs	Pathogenic
NOC150	TET2	chr4	105236490	SNV	C	G	1118	28	0.025	nonsynonymous_SNV	NM_001127208	exon3	c.2548C>G	p.H850D	VUS
NOC150	TET2	chr4	105237249	INDEL	AATTTTATA	T	1209	16	0.0132	frameshift_substitution	NM_001127208	exon3	c.3307_3315T	p.N1103X	Pathogenic
NOC150	EP300	chr22	41176460	SNV	C	T	1306	32	0.0245	nonsynonymous_SNV	NM_001429	exon30	c.4993C>T	p.R1665C	Pathogenic
NOC151	JAK2	chr9	5073770	SNV	G	T	849	9	0.0106	nonsynonymous_SNV	NM_004972	exon14	c.1849G>T	p.V617F	Pathogenic
NOC151	ASXL1	chr20	32436833	INDEL	T	TG	1242	20	0.0161	frameshift_insertion	NM_015338	exon12	c.4122dupG	p.V1374fs	Pathogenic
NOC152	KMT2C	chr7	152265054	SNV	C	A	1882	85	0.044	nonsynonymous_SNV	NM_170606	exon8	c.1168G>T	p.V390L	Pathogenic
NOC152	SMC1A	chrX	53413086	SNV	A	G	1186	15	0.0126	nonsynonymous_SNV	NM_006306	exon5	c.668T>C	p.L223P	Pathogenic
NOC158	DNMT3A	chr2	25243919	SNV	G	A	627	12	0.0191	nonsynonymous_SNV	NM_022552	exon16	c.1915C>T	p.L639F	Pathogenic
NOC160	KMT2D	chr12	49051324	INDEL	GTGGCTCCTC AGGCACAGC GCATAGGCA	G	1158	273	0.2358	nonframeshift_deletion	NM_003482	exon10	c.2332_2358del	p.778_786del	VUS
NOC162	DNMT3A	chr2	25247142	SNV	A	G	1076	12	0.0112	nonsynonymous_SNV	NM_022552	exon9	c.1031T>C	p.L344P	Pathogenic
NOC163	PRPF8	chr17	1679661	SNV	G	T	1389	26	0.0187	nonsynonymous_SNV	NM_006445	exon9	c.1237C>A	p.L413I	Pathogenic
NOC164	TET2	chr4	105243757	SNV	G	T	899	10	0.0111	nonsynonymous_SNV	NM_001127208	exon6	c.3782G>T	p.R1261L	Pathogenic
NOC168	DNMT3A	chr2	25244289	SNV	G	A	978	51	0.0521	stopgain	NM_022552	exon15	c.1717C>T	p.Q573X	Pathogenic
NOC170	TET2	chr4	105269685	SNV	T	C	953	25	0.0262	nonsynonymous_SNV	NM_001127208	exon9	c.4120T>C	p.C1374R	Pathogenic
NOC171	DNMT3A	chr2	25244265	INDEL	A	AG	505	116	0.2297	frameshift_insertion	NM_022552	exon15	c.1740dupC	p.W581fs	Pathogenic
NOC174	SF3B1	chr2	197402635	SNV	C	G	763	117	0.1533	nonsynonymous_SNV	NM_012433	exon14	c.1998G>C	p.K666N	Pathogenic
NOC176	PPM1D	chr17	60663092	SNV	C	A	732	31	0.0423	stopgain	NM_003620	exon6	c.1358C>A	p.S453X	Pathogenic
NOC176	PPM1D	chr17	60663365	INDEL	G	GC	870	159	0.1828	frameshift_insertion	NM_003620	exon6	c.1632dupC	p.G544fs	Pathogenic
NOC177	DNMT3A	chr2	25235724	SNV	C	T	492	8	0.0163	stopgain	NM_022552	exon22	c.2580G>A	p.W860X	Pathogenic
NOC179	TET2	chr4	105261769	SNV	T	A	424	26	0.0613	nonsynonymous_SNV	NM_001127208	exon8	c.3965T>A	p.L1322Q	Pathogenic
NOC180	ASXL1	chr20	32433562	INDEL	AGACT	A	615	47	0.0764	frameshift_deletion	NM_015338	exon11	c.1365_1368del	p.K455fs	Pathogenic
NOC181	CSF3R	chr1	36468090	SNV	G	C	589	16	0.0272	nonsynonymous_SNV	NM_000760	exon13	c.1708C>G	p.Q570E	Pathogenic
NOC185	TET2	chr4	105259734	INDEL	AG	A	1032	98	0.095	frameshift_deletion	NM_001127208	exon7	c.3920delG	p.R1307fs	Pathogenic
NOC188	KMT2A	chr11	118501091	SNV	G	A	472	5	0.0106	nonsynonymous_SNV	NM_001197104	exon25	c.6263G>A	p.S2088N	VUS
NOC189	MTA2	chr11	62596641	SNV	T	A	467	47	0.1006	nonsynonymous_SNV	NM_004739	exon9	c.878A>T	p.D293V	Pathogenic
NOC189	DNMT3A	chr2	25234400	SNV	T	C	502	9	0.0179	nonsynonymous_SNV	NM_022552	exon23	c.2618A>G	p.H873R	Pathogenic
NOC189	NF1	chr17	31337881	SNV	G	A	445	6	0.0135	splicing	NM_001042492	exon44	c.6704+1G>A	p.R2235X	Pathogenic
NOC190	DNMT3A	chr2	25244603	SNV	G	A	561	6	0.0107	nonsynonymous_SNV	NM_022552	exon14	c.1604C>T	p.S535F	Pathogenic
NOC195	STAG2	chrX	124061858	SNV	G	T	378	7	0.0185	nonsynonymous_SNV	NM_006603	exon16	c.1622G>T	p.R541I	VUS
NOC196	GNAS	chr20	58854799	SNV	C	T	730	8	0.011	nonsynonymous_SNV	NM_080425	exon1	c.1534C>T	p.R512W	VUS
NOC197	DNMT3A	chr2	25234373	SNV	C	T	410	9	0.022	nonsynonymous_SNV	NM_022552	exon23	c.2645G>A	p.R882H	Pathogenic
NOC197	GATA2	chr3	128485976	SNV	T	C	493	9	0.0183	nonsynonymous_SNV	NM_032638	exon3	c.622A>G	p.K208E	VUS

A.3 Flow cytometry antibodies

Table A.4: Antibodies used for immunophenotypic analysis of BM HSPCs.

Antigen	Clone	Fluorochrome	Manufacturer	Catalogue No.	Dilution
CD38	HIT2	BV421	BioLegend	303526	1:20
CD10	HI10a	BV605	BioLegend	312222	1:40
CD49f	GoH3	BV650	BioLegend	313629	1:40
CD117	104D2	BV785	BioLegend	313238	1:40
CD45RA	HI100	BB515	BD	564552	1:40
CD123	6H6	PE	BioLegend	306006	1:40
CD90	5E10	PE/Cy7	BioLegend	328124	1:20
CD34	581	APC	BioLegend	343510	1:160
CD2	RPA-2.10	PE/Cy5	BioLegend	300210	1:160
CD3	HIT3a	PE/Cy5	BioLegend	300310	1:320
CD4	RPA-T4	PE/Cy5	BioLegend	300510	1:160
CD8a	RPA-T8	PE/Cy5	BioLegend	301010	1:320
CD11b	ICRF44	PE/Cy5	BioLegend	301308	1:160
CD14	61D3	PE/Cy5	eBioscience	15-0149-42	1:160
CD19	HIB19	PE/Cy5	BioLegend	302210	1:160
CD20	2H7	PE/Cy5	BioLegend	302308	1:160
CD56	MEM188	PE/Cy5	BioLegend	304608	1:80
CD235ab	HIR2	PE/Cy5	BioLegend	306606	1:320
Live/dead	-	Propidium Iodide	BioLegend	421301	1:500 (3µM)

Table A.5: Antibodies used for sorting BM samples for TARGET-seq+.

Antigen	Clone	Fluorochrome	Manufacturer	Catalogue No.	Dilution
CD45RA	HI100	BB515	BD	564552	1:40
CD123	6H6	PE	BioLegend	306006	1:40
CD49f	GoH3	PE/Dazzle 594	BioLegend	313626	1:160
CD90	5E10	PE/Cy7	BioLegend	328124	1:20
CD38	HIT2	BV421	BioLegend	303526	1:20
CD10	HI10a	BV605	BioLegend	312222	1:40
CD34	581	APC	BioLegend	343510	1:160
CD117	104D2	BV785	BioLegend	313238	1:40
CD2	RPA-2.10	PE/Cy5	BioLegend	300210	1:160
CD3	HIT3a	PE/Cy5	BioLegend	300310	1:320
CD4	RPA-T4	PE/Cy5	BioLegend	300510	1:160
CD8a	RPA-T8	PE/Cy5	BioLegend	301010	1:320
CD11b	ICRF44	PE/Cy5	BioLegend	301308	1:160
CD14	61D3	PE/Cy5	eBioscience	15-0149-42	1:160
CD19	HIB19	PE/Cy5	BioLegend	302210	1:160
CD20	2H7	PE/Cy5	BioLegend	302308	1:160
CD56	MEM188	PE/Cy5	BioLegend	304608	1:80
CD235ab	HIR2	PE/Cy5	BioLegend	306606	1:320
Live/dead	-	7-AAD	BioLegend	420404	1:100

A.4 Reagents and conditions used for cDNA synthesis in TARGET-seq

Table A.6: TARGET-seq lysis buffer. Reagents used for cell lysis in TARGET-seq.

Reagent	Volume (µL)	Reaction concentration
Triton X-100 (0.4% solution)	1.9	0.18%
dNTPs (10 mM each)	1.0	1.0 mM each
RNAse Inhibitor (40 U/µL)	0.1	1.0 U/µL
Protease (1.09 AU/mL)	0.1	2.7E-05 AU/µL
ERCC mix 1:400,000	0.02	1:8,000,000
Total without barcode	**3.1**	
OligodT-ISPCR barcode (10 µM)	1.0	1.0 µM
Total with barcode	**4.1**	

Table A.7: TARGET-seq RT mix. Reagents used for reverse transcription and mRNA targeting in TARGET-seq.

Reagent	Volume (µL)	Final concentration
5X First Strand Buffer (SMARTScribe)	2.0	1 X
RNase Inhibitor (40 U/µL)	0.25	1.0 U/µL
TSO-LNA 100 µM	0.1	1.0 µM
SMARTScribe (100 U/µL)	1.0	10 U/µL
mRNA primers (25 µM)	0.28	700 nM
Nuclease free water	Variable	-
Total	**5.6**	
Total (Cumulative)	**9.7**	

Table A.8: TARGET-seq pre-amplification PCR buffer. Reagents used for pre-amplification PCR and gDNA/cDNA targeting in TARGET-seq.

Reagent	Volume (µL)	Final concentration
2X Buffer SeqAmp	12.5	1 X
ISPCR Primer (10 µM)	0.125	50 nM
SeqAmp (100 U/µL)	0.5	2 U/µL
cDNA primers (25 µM, same as mRNA primers)	0.028	28 nM
gDNA primers (100 µM)	0.2	800 nM
Nuclease free water	Variable	-
Total	**15.0**	
Total (Cumulative)	**24.7**	

A.5 Reagents and conditions used for cDNA synthesis in TARGET-seq+

Table A.9: TARGET-seq+ lysis buffer. Reagents used for cell lysis in TARGET-seq+.

Reagent	Volume (µL)	Reaction concentration
Triton X-100 (10% solution)	0.03	0.1%
dNTPs (10 mM each)	0.2	0.5 mM each
Poly-ethylene Glycol 8000 (40% solution)	0.5	5%
RNAse Inhibitor (40 U/µL)	0.04	0.5 U/µL
Protease (1.09 AU/mL)	0.075	2.7E-05 AU/µL
ERCC mix 1:400,000	0.015	1:8,000,000
Nuclease Free Water	1.94	-
Total without barcode	**2.8**	
OligodT-ISPCR barcode (10 µM)	0.2	0.5 µM
Total with barcode	**3.0**	

Table A.10: TARGET-seq+ RT mix. Reagents used for reverse transcription and mRNA targeting in TARGET-seq+.

Reagent	Volume (µL)	Final concentration
Tris-HCl pH 8.3 (1 M)	0.1	25 mM
NaCl (1 M)	0.12	30 mM
MgCl2 (100 mM)	0.1	2.5 mM
GTP (100 mM)	0.04	1 mM
DTT (100 mM)	0.32	8 mM
RNase Inhibitor (40 U/µL)	0.05	0.5 U/µL
TSO-LNA (100 µM)	0.08	2 µM
Maxima H-minus RT enzyme (200 U/µL)	0.04	2 U/µL
mRNA primers (25 µM)	0.0112	70 nM
Nuclease free water	Variable	-
Total	**1.0**	
Total (Cumulative)	**4.0**	

Table A.11: TARGET-seq+ pre-amplification PCR buffer. Reagents used for pre-amplification PCR and gDNA/cDNA targeting in TARGET-seq+.

Reagent	Volume (µL)	Final concentration
Kapa HiFi HotStart Ready Mix (2 X)	5.0	1 X
ISPCR Primer (10 µM)	0.05	50 nM
cDNA primers (same as mRNA primers)	0.0112	28 nM
gDNA primers (100 µM)	0.04	400 nM
Nuclease free water	Variable	-
Total	**6.0**	
Total (Cumulative)	**10.0**	

A.6 Oligonucleotides used in TARGET-seq+

Table A.12: Common oligonucleotides used for TARGET-seq+. List of common oligonucelotide primers used for TARGET-seq+ library preparation. For i7 and i5 index primers, the generic sequence is shown with the location of the index sequence indicated by the square brackets. r: ribonucleotide base, +: locked nucleic acid (LNA), *: Phosphorothioated DNA base.

Oligo name	Step	Type	Purification	Sequence
OligodT-ISPCR (non-barcoded)	Lysis buffer	DNA	HPLC	AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
TSO-LNA	RT	RNA LNA	RNase free HPLC	AAGCAGTGGTATCAACGCAGAGTACATrGrG+G
ISPCR primer	Pre-amplification PCR	DNA	HPLC	AAGCAGTGGTATCAACGCAGAGT
i7-index primer (Nextera)	Transcriptome library PCR	DNA	HPLC	CAAGCAGAAGACGGCATACGAGAT**[i7]**GTCTCGTGGGCTCGG
i5-index primer (custom)	Transcriptome library PCR	DNA	HPLC	AATGATACGGCGACCACCGAGATCTACAC**[i5]**GCCTGTCCGCG GAAGCAGTGGTATCAACGCAGAGT*T*G
P5-SEQ (Read1)	Transcriptome library sequencing	DNA	PAGE	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTTGC*T
I5-SEQ (Index2)	Transcriptome library sequencing	DNA	PAGE	AGCAACTCTGCGTTGATACCACTGCTTCCGCGGACAGG*C
LCS1 (Read1)	Genotyping library sequencing	DNA	HPLC	GGCGACCACCGAGATCTACACTGACGACATGGTTCTACA
CS2 (Read2)	Genotyping library sequencing	DNA LNA	HPLC	T+AC+GGT+AGCAGAGACTTGGTCT
CS2rc (Index)	Genotyping library sequencing	DNA LNA	HPLC	A+GAC+CA+AGTCTCTGCTACCGTA

Table A.13: Custom i5 transcriptome index primers. Sequences of custom i5 index primers used for amplification of 3' transcriptome libraries. The index sequence is shown underlined in bold.

Index name	Sequence
P5_S502	AATGATACGGCGACCACCGAGATCTACAC**CTCTCTAT**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S503	AATGATACGGCGACCACCGAGATCTACAC**TATCCTCT**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S504	AATGATACGGCGACCACCGAGATCTACAC**AGAGTAGA**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S505	AATGATACGGCGACCACCGAGATCTACAC**GTAAGGAG**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S507	AATGATACGGCGACCACCGAGATCTACAC**AAGGAGTA**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S508	AATGATACGGCGACCACCGAGATCTACAC**CTAAGCCT**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S513	AATGATACGGCGACCACCGAGATCTACAC**TCGACTAG**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G
P5_S517	AATGATACGGCGACCACCGAGATCTACAC**GCGTAAGA**GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*T*G

Table A.14: Barcoded OligodT-ISPCR primers.
Sequences of barcoded oligodT-ISPCR primers used in the amplification of 3' transcriptome libraries for TARGET-seq+ experiments. Capitalised bases indicate the cell barcode.

Barcode Name	Sequence
HT001	aagcagtggtatcaacgcagagttgctCATATTCCTGGTGGvttttttttttttttttttttttttttttt
HT002	aagcagtggtatcaacgcagagttgctACAACAAGGTTCGGvttttttttttttttttttttttttttttt
HT003	aagcagtggtatcaacgcagagttgctTCGGTGAGGATTGGvttttttttttttttttttttttttttttt
HT004	aagcagtggtatcaacgcagagttgctGACAACATCAGCGGvttttttttttttttttttttttttttttt
HT005	aagcagtggtatcaacgcagagttgctATATTGCAACCTGGvttttttttttttttttttttttttttttt
HT006	aagcagtggtatcaacgcagagttgctTTGGAACCTTGCCGvttttttttttttttttttttttttttttt
HT007	aagcagtggtatcaacgcagagttgctGTTATAGCTTGTCGvttttttttttttttttttttttttttttt
HT008	aagcagtggtatcaacgcagagttgctGAGGTATGGTCTTGvttttttttttttttttttttttttttttt
HT009	aagcagtggtatcaacgcagagttgctAGCTCATAAGTGAGvttttttttttttttttttttttttttttt
HT010	aagcagtggtatcaacgcagagttgctGAGAAGTAATCCAGvttttttttttttttttttttttttttttt
HT011	aagcagtggtatcaacgcagagttgctGGAATTCGCTTCGGvttttttttttttttttttttttttttttt
HT012	aagcagtggtatcaacgcagagttgctTGATTCCTAGGCGGvttttttttttttttttttttttttttttt
HT013	aagcagtggtatcaacgcagagttgctCCTTCTCCATCCTGvttttttttttttttttttttttttttttt
HT014	aagcagtggtatcaacgcagagttgctCGCGAATTGATTAGvttttttttttttttttttttttttttttt
HT015	aagcagtggtatcaacgcagagttgctTTAGGCGTTGGTGGvttttttttttttttttttttttttttttt
HT016	aagcagtggtatcaacgcagagttgctACAAGTCTCACGTGvttttttttttttttttttttttttttttt
HT017	aagcagtggtatcaacgcagagttgctAGACGGAGTTGTGGvttttttttttttttttttttttttttttt
HT018	aagcagtggtatcaacgcagagttgctACATGAGTGTTCTGvttttttttttttttttttttttttttttt
HT019	aagcagtggtatcaacgcagagttgctAATCCTTCGTGTGGvttttttttttttttttttttttttttttt
HT020	aagcagtggtatcaacgcagagttgctCTCTCCAAGACTGGvttttttttttttttttttttttttttttt
HT021	aagcagtggtatcaacgcagagttgctCCAAGCGAACCTTGvttttttttttttttttttttttttttttt
HT022	aagcagtggtatcaacgcagagttgctTACTACGGTTACGGvttttttttttttttttttttttttttttt
HT023	aagcagtggtatcaacgcagagttgctCACGTGCATACTGGvttttttttttttttttttttttttttttt
HT024	aagcagtggtatcaacgcagagttgctTATAGGCAACGCGGvttttttttttttttttttttttttttttt
HT025	aagcagtggtatcaacgcagagttgctATATGAAGGCTTGGvttttttttttttttttttttttttttttt
HT026	aagcagtggtatcaacgcagagttgctCCTTAGATCCTCCGvttttttttttttttttttttttttttttt
HT027	aagcagtggtatcaacgcagagttgctTACATAAGCCTTCGvttttttttttttttttttttttttttttt
HT028	aagcagtggtatcaacgcagagttgctGCCAGGAAGTCTTGvttttttttttttttttttttttttttttt
HT029	aagcagtggtatcaacgcagagttgctAGCAAGACTAGCGGvttttttttttttttttttttttttttttt
HT030	aagcagtggtatcaacgcagagttgctAGCCTTGGATCTGGvttttttttttttttttttttttttttttt
HT031	aagcagtggtatcaacgcagagttgctTGAGATCCACCTCGvttttttttttttttttttttttttttttt
HT032	aagcagtggtatcaacgcagagttgctGTCCAGTAGTATGGvttttttttttttttttttttttttttttt
HT033	aagcagtggtatcaacgcagagttgctTTATACTCCTTGCGvttttttttttttttttttttttttttttt
HT034	aagcagtggtatcaacgcagagttgctTCCTTGGTACGTGGvttttttttttttttttttttttttttttt
HT035	aagcagtggtatcaacgcagagttgctAGGCAATCAGGTGGvttttttttttttttttttttttttttttt
HT036	aagcagtggtatcaacgcagagttgctTCTAAGCGGTCTTGvttttttttttttttttttttttttttttt
HT037	aagcagtggtatcaacgcagagttgctGAGGCTTAACTTGGvttttttttttttttttttttttttttttt
HT038	aagcagtggtatcaacgcagagttgctTGACAGGACTTCGGvttttttttttttttttttttttttttttt
HT039	aagcagtggtatcaacgcagagttgctGGAGCACTAGCTTGvttttttttttttttttttttttttttttt
HT040	aagcagtggtatcaacgcagagttgctATGTTATCTCGTGGvttttttttttttttttttttttttttttt
HT041	aagcagtggtatcaacgcagagttgctAGCTAATGTTGTCGvttttttttttttttttttttttttttttt
HT042	aagcagtggtatcaacgcagagttgctGGAACGATATTCAGvttttttttttttttttttttttttttttt
HT043	aagcagtggtatcaacgcagagttgctTATGGTGGACCTGGvttttttttttttttttttttttttttttt
HT044	aagcagtggtatcaacgcagagttgctGCATGTATTACCGGvttttttttttttttttttttttttttttt
HT045	aagcagtggtatcaacgcagagttgctACACGCTACTTCGGvttttttttttttttttttttttttttttt
HT046	aagcagtggtatcaacgcagagttgctTTGTCAAGTGGCGGvttttttttttttttttttttttttttttt
HT047	aagcagtggtatcaacgcagagttgctAGGTATGTATACGGvttttttttttttttttttttttttttttt
HT048	aagcagtggtatcaacgcagagttgctGCAAGTACACATAGvttttttttttttttttttttttttttttt
HT049	aagcagtggtatcaacgcagagttgctTTGCGAGTCTCCTGvttttttttttttttttttttttttttttt
HT050	aagcagtggtatcaacgcagagttgctCAATCCACACTCCGvttttttttttttttttttttttttttttt
HT051	aagcagtggtatcaacgcagagttgctACACTTCGCCTTGGvttttttttttttttttttttttttttttt
HT052	aagcagtggtatcaacgcagagttgctGTATATAACTCCGGvttttttttttttttttttttttttttttt
HT053	aagcagtggtatcaacgcagagttgctCGTCATTCTTGCGGvttttttttttttttttttttttttttttt
HT054	aagcagtggtatcaacgcagagttgctACAGAATCGTCCGGvttttttttttttttttttttttttttttt
HT055	aagcagtggtatcaacgcagagttgctTTACTGGCTAGCGGvttttttttttttttttttttttttttttt
HT056	aagcagtggtatcaacgcagagttgctGTAATGCCGTGTAGvttttttttttttttttttttttttttttt
HT057	aagcagtggtatcaacgcagagttgctCGCGATAACGATCGvttttttttttttttttttttttttttttt
HT058	aagcagtggtatcaacgcagagttgctTCTCTGGTAGCTAGvttttttttttttttttttttttttttttt
HT059	aagcagtggtatcaacgcagagttgctATTGGTTATGACGGvttttttttttttttttttttttttttttt
HT060	aagcagtggtatcaacgcagagttgctCAACCATTGCTTGGvttttttttttttttttttttttttttttt
HT061	aagcagtggtatcaacgcagagttgctGCTAATTCCACTCGvttttttttttttttttttttttttttttt
HT062	aagcagtggtatcaacgcagagttgctACTAGTCCTTGGCGvttttttttttttttttttttttttttttt
HT063	aagcagtggtatcaacgcagagttgctGACAATCTTCTTGGvttttttttttttttttttttttttttttt
HT064	aagcagtggtatcaacgcagagttgctAGGTCACTGCCTTGvttttttttttttttttttttttttttttt

(Continued on next page)

Barcode Name	Sequence
HT065	aagcagtggtatcaacgcagagttgctTTCTGCGCCATTCGvttttttttttttttttttttttttttttt
HT066	aagcagtggtatcaacgcagagttgctCCTCGAGAATGTGGvttttttttttttttttttttttttttttt
HT067	aagcagtggtatcaacgcagagttgctGAGTTCTTGTGCGGvttttttttttttttttttttttttttttt
HT068	aagcagtggtatcaacgcagagttgctACCATGTCGTTCCGvttttttttttttttttttttttttttttt
HT069	aagcagtggtatcaacgcagagttgctTATAATTGCCGCGGvttttttttttttttttttttttttttttt
HT070	aagcagtggtatcaacgcagagttgctTAGAAGATTGTCGGvttttttttttttttttttttttttttttt
HT071	aagcagtggtatcaacgcagagttgctGATGTGGACTTCCGvttttttttttttttttttttttttttttt
HT072	aagcagtggtatcaacgcagagttgctAACCTGGAACGCTGvttttttttttttttttttttttttttttt
HT073	aagcagtggtatcaacgcagagttgctCTGAACGTGGCTTGvttttttttttttttttttttttttttttt
HT074	aagcagtggtatcaacgcagagttgctGAGGATATACTGAGvttttttttttttttttttttttttttttt
HT075	aagcagtggtatcaacgcagagttgctGTTCAACTATACGGvttttttttttttttttttttttttttttt
HT076	aagcagtggtatcaacgcagagttgctCAGCATTGTACGTGvttttttttttttttttttttttttttttt
HT077	aagcagtggtatcaacgcagagttgctCCAACTTCGGTCTGvttttttttttttttttttttttttttttt
HT078	aagcagtggtatcaacgcagagttgctTCTTGCATCTTCGGvttttttttttttttttttttttttttttt
HT079	aagcagtggtatcaacgcagagttgctTTGCTAAGATACGGvttttttttttttttttttttttttttttt
HT080	aagcagtggtatcaacgcagagttgctAGAACACCATCCGGvttttttttttttttttttttttttttttt
HT081	aagcagtggtatcaacgcagagttgctCAGTCAGAACATGGvttttttttttttttttttttttttttttt
HT082	aagcagtggtatcaacgcagagttgctATTGGCTGCTGTGGvttttttttttttttttttttttttttttt
HT083	aagcagtggtatcaacgcagagttgctAACACCAATTACGGvttttttttttttttttttttttttttttt
HT084	aagcagtggtatcaacgcagagttgctAACAGTACCTGCGGvttttttttttttttttttttttttttttt
HT085	aagcagtggtatcaacgcagagttgctGATAGTTGATACGGvttttttttttttttttttttttttttttt
HT086	aagcagtggtatcaacgcagagttgctATAAGGCATAGTGGvttttttttttttttttttttttttttttt
HT087	aagcagtggtatcaacgcagagttgctTTGTTGACACTTGGvttttttttttttttttttttttttttttt
HT088	aagcagtggtatcaacgcagagttgctCATTACCGCTTCGGvttttttttttttttttttttttttttttt
HT089	aagcagtggtatcaacgcagagttgctGTATCACGCGGTTGvttttttttttttttttttttttttttttt
HT090	aagcagtggtatcaacgcagagttgctAGCATCCAACATGGvttttttttttttttttttttttttttttt
HT091	aagcagtggtatcaacgcagagttgctCGTCTTCGACTTCCvttttttttttttttttttttttttttttt
HT092	aagcagtggtatcaacgcagagttgctTCAACTCGACAGCCvttttttttttttttttttttttttttttt
HT093	aagcagtggtatcaacgcagagttgctAAGGCGACAGTCTCvttttttttttttttttttttttttttttt
HT094	aagcagtggtatcaacgcagagttgctGACTGAAGATCTCCvttttttttttttttttttttttttttttt
HT095	aagcagtggtatcaacgcagagttgctCTAGCTAATGCTCCvttttttttttttttttttttttttttttt
HT096	aagcagtggtatcaacgcagagttgctCAGGTACTTATGTCvttttttttttttttttttttttttttttt
HT097	aagcagtggtatcaacgcagagttgctTGCTGTCAGATTGCvttttttttttttttttttttttttttttt
HT098	aagcagtggtatcaacgcagagttgctCTATTGCCTTCGCCvttttttttttttttttttttttttttttt
HT099	aagcagtggtatcaacgcagagttgctCAGAGGCTGTTGTCvttttttttttttttttttttttttttttt
HT100	aagcagtggtatcaacgcagagttgctATGAGTAGACCTTCvttttttttttttttttttttttttttttt
HT101	aagcagtggtatcaacgcagagttgctACCGTTAAGTCGCCvttttttttttttttttttttttttttttt
HT102	aagcagtggtatcaacgcagagttgctATCTGTGGTTCTACvttttttttttttttttttttttttttttt
HT103	aagcagtggtatcaacgcagagttgctACCAAGGAACATACvttttttttttttttttttttttttttttt
HT104	aagcagtggtatcaacgcagagttgctTTCGTTAACTGTCCvttttttttttttttttttttttttttttt
HT105	aagcagtggtatcaacgcagagttgctATGCTCTAACCTCCvttttttttttttttttttttttttttttt
HT106	aagcagtggtatcaacgcagagttgctTGACTCAAGCGTGCvttttttttttttttttttttttttttttt
HT107	aagcagtggtatcaacgcagagttgctTTAACCGGAGTTCCvttttttttttttttttttttttttttttt
HT108	aagcagtggtatcaacgcagagttgctGACACGTGTTCTCCvttttttttttttttttttttttttttttt
HT109	aagcagtggtatcaacgcagagttgctTGTGCGGTTCGTTCvttttttttttttttttttttttttttttt
HT110	aagcagtggtatcaacgcagagttgctAAGCGCCATGGTTCvttttttttttttttttttttttttttttt
HT111	aagcagtggtatcaacgcagagttgctTCAGAGGTAGTTACvttttttttttttttttttttttttttttt
HT112	aagcagtggtatcaacgcagagttgctGTTGGTTAGTCTTCvttttttttttttttttttttttttttttt
HT113	aagcagtggtatcaacgcagagttgctCATTACTATACCGCvttttttttttttttttttttttttttttt
HT114	aagcagtggtatcaacgcagagttgctCGACAACTACTGCCvttttttttttttttttttttttttttttt
HT115	aagcagtggtatcaacgcagagttgctATACCTCAAGGTCCvttttttttttttttttttttttttttttt
HT116	aagcagtggtatcaacgcagagttgctCAAGGTCTTCCTCCvttttttttttttttttttttttttttttt
HT117	aagcagtggtatcaacgcagagttgctCCAATTGAGCGCTCvttttttttttttttttttttttttttttt
HT118	aagcagtggtatcaacgcagagttgctACACTACCGTGTCCvttttttttttttttttttttttttttttt
HT119	aagcagtggtatcaacgcagagttgctAATAGTTGCGTTGCvttttttttttttttttttttttttttttt
HT120	aagcagtggtatcaacgcagagttgctCAATTCAGTAGCACvttttttttttttttttttttttttttttt
HT121	aagcagtggtatcaacgcagagttgctAACCACGATAGGCCvttttttttttttttttttttttttttttt
HT122	aagcagtggtatcaacgcagagttgctGCTAGAGACTCTTCvttttttttttttttttttttttttttttt
HT123	aagcagtggtatcaacgcagagttgctGTTAGGAAGTTCACvttttttttttttttttttttttttttttt
HT124	aagcagtggtatcaacgcagagttgctTTGATCCGTTCTTCvttttttttttttttttttttttttttttt
HT125	aagcagtggtatcaacgcagagttgctGGTTAGTAGATTCCvttttttttttttttttttttttttttttt
HT126	aagcagtggtatcaacgcagagttgctCCGGACTAACCTTCvttttttttttttttttttttttttttttt
HT127	aagcagtggtatcaacgcagagttgctTAGTTGGTACCTTCvttttttttttttttttttttttttttttt
HT128	aagcagtggtatcaacgcagagttgctAGAGCGATCGTTGCvttttttttttttttttttttttttttttt

(Continued on next page)

Barcode Name	Sequence
HT129	aagcagtggtatcaacgcagagttgctAAGAGTTCTACCACvttttttttttttttttttttttttttttttt
HT130	aagcagtggtatcaacgcagagttgctATAATAAGCACGCCvttttttttttttttttttttttttttttttt
HT131	aagcagtggtatcaacgcagagttgctGACCTAACTAGCGCvttttttttttttttttttttttttttttttt
HT132	aagcagtggtatcaacgcagagttgctCTTCTAGAGGTTCCvttttttttttttttttttttttttttttttt
HT133	aagcagtggtatcaacgcagagttgctAGCATGGACTCTCCvttttttttttttttttttttttttttttttt
HT134	aagcagtggtatcaacgcagagttgctTTCCAGATACCTTCvttttttttttttttttttttttttttttttt
HT135	aagcagtggtatcaacgcagagttgctATAAGACACTCTGCvttttttttttttttttttttttttttttttt
HT136	aagcagtggtatcaacgcagagttgctGAATTCTCGATTCCvttttttttttttttttttttttttttttttt
HT137	aagcagtggtatcaacgcagagttgctGTCAATTCACCTACvttttttttttttttttttttttttttttttt
HT138	aagcagtggtatcaacgcagagttgctGGTATTAAGACGTCvttttttttttttttttttttttttttttttt
HT139	aagcagtggtatcaacgcagagttgctTCTCACTTCACTCCvttttttttttttttttttttttttttttttt
HT140	aagcagtggtatcaacgcagagttgctTGTTCACCTGCGTCvttttttttttttttttttttttttttttttt
HT141	aagcagtggtatcaacgcagagttgctTGTAACTAGTTGCCvttttttttttttttttttttttttttttttt
HT142	aagcagtggtatcaacgcagagttgctGGTCTACTGTTCGCvttttttttttttttttttttttttttttttt
HT143	aagcagtggtatcaacgcagagttgctTTCCTCTAATCGACvttttttttttttttttttttttttttttttt
HT144	aagcagtggtatcaacgcagagttgctGAGTCCGATCTCTCvttttttttttttttttttttttttttttttt
HT145	aagcagtggtatcaacgcagagttgctGGTGGATAAGTTCCvttttttttttttttttttttttttttttttt
HT146	aagcagtggtatcaacgcagagttgctCCTCCTGAATAGACvttttttttttttttttttttttttttttttt
HT147	aagcagtggtatcaacgcagagttgctTAATTGCAATCCGCvttttttttttttttttttttttttttttttt
HT148	aagcagtggtatcaacgcagagttgctAATACTGACGCTCCvttttttttttttttttttttttttttttttt
HT149	aagcagtggtatcaacgcagagttgctAAGTATTCCACTCCvttttttttttttttttttttttttttttttt
HT150	aagcagtggtatcaacgcagagttgctAGGCAGCTAAGTCCvttttttttttttttttttttttttttttttt
HT151	aagcagtggtatcaacgcagagttgctGTATACTCTCAGTCvttttttttttttttttttttttttttttttt
HT152	aagcagtggtatcaacgcagagttgctTATCGGTTGCCTCCvttttttttttttttttttttttttttttttt
HT153	aagcagtggtatcaacgcagagttgctTGGCCTGTACTGTCvttttttttttttttttttttttttttttttt
HT154	aagcagtggtatcaacgcagagttgctTACACCTTATTGCCvttttttttttttttttttttttttttttttt
HT155	aagcagtggtatcaacgcagagttgctATACATGACTTGCCvttttttttttttttttttttttttttttttt
HT156	aagcagtggtatcaacgcagagttgctTCCGCTTGGTTGTCvttttttttttttttttttttttttttttttt
HT157	aagcagtggtatcaacgcagagttgctTCCATGCAATTGTCvttttttttttttttttttttttttttttttt
HT158	aagcagtggtatcaacgcagagttgctTCCATACGGAGTCCvttttttttttttttttttttttttttttttt
HT159	aagcagtggtatcaacgcagagttgctTGACTCTTCTCGCCvttttttttttttttttttttttttttttttt
HT160	aagcagtggtatcaacgcagagttgctACTTCAATTCTGCCvttttttttttttttttttttttttttttttt
HT161	aagcagtggtatcaacgcagagttgctTTCACTATTGAGCCvttttttttttttttttttttttttttttttt
HT162	aagcagtggtatcaacgcagagttgctAGAGGAACCGCTTCvttttttttttttttttttttttttttttttt
HT163	aagcagtggtatcaacgcagagttgctAGAAGTCAAGGCGCvttttttttttttttttttttttttttttttt
HT164	aagcagtggtatcaacgcagagttgctATCGGATAAGATGCvttttttttttttttttttttttttttttttt
HT165	aagcagtggtatcaacgcagagttgctGCGCAATAGTCTCCvttttttttttttttttttttttttttttttt
HT166	aagcagtggtatcaacgcagagttgctAATACGCCATATCCvttttttttttttttttttttttttttttttt
HT167	aagcagtggtatcaacgcagagttgctACCTCGTAGCTCTCvttttttttttttttttttttttttttttttt
HT168	aagcagtggtatcaacgcagagttgctAACCAGACTTCTCCvttttttttttttttttttttttttttttttt
HT169	aagcagtggtatcaacgcagagttgctTAGGTTCGAGCTCCvttttttttttttttttttttttttttttttt
HT170	aagcagtggtatcaacgcagagttgctCGATCGGATCATCCvttttttttttttttttttttttttttttttt
HT171	aagcagtggtatcaacgcagagttgctCGTCAAGCATCTCCvttttttttttttttttttttttttttttttt
HT172	aagcagtggtatcaacgcagagttgctGGATTATTGGCGCCvttttttttttttttttttttttttttttttt
HT173	aagcagtggtatcaacgcagagttgctGCGAGGTCGTATTCvttttttttttttttttttttttttttttttt
HT174	aagcagtggtatcaacgcagagttgctCTGGCGGAACATACvttttttttttttttttttttttttttttttt
HT175	aagcagtggtatcaacgcagagttgctCATGGCCAATAGCCvttttttttttttttttttttttttttttttt
HT176	aagcagtggtatcaacgcagagttgctTCTCGAGGATATTCvttttttttttttttttttttttttttttttt
HT177	aagcagtggtatcaacgcagagttgctCTGTGGACCATGTCvttttttttttttttttttttttttttttttt
HT178	aagcagtggtatcaacgcagagttgctTAACTTAAGACCGCvttttttttttttttttttttttttttttttt
HT179	aagcagtggtatcaacgcagagttgctGTTGCCTTATGGCCvttttttttttttttttttttttttttttttt
HT180	aagcagtggtatcaacgcagagttgctGTGATCAATGCCTCvttttttttttttttttttttttttttttttt
HT181	aagcagtggtatcaacgcagagttgctCAAGGATCTGCGAAvttttttttttttttttttttttttttttttt
HT182	aagcagtggtatcaacgcagagttgctGCTTCACACTCCAAvttttttttttttttttttttttttttttttt
HT183	aagcagtggtatcaacgcagagttgctAGGAAGAACCTTGAvttttttttttttttttttttttttttttttt
HT184	aagcagtggtatcaacgcagagttgctCTCGCTAACAAGAAvttttttttttttttttttttttttttttttt
HT185	aagcagtggtatcaacgcagagttgctTCTGGACTCTGTTAvttttttttttttttttttttttttttttttt
HT186	aagcagtggtatcaacgcagagttgctTATTGAGAGCGGAAvttttttttttttttttttttttttttttttt
HT187	aagcagtggtatcaacgcagagttgctAACTCTCCGGAGAAvttttttttttttttttttttttttttttttt
HT188	aagcagtggtatcaacgcagagttgctGAACAGTGGACCAAvttttttttttttttttttttttttttttttt
HT189	aagcagtggtatcaacgcagagttgctAGGTGTAATGTGGAvttttttttttttttttttttttttttttttt
HT190	aagcagtggtatcaacgcagagttgctGCGCTTGTCGACAAvttttttttttttttttttttttttttttttt
HT191	aagcagtggtatcaacgcagagttgctGCTCCGTCACACAAvttttttttttttttttttttttttttttttt
HT192	aagcagtggtatcaacgcagagttgctGTCATTGGTCCGAAvttttttttttttttttttttttttttttttt

(Continued on next page)

Barcode Name	Sequence
HT193	aagcagtggtatcaacgcagagttgctAATATCTTCGGCCAvttttttttttttttttttttttttttttttt
HT194	aagcagtggtatcaacgcagagttgctAGTGGCGTTAACAAvttttttttttttttttttttttttttttttt
HT195	aagcagtggtatcaacgcagagttgctGTGTCCGGCATTGAvttttttttttttttttttttttttttttttt
HT196	aagcagtggtatcaacgcagagttgctCAACCGGAGAAGAAvttttttttttttttttttttttttttttttt
HT197	aagcagtggtatcaacgcagagttgctCTCAGGACTTAGTAvttttttttttttttttttttttttttttttt
HT198	aagcagtggtatcaacgcagagttgctGCTTAGACTAGCTAvttttttttttttttttttttttttttttttt
HT199	aagcagtggtatcaacgcagagttgctGCAGATTAGTGTCAvttttttttttttttttttttttttttttttt
HT200	aagcagtggtatcaacgcagagttgctGGTTGGTGGCTTAAvttttttttttttttttttttttttttttttt
HT201	aagcagtggtatcaacgcagagttgctCTAAGAACACCGAAvttttttttttttttttttttttttttttttt
HT202	aagcagtggtatcaacgcagagttgctTTAGACGTCTACGAvttttttttttttttttttttttttttttttt
HT203	aagcagtggtatcaacgcagagttgctGGACCTCGTGAGAAvttttttttttttttttttttttttttttttt
HT204	aagcagtggtatcaacgcagagttgctGGAAGTTGAGGTAAvttttttttttttttttttttttttttttttt
HT205	aagcagtggtatcaacgcagagttgctACATTGTCCACCAAvttttttttttttttttttttttttttttttt
HT206	aagcagtggtatcaacgcagagttgctATCGCGCCAACCAAvttttttttttttttttttttttttttttttt
HT207	aagcagtggtatcaacgcagagttgctGCGAAGTCCGAGAAvttttttttttttttttttttttttttttttt
HT208	aagcagtggtatcaacgcagagttgctATTAGGAGCTCGAAvttttttttttttttttttttttttttttttt
HT209	aagcagtggtatcaacgcagagttgctGCGTGACGCAGTAAvttttttttttttttttttttttttttttttt
HT210	aagcagtggtatcaacgcagagttgctGGATCCTTAGACAAvttttttttttttttttttttttttttttttt
HT211	aagcagtggtatcaacgcagagttgctCATCAATGCGGCAAvttttttttttttttttttttttttttttttt
HT212	aagcagtggtatcaacgcagagttgctAAGGCCACTAACAAvttttttttttttttttttttttttttttttt
HT213	aagcagtggtatcaacgcagagttgctGGTACCGAAGAGAAvttttttttttttttttttttttttttttttt
HT214	aagcagtggtatcaacgcagagttgctCGATCAGGTCGGAAvttttttttttttttttttttttttttttttt
HT215	aagcagtggtatcaacgcagagttgctACAACATGCTCCTAvttttttttttttttttttttttttttttttt
HT216	aagcagtggtatcaacgcagagttgctGAGACAAGCGATAAvttttttttttttttttttttttttttttttt
HT217	aagcagtggtatcaacgcagagttgctGAAGCATAGGAGAAvttttttttttttttttttttttttttttttt
HT218	aagcagtggtatcaacgcagagttgctCACCTCCATGTGTAvttttttttttttttttttttttttttttttt
HT219	aagcagtggtatcaacgcagagttgctCGCATGCTATAGAAvttttttttttttttttttttttttttttttt
HT220	aagcagtggtatcaacgcagagttgctTGGTCGATCTTGAAvttttttttttttttttttttttttttttttt
HT221	aagcagtggtatcaacgcagagttgctTAGATCTCGCAGAAvttttttttttttttttttttttttttttttt
HT222	aagcagtggtatcaacgcagagttgctTAGTGAGCTAGTCAvttttttttttttttttttttttttttttttt
HT223	aagcagtggtatcaacgcagagttgctCCTGTGAGCCGTTAvttttttttttttttttttttttttttttttt
HT224	aagcagtggtatcaacgcagagttgctCGACGACTCAGGAAvttttttttttttttttttttttttttttttt
HT225	aagcagtggtatcaacgcagagttgctATCTTGACGGCGAAvttttttttttttttttttttttttttttttt
HT226	aagcagtggtatcaacgcagagttgctCTCTTGGTGTTCAAvttttttttttttttttttttttttttttttt
HT227	aagcagtggtatcaacgcagagttgctTCGGTACTAAGGAAvttttttttttttttttttttttttttttttt
HT228	aagcagtggtatcaacgcagagttgctACCTGGTGATACAAvttttttttttttttttttttttttttttttt
HT229	aagcagtggtatcaacgcagagttgctATGAGCGCATATCAvttttttttttttttttttttttttttttttt
HT230	aagcagtggtatcaacgcagagttgctCTATACAGTGCGTAvttttttttttttttttttttttttttttttt
HT231	aagcagtggtatcaacgcagagttgctGAGAGAAGGTTGTAvttttttttttttttttttttttttttttttt
HT232	aagcagtggtatcaacgcagagttgctCTGGAGAGGTACAAvttttttttttttttttttttttttttttttt
HT233	aagcagtggtatcaacgcagagttgctTTATACGATCGCCAvttttttttttttttttttttttttttttttt
HT234	aagcagtggtatcaacgcagagttgctCTCGAATCCTACAAvttttttttttttttttttttttttttttttt
HT235	aagcagtggtatcaacgcagagttgctAAGAGGTCGGCTAAvttttttttttttttttttttttttttttttt
HT236	aagcagtggtatcaacgcagagttgctAACAAGCTCCGCAAvttttttttttttttttttttttttttttttt
HT237	aagcagtggtatcaacgcagagttgctCAGGTCCTGAAGAAvttttttttttttttttttttttttttttttt
HT238	aagcagtggtatcaacgcagagttgctATTCTTGCTCCTGAvttttttttttttttttttttttttttttttt
HT239	aagcagtggtatcaacgcagagttgctATGAAGGTCTGTGAvttttttttttttttttttttttttttttttt
HT240	aagcagtggtatcaacgcagagttgctTTGTCGGCCGAGAAvttttttttttttttttttttttttttttttt
HT241	aagcagtggtatcaacgcagagttgctGGCTTAGCGGAGAAvttttttttttttttttttttttttttttttt
HT242	aagcagtggtatcaacgcagagttgctTGGACGTCGATTAAvttttttttttttttttttttttttttttttt
HT243	aagcagtggtatcaacgcagagttgctGTGGTAGCCTAGAAvttttttttttttttttttttttttttttttt
HT244	aagcagtggtatcaacgcagagttgctTCTTCGTATCGCAAvttttttttttttttttttttttttttttttt
HT245	aagcagtggtatcaacgcagagttgctAAGTGGTGTTCCTAvttttttttttttttttttttttttttttttt
HT246	aagcagtggtatcaacgcagagttgctGGAAGGTCATACAAvttttttttttttttttttttttttttttttt
HT247	aagcagtggtatcaacgcagagttgctGATTCGGTCCGCTAvttttttttttttttttttttttttttttttt
HT248	aagcagtggtatcaacgcagagttgctACACGTATGTGTGAvttttttttttttttttttttttttttttttt
HT249	aagcagtggtatcaacgcagagttgctGCAGTTGCTAGTAAvttttttttttttttttttttttttttttttt
HT250	aagcagtggtatcaacgcagagttgctGTGTTGTCGCGGAAvttttttttttttttttttttttttttttttt
HT251	aagcagtggtatcaacgcagagttgctCGCTGGTAATGTAAvttttttttttttttttttttttttttttttt
HT252	aagcagtggtatcaacgcagagttgctCAGCTGTGAATCAAvttttttttttttttttttttttttttttttt
HT253	aagcagtggtatcaacgcagagttgctTCGCGTTGCCAGAAvttttttttttttttttttttttttttttttt
HT254	aagcagtggtatcaacgcagagttgctTGCACTCTACCTTAvttttttttttttttttttttttttttttttt
HT255	aagcagtggtatcaacgcagagttgctCTGAGCTTGTTCAAvttttttttttttttttttttttttttttttt
HT256	aagcagtggtatcaacgcagagttgctTATCGGACGTACAAvttttttttttttttttttttttttttttttt

(Continued on next page)

Barcode Name	Sequence
HT257	aagcagtggtatcaacgcagagttgctCAGCCAATAAGGAAvttttttttttttttttttttttttttttttt
HT258	aagcagtggtatcaacgcagagttgctGGTTCTATAGCCTAvttttttttttttttttttttttttttttttt
HT259	aagcagtggtatcaacgcagagttgctCTTACTGGTGCCAAvttttttttttttttttttttttttttttttt
HT260	aagcagtggtatcaacgcagagttgctACATCTTGAGCTGAvttttttttttttttttttttttttttttttt
HT261	aagcagtggtatcaacgcagagttgctGATCTAACCGAGAAvttttttttttttttttttttttttttttttt
HT262	aagcagtggtatcaacgcagagttgctGAAGCCTAATCCAAvttttttttttttttttttttttttttttttt
HT263	aagcagtggtatcaacgcagagttgctAGTAACGGTTGGTAvttttttttttttttttttttttttttttttt
HT264	aagcagtggtatcaacgcagagttgctCCTCTTCCTATCAAvttttttttttttttttttttttttttttt
HT265	aagcagtggtatcaacgcagagttgctGAGAGGCGACACAAvttttttttttttttttttttttttttttttt
HT266	aagcagtggtatcaacgcagagttgctCGAATCGCAAGGAAvttttttttttttttttttttttttttttttt
HT267	aagcagtggtatcaacgcagagttgctTGACTCGTTGGTTAvttttttttttttttttttttttttttttttt
HT268	aagcagtggtatcaacgcagagttgctTCCTAGGAGAAGAAvttttttttttttttttttttttttttttttt
HT269	aagcagtggtatcaacgcagagttgctTCGTGATGTGTGTAvttttttttttttttttttttttttttttttt
HT270	aagcagtggtatcaacgcagagttgctTCTGTGCACCAGAAvttttttttttttttttttttttttttttttt
HT271	aagcagtggtatcaacgcagagttgctGTCTTACGCCGCTAvttttttttttttttttttttttttttttttt
HT272	aagcagtggtatcaacgcagagttgctCTTGTACTTCCAGAvttttttttttttttttttttttttttttt
HT273	aagcagtggtatcaacgcagagttgctATAGTGGTCGTAGAvttttttttttttttttttttttttttttttt
HT274	aagcagtggtatcaacgcagagttgctGCACAACTTGTAGAvttttttttttttttttttttttttttttttt
HT275	aagcagtggtatcaacgcagagttgctAAGAGCCTTGTAGAvttttttttttttttttttttttttttttttt
HT276	aagcagtggtatcaacgcagagttgctGACCTGTCATTAGAvttttttttttttttttttttttttttttttt
HT277	aagcagtggtatcaacgcagagttgctCTGTCGTAATGACAvttttttttttttttttttttttttttttttt
HT278	aagcagtggtatcaacgcagagttgctTATCTCGGTACACAvttttttttttttttttttttttttttttttt
HT279	aagcagtggtatcaacgcagagttgctTGATTGCTTGCACAvttttttttttttttttttttttttttttttt
HT280	aagcagtggtatcaacgcagagttgctATTACCTCTCCACAvttttttttttttttttttttttttttttt
HT281	aagcagtggtatcaacgcagagttgctGGTATCCAATCACAvttttttttttttttttttttttttttttttt
HT282	aagcagtggtatcaacgcagagttgctAGACGAGACTCACAvttttttttttttttttttttttttttttttt
HT283	aagcagtggtatcaacgcagagttgctACCTTCTTCTCACAvttttttttttttttttttttttttttttt
HT284	aagcagtggtatcaacgcagagttgctGCATCAAGAGTACAvttttttttttttttttttttttttttttttt
HT285	aagcagtggtatcaacgcagagttgctTAGTCCTTCCTACAvttttttttttttttttttttttttttttt
HT286	aagcagtggtatcaacgcagagttgctCTTGGCCTATTACAvttttttttttttttttttttttttttttttt
HT287	aagcagtggtatcaacgcagagttgctCGCGACAAGTTACAvttttttttttttttttttttttttttttttt
HT288	aagcagtggtatcaacgcagagttgctGTGCTTGAGTTACAvttttttttttttttttttttttttttttttt
HT289	aagcagtggtatcaacgcagagttgctAACAGACCGTTACAvttttttttttttttttttttttttttttt
HT290	aagcagtggtatcaacgcagagttgctGTGCCTCACCGATAvttttttttttttttttttttttttttttttt
HT291	aagcagtggtatcaacgcagagttgctTTGCGCCTCACATAvttttttttttttttttttttttttttttttt
HT292	aagcagtggtatcaacgcagagttgctACCGCGAAGATATAvttttttttttttttttttttttttttttttt
HT293	aagcagtggtatcaacgcagagttgctCAGGACCGTATATAvttttttttttttttttttttttttttttttt
HT294	aagcagtggtatcaacgcagagttgctTTATCTAGACGAGGvttttttttttttttttttttttttttttttt
HT295	aagcagtggtatcaacgcagagttgctTAACCGTAATGAGGvttttttttttttttttttttttttttttttt
HT296	aagcagtggtatcaacgcagagttgctACTAAGCTGACAGGvttttttttttttttttttttttttttttttt
HT297	aagcagtggtatcaacgcagagttgctGTTATGCTACCAGGvttttttttttttttttttttttttttttttt
HT298	aagcagtggtatcaacgcagagttgctCCATGAGTAATAGGvttttttttttttttttttttttttttttttt
HT299	aagcagtggtatcaacgcagagttgctACGGTTCAGATAGGvttttttttttttttttttttttttttttttt
HT300	aagcagtggtatcaacgcagagttgctTGACGGAGGATAGGvttttttttttttttttttttttttttttttt
HT301	aagcagtggtatcaacgcagagttgctACTATCGCGCTAGGvttttttttttttttttttttttttttttttt
HT302	aagcagtggtatcaacgcagagttgctATTGGAGCATTAGGvttttttttttttttttttttttttttttttt
HT303	aagcagtggtatcaacgcagagttgctATGCAGAGCTGACGvttttttttttttttttttttttttttttttt
HT304	aagcagtggtatcaacgcagagttgctATTGTATCACCACGvttttttttttttttttttttttttttttttt
HT305	aagcagtggtatcaacgcagagttgctGTAAGTCCTATACGvttttttttttttttttttttttttttttttt
HT306	aagcagtggtatcaacgcagagttgctTTCATGACCAGATGvttttttttttttttttttttttttttttttt
HT307	aagcagtggtatcaacgcagagttgctACCGACTGATTATGvttttttttttttttttttttttttttttttt
HT308	aagcagtggtatcaacgcagagttgctAATAACCTAGCAGCvttttttttttttttttttttttttttttttt
HT309	aagcagtggtatcaacgcagagttgctTAAGACAAGACACCvttttttttttttttttttttttttttttttt
HT310	aagcagtggtatcaacgcagagttgctGTATTCTTACCACCvttttttttttttttttttttttttttttttt
HT311	aagcagtggtatcaacgcagagttgctAGACTTCAGATACCvttttttttttttttttttttttttttttttt
HT312	aagcagtggtatcaacgcagagttgctCGAACGCTCACATCvttttttttttttttttttttttttttttttt
HT313	aagcagtggtatcaacgcagagttgctGACAGGAGAGAGATvttttttttttttttttttttttttttttttt
HT314	aagcagtggtatcaacgcagagttgctCCAGTCCAACAGATvttttttttttttttttttttttttttttttt
HT315	aagcagtggtatcaacgcagagttgctCAGAATGGACAGATvttttttttttttttttttttttttttttttt
HT316	aagcagtggtatcaacgcagagttgctGTTCGCAGGCAGATvttttttttttttttttttttttttttttttt
HT317	aagcagtggtatcaacgcagagttgctCATGCGAGTCGGATvttttttttttttttttttttttttttttttt
HT318	aagcagtggtatcaacgcagagttgctGGAACGAGGACGATvttttttttttttttttttttttttttttttt
HT319	aagcagtggtatcaacgcagagttgctGTCAGTTCGATGATvttttttttttttttttttttttttttttttt
HT320	aagcagtggtatcaacgcagagttgctGGTCGGTCCATCATvttttttttttttttttttttttttttttttt

(Continued on next page)

Barcode Name	Sequence
HT321	aagcagtggtatcaacgcagagttgctTCGCTGAACCATATvttttttttttttttttttttttttttttt
HT322	aagcagtggtatcaacgcagagttgctCGCCGGCACTATATvttttttttttttttttttttttttttttt
HT323	aagcagtggtatcaacgcagagttgctGGCACCGCGATTATvttttttttttttttttttttttttttttt
HT324	aagcagtggtatcaacgcagagttgctCTGCATCAGGTTATvttttttttttttttttttttttttttttt
HT325	aagcagtggtatcaacgcagagttgctCGGAAGCCACTTATvttttttttttttttttttttttttttttt
HT326	aagcagtggtatcaacgcagagttgctAGTCCACTAGGCGTvttttttttttttttttttttttttttttt
HT327	aagcagtggtatcaacgcagagttgctTTGGTCTGGTGCGTvttttttttttttttttttttttttttttt
HT328	aagcagtggtatcaacgcagagttgctCTCACCGTGATCGTvttttttttttttttttttttttttttttt
HT329	aagcagtggtatcaacgcagagttgctTAACAAGTGCTCGTvttttttttttttttttttttttttttttt
HT330	aagcagtggtatcaacgcagagttgctGGATTCCAGTTCGTvttttttttttttttttttttttttttttt
HT331	aagcagtggtatcaacgcagagttgctAGTTGATGAGGTGTvttttttttttttttttttttttttttttt
HT332	aagcagtggtatcaacgcagagttgctGCTTGAAGGCGTGTvttttttttttttttttttttttttttttt
HT333	aagcagtggtatcaacgcagagttgctGCCATAGAACCTGTvttttttttttttttttttttttttttttt
HT334	aagcagtggtatcaacgcagagttgctCGAATCTTACCTGTvttttttttttttttttttttttttttttt
HT335	aagcagtggtatcaacgcagagttgctCTGCCAGAAGTTGTvttttttttttttttttttttttttttttt
HT336	aagcagtggtatcaacgcagagttgctGGTCCAACCGTTGTvttttttttttttttttttttttttttttt
HT337	aagcagtggtatcaacgcagagttgctCATTCCTCTGTTGTvttttttttttttttttttttttttttttt
HT338	aagcagtggtatcaacgcagagttgctGCTCCTCGACTTGTvttttttttttttttttttttttttttttt
HT339	aagcagtggtatcaacgcagagttgctCACCAACACCTTGTvttttttttttttttttttttttttttttt
HT340	aagcagtggtatcaacgcagagttgctACACCGGTTCTTGTvttttttttttttttttttttttttttttt
HT341	aagcagtggtatcaacgcagagttgctACAGCAGGCAGACTvttttttttttttttttttttttttttttt
HT342	aagcagtggtatcaacgcagagttgctATGGCGTGATCGCTvttttttttttttttttttttttttttttt
HT343	aagcagtggtatcaacgcagagttgctGTCGTCGTAGTGCTvttttttttttttttttttttttttttttt
HT344	aagcagtggtatcaacgcagagttgctGTGTCGTCACACCTvttttttttttttttttttttttttttttt
HT345	aagcagtggtatcaacgcagagttgctCTTGACGAACGCCTvttttttttttttttttttttttttttttt
HT346	aagcagtggtatcaacgcagagttgctTCTTGAAGCCTCCTvttttttttttttttttttttttttttttt
HT347	aagcagtggtatcaacgcagagttgctCACATAACCAGTCTvttttttttttttttttttttttttttttt
HT348	aagcagtggtatcaacgcagagttgctGAATTGCGGCGTCTvttttttttttttttttttttttttttttt
HT349	aagcagtggtatcaacgcagagttgctGACGTTCAGCCTCTvttttttttttttttttttttttttttttt
HT350	aagcagtggtatcaacgcagagttgctTTACCGTGGCCTCTvttttttttttttttttttttttttttttt
HT351	aagcagtggtatcaacgcagagttgctGGCATACACGTTCTvttttttttttttttttttttttttttttt
HT352	aagcagtggtatcaacgcagagttgctCGTATCGTTGTTCTvttttttttttttttttttttttttttttt
HT353	aagcagtggtatcaacgcagagttgctCCATGGCCACTTCTvttttttttttttttttttttttttttttt
HT354	aagcagtggtatcaacgcagagttgctCTTCGGTCTCTTCTvttttttttttttttttttttttttttttt
HT355	aagcagtggtatcaacgcagagttgctAGTTGTGCCACATTvttttttttttttttttttttttttttttt
HT356	aagcagtggtatcaacgcagagttgctCCGGCGCCAATATTvttttttttttttttttttttttttttttt
HT357	aagcagtggtatcaacgcagagttgctGGCGAGAGGTTATTvttttttttttttttttttttttttttttt
HT358	aagcagtggtatcaacgcagagttgctCCACCGAAGGAGTTvttttttttttttttttttttttttttttt
HT359	aagcagtggtatcaacgcagagttgctGCCTTCCAATAGTTvttttttttttttttttttttttttttttt
HT360	aagcagtggtatcaacgcagagttgctTTGCCGTACCGGTTvttttttttttttttttttttttttttttt
HT361	aagcagtggtatcaacgcagagttgctGGTAGCCTGTGGTTvttttttttttttttttttttttttttttt
HT362	aagcagtggtatcaacgcagagttgctGCTCGTCATGCGTTvttttttttttttttttttttttttttttt
HT363	aagcagtggtatcaacgcagagttgctGTATTGTAGCCGTTvttttttttttttttttttttttttttttt
HT364	aagcagtggtatcaacgcagagttgctTGGTTCATGTCGTTvttttttttttttttttttttttttttttt
HT365	aagcagtggtatcaacgcagagttgctAGGAGGTGCATGTTvttttttttttttttttttttttttttttt
HT366	aagcagtggtatcaacgcagagttgctAAGGAGCAGGTGTTvttttttttttttttttttttttttttttt
HT367	aagcagtggtatcaacgcagagttgctGAGAACCGTGTGTTvttttttttttttttttttttttttttttt
HT368	aagcagtggtatcaacgcagagttgctCCGGATGCACTGTTvttttttttttttttttttttttttttttt
HT369	aagcagtggtatcaacgcagagttgctGGACGGAGACACTTvttttttttttttttttttttttttttttt
HT370	aagcagtggtatcaacgcagagttgctCCTCACCTACGCTTvttttttttttttttttttttttttttttt
HT371	aagcagtggtatcaacgcagagttgctATTGTGGCGCGCTTvttttttttttttttttttttttttttttt
HT372	aagcagtggtatcaacgcagagttgctAAGTCTGTCTGCTTvttttttttttttttttttttttttttttt
HT373	aagcagtggtatcaacgcagagttgctGGAGTTAACACCTTvttttttttttttttttttttttttttttt
HT374	aagcagtggtatcaacgcagagttgctGCGAGATGTGCCTTvttttttttttttttttttttttttttttt
HT375	aagcagtggtatcaacgcagagttgctTCGAGCACATCCTTvttttttttttttttttttttttttttttt
HT376	aagcagtggtatcaacgcagagttgctGCATCCTTATCCTTvttttttttttttttttttttttttttttt
HT377	aagcagtggtatcaacgcagagttgctCATCCGGAGATCTTvttttttttttttttttttttttttttttt
HT378	aagcagtggtatcaacgcagagttgctGTTACACCGATCTTvttttttttttttttttttttttttttttt
HT379	aagcagtggtatcaacgcagagttgctAGTTGACGCATCTTvttttttttttttttttttttttttttttt
HT380	aagcagtggtatcaacgcagagttgctCTTGTCACAGTCTTvttttttttttttttttttttttttttttt
HT381	aagcagtggtatcaacgcagagttgctGCTGTCGAGGTCTTvttttttttttttttttttttttttttttt
HT382	aagcagtggtatcaacgcagagttgctCTCCACTGACTCTTvttttttttttttttttttttttttttttt
HT383	aagcagtggtatcaacgcagagttgctGAACATTGGCTCTTvttttttttttttttttttttttttttttt
HT384	aagcagtggtatcaacgcagagttgctGGCGAGTACCTCTTvttttttttttttttttttttttttttttt

Table A.15: Targeted pre-amplification genotyping primers. Sequences of targeted primers used for pre-amplification of gDNA and mRNA/cDNA amplicons during RT-PCR steps for TARGET-seq+ experiments. mRNA/cDNA primers were added in the RT and PCR steps, gDNA primers in the PCR step only.

Sample	Gene	Exon	Target	Amplicon length (bp)	Forward primer sequence	Reverse primer sequence
NOC108	DNMT3A	8-10	mRNA	599	GGGACAAGAATGCCACCAAAG	CTCCTTGACCTTGGGCTTCTC
NOC108	DNMT3A	9	gDNA	468	TCTGAGTAGTGAGGGTGGCA	TGGACTCCTCTTTGCATCGG
NOC132	DNMT3A	18-20	mRNA	529	GTGGACCGCTACATTGCCTC	TGGTCCTCACTTTGCTGAACT
NOC132	DNMT3A	18-19	gDNA	743	GGTTCTTTCCGAGGTAGGCC	CTGGGTCTCCTCTCTTTCGTG
NOC171	DNMT3A	14-16	mRNA	491	TGGGAGCCTCAATGTTACCC	TGTAGCGGTCCACCTGAATG
NOC171	DNMT3A	15-16	gDNA	546	GTTTTGCCAGAGTTGCCCAC	TCTGCTCACTGGGTCTCCTT
NOC131	DNMT3A	14-16	mRNA	491	TGGGAGCCTCAATGTTACCC	TGTAGCGGTCCACCTGAATG
NOC131	DNMT3A	18-20	mRNA	529	GTGGACCGCTACATTGCCTC	TGGTCCTCACTTTGCTGAACT
NOC131	DNMT3A	15-16	gDNA	546	GTTTTGCCAGAGTTGCCCAC	TCTGCTCACTGGGTCTCCTT
NOC131	DNMT3A	20	gDNA	430	GCTCGGGACACATCACCAAT	CTTGGCTCATCTTCAAACCGTC
NOC062	DNMT3A	23	mRNA	408	ATGATGTCCAACCCTTTTCGC	GGGCAAAGACCAGCATTTTCC
NOC062	SF3B1	15	mRNA	188	TGTGCCATCTTGCCACATCT	CTGTGTTGGCGGATACCCTT
NOC062	DNMT3A	23	gDNA	411	ATGATGTCCAACCCTTTTCGC	TGTGTGGTTAGACGGCTTCC
NOC062	BCORL1	3	gDNA	521	GAACCTGCCTAAGATGCCTGA	GGGTCACAGGAGGTTGCTAT
NOC062	SF3B1	15	gDNA	296	ACAGGCTATGGTTCATGTTTTGC	CTGTGTTGGCGGATACCCTT
NOC031	TET2	3	gDNA	763	ACCAGGTGCACTTCTCCAAAA	GCAGCTTGTTGACCAGACAT
NOC117	TET2	5	gDNA	485	CCAGTTTGCTTGGCGTAGAC	GGCACTAGAATGGACCTCTTC
NOC117	TET2	3	gDNA	765	GCAGGAGGGGAAAAGTGCTAAT	TTGGGGTAGTGGTGGTGTTC
NOC002	TET2	6	mRNA	389	AGGCAAAAGTTCTCAGGGATGT	GGTCATCCCCAAGCAGCTTA
NOC002	ATM	43	mRNA	327	ATCCCCTCATCAACACGCC	AAGGCTGCGCTTACACATCT
NOC002	ATM	43	gDNA	462	ACAAAATAACTCCTGTTTAGGCCT	TCAAGCCTAGGTGCATTAACTG
NOC002	TET2	6	gDNA	865	TACCCAGATCAAAGACTCATGC	AAGTGCACGCTGAACTCTCTTC
NOC002	TET2	3	gDNA	807	CCAAGGTACAGTGGACCAACA	GCAGCTTGTTGACCAGACAT
NOC115	DNMT3A	17-20	mRNA	593	GGTGCTGTCTCTCTTTGATGGA	TGGTCCTCACTTTGCTGAACT
NOC115	TET2	6	mRNA	389	AGGCAAAAGTTCTCAGGGATGT	GGTCATCCCCAAGCAGCTTA
NOC115	DNMT3A	17	gDNA	365	ATTTGCCCTTTACCCTCTCAAGA	AGATGGCTCCAAGTAACGGTG
NOC115	TET2	6	gDNA	865	TACCCAGATCAAAGACTCATGC	AAGTGCACGCTGAACTCTCTTC
NOC137	TET2	6	mRNA	389	AGGCAAAAGTTCTCAGGGATGT	GGTCATCCCCAAGCAGCTTA
NOC137	TET2	6	gDNA	865	TACCCAGATCAAAGACTCATGC	AAGTGCACGCTGAACTCTCTTC
NOC137	ASXL1	13	gDNA	613	CACACCTGAATCCTCACCGA	ACGCATGTCACCATTCACCT

Table A.16: Genotyping barcoding primers.
Sequences of primers used for the first barcoding PCR (genotyping PCR1) step in single-cell genotyping. Primers are designed to anneal nested within the pre-amplified gDNA and cDNA amplicons. The CS1/CS2 adapter sequence is at the 5' end of each primer, lower case bases indicate the plate barcode, and the target-specific sequence is at the 3' end of each primer.

Sample	Gene	Exon	Target	Insert length without adaptors (bp)	Forward primer sequence	Reverse primer sequence
NOC108	DNMT3A	8-10	cDNA	205	ACACTGACGACATGGTTCTACAaactcgTGGGTCATGTGGTTCGGAGA	TACGGTAGCAGAGACTTGGTCTaagaccTGTCACTCTCATCGCTGTCG
					ACACTGACGACATGGTTCTACAcctcaaTGGGTCATGTGGTTCGGAGA	TACGGTAGCAGAGACTTGGTCTccagttTGTCACTCTCATCGCTGTCG
					ACACTGACGACATGGTTCTACAtgactgTGGGTCATGTGGTTCGGAGA	TACGGTAGCAGAGACTTGGTCTtggcaaTGTCACTCTCATCGCTGTCG
					ACACTGACGACATGGTTCTACAgttggaTGGGTCATGTGGTTCGGAGA	TACGGTAGCAGAGACTTGGTCTgaacagTGTCACTCTCATCGCTGTCG
					ACACTGACGACATGGTTCTACAcaatccTGGGTCATGTGGTTCGGAGA	TACGGTAGCAGAGACTTGGTCTacttccTGTCACTCTCATCGCTGTCG
					ACACTGACGACATGGTTCTACAagtacgTGGGTCATGTGGTTCGGAGA	TACGGTAGCAGAGACTTGGTCTtccttgTGTCACTCTCATCGCTGTCG
NOC108	DNMT3A	9	gDNA	209	ACACTGACGACATGGTTCTACAaactcgCTTCCAGGCCTCCTAGTGCT	TACGGTAGCAGAGACTTGGTCTaagaccCAAGTGTAAGCCTCGGCAAA
					ACACTGACGACATGGTTCTACAcctcaaCTTCCAGGCCTCCTAGTGCT	TACGGTAGCAGAGACTTGGTCTccagttCAAGTGTAAGCCTCGGCAAA
					ACACTGACGACATGGTTCTACAtgactgCTTCCAGGCCTCCTAGTGCT	TACGGTAGCAGAGACTTGGTCTtggcaaCAAGTGTAAGCCTCGGCAAA
					ACACTGACGACATGGTTCTACAgttggaCTTCCAGGCCTCCTAGTGCT	TACGGTAGCAGAGACTTGGTCTgaacagCAAGTGTAAGCCTCGGCAAA
					ACACTGACGACATGGTTCTACAcaatccCTTCCAGGCCTCCTAGTGCT	TACGGTAGCAGAGACTTGGTCTacttccCAAGTGTAAGCCTCGGCAAA
					ACACTGACGACATGGTTCTACAagtacgCTTCCAGGCCTCCTAGTGCT	TACGGTAGCAGAGACTTGGTCTtccttgCAAGTGTAAGCCTCGGCAAA
NOC132	DNMT3A	18-19	cDNA	307	ACACTGACGACATGGTTCTACAacatggAGATCATGTACGTCGGGGAC	TACGGTAGCAGAGACTTGGTCTctgtgaGCATCAATCATCACAGGGTTGG
					ACACTGACGACATGGTTCTACAggattcAGATCATGTACGTCGGGGAC	TACGGTAGCAGAGACTTGGTCTatgcagGCATCAATCATCACAGGGTTGG
					ACACTGACGACATGGTTCTACAaagaccAGATCATGTACGTCGGGGAC	TACGGTAGCAGAGACTTGGTCTaactcgGCATCAATCATCACAGGGTTGG
NOC132	DNMT3A	18-19	gDNA	489	ACACTGACGACATGGTTCTACAacatggCCATTAGTGAGCTGGCCAAAC	TACGGTAGCAGAGACTTGGTCTctgtgaACGTTGCCTTTATCCTCCCA
					ACACTGACGACATGGTTCTACAggattcCCATTAGTGAGCTGGCCAAAC	TACGGTAGCAGAGACTTGGTCTatgcagACGTTGCCTTTATCCTCCCA
					ACACTGACGACATGGTTCTACAaagaccCCATTAGTGAGCTGGCCAAAC	TACGGTAGCAGAGACTTGGTCTaactcgACGTTGCCTTTATCCTCCCA
NOC171	DNMT3A	15	cDNA	250	ACACTGACGACATGGTTCTACAcaatccCGGAAACAACAACTGCTGCA	TACGGTAGCAGAGACTTGGTCTacttccCTTCCTCTTCTCAGCTGGGAC
					ACACTGACGACATGGTTCTACAagtacgCGGAAACAACAACTGCTGCA	TACGGTAGCAGAGACTTGGTCTtccttgCTTCCTCTTCTCAGCTGGGAC
NOC171	DNMT3A	15	gDNA	353	ACACTGACGACATGGTTCTACAcaatccACAACAGGTCAGATGCAGGC	TACGGTAGCAGAGACTTGGTCTacttccCTCCTTCCAGGTGCTTTTGC
					ACACTGACGACATGGTTCTACAagtacgACAACAGGTCAGATGCAGGC	TACGGTAGCAGAGACTTGGTCTtccttgCTCCTTCCAGGTGCTTTTGC
NOC131	DNMT3A	15	cDNA	250	ACACTGACGACATGGTTCTACAcctcaaCGGAAACAACAACTGCTGCA	TACGGTAGCAGAGACTTGGTCTccagttCTTCCTCTTCTCAGCTGGGAC
					ACACTGACGACATGGTTCTACAtgactgCGGAAACAACAACTGCTGCA	TACGGTAGCAGAGACTTGGTCTtggcaaCTTCCTCTTCTCAGCTGGGAC
					ACACTGACGACATGGTTCTACAgttggaCGGAAACAACAACTGCTGCA	TACGGTAGCAGAGACTTGGTCTgaacagCTTCCTCTTCTCAGCTGGGAC
NOC131	DNMT3A	20	cDNA	189	ACACTGACGACATGGTTCTACAcctcaaCGTTAGTGACAAGAGGGACATCT	TACGGTAGCAGAGACTTGGTCTccagttGCTATCCTGCCATGCTCCAG
					ACACTGACGACATGGTTCTACAtgactgCGTTAGTGACAAGAGGGACATCT	TACGGTAGCAGAGACTTGGTCTtggcaaGCTATCCTGCCATGCTCCAG
					ACACTGACGACATGGTTCTACAgttggaCGTTAGTGACAAGAGGGACATCT	TACGGTAGCAGAGACTTGGTCTgaacagGCTATCCTGCCATGCTCCAG
NOC131	DNMT3A	15	gDNA	277	ACACTGACGACATGGTTCTACAcctcaaACAACAGGTCAGATGCAGGC	TACGGTAGCAGAGACTTGGTCTccagttAGACCCCTGGAACTGCTACA
					ACACTGACGACATGGTTCTACAtgactgACAACAGGTCAGATGCAGGC	TACGGTAGCAGAGACTTGGTCTtggcaaAGACCCCTGGAACTGCTACA
					ACACTGACGACATGGTTCTACAgttggaACAACAGGTCAGATGCAGGC	TACGGTAGCAGAGACTTGGTCTgaacagAGACCCCTGGAACTGCTACA
NOC131	DNMT3A	20	gDNA	381	ACACTGACGACATGGTTCTACAcctcaaAGCACAGGACTCTTCTTTGGT	TACGGTAGCAGAGACTTGGTCTccagttTGGCTCATCTTCAAACCGTCT
					ACACTGACGACATGGTTCTACAtgactgAGCACAGGACTCTTCTTTGGT	TACGGTAGCAGAGACTTGGTCTtggcaaTGGCTCATCTTCAAACCGTCT
					ACACTGACGACATGGTTCTACAgttggaAGCACAGGACTCTTCTTTGGT	TACGGTAGCAGAGACTTGGTCTgaacagTGGCTCATCTTCAAACCGTCT
NOC062	DNMT3A	23	cDNA	198	ACACTGACGACATGGTTCTACAccagttGAGGACATCTTATGGTGCACTG	TACGGTAGCAGAGACTTGGTCTcctcaaTACCTCAGTTTGCCCCCATG
					ACACTGACGACATGGTTCTACAtggcaaGAGGACATCTTATGGTGCACTG	TACGGTAGCAGAGACTTGGTCTtgactgTACCTCAGTTTGCCCCCATG
NOC062	SF3B1	15	cDNA	182	ACACTGACGACATGGTTCTACAccagttTGCCATCTTGCCACATCTTAGA	TACGGTAGCAGAGACTTGGTCTcctcaaGTTGGCGGATACCCTTCCATA
					ACACTGACGACATGGTTCTACAtggcaaTGCCATCTTGCCACATCTTAGA	TACGGTAGCAGAGACTTGGTCTtgactgGTTGGCGGATACCCTTCCATA
NOC062	DNMT3A	23	gDNA	199	ACACTGACGACATGGTTCTACAccagttTGCCCTCTCTGCCTTTTCTC	TACGGTAGCAGAGACTTGGTCTcctcaaGTGTCGCTACCTCAGTTTGC
					ACACTGACGACATGGTTCTACAtggcaaTGCCCTCTCTGCCTTTTCTC	TACGGTAGCAGAGACTTGGTCTtgactgGTGTCGCTACCTCAGTTTGC
NOC062	BCORL1	3	gDNA	339	ACACTGACGACATGGTTCTACAccagttAAGGTGGATGGTGATGTGGTC	TACGGTAGCAGAGACTTGGTCTcctcaaCGAAGTAGACCAGCGGCATC
					ACACTGACGACATGGTTCTACAtggcaaAAGGTGGATGGTGATGTGGTC	TACGGTAGCAGAGACTTGGTCTtgactgCGAAGTAGACCAGCGGCATC
NOC062	SF3B1	15	gDNA	239	ACACTGACGACATGGTTCTACAccagttTTCTGCAGTTTGGCTGAATAGTTG	TACGGTAGCAGAGACTTGGTCTcctcaaGTTGGCGGATACCCTTCCATA
					ACACTGACGACATGGTTCTACAtggcaaTTCTGCAGTTTGGCTGAATAGTTG	TACGGTAGCAGAGACTTGGTCTtgactgGTTGGCGGATACCCTTCCATA

(Continued on next page)

Sample	Gene	Exon	Target	Insert length without adaptors (bp)	Forward primer sequence	Reverse primer sequence
NOC031	TET2	3	gDNA	208	ACACTGACGACATGGTTCTACAaggtgtAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTcgatagTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACAtcacacAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTttcggtTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACAcacgaaAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTcgtcttTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACAtagaggAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTgcgtaaTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACAgtccaaAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTttagccTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACActgtgaAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTacatggTGCTGATCATTGTTGCTTTGGG
NOC117	TET2	5	gDNA	355	ACACTGACGACATGGTTCTACAttcggtTGTGTGTGTGTTTCTGTGGGT	TACGGTAGCAGAGACTTGGTCTtcacacGGCACTAGAATGGACCTCTTC
					ACACTGACGACATGGTTCTACAcgtcttTGTGTGTGTGTTTCTGTGGGT	TACGGTAGCAGAGACTTGGTCTcacgaaGGCACTAGAATGGACCTCTTC
					ACACTGACGACATGGTTCTACAgcgtaaTGTGTGTGTGTTTCTGTGGGT	TACGGTAGCAGAGACTTGGTCTtagaggGGCACTAGAATGGACCTCTTC
					ACACTGACGACATGGTTCTACAttagccTGTGTGTGTGTTTCTGTGGGT	TACGGTAGCAGAGACTTGGTCTgtccaaGGCACTAGAATGGACCTCTTC
NOC117	TET2	3	gDNA	516	ACACTGACGACATGGTTCTACAttcggtACTTAAAAACAAGGCAGTGCTAA	TACGGTAGCAGAGACTTGGTCTtcacacAAATACCGTTCAGAGCTGCCA
					ACACTGACGACATGGTTCTACAcgtcttACTTAAAAACAAGGCAGTGCTAA	TACGGTAGCAGAGACTTGGTCTcacgaaAAATACCGTTCAGAGCTGCCA
					ACACTGACGACATGGTTCTACAgcgtaaACTTAAAAACAAGGCAGTGCTAA	TACGGTAGCAGAGACTTGGTCTtagaggAAATACCGTTCAGAGCTGCCA
					ACACTGACGACATGGTTCTACAttagccACTTAAAAACAAGGCAGTGCTAA	TACGGTAGCAGAGACTTGGTCTgtccaaAAATACCGTTCAGAGCTGCCA
NOC002	TET2	6	cDNA	307	ACACTGACGACATGGTTCTACAgtccaaGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTttagccCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAacatggGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTctgtgaCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAggattcGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTatgcagCAACCAAAAGAGAAGGAGGCAC
NOC002	ATM	43	cDNA	309	ACACTGACGACATGGTTCTACAgtccaaCGCCAGGCAGGAATCATTCA	TACGGTAGCAGAGACTTGGTCTttagccGCTGCGCTTACACATCTCTTC
					ACACTGACGACATGGTTCTACAacatggCGCCAGGCAGGAATCATTCA	TACGGTAGCAGAGACTTGGTCTctgtgaGCTGCGCTTACACATCTCTTC
					ACACTGACGACATGGTTCTACAggattcCGCCAGGCAGGAATCATTCA	TACGGTAGCAGAGACTTGGTCTatgcagGCTGCGCTTACACATCTCTTC
NOC002	ATM	43	gDNA	217	ACACTGACGACATGGTTCTACAgtccaaAACTCCTGTTTAGGCCTTGC	TACGGTAGCAGAGACTTGGTCTttagccGTTGTTTAGAATGAGGAGAGAGGC
					ACACTGACGACATGGTTCTACAacatggAACTCCTGTTTAGGCCTTGC	TACGGTAGCAGAGACTTGGTCTctgtgaGTTGTTTAGAATGAGGAGAGAGGC
					ACACTGACGACATGGTTCTACAggattcAACTCCTGTTTAGGCCTTGC	TACGGTAGCAGAGACTTGGTCTatgcagGTTGTTTAGAATGAGGAGAGAGGC
NOC002	TET2	6	gDNA	287	ACACTGACGACATGGTTCTACAgtccaaGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTttagccTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAacatggGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTctgtgaTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAggattcGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTatgcagTCCTATCAGTGGCCGCAAAG
NOC002	TET2	3	gDNA	208	ACACTGACGACATGGTTCTACAgtccaaAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTttagccTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACAacatggAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTctgtgaTGCTGATCATTGTTGCTTTGGG
					ACACTGACGACATGGTTCTACAggattcAGCACTTGAATCAACAGGCT	TACGGTAGCAGAGACTTGGTCTatgcagTGCTGATCATTGTTGCTTTGGG
NOC115	DNMT3A	17-19	cDNA	440	ACACTGACGACATGGTTCTACAtagaggGTGCTGTCTCTCTTTGATGGAATC	TACGGTAGCAGAGACTTGGTCTgcgtaaGCATCAATCATCACAGGGTTGG
					ACACTGACGACATGGTTCTACAgtccaaGTGCTGTCTCTCTTTGATGGAATC	TACGGTAGCAGAGACTTGGTCTttagccGCATCAATCATCACAGGGTTGG
					ACACTGACGACATGGTTCTACAacatggGTGCTGTCTCTCTTTGATGGAATC	TACGGTAGCAGAGACTTGGTCTctgtgaGCATCAATCATCACAGGGTTGG
					ACACTGACGACATGGTTCTACAggattcGTGCTGTCTCTCTTTGATGGAATC	TACGGTAGCAGAGACTTGGTCTatgcagGCATCAATCATCACAGGGTTGG
NOC115	TET2	6	cDNA	307	ACACTGACGACATGGTTCTACAtagaggGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTgcgtaaCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAgtccaaGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTttagccCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAacatggGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTctgtgaCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAggattcGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTatgcagCAACCAAAAGAGAAGGAGGCAC
NOC115	DNMT3A	17	gDNA	298	ACACTGACGACATGGTTCTACAtagaggCTCCAAGTAACGGTGCTGTC	TACGGTAGCAGAGACTTGGTCTgcgtaaGGGCAAAGGGTGAAGAGAAAG
					ACACTGACGACATGGTTCTACAgtccaaCTCCAAGTAACGGTGCTGTC	TACGGTAGCAGAGACTTGGTCTttagccGGGCAAAGGGTGAAGAGAAAG
					ACACTGACGACATGGTTCTACAacatggCTCCAAGTAACGGTGCTGTC	TACGGTAGCAGAGACTTGGTCTctgtgaGGGCAAAGGGTGAAGAGAAAG
					ACACTGACGACATGGTTCTACAggattcCTCCAAGTAACGGTGCTGTC	TACGGTAGCAGAGACTTGGTCTatgcagGGGCAAAGGGTGAAGAGAAAG
NOC115	TET2	6	gDNA	287	ACACTGACGACATGGTTCTACAtagaggGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTgcgtaaTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAgtccaaGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTttagccTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAacatggGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTctgtgaTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAggattcGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTatgcagTCCTATCAGTGGCCGCAAAG
NOC137	TET2	6	cDNA	307	ACACTGACGACATGGTTCTACAaggtgtGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTcgatagCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAtcacacGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTttcggtCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAcacgaaGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTcgtcttCAACCAAAAGAGAAGGAGGCAC
					ACACTGACGACATGGTTCTACAtagaggGCAAAAGTTCTCAGGGATGTCC	TACGGTAGCAGAGACTTGGTCTgcgtaaCAACCAAAAGAGAAGGAGGCAC
NOC137	TET2	6	gDNA	287	ACACTGACGACATGGTTCTACAaggtgtGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTcgatagTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAtcacacGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTttcggtTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAcacgaaGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTcgtcttTCCTATCAGTGGCCGCAAAG
					ACACTGACGACATGGTTCTACAtagaggGGTGTTTGGGATGGAATGGTG	TACGGTAGCAGAGACTTGGTCTgcgtaaTCCTATCAGTGGCCGCAAAG
NOC137	ASXL1	13	gDNA	285	ACACTGACGACATGGTTCTACAaggtgtGCATTGCCTGGGGATTTGAC	TACGGTAGCAGAGACTTGGTCTcgatagCTCATCCTTGTCCACCGAAGA
					ACACTGACGACATGGTTCTACAtcacacGCATTGCCTGGGGATTTGAC	TACGGTAGCAGAGACTTGGTCTttcggtCTCATCCTTGTCCACCGAAGA
					ACACTGACGACATGGTTCTACAcacgaaGCATTGCCTGGGGATTTGAC	TACGGTAGCAGAGACTTGGTCTcgtcttCTCATCCTTGTCCACCGAAGA
					ACACTGACGACATGGTTCTACAtagaggGCATTGCCTGGGGATTTGAC	TACGGTAGCAGAGACTTGGTCTgcgtaaCTCATCCTTGTCCACCGAAGA

References

[1] John S. Welch *et al.* "The Origin and Evolution of Mutations in Acute Myeloid Leukemia". In: *Cell* 150.2 (July 2012), pp. 264–278. URL: https://doi.org/10.1016/j.cell.2012.06.023.

[2] Isabel Beerman *et al.* "Quiescent Hematopoietic Stem Cells Accumulate DNA Damage during Aging that Is Repaired upon Entry into Cell Cycle". In: *Cell Stem Cell* 15.1 (2014), pp. 37–50. URL: https://www.sciencedirect.com/science/article/pii/S1934590914001532.

[3] Francis Blokzijl *et al.* "Tissue-specific mutation accumulation in human adult stem cells during life". In: *Nature* 538.7624 (2016), pp. 260–264. URL: https://doi.org/10.1038/nature19768.

[4] Mary Armanios *et al.* "Short Telomeres are Sufficient to Cause the Degenerative Defects Associated with Aging". In: *The American Journal of Human Genetics* 85.6 (2009), pp. 823–832. URL: https://www.sciencedirect.com/science/article/pii/S0002929709005187.

[5] Mary Armanios and Elizabeth H Blackburn. "The telomere syndromes". In: *Nature Reviews Genetics* 13.10 (2012), pp. 693–704. URL: https://doi.org/10.1038/nrg3246.

[6] Steve Horvath. "DNA methylation age of human tissues and cell types". In: *Genome Biology* 14.10 (2013), p. 3156. URL: https://doi.org/10.1186/gb-2013-14-10-r115.

[7] Douglas R Green, Lorenzo Galluzzi, and Guido Kroemer. "Mitochondria and the Autophagy–Inflammation–Cell Death Axis in Organismal Aging". In: *Science* 333.6046 (Aug. 2011), pp. 1109–1112. URL: https://doi.org/10.1126/science.1201940.

[8] Carlos López-Otín *et al.* "The Hallmarks of Aging". In: *Cell* 153.6 (2013), pp. 1194–1217. URL: https://www.sciencedirect.com/science/article/pii/S0092867413006454.

[9] Robert A.J. Signer and Sean J. Morrison. "Mechanisms that Regulate Stem Cell Aging and Life Span". In: *Cell Stem Cell* 12.2 (2013), pp. 152–165. URL: https://dx.doi.org/10.1016/j.stem.2013.01.001.

[10] Margaret A Goodell and Thomas A Rando. "Stem cells and healthy aging". In: *Science* 350.6265 (Dec. 2015), pp. 1199–1204. URL: http://science.sciencemag.org/content/350/6265/1199.abstract.

[11] Irving L. Weissman. "Stem cells are units of natural selection for tissue formation, for germline development, and in cancer development". In: *Proceedings of the National Academy of Sciences* 112.29 (July 2015), pp. 8922–8928. URL: https://doi.org/10.1073/pnas.1505464112.

[12] Walter F Bodmer and Daniel J M Crouch. "Somatic selection of poorly differentiating variant stem cell clones could be a key to human ageing". In: *Journal of Theoretical Biology* 489 (2020), p. 110153. URL: https://doi.org/10.1016/j.jtbi.2020.110153%200022-5193.

[13] E Premkumar Reddy *et al.* "A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene". In: *Nature* 300.5888 (1982), pp. 149–152. URL: https://doi.org/10.1038/300149a0.

[14] Clifford J Tabin *et al.* "Mechanism of activation of a human oncogene". In: *Nature* 300.5888 (1982), pp. 143–149. URL: https://doi.org/10.1038/300143a0.

[15] 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. URL: https://doi.org/10.1038/nature12213.

[16] Iñigo Martincorena *et al.* "Universal Patterns of Selection in Cancer and Somatic Tissues". In: *Cell* 171.5 (2017), 1029–1041.e21. URL: https://www.sciencedirect.com/science/article/pii/S0092867417311364.

[17] Matthew H Bailey *et al.* "Comprehensive Characterization of Cancer Driver Genes and Mutations". In: *Cell* 173.2 (2018), 371–385.e18. URL: https://www.sciencedirect.com/science/article/pii/S009286741830237X.

[18] Peter J Campbell *et al.* "Pan-cancer analysis of whole genomes". In: *Nature* 578.7793 (2020), pp. 82–93. URL: https://doi.org/10.1038/s41586-020-1969-6.

[19] Lynn Quek *et al.* "Clonal heterogeneity of acute myeloid leukemia treated with the IDH2 inhibitor enasidenib". In: *Nature Medicine* 24.8 (2018), pp. 1167–1177. URL: https://doi.org/10.1038/s41591-018-0115-6.

[20] Martin F Fey *et al.* "Clonality and X-Inactivation Patterns in Hematopoietic Cell Populations Detected by the Highly Informative M27β DNA Probe". In: *Blood* 83.4 (1994), pp. 931–938. URL: https://www.sciencedirect.com/science/article/pii/S0006497120757228.

[21] L Busque *et al.* "Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age". In: *Blood* 88.1 (July 1996), pp. 59–65. URL: http://www.bloodjournal.org/content/88/1/59.abstract.

[22] Rosemary E Gale *et al.* "Acquired skewing of X-chromosome inactivation patterns in myeloid cells of the elderly suggests stochastic clonal loss with age". In: *British Journal of Haematology* 98.3 (Sept. 1997), pp. 512–519. URL: http://dx.doi.org/10.1046/j.1365-2141.1997.2573078.x.

[23] Lambert Busque *et al.* "Recurrent Somatic TET2 Mutations in Normal Elderly Individuals With Clonal Hematopoiesis". In: *Nature Genetics* 44.11 (Nov. 2012), pp. 1179–81. arXiv: NIHMS150003. URL: https://www.nature.com/articles/ng.2413.

[24] Siddhartha Jaiswal *et al.* "Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes". In: *New England Journal of Medicine* 371.26 (Dec. 2014), pp. 2488–2498. arXiv: NIHMS150003. URL: https://www.nejm.org/doi/10.1056/NEJMoa1408617.

[25] Giulio Genovese *et al.* "Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence". In: *New England Journal of Medicine* 371.26 (Dec. 2014), pp. 2477–2487. arXiv: NIHMS150003. URL: https://www.nejm.org/doi/10.1056/NEJMoa1409405.

[26] Mingchao Xie *et al.* "Age-related mutations associated with clonal hematopoietic expansion and malignancies". In: *Nature Medicine* 20.12 (Dec. 2014), pp. 1472–1478. arXiv: NIHMS150003. URL: https://www.nature.com/articles/nm.3733.

[27] H Nakazawa *et al.* "UV and skin cancer: specific p53 gene mutation in normal skin as a biologically relevant exposure measurement." In: *Proceedings of the National Academy of Sciences* 91.1 (Jan. 1994), pp. 360–364. URL: https://doi.org/10.1073/pnas.91.1.360.

[28] Annemarie Ziegler *et al.* "Sunburn and p53 in the onset of skin cancer". In: *Nature* 372.6508 (1994), pp. 773–776. URL: https://doi.org/10.1038/372773a0.

[29] Alan S. Jonason *et al.* "Frequent clones of p53-mutated keratinocytes in normal human skin". In: *Proceedings of the National Academy of Sciences* 93.24 (Nov. 1996), pp. 14025–14029. URL: https://doi.org/10.1073/pnas.93.24.14025.

[30] Iñigo Martincorena *et al.* "High burden and pervasive positive selection of somatic mutations in normal human skin". In: *Science* 348.6237 (May 2015), pp. 880–886. URL: http://science.sciencemag.org/content/348/6237/880.abstract.

[31] Iñigo Martincorena *et al.* "Somatic mutant clones colonize the human esophagus with age". In: *Science* 362.6417 (Nov. 2018), pp. 911–917. URL: http://science.sciencemag.org/content/362/6417/911.abstract.

[32] Akira Yokoyama *et al.* "Age-related remodelling of oesophageal epithelia by mutated cancer drivers". In: *Nature* (2019). URL: https://doi.org/10.1038/s41586-018-0811-x.

[33] Keren Yizhak *et al.* "RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues". In: *Science* 364.6444 (June 2019), eaaw0726. URL: http://science.sciencemag.org/content/364/6444/eaaw0726.abstract.

[34] Henry Lee-Six *et al.* "The landscape of somatic mutation in normal colorectal epithelial cells". In: *Nature* 574.7779 (2019), pp. 532–537. URL: https://doi.org/10.1038/s41586-019-1672-7.

[35] Nobuyuki Kakiuchi *et al.* "Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis". In: *Nature* 577.7789 (2020), pp. 260–265. URL: https://doi.org/10.1038/s41586-019-1856-1.

[36] Kosaku Nanki *et al.* "Somatic inflammatory gene mutations in human ulcerative colitis epithelium". In: *Nature* 577.7789 (2020), pp. 254–259. URL: https://doi.org/10.1038/s41586-019-1844-5.

[37] Kenichi Yoshida *et al.* "Tobacco smoking and somatic mutations in human bronchial epithelium". In: *Nature* 578.7794 (2020), pp. 266–272. URL: https://doi.org/10.1038/s41586-020-1961-1.

[38] Simon F Brunner *et al.* "Somatic mutations and clonal dynamics in healthy and cirrhotic human liver". In: *Nature* 574.7779 (2019), pp. 538–542. URL: https://doi.org/10.1038/s41586-019-1670-9.

[39] Soo Ki Kim *et al.* "Comprehensive analysis of genetic aberrations linked to tumorigenesis in regenerative nodules of liver cirrhosis". In: *Journal of Gastroenterology* 54.7 (2019), pp. 628–640. URL: https://doi.org/10.1007/s00535-019-01555-z.

[40] Luiza Moore *et al.* "The mutational landscape of normal human endometrial epithelium". In: *Nature* 580.7805 (2020), pp. 640–646. URL: https://doi.org/10.1038/s41586-020-2214-z.

[41] Ruoyan Li *et al.* "Macroscopic somatic clonal expansion in morphologically normal human urothelium". In: *Science* 370.6512 (Oct. 2020), pp. 82–89. URL: https://doi.org/10.1126/science.aba7300.

[42] Andrew R J Lawson *et al.* "Extensive heterogeneity in somatic mutation and selection in the human bladder". In: *Science* 370.6512 (Oct. 2020), pp. 75–82. URL: https://doi.org/10.1126/science.aba8347.

[43] Max Jan *et al.* "Clonal Evolution of Preleukemic Hematopoietic Stem Cells Precedes Human Acute Myeloid Leukemia". en. In: *Science Translational Medicine* 4.149 (Aug. 2012), 149ra118. URL: http://stm.sciencemag.org/content/4/149/149ra118.

[44] M. Ryan Corces-Zimmerman *et al.* "Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission". en. In: *Proceedings of the National Academy of Sciences* 111.7 (Feb. 2014), pp. 2548–2553. URL: https://doi.org/10.1073/pnas.1324297111.

[45] Liran I. Shlush *et al.* "Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia". en. In: *Nature* 506.7488 (Feb. 2014), pp. 328–333. URL: http://www.nature.com/nature/journal/v506/n7488/full/nature13038.html.

[46] Douglas E Wright *et al.* "Physiological Migration of Hematopoietic Stem and Progenitor Cells". In: *Science* 294.5548 (Nov. 2001), pp. 1933–1936. URL: https://doi.org/10.1126/science.1064081.

[47] Toshihiro Miyamoto, Irving L Weissman, and Koichi Akashi. "AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation". In: *Proceedings of the National Academy of Sciences* 97.13 (June 2000), pp. 7521–7526. URL: https://doi.org/10.1073/pnas.97.13.7521.

[48] Petter S. Woll *et al.* "Myelodysplastic Syndromes Are Propagated by Rare and Distinct Human Cancer Stem Cells In Vivo". English. In: *Cancer Cell* 25.6 (June 2014), pp. 794–808. URL: https://doi.org/10.1016/j.ccr.2014.03.036.

[49] Nicolas Goardon *et al.* "Coexistence of LMPP-like and GMP-like Leukemia Stem Cells in Acute Myeloid Leukemia". English. In: *Cancer Cell* 19.1 (Jan. 2011), pp. 138–152. URL: https://doi.org/10.1016/j.ccr.2010.12.012.

[50] Lynn Quek *et al.* "Genetically distinct leukemic stem cells in human CD34- acute myeloid leukemia are arrested at a hemopoietic precursor-like stage". eng. In: *The Journal of Experimental Medicine* 213.8 (July 2016), pp. 1513–1535. URL: https://doi.org/10.1084/jem.20151775.

[51] Thomas McKerrell and George S Vassiliou. "Aging as a driver of leukemogenesis". In: *Science Translational Medicine* 7.306 (Sept. 2015), 306fs38. URL: http://stm.sciencemag.org/content/7/306/306fs38.abstract.

[52] Andrew L Young *et al.* "Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults". In: *Nature Communications* 7 (2016), p. 12484. URL: http://dx.doi.org/10.1038/ncomms12484.

[53] Rocio Acuna-Hidalgo *et al.* "Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life". In: *The American Journal of Human Genetics* 101.1 (July 2017), pp. 50–64. URL: http://www.sciencedirect.com/science/article/pii/S0002929717302008.

[54] Bruce K Duncan and Jeffrey H Miller. "Mutagenic deamination of cytosine residues in DNA". In: *Nature* 287.5782 (1980), pp. 560–561. URL: https://doi.org/10.1038/287560a0.

[55] Ludmil B Alexandrov *et al.* "Signatures of mutational processes in human cancer". In: *Nature* 500.7463 (2013), pp. 415–421. URL: https://doi.org/10.1038/nature12477.

[56] Manuel Buscarlet *et al.* "DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions". In: *Blood* 130.6 (June 2017), pp. 753–762. URL: http://www.bloodjournal.org/content/130/6/753.abstract.

[57] Florian Zink *et al.* "Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly". In: *Blood* 130.6 (Aug. 2017), pp. 742–752. URL: https://doi.org/10.1182/blood-2017-02-769869.

[58] Pinkal Desai *et al.* "Somatic mutations precede acute myeloid leukemia years before diagnosis". In: *Nature Medicine* 24.7 (2018), pp. 1015–1023. URL: http://www.nature.com/articles/s41591-018-0081-z.

[59] Sagi Abelson *et al.* "Prediction of acute myeloid leukaemia risk in healthy individuals". In: *Nature* 559.7714 (2018), pp. 400–404. URL: https://doi.org/10.1038/s41586-018-0317-6.

[60] Michael D Kessler *et al.* "Exome sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes". In: *medRxiv* (Jan. 2022), p. 2021.12.29.21268342. URL: http://medrxiv.org/content/early/2022/01/01/2021.12.29.21268342.abstract.

[61] Catherine 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 (Aug. 2017), 374–382.e4. URL: http://dx.doi.org/10.1016/j.stem.2017.07.010.

[62] Kelly L Bolton *et al.* "Cancer therapy shapes the fitness landscape of clonal hematopoiesis". In: *Nature Genetics* 52.11 (2020), pp. 1219–1226. URL: https://doi.org/10.1038/s41588-020-00710-0.

[63] Kevin B Jacobs *et al.* "Detectable clonal mosaicism and its relationship to aging and cancer". In: *Nature Genetics* 44.6 (2012), pp. 651–658. URL: https://doi.org/10.1038/ng.2270.

[64] Cathy C Laurie *et al.* "Detectable clonal mosaicism from birth to old age and its relationship to cancer". In: *Nature Genetics* 44.6 (2012), pp. 642–650. URL: https://doi.org/10.1038/ng.2271.

[65] Po-Ru Loh *et al.* "Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations". In: *Nature* 559.7714 (2018), pp. 350–355. URL: https://doi.org/10.1038/s41586-018-0321-x.

[66] Po-Ru Loh, Giulio Genovese, and Steven A McCarroll. "Monogenic and polygenic inheritance become instruments for clonal selection". In: *Nature* 584.7819 (2020), pp. 136–141. URL: https://doi.org/10.1038/s41586-020-2430-6.

[67] Chikashi Terao *et al.* "Chromosomal alterations among age-related haematopoietic clones in Japan". In: *Nature* 584.7819 (2020), pp. 130–135. URL: https://doi.org/10.1038/s41586-020-2426-2.

[68] Ruli Gao *et al.* "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes". In: *Nature Biotechnology* 39.5 (2021), pp. 599–608. URL: https://doi.org/10.1038/s41587-020-00795-2.

[69] Henne Holstege *et al.* "Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis". In: *Genome Research* 24.5 (Apr. 2014), pp. 733–742. URL: http://www.genome.org/cgi/doi/10.1101/gr.162131.113.

[70] Verena Körber and Thomas Höfer. "Inferring growth and genetic evolution of tumors from genome sequences". In: *Current Opinion in Systems Biology* 16 (2019), pp. 1–9. URL: https://doi.org/10.1016/j.coisb.2019.10.015.

[71] Gladys Y P Poon *et al.* "Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues". In: *Nature Genetics* 53.11 (2021), pp. 1597–1605. URL: https://doi.org/10.1038/s41588-021-00957-1.

[72] Alexander G. Bick *et al.* "Inherited causes of clonal haematopoiesis in 97,691 whole genomes". In: *Nature* 586.7831 (Oct. 2020), pp. 763–768. URL: http://www.nature.com/articles/s41586-020-2819-2.

[73] Alexey Palamarchuk *et al.* "Tcl1 protein functions as an inhibitor of de novo DNA methylation in B-cell chronic lymphocytic leukemia (CLL)". In: *Proceedings of the National Academy of Sciences* 109.7 (Feb. 2012), pp. 2555–2560. URL: https://doi.org/10.1073/pnas.1200003109.

[74] Lucile Couronné, Christian Bastard, and Olivier A Bernard. "TET2 and DNMT3A Mutations in Human T-Cell Lymphoma". In: *New England Journal of Medicine* 366.1 (Jan. 2012), pp. 95–96. URL: http://dx.doi.org/10.1056/NEJMc1111708.

[75] Fazila Asmar *et al.* "Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma". In: *Haematologica* 98.12 (Dec. 2013), pp. 1912–1920. URL: https://haematologica.org/article/view/6867.

[76] David P. Steensma *et al.* "Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes". en. In: *Blood* 126.1 (July 2015), pp. 9–16. URL: https://doi.org/10.1182/blood-2015-03-631747.

[77] Michael Heuser, Felicitas Thol, and Arnold Ganser. "Clonal Hematopoiesis of Indeterminate Potential". eng. In: *Deutsches Arzteblatt international* 113.18 (May 2016), pp. 317–322. URL: https://pubmed.ncbi.nlm.nih.gov/27215596.

[78] Luca Malcovati *et al.* "Clinical significance of somatic mutation in unexplained blood cytopenia". In: *Blood* 129.25 (June 2017), pp. 3371–3378. URL: https://doi.org/10.1182/blood-2017-01-763425.

[79] Adam S. Sperling, Christopher J. Gibson, and Benjamin L. Ebert. "The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia". In: *Nature Reviews Cancer* 17.1 (2017), pp. 5–19. URL: https://www.nature.com/articles/nrc.2016.112.

[80] Hartmut Döhner, Daniel J Weisdorf, and Clara D Bloomfield. "Acute Myeloid Leukemia". In: *New England Journal of Medicine* 373.12 (Sept. 2015), pp. 1136–1152. URL: https://dx.doi.org/10.1056/NEJMra1406184.

[81] Mya S Thein *et al.* "Outcome of older patients with acute myeloid leukemia". In: *Cancer* 119.15 (Aug. 2013), pp. 2720–2727. URL: http://dx.doi.org/10.1002/cncr.28129.

[82] R Coleman Lindsley *et al.* "Acute myeloid leukemia ontogeny is defined by distinct somatic mutations". In: *Blood* 125.9 (Feb. 2015), pp. 1367–1376. URL: http://www.bloodjournal.org/content/125/9/1367.abstract.

[83] The Cancer Genome Atlas Research Network. "Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia". In: *New England Journal of Medicine* 368.22 (May 2013), pp. 2059–2074. URL: http://www.nejm.org/doi/pdf/10.1056/NEJMoa1301689.

[84] Elli Papaemmanuil *et al.* "Genomic Classification and Prognosis in Acute Myeloid Leukemia". In: *New England Journal of Medicine* 374.23 (June 2016), pp. 2209–2221. URL: https://www.nejm.org/doi/full/10.1056/nejmoa1516192.

[85] F Thol *et al.* "Acute myeloid leukemia derived from lympho-myeloid clonal hematopoiesis". In: *Leukemia* 31.6 (June 2017), pp. 1286–1295. URL: http://dx.doi.org/10.1038/leu.2016.345.

[86] Maja Rothenberg-Thurley *et al.* "Persistence of pre-leukemic clones during first remission and risk of relapse in acute myeloid leukemia". In: *Leukemia* 32.7 (2018), pp. 1598–1608. URL: https://doi.org/10.1038/s41375-018-0034-z.

[87] Gro G. Pløen *et al.* "Persistence of DNMT3A mutations at long-term remission in adult patients with AML". en. In: *British Journal of Haematology* 167.4 (Nov. 2014), pp. 478–486. URL: https://doi.org/10.1111/bjh.13062.

[88] V I Gaidzik *et al.* "DNMT3A mutant transcript levels persist in remission and do not predict outcome in patients with acute myeloid leukemia". In: *Leukemia* 32.1 (2018), pp. 30–37. URL: https://doi.org/10.1038/leu.2017.200.

[89] Mojca Jongen-Lavrencic *et al.* "Molecular Minimal Residual Disease in Acute Myeloid Leukemia". In: *New England Journal of Medicine* 378.13 (2018), pp. 1189–1199. URL: http://www.nejm.org/doi/10.1056/NEJMoa1716863.

[90] Siddhartha Jaiswal *et al.* "Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease". In: *New England Journal of Medicine* 377.2 (June 2017), pp. 111–121. URL: http://www.nejm.org/doi/10.1056/NEJMoa1701719.

[91] Siddhartha Jaiswal and Benjamin L Ebert. "Clonal hematopoiesis in human aging and disease". In: *Science* 366.6465 (Nov. 2019), eaan4673. URL: http://science.sciencemag.org/content/366/6465/eaan4673.abstract.

[92] José J Fuster *et al.* "Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice". In: *Science* 355.6327 (Feb. 2017), pp. 842–847. URL: http://science.sciencemag.org/content/355/6327/842.abstract.

[93] Alexander Heyde *et al.* "Increased stem cell proliferation in atherosclerosis accelerates clonal hematopoiesis". In: *Cell* 184.5 (Mar. 2021), 1348–1361.e22. URL: https://doi.org/10.1016/j.cell.2021.01.049.

[94] Judith S Hecker *et al.* "CHIP and hips: clonal hematopoiesis is common in patients undergoing hip arthroplasty and is associated with autoimmune disease". In: *Blood* 138.18 (2021), pp. 1727–1732. URL: https://doi.org/10.1182/blood.2020010163%7B%5C%%7D0A.

[95] Tetsuichi Yoshizato *et al.* "Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia". In: *New England Journal of Medicine* 373.1 (July 2015), pp. 35–47. URL: https://dx.doi.org/10.1056/NEJMoa1414799.

[96] Paula Savola *et al.* "Clonal hematopoiesis in patients with rheumatoid arthritis". In: *Blood Cancer Journal* 8.8 (2018), p. 69. URL: http://www.nature.com/articles/s41408-018-0107-2.

[97] Nila J Dharan *et al.* "HIV is associated with an increased risk of age-related clonal hematopoiesis among older adults". In: *Nature Medicine* 27.6 (2021), pp. 1006–1011. URL: https://doi.org/10.1038/s41591-021-01357-y.

[98] 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. URL: https://doi.org/10.1038/s41591-021-01371-0.

[99] Kelly L Bolton *et al.* "Clonal hematopoiesis is associated with risk of severe Covid-19". In: *medRxiv* (Jan. 2020), p. 2020.11.25.20233163. URL: http://medrxiv.org/content/early/2020/11/27/2020.11.25.20233163.abstract.

[100] Hind Bouzid *et al.* "Clonal hematopoiesis is associated with protection from Alzheimer's disease". In: *medRxiv* (Jan. 2021), p. 2021.12.10.21267552. URL: http://medrxiv.org/content/early/2021/12/13/2021.12.10.21267552.abstract.

[101] Henry Lee-Six *et al.* "Population dynamics of normal human blood inferred from somatic mutations". In: *Nature* 561.7724 (2018), pp. 473–478. URL: https://doi.org/10.1038/s41586-018-0497-0.

[102] Emily Mitchell *et al.* "Clonal dynamics of haematopoiesis across the human lifespan". In: *bioRxiv* (Jan. 2021), p. 2021.08.16.456475. URL: http://biorxiv.org/content/early/2021/08/16/2021.08.16.456475.abstract.

[103] Margarete A Fabre *et al.* "The longitudinal dynamics and natural history of clonal haematopoiesis". In: *bioRxiv* (Jan. 2021), p. 2021.08.12.455048. URL: http://biorxiv.org/content/early/2021/08/12/2021.08.12.455048.abstract.

[104] Caroline J Watson *et al.* "The evolutionary dynamics and fitness landscape of clonal hematopoiesis". In: *Science* 367.6485 (Mar. 2020), pp. 1449–1454. URL: http://science.sciencemag.org/content/367/6485/1449.abstract.

[105] Dirk Schübeler. "Function and information content of DNA methylation". In: *Nature* 517.7534 (2015), pp. 321–326. URL: https://doi.org/10.1038/nature14192.

[106] Humaira Gowher and Albert Jeltsch. "Mammalian DNA methyltransferases: new discoveries and open questions". In: *Biochemical Society Transactions* 46.5 (Aug. 2018), pp. 1191–1202. URL: https://doi.org/10.1042/BST20170574.

[107] Colum P Walsh, J Richard Chaillet, and Timothy H Bestor. "Transcription of IAP endogenous retroviruses is constrained by cytosine methylation". In: *Nature Genetics* 20.2 (1998), pp. 116–117. URL: https://doi.org/10.1038/2413.

[108] Déborah Bourc'his and Timothy H Bestor. "Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L". In: *Nature* 431.7004 (2004), pp. 96–99. URL: https://doi.org/10.1038/nature02886.

[109] Masaki Okano, Shaoping Xie, and En Li. "Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases". en. In: *Nature Genetics* 19.3 (July 1998), pp. 219–220. URL: https://www.nature.com/articles/ng0798%7B%5C_%7D219.

[110] Shaoping Xie *et al.* "Cloning, expression and chromosome locations of the human DNMT3 gene family". In: *Gene* 236.1 (Aug. 1999), pp. 87–95. URL: https://doi.org/10.1016/S0378-1119(99)00252-8.

[111] Masaki Okano *et al.* "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development". In: *Cell* 99.3 (Dec. 1999), pp. 247–257. URL: http://dx.doi.org/10.1016/S0092-8674(00)81656-6.

[112] Chan-Wang J Lio, Hiroshi Yuita, and Anjana Rao. "Dysregulation of the TET family of epigenetic regulators in lymphoid and myeloid malignancies". In: *Blood* 134.18 (Oct. 2019), pp. 1487–1497. URL: https://doi.org/10.1182/blood.2019791475.

[113] Mamta Tahiliani *et al.* "Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1". In: *Science* 324.5929 (May 2009), pp. 930–935. URL: http://science.sciencemag.org/content/324/5929/930.abstract.

[114] Lakshminarayan M Iyer *et al.* "Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids". eng. In: *Cell cycle (Georgetown, Tex.)* 8.11 (June 2009), pp. 1698–1710. URL: https://pubmed.ncbi.nlm.nih.gov/19411852.

[115] Shinsuke Ito *et al.* "Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine". In: *Science* 333.6047 (Sept. 2011), pp. 1300–1303. URL: https://doi.org/10.1126/science.1210597.

[116] Hideharu Hashimoto *et al.* "Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation". In: *Nucleic Acids Research* 40.11 (June 2012), pp. 4841–4849. URL: https://doi.org/10.1093/nar/gks155.

[117] Junji Otani *et al.* "Cell Cycle-Dependent Turnover of 5-Hydroxymethyl Cytosine in Mouse Embryonic Stem Cells". In: *PLOS ONE* 8.12 (Dec. 2013), e82961. URL: https://doi.org/10.1371/journal.pone.0082961.

[118] Skirmantas Kriaucionis and Nathaniel Heintz. "The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain". In: *Science* 324.5929 (May 2009), pp. 929–930. URL: http://science.sciencemag.org/content/324/5929/929.abstract.

[119] Marian Mellén *et al.* "MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System". In: *Cell* 151.7 (Dec. 2012), pp. 1417–1430. URL: https://doi.org/10.1016/j.cell.2012.11.022.

[120] Ageliki Tsagaratou *et al.* "Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.32 (Aug. 2014), E3306–E3315. URL: https://doi.org/10.1073/pnas.1412327111.

[121] Lorenzo Rinaldi *et al.* "Dnmt3a and Dnmt3b Associate with Enhancers to Regulate Human Epidermal Stem Cell Homeostasis". English. In: *Cell Stem Cell* 19.4 (Oct. 2016), pp. 491–501. URL: https://doi.org/10.1016/j.stem.2016.06.020.

[122] Timothy H Bestor, John R Edwards, and Mathieu Boulard. "Notes on the role of dynamic DNA methylation in mammalian development". In: *Proceedings of the National Academy of Sciences* 112.22 (June 2015), pp. 6796–6799. URL: https://doi.org/10.1073/pnas.1415301111.

[123] Heng Zhu, Guohua Wang, and Jiang Qian. "Transcription factors as readers and effectors of DNA methylation". In: *Nature Reviews Genetics* 17.9 (2016), pp. 551–565. URL: https://doi.org/10.1038/nrg.2016.83.

[124] Michael Brandeis *et al.* "Spl elements protect a CpG island from de novo methylation". In: *Nature* 371.6496 (1994), pp. 435–438. URL: https://doi.org/10.1038/371435a0.

[125] Han Li, Lin Iping G., and Hsieh Chih-Lin. "Protein Binding Protects Sites on Stable Episomes and in the Chromosome from De Novo Methylation". In: *Molecular and Cellular Biology* 21.10 (May 2001), pp. 3416–3424. URL: https://doi.org/10.1128/MCB.21.10.3416-3424.2001.

[126] Yimeng Yin *et al.* "Impact of cytosine methylation on DNA binding specificities of human transcription factors". In: *Science* 356.6337 (May 2017). URL: http://science.sciencemag.org/content/356/6337/eaaj2239.abstract.

[127] Kelly R. Barnett *et al.* "ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions". In: *Molecular Cell* 77.6 (2020), 1350–1364.e6. URL: https://doi.org/10.1016/j.molcel.2020.01.004.

[128] Taiping Chen *et al.* "A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation". en. In: *Journal of Biological Chemistry* 277.41 (Oct. 2002), pp. 38746–38754. URL: http://www.jbc.org/content/277/41/38746.

[129] E Jost *et al.* "Epimutations mimic genomic mutations of DNMT3A in acute myeloid leukemia". In: *Leukemia* 28.6 (2014), pp. 1227–1234. URL: https://doi.org/10.1038/leu.2013.362.

[130] Tanja Božić *et al.* "Variants of DNMT3A cause transcript-specific DNA methylation patterns and affect hematopoiesis". In: *Life Science Alliance* 1.6 (2018), e201800153. URL: https://doi.org/10.26508/lsa.201800153.

[131] Daniel J. Weisenberger *et al.* "Identification and characterization of alternatively spliced variants of DNA methyltransferase 3a in mammalian cells". In: *Gene* 298.1 (Sept. 2002), pp. 91–99. URL: https://doi.org/10.1016/S0378-1119(02)00976-9.

[132] Taiping Chen, Naomi Tsujimoto, and En Li. "The PWWP Domain of Dnmt3a and Dnmt3b Is Required for Directing DNA Methylation to the Major Satellite Repeats at Pericentric Heterochromatin". In: *Molecular and Cellular Biology* 24.20 (Oct. 2004), pp. 9048–9058. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC517890/.

[133] Arunkumar Dhayalan *et al.* "The Dnmt3a PWWP Domain Reads Histone 3 Lysine 36 Trimethylation and Guides DNA Methylation". In: *Journal of Biological Chemistry* 285.34 (Aug. 2010), pp. 26114–26120. URL: http://www.jbc.org/content/285/34/26114.abstract.

[134] Patricia Heyn *et al.* "Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions". In: *Nature Genetics* 51.1 (2019), pp. 96–105. URL: https://doi.org/10.1038/s41588-018-0274-x.

[135] Daniel N Weinberg *et al.* "The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape". In: *Nature* 573.7773 (2019), pp. 281–286. URL: https://doi.org/10.1038/s41586-019-1534-3.

[136] Daniel N Weinberg *et al.* "Two competing mechanisms of DNMT3A recruitment regulate the dynamics of de novo DNA methylation at PRC1-targeted CpG islands". In: *Nature Genetics* 53.6 (2021), pp. 794–800. URL: https://doi.org/10.1038/s41588-021-00856-5.

[137] Zhi-Min Zhang *et al.* "Structural basis for DNMT3A-mediated de novo DNA methylation". In: *Nature* 554.7692 (2018), pp. 387–391. URL: https://doi.org/10.1038/nature25477.

[138] Celeste Holz-Schietinger, Doug M Matje, and Norbert O Reich. "Mutations in DNA methyltransferase (DNMT3A) observed in acute myeloid leukemia patients disrupt processive methylation." In: *The Journal of Biological Chemistry* 287.37 (Sept. 2012), pp. 30941–51. URL: http://www.ncbi.nlm.nih.gov/pubmed/22722925.

[139] Yingying Zhang *et al.* "Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail". In: *Nucleic Acids Research* 38.13 (July 2010), pp. 4246–4253. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2910041/.

[140] Xue Guo *et al.* "Structural insight into autoinhibition and histone H3-induced activation of DNMT3A". In: *Nature* 517.7536 (Jan. 2015), pp. 640–644. URL: https://dx.doi.org/10.1038/nature13899.

[141] Amy L Hughes, Jessica R Kelley, and Robert J Klose. "Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863.8 (2020), p. 194567. URL: https://doi.org/10.1016/j.bbagrm.2020.194567.

[142] Emmanuelle Vire *et al.* "The Polycomb group protein EZH2 directly controls DNA methylation". In: *Nature* 439.7078 (Feb. 2006), pp. 871–874. URL: http://dx.doi.org/10.1038/nature04431.

[143] Renata Zofia Jurkowska, Tomasz Piotr Jurkowski, and Albert Jeltsch. "Structure and Function of Mammalian DNA Methyltransferases". In: *ChemBioChem* 12.2 (Jan. 2011), pp. 206–222. URL: http://dx.doi.org/10.1002/cbic.201000195.

[144] Jiamu Du *et al.* "DNA methylation pathways and their crosstalk with histone methylation". In: *Nature Reviews Molecular Cell Biology* 16.9 (2015), pp. 519–532. URL: http://dx.doi.org/10.1038/nrm4043.

[145] Liubin Yang, Rachel Rau, and Margaret A Goodell. "DNMT3A in haematological malignancies". en. In: *Nature Reviews Cancer* 15.3 (Mar. 2015), pp. 152–165. URL: http://www.nature.com/nrc/journal/v15/n3/full/nrc3895.html?cookies=accepted.

[146] Katrina Tatton-Brown *et al.* "Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability". en. In: *Nature Genetics* 46.4 (Apr. 2014), pp. 385–388. URL: http://www.nature.com/ng/journal/v46/n4/full/ng.2917.html.

[147] Timothy J Ley *et al.* "DNMT3A Mutations in Acute Myeloid Leukemia". In: *New England Journal of Medicine* 363.25 (Nov. 2010), pp. 2424–2433. URL: http://dx.doi.org/10.1056/NEJMoa1005143.

[148] Felicitas Thol *et al.* "Incidence and Prognostic Influence of DNMT3A Mutations in Acute Myeloid Leukemia". In: *Journal of Clinical Oncology* 29.21 (July 2011), pp. 2889–2896. URL: https://doi.org/10.1200/JCO.2011.35.4894.

[149] Xiao-Jing Yan *et al.* "Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia". In: *Nature Genetics* 43.4 (Mar. 2011), pp. 309–315. URL: http://www.nature.com/doifinder/10.1038/ng.788.

[150] Elli Papaemmanuil *et al.* "Clinical and biological implications of driver mutations in myelodysplastic syndromes". In: *Blood* 122.22 (Nov. 2013), pp. 3616–3627. URL: http://www.bloodjournal.org/content/122/22/3616.abstract.

[151] M J Walter *et al.* "Recurrent DNMT3A mutations in patients with myelodysplastic syndromes". In: *Leukemia* 25.7 (July 2011), pp. 1153–1158. URL: http://www.nature.com/doifinder/10.1038/leu.2011.44.

[152] Anna M Jankowska *et al.* "Mutational spectrum analysis of chronic myelomonocytic leukemia includes genes associated with epigenetic regulation: UTX, EZH2, and DNMT3A". In: *Blood* 118.14 (Oct. 2011), pp. 3932–3941. URL: https://doi.org/10.1182/blood-2010-10-311019.

[153] Vera Grossmann *et al.* "The molecular profile of adult T-cell acute lymphoblastic leukemia: Mutations in RUNX1 and DNMT3A are associated with poor prognosis in T-ALL". In: *Genes, Chromosomes and Cancer* 52.4 (Apr. 2013), pp. 410–422. URL: http://dx.doi.org/10.1002/gcc.22039.

[154] A Roller *et al.* "Landmark analysis of DNMT3A mutations in hematological malignancies". In: *Leukemia* 27.7 (July 2013), pp. 1573–1578. URL: http://dx.doi.org/10.1038/leu.2013.65.

[155] Pieter Van Vlierberghe *et al.* "ETV6 mutations in early immature human T cell leukemias". In: *The Journal of Experimental Medicine* 208.13 (Dec. 2011), pp. 2571–2579. URL: http://jem.rupress.org/content/208/13/2571.abstract.

[156] Rosemary E Gale *et al.* "Simpson's Paradox and the Impact of Different DNMT3A Mutations on Outcome in Younger Adults With Acute Myeloid Leukemia". In: *Journal of Clinical Oncology* 33.18 (June 2015), pp. 2072–2083. URL: https://doi.org/10.1200/JCO.2014.59.2022.

[157] Soo Jin Kim *et al.* "A DNMT3A mutation common in AML exhibits dominant-negative effects in murine ES cells". In: *Blood* 122.25 (Dec. 2013), pp. 4086–4089. URL: http://www.bloodjournal.org/content/122/25/4086.abstract.

[158] David A. Russler-Germain *et al.* "The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers". English. In: *Cancer Cell* 25.4 (Apr. 2014), pp. 442–454. URL: https://doi.org/10.1016/j.ccr.2014.02.010.

[159] Junji Koya *et al.* "DNMT3A R882 mutants interact with polycomb proteins to block haematopoietic stem and leukaemic cell differentiation". en. In: *Nature Communications* 7 (Mar. 2016), p. 10924. URL: http://www.nature.com/ncomms/2016/160324/ncomms10924/full/ncomms10924.html.

[160] Max Emperle *et al.* "Mutations of R882 change flanking sequence preferences of the DNA methyltransferase DNMT3A and cellular methylation patterns". In: *Nucleic Acids Research* 47.21 (Dec. 2019), pp. 11355–11367. URL: https://doi.org/10.1093/nar/gkz911.

[161] Hiwot Anteneh, Jian Fang, and Jikui Song. "Structural basis for impairment of DNA methylation by the DNMT3A R882H mutation". In: *Nature Communications* 11.1 (2020), p. 2294. URL: https://doi.org/10.1038/s41467-020-16213-9.

[162] Yung-Hsin Huang *et al.* "Systematic Profiling of DNMT3A Variants Reveals Protein Instability Mediated by the DCAF8 E3 Ubiquitin Ligase Adaptor". In: *Cancer Discovery* 12.1 (Jan. 2022), pp. 220–235. URL: http://cancerdiscovery.aacrjournals.org/content/12/1/220.abstract.

[163] Jeffrey W. Tyner *et al.* "Functional genomic landscape of acute myeloid leukaemia". In: *Nature* 562.7728 (2018), pp. 526–531. URL: https://doi.org/10.1038/s41586-018-0623-z.

[164] Verena I Gaidzik *et al.* "Clinical impact of DNMT3A mutations in younger adult patients with acute myeloid leukemia: results of the AML Study Group (AMLSG)". In: *Blood* 121.23 (June 2013), pp. 4769–4777. URL: http://www.bloodjournal.org/content/121/23/4769.abstract.

[165] Grant Challen and Margaret A Goodell. "Clonal Hematopoiesis: Mechanisms Driving Dominance of Stem Cell Clones". In: *Blood* 136.1590 (2020), pp. 1590–1598.

[166] Grant A Challen *et al.* "Dnmt3a is essential for hematopoietic stem cell differentiation". In: *Nature Genetics* 44.1 (Jan. 2012), pp. 23–31. URL: http://dx.doi.org/10.1038/ng.1009.

[167] Mira Jeong *et al.* "Loss of Dnmt3a Immortalizes Hematopoietic Stem Cells In Vivo". In: *Cell Reports* 23.1 (2018), pp. 1–10. URL: https://doi.org/10.1016/j.celrep.2018.03.025.

[168] Hamza Celik *et al.* "Enforced differentiation of Dnmt3a-null bone marrow leads to failure with c-Kit mutations driving leukemic transformation". In: *Blood* 125.4 (2015), pp. 619–628. URL: http://www.bloodjournal.org/content/125/4/619.long.

[169] Allison Mayle *et al.* "Dnmt3a loss predisposes murine hematopoietic stem cells to malignant transformation". en. In: *Blood* 125.4 (Jan. 2015), pp. 629–638. URL: http://www.bloodjournal.org/content/125/4/629.

[170] O A Guryanova *et al.* "Dnmt3a regulates myeloproliferation and liver-specific expansion of hematopoietic stem and progenitor cells". In: *Leukemia* 30.5 (May 2016), pp. 1133–1142. URL: http://www.nature.com/doifinder/10.1038/leu.2015.358.

[171] Liubin Yang *et al.* "DNMT3A Loss Drives Enhancer Hypomethylation in FLT3-ITD-Associated Leukemias". In: *Cancer Cell* 29.6 (June 2016), pp. 922–934. URL: https://doi.org/10.1016/j.ccell.2016.05.003.

[172] Grant A. Challen *et al.* "Dnmt3a and Dnmt3b Have Overlapping and Distinct Functions in Hematopoietic Stem Cells". In: *Cell Stem Cell* 15.3 (Sept. 2014), pp. 350–364. arXiv: NIHMS150003. URL: https://doi.org/10.1016/j.stem.2014.06.018.

[173] Mira Jeong *et al.* "Large conserved domains of low DNA methylation maintained by Dnmt3a". In: *Nature Genetics* 46.1 (Nov. 2013), pp. 17–23. arXiv: NIHMS150003. URL: http://www.nature.com/doifinder/10.1038/ng.2836.

[174] Rong Li *et al.* "TET2 Loss Dysregulates the Behavior of Bone Marrow Mesenchymal Stromal Cells and Accelerates Tet2-/–Driven Myeloid Malignancy Progression". In: *Stem Cell Reports* 10 (Jan. 2018). URL: https://doi.org/10.1016/j.stemcr.2017.11.019.

[175] Christopher B. Cole *et al.* "Haploinsufficiency for DNA methyltransferase 3A predisposes hematopoietic cells to myeloid malignancies". In: *Journal of Clinical Investigation* 127.10 (Sept. 2017), pp. 3657–3674. URL: https://www.jci.org/articles/view/93041.

[176] Olga A Guryanova *et al.* "DNMT3A mutations promote anthracycline resistance in acute myeloid leukemia via impaired nucleosome remodeling". In: *Nature Medicine* 22.12 (Dec. 2016), pp. 1488–1495. URL: http://dx.doi.org/10.1038/nm.4210.

[177] Rui Lu *et al.* "Epigenetic Perturbations by Arg882-Mutated DNMT3A Potentiate Aberrant Stem Cell Gene-Expression Program and Acute Leukemia Development". In: *Cancer Cell* 30.1 (July 2016), pp. 92–107. URL: https://dx.doi.org/10.1016/j.ccell.2016.05.008.

[178] Megan T Baldridge *et al.* "Quiescent haematopoietic stem cells are activated by IFN-γ in response to chronic infection". In: *Nature* 465.7299 (2010), pp. 793–797. URL: https://doi.org/10.1038/nature09135.

[179] Eric M Pietras *et al.* "Re-entry into quiescence protects hematopoietic stem cells from the killing effect of chronic exposure to type I interferons". In: *Journal of Experimental Medicine* 211.2 (Feb. 2014), pp. 245–262. URL: https://doi.org/10.1084/jem.20131043.

[180] 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. URL: https://doi.org/10.1016/j.celrep.2016.11.031.

[181] Daniel Hormaechea-Agulla *et al.* "Chronic infection drives Dnmt3a-loss-of-function clonal hematopoiesis via IFNγ signaling". In: *Cell Stem Cell* 28.8 (Aug. 2021), 1428–1442.e6. URL: https://doi.org/10.1016/j.stem.2021.03.002.

[182] Xiaotian Zhang *et al.* "DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells". In: *Nature Genetics* 48.9 (Sept. 2016), pp. 1014–1023. URL: https://dx.doi.org/10.1038/ng.3610.

[183] Franco Izzo *et al.* "DNA methylation disruption reshapes the hematopoietic differentiation landscape". In: *Nature Genetics* 52.4 (2020), pp. 378–387. URL: https://doi.org/10.1038/s41588-020-0595-4.

[184] 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* (Jan. 2022), p. 2022.01.14.476225. URL: https://doi.org/10.1101/2022.01.14.476225.

[185] Hong Lou *et al.* "The Human TET2 Gene Contains Three Distinct Promoter Regions With Differing Tissue and Developmental Specificities". In: *Frontiers in Cell and Developmental Biology* 7 (2019), p. 99.

[186] Lulu Hu *et al.* "Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation". In: *Cell* 155.7 (Dec. 2013), pp. 1545–1555. URL: https://doi.org/10.1016/j.cell.2013.11.020.

[187] Miao Yu *et al.* "Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome". In: *Cell* 149.6 (June 2012), pp. 1368–1380. URL: https://doi.org/10.1016/j.cell.2012.04.027.

[188] Lorenzo de la Rica *et al.* "PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation". In: *Genome Biology* 14.9 (2013), R99. URL: https://doi.org/10.1186/gb-2013-14-9-r99.

[189] Yiping Wang *et al.* "WT1 Recruits TET2 to Regulate Its Target Gene Expression and Suppress Leukemia Cell Proliferation". In: *Molecular Cell* 57.4 (Feb. 2015), pp. 662–673. URL: https://doi.org/10.1016/j.molcel.2014.12.023.

[190] Chengli Song *et al.* "PML Recruits TET2 to Regulate DNA Modification and Cell Proliferation in Response to Chemotherapeutic Agent". In: *Cancer Research* 78.10 (May 2018), pp. 2475–2489. URL: https://doi.org/10.1158/0008-5472.CAN-17-3091.

[191] Xi Chen *et al.* "A rapid and robust method for single cell chromatin accessibility profiling". In: *Nature Communications* 9.1 (2018), p. 5345. URL: https://doi.org/10.1038/s41467-018-07771-0%20http://www.nature.com/articles/s41467-018-07771-0.

[192] Jose Luis Sardina *et al.* "Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate". In: *Cell Stem Cell* 23.5 (Sept. 2018), 727–741.e9. URL: https://doi.org/10.1016/j.stem.2018.08.016.

[193] Kathryn Blaschke *et al.* "Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells". In: *Nature* 500.7461 (2013), pp. 222–226. URL: https://doi.org/10.1038/nature12362.

[194] Michalis Agathocleous *et al.* "Ascorbate regulates haematopoietic stem cell function and leukaemogenesis". In: *Nature* 549.7673 (Aug. 2017), pp. 476–481. URL: https://doi.org/10.1038/nature23876.

[195] Luisa Cimmino *et al.* "Restoration of TET2 Function Blocks Aberrant Self-Renewal and Leukemia Progression". In: *Cell* 170.6 (Sept. 2017), 1079–1095.e20. URL: https://dx.doi.org/10.1016/j.cell.2017.07.032.

[196] Jie Sun *et al.* "SIRT1 Activation Disrupts Maintenance of Myelodysplastic Syndrome Stem and Progenitor Cells by Restoring TET2 Function". In: *Cell Stem Cell* 23.3 (Sept. 2018), 355–369.e9. URL: https://doi.org/10.1016/j.stem.2018.07.018.

[197] Su Jung Song *et al.* "The Oncogenic MicroRNA miR-22 Targets the TET2 Tumor Suppressor to Promote Hematopoietic Stem Cell Self-Renewal and Transformation". In: *Cell Stem Cell* 13.1 (July 2013), pp. 87–101. arXiv: NIHMS150003. URL: https://dx.doi.org/10.1016/j.stem.2013.06.003.

[198] Myunggon Ko *et al.* "Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX". In: *Nature* 497.7447 (2013), pp. 122–126. URL: https://doi.org/10.1038/nature12052.

[199] Lijuan Fu *et al.* "Tet-Mediated Formation of 5-Hydroxymethylcytosine in RNA". In: *Journal of the American Chemical Society* 136.33 (Aug. 2014), pp. 11582–11585. URL: https://doi.org/10.1021/ja505305z.

[200] Qicong Shen *et al.* "Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation". In: *Nature* 554.7690 (2018), pp. 123–127. URL: https://doi.org/10.1038/nature25434.

[201] Chongsheng He *et al.* "TET2 chemically modifies tRNAs and regulates tRNA fragment levels". In: *Nature Structural & Molecular Biology* 28.1 (2021), pp. 62–70. URL: https://doi.org/10.1038/s41594-020-00526-w.

[202] Jie Lan *et al.* "Functional role of Tet-mediated RNA hydroxymethylcytosine in mouse ES cells and during differentiation". In: *Nature Communications* 11.1 (2020), p. 4956. URL: https://doi.org/10.1038/s41467-020-18729-6.

[203] François Delhommeau *et al.* "Mutation in TET2 in Myeloid Cancers". In: *New England Journal of Medicine* 360.22 (May 2009), pp. 2289–2301. URL: https://doi.org/10.1056/NEJMoa0810069.

[204] T Haferlach *et al.* "Landscape of genetic lesions in 944 patients with myelodysplastic syndromes". In: *Leukemia* 28 (Nov. 2013), p. 241. URL: https://dx.doi.org/10.1038/leu.2013.336.

[205] François Lemonnier *et al.* "Recurrent TET2 mutations in peripheral T-cell lymphomas correlate with TFH-like features and adverse clinical parameters". In: *Blood* 120.7 (2012), pp. 1466–1470.

[206] Mamiko Sakata-Yanagimoto *et al.* "Somatic RHOA mutation in angioimmunoblastic T cell lymphoma". In: *Nature Genetics* 46.2 (2014), pp. 171–175. URL: https://doi.org/10.1038/ng.2872.

[207] Anupama Reddy *et al.* "Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma". In: *Cell* 171.2 (2017), 481–494.e15. URL: https://doi.org/10.1016/j.cell.2017.09.027.

[208] Roland Schmitz *et al.* "Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma". In: *New England Journal of Medicine* 378.15 (Apr. 2018), pp. 1396–1407. URL: https://doi.org/10.1056/NEJMoa1801445.

[209] Myunggon Ko *et al.* "Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2". en. In: *Nature* 468.7325 (Dec. 2010), pp. 839–843. URL: https://doi.org/10.1038/nature09586.

[210] S Weissmann *et al.* "Landscape of TET2 mutations in acute myeloid leukemia". In: *Leukemia* 26.5 (2012), pp. 934–942. URL: https://doi.org/10.1038/leu.2011.326.

[211] Hassan Awada *et al.* "Invariant phenotype and molecular association of biallelic TET2 mutant myeloid neoplasia". In: *Blood Advances* 3.3 (Feb. 2019), pp. 339–349. URL: https://doi.org/10.1182/bloodadvances.2018024216.

[212] Myunggon Ko *et al.* "Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice". In: *Proceedings of the National Academy of Sciences* 108.35 (Aug. 2011), pp. 14566–14571. URL: http://www.pnas.org/content/108/35/14566.abstract.

[213] Cyril Quivoron *et al.* "TET2 Inactivation Results in Pleiotropic Hematopoietic Abnormalities in Mouse and Is a Recurrent Event during Human Lymphomagenesis". In: *Cancer Cell* 20.1 (July 2011), pp. 25–38. URL: https://doi.org/10.1016/j.ccr.2011.06.003.

[214] Kelly Moran-Crusio *et al.* "TET2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation". In: *Cancer Cell* 20.1 (Nov. 2011), pp. 11–24. URL: https://doi.org/10.1016/j.ccr.2011.06.001.

[215] Zhe Li *et al.* "Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies". In: *Blood* 118.17 (Oct. 2011), pp. 4509–4518. URL: http://www.bloodjournal.org/content/118/17/4509.abstract.

[216] Elizabeth L Ostrander *et al.* "Divergent Effects of Dnmt3a and Tet2 Mutations on Hematopoietic Progenitor Cell Fitness". In: *Stem Cell Reports* 3.4 (Mar. 2020), pp. 539–547. URL: https://doi.org/10.1016/j.stemcr.2020.02.011.

[217] Elodie Pronier *et al.* "Inhibition of TET2-mediated conversion of 5-methylcytosine to 5-hydroxymethylcytosine disturbs erythroid and granulomonocytic differentiation of human hematopoietic progenitors". In: *Blood* 118.9 (Sept. 2011), pp. 2551–2555. URL: http://www.bloodjournal.org/content/118/9/2551.abstract.

[218] Xiaoli Qu *et al.* "TET2 deficiency leads to stem cell factor–dependent clonal expansion of dysfunctional erythroid progenitors". In: *Blood* 132.22 (Nov. 2018), pp. 2406–2417. URL: http://www.bloodjournal.org/content/132/22/2406.abstract.

[219] Kyoko Ito *et al.* "Non-catalytic Roles of Tet2 Are Essential to Regulate Hematopoietic Stem and Progenitor Cell Homeostasis". In: *Cell Reports* 28.10 (2019), 2480–2490.e4. URL: https://doi.org/10.1016/j.celrep.2019.07.094.

[220] Marlies Meisel *et al.* "Microbial signals drive pre-leukaemic myeloproliferation in a Tet2-deficient host". In: *Nature* 557.7706 (2018), pp. 580–584. URL: https://doi.org/10.1038/s41586-018-0125-z.

[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. URL: https://doi.org/10.1016/j.stem.2018.10.013.

[222] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760. URL: https://doi.org/10.1093/bioinformatics/btp324.

[223] Zhongwu Lai *et al.* "VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research". In: *Nucleic Acids Research* 44.11 (June 2016), e108–e108. URL: http://dx.doi.org/10.1093/nar/gkw227.

[224] Kristian Cibulskis *et al.* "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples". In: *Nature Biotechnology* 31.3 (Feb. 2013), p. 213. arXiv: NIHMS150003. URL: https://www.nature.com/articles/nbt.2514.

[225] Hui Yang and Kai Wang. "Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR". In: *Nature Protocols* 10 (Sept. 2015), p. 1556. URL: https://www.nature.com/articles/nprot.2015.105.

[226] Alba Rodriguez-Meira *et al.* "Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing". In: *Molecular Cell* 73.6 (Mar. 2019), 1292–1305.e8. URL: https://doi.org/10.1016/j.molcel.2019.01.009.

[227] Alba Rodriguez-Meira *et al.* "TARGET-Seq: A Protocol for High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing". In: *STAR Protocols* 1.3 (2020), p. 100125. URL: https://doi.org/10.1016/j.xpro.2020.100125.

[228] Andreas Untergasser *et al.* "Primer3–new capabilities and interfaces". eng. In: *Nucleic acids research* 40.15 (Aug. 2012), e115–e115. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3424584/.

[229] 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. URL: https://doi.org/10.1186/1471-2105-13-134.

[230] 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 (Sept. 2021), p. 109698. URL: https://doi.org/10.1016/j.celrep.2021.109698.

[231] Aaron TL 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. URL: https://doi.org/10.1186/s13059-016-0947-7.

[232] Tallulah S Andrews and Martin Hemberg. "M3Drop: dropout-based feature selection for scRNASeq". In: *Bioinformatics* 35.16 (Aug. 2019), pp. 2865–2867. URL: https://doi.org/10.1093/bioinformatics/bty1044.

[233] Philip Brennecke *et al.* "Accounting for technical noise in single-cell RNA-seq experiments". In: *Nature Methods* 10.11 (2013), pp. 1093–1095. URL: https://doi.org/10.1038/nmeth.2645.

[234] Etienne Becht *et al.* "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature Biotechnology* 37.1 (2019), pp. 38–44. URL: https://doi.org/10.1038/nbt.4314.

[235] Ilya Korsunsky *et al.* "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature Methods* 16.12 (2019), pp. 1289–1296. URL: http://dx.doi.org/10.1038/s41592-019-0619-0.

[236] Itay Tirosh *et al.* "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". In: *Science* 352.6282 (Apr. 2016), pp. 189–196. URL: https://doi.org/10.1126/science.aad0501.

[237] Gennady Korotkevich *et al.* "Fast gene set enrichment analysis". In: *bioRxiv* (Jan. 2021), p. 60012. URL: http://biorxiv.org/content/early/2021/02/01/060012.abstract.

[238] Lu Chen *et al.* "Transcriptional diversity during lineage commitment of human blood progenitors". en. In: *Science* 345.6204 (Sept. 2014), p. 1251033. URL: https://www.science.org/doi/10.1126/science.1251033.

[239] Lars Velten *et al.* "Human haematopoietic stem cell lineage commitment is a continuous process". In: *Nature Cell Biology* 19.4 (Apr. 2017), pp. 271–281. URL: http://dx.doi.org/10.1038/ncb3493.

[240] Dorin-Mirel Popescu *et al.* "Decoding human fetal liver haematopoiesis". In: *Nature* 574.7778 (2019), pp. 365–371. URL: https://doi.org/10.1038/s41586-019-1652-y.

[241] Roy Drissen *et al.* "Identification of two distinct pathways of human myelopoiesis". In: *Science Immunology* 4.35 (2019), eaau7148. URL: https://doi.org/10.1126/sciimmunol.aau7148.

[242] 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. URL: http://dx.doi.org/10.1038/s41590-017-0001-2.

[243] Wenjun Ju *et al.* "Defining cell-type specificity at the transcriptional level in human disease". In: *Genome Research* 23.11 (Nov. 2013), pp. 1862–1873. URL: http://genome.cshlp.org/content/23/11/1862.abstract.

[244] Bethan Psaila *et al.* "Single-Cell Analyses Reveal Megakaryocyte-Biased Hematopoiesis in Myelofibrosis and Identify Mutant Clone-Specific Targets". In: *Molecular Cell* 78.3 (May 2020), 477–492.e8. URL: https://doi.org/10.1016/j.molcel.2020.04.008.

[245] Noa Novershtern *et al.* "Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis". In: *Cell* 144.2 (Jan. 2011), pp. 296–309. URL: https://doi.org/10.1016/j.cell.2011.01.004.

[246] Emma Dann *et al.* "Differential abundance testing on single-cell data using k-nearest neighbor graphs". In: *Nature Biotechnology* 40 (2022), pp. 245–253. URL: https://doi.org/10.1038/s41587-021-01033-z.

[247] Kelly Street *et al.* "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC Genomics* 19.1 (2018), p. 477. URL: https://doi.org/10.1186/s12864-018-4772-0.

[248] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (June 2019), e8746. URL: https://doi.org/10.15252/msb.20188746.

[249] Beate Vieth *et al.* "A systematic evaluation of single cell RNA-seq analysis pipelines". In: *Nature Communications* 10.1 (2019), p. 4667. URL: https://doi.org/10.1038/s41467-019-12266-7.

[250] Matthew E Ritchie *et al.* "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (Apr. 2015), e47–e47. URL: https://doi.org/10.1093/nar/gkv007.

[251] Stanley W K Ng *et al.* "A 17-gene stemness score for rapid determination of risk in acute leukaemia". In: *Nature* 540.7633 (2016), pp. 433–437. URL: https://doi.org/10.1038/nature20598.

[252] Yoriko Saito *et al.* "Identification of Therapeutic Targets for Quiescent, Chemotherapy-Resistant Human Leukemia Stem Cells". In: *Science Translational Medicine* 2.17 (Feb. 2010), 17ra9–17ra9. URL: https://doi.org/10.1126/scitranslmed.3000349.

[253] Stephanie Z Xie *et al.* "Sphingolipid Modulation Activates Proteostasis Programs to Govern Human Hematopoietic Stem Cell Self-Renewal". In: *Cell Stem Cell* 25.5 (Oct. 2019), 639–653.e7. URL: https://doi.org/10.1016/j.stem.2019.09.008.

[254] Laura García-Prat *et al.* "TFEB-mediated endolysosomal activity controls human hematopoietic stem cell fate". In: *Cell Stem Cell* 28.10 (Aug. 2021), 1838–1850.e10. URL: https://doi.org/10.1016/j.stem.2021.07.003.

[255] Nina Cabezas-Wallscheid *et al.* "Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy". In: *Cell* 169.5 (2017), 807–823.e19. URL: http://dx.doi.org/10.1016/j.cell.2017.04.018.

[256] Alejo E Rodriguez-Fraticelli *et al.* "Clonal analysis of lineage fate in native haematopoiesis". In: *Nature* 553 (Jan. 2018), pp. 212–216. URL: https://dx.doi.org/10.1038/nature25168.

[257] 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. URL: https://doi.org/10.1038/s41556-018-0121-4.

[258] Weike Pei *et al.* "Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding". In: *Cell Stem Cell* 27 (Aug. 2020), pp. 383–395. URL: https://doi.org/10.1016/j.stem.2020.07.018.

[259] Sara Aibar *et al.* "SCENIC: single-cell regulatory network inference and clustering". In: *Nature Methods* 14.11 (2017), pp. 1083–1086. URL: https://doi.org/10.1038/nmeth.4463.

[260] Johann-Christoph Jann *et al.* "Comparative Analysis of Clonal Hematopoiesis of Multipotent Stem Cells in Healthy Elderly in Blood and Bone Marrow". In: *Leukemia Research* 82 (2019), pp. 15–18.

[261] Wendy W Pang *et al.* "Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age". In: *Proceedings of the National Academy of Sciences* 108.50 (Dec. 2011), pp. 20012–20017. URL: http://www.pnas.org/content/108/50/20012.abstract.

[262] Klaudia Kuranda *et al.* "Age-related changes in human hematopoietic stem/progenitor cells". In: *Aging Cell* 10.3 (2011), pp. 542–546.

[263] Simone Picelli *et al.* "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature Protocols* 9.1 (2014), pp. 171–181. arXiv: NIHMS150003. URL: https://www.nature.com/articles/nprot.2014.006.

[264] Anna S. Nam *et al.* "Somatic mutations and cell identity linked by Genotyping of Transcriptomes". In: *Nature* 571.7765 (2019), pp. 355–360. URL: https://doi.org/10.1038/s41586-019-1367-0.

[265] Michael Hagemann-Jensen *et al.* "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". In: *Nature Biotechnology* 38.6 (June 2020), pp. 708–714. URL: http://www.nature.com/articles/s41587-020-0497-0.

[266] Junjie U Guo and David P Bartel. "RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria". In: *Science* 353.6306 (Sept. 2016), aaf5371. URL: https://doi.org/10.1126/science.aaf5371.

[267] Alice Giustacchini *et al.* "Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia". In: *Nature Medicine* 23.6 (June 2017), pp. 692–702. URL: http://dx.doi.org/10.1038/nm.4336.

[268] Rahul Sinha *et al.* "Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing". In: *bioRxiv* (Jan. 2017), p. 125724. URL: http://biorxiv.org/content/early/2017/04/09/125724.abstract.

[269] Maura Costello *et al.* "Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms". In: *BMC Genomics* 19.1 (2018), p. 332. URL: https://doi.org/10.1186/s12864-018-4703-0.

[270] Christoph Ziegenhain *et al.* "Comparative Analysis of Single-Cell RNA Sequencing Methods". In: *Molecular Cell* 65.4 (2017), 631–643.e4. URL: https://doi.org/10.1016/j.molcel.2017.01.023.

[271] Max Jan, Benjamin L Ebert, and Siddhartha Jaiswal. "Clonal hematopoiesis". In: *Seminars in Hematology* 54.1 (Jan. 2017), pp. 43–50. URL: https://doi.org/10.1053/j.seminhematol.2016.10.002.

[272] F. Notta *et al.* "Isolation of Single Human Hematopoietic Stem Cells Capable of Long-Term Multilineage Engraftment". In: *Science* 333.6039 (July 2011), pp. 218–221. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1201219.

[273] T. Tickle *et al. inferCNV of the Trinity CTAT Project.* 2019. URL: https://github.com/broadinstitute/inferCNV.

[274] Kristian Reckzeh *et al.* "Human adult HSCs can be discriminated from lineage-committed HPCs by the expression of endomucin". In: *Blood Advances* 2.13 (July 2018), pp. 1628–1632. URL: https://doi.org/10.1182/bloodadvances.2018015743.

[275] Pia Sommerkamp *et al.* "The long non-coding RNA Meg3 is dispensable for hematopoietic stem cells". In: *Scientific Reports* 9.1 (2019), p. 2110. URL: https://doi.org/10.1038/s41598-019-38605-8.

[276] Bernhard Lehnertz *et al.* "HLF expression defines the human hematopoietic stem cell state". In: *Blood* 138.25 (Dec. 2021), pp. 2642–2654. URL: https://doi.org/10.1182/blood.2021010745.

[277] Christopher Maximilian Arends *et al.* "Hematopoietic lineage distribution and evolutionary dynamics of clonal hematopoiesis". In: *Leukemia* 32 (2018), pp. 1908–1919. URL: https://doi.org/10.1038/s41375-018-0047-7.

[278] Manuel Buscarlet *et al.* "Lineage restriction analyses in CHIP indicate myeloid bias for TET2 and multipotent stem cell origin for DNMT3A". In: *Blood* 132.3 (2018), pp. 277–280. URL: https://doi.org/10.1182/blood-2018-01-829937.

[279] Kerstin B Kaufmann *et al.* "A latent subset of human hematopoietic stem cells resists regenerative stress to preserve stemness". In: *Nature Immunology* 22.6 (2021), pp. 723–734. URL: https://doi.org/10.1038/s41590-021-00925-1.

[280] Denis Clay *et al.* "CD9 and megakaryocyte differentiation". In: *Blood* 97.7 (Apr. 2001), pp. 1982–1989. URL: https://doi.org/10.1182/blood.V97.7.1982.

[281] Elisa Laurenti *et al.* "CDK6 Levels Regulate Quiescence Exit in Human Hematopoietic Stem Cells". In: *Cell Stem Cell* 16.3 (Mar. 2015), pp. 302–313. URL: http://dx.doi.org/10.1016/j.stem.2015.01.017.

[282] Matej Fabisik *et al. Regulation of Inflammatory Response by Transmembrane Adaptor Protein LST1*. 2021. URL: https://www.frontiersin.org/article/10.3389/fimmu.2021.618332.

[283] Qi Liu *et al.* "Overexpression of DOC-1R Inhibits Cell Cycle G1/S Transition by Repressing CDK2 Expression and Activation". In: *International Journal of Biological Sciences* 9.6 (2013), pp. 541–549. URL: https://www.ijbs.com/v09p0541.htm.

[284] Claudio Mauro *et al.* "ABIN-1 Binds to NEMO/IKKγ and Co-operates with A20 in Inhibiting NF-κB*". In: *Journal of Biological Chemistry* 281.27 (July 2006), pp. 18482–18488. URL: https://doi.org/10.1074/jbc.M601502200.

[285] S Kinoshita, S Akira, and T Kishimoto. "A member of the C/EBP family, NF-IL6 beta, forms a heterodimer and transcriptionally synergizes with NF-IL6." In: *Proceedings of the National Academy of Sciences* 89.4 (Feb. 1992), pp. 1473–1476. URL: https://doi.org/10.1073/pnas.89.4.1473.

[286] Qian-fei Wang and Alan D Friedman. "CCAAT/enhancer-binding proteins are required for granulopoiesis independent of their induction of the granulocyte colony-stimulating factor receptor". In: *Blood* 99.8 (Apr. 2002), pp. 2776–2785. URL: https://doi.org/10.1182/blood.V99.8.2776.

[287] Shuchi Agrawal *et al.* "The C/EBPδ tumor suppressor is silenced by hypermethylation in acute myeloid leukemia". In: *Blood* 109.9 (Jan. 2007), pp. 3895–3905. URL: https://doi.org/10.1182/blood-2006-08-040147.

[288] Mark W Feinberg *et al.* "The Kruppel-like factor KLF4 is a critical regulator of monocyte differentiation". In: *The EMBO Journal* 26.18 (Sept. 2007), pp. 4138–4148. URL: https://doi.org/10.1038/sj.emboj.7601824.

[289] Dorine Sichien *et al.* "IRF8 Transcription Factor Controls Survival and Function of Terminally Differentiated Conventional and Plasmacytoid Dendritic Cells, Respectively". In: *Immunity* 45.3 (2016), pp. 626–640. URL: https://www.sciencedirect.com/science/article/pii/S1074761316303375.

[290] Felicia Kathrine Bratt Lauridsen *et al.* "Differences in Cell Cycle Status Underlie Transcriptional Heterogeneity in the HSC Compartment". In: *Cell Reports* 24.3 (July 2018), pp. 766–780. URL: https://doi.org/10.1016/j.celrep.2018.06.057.

[291] Harold S Bernstein and Shaun R Coughlin. "A Mammalian Homolog of Fission Yeast Cdc5 Regulates G2 Progression and Mitotic Entry". In: *Journal of Biological Chemistry* 273.8 (1998), pp. 4666–4671. URL: https://doi.org/10.1074/jbc.273.8.4666.

[292] Michael Grote *et al.* "Molecular Architecture of the Human Prp19/CDC5L Complex". In: *Molecular and Cellular Biology* 30.9 (May 2010), pp. 2105–2119. URL: https://doi.org/10.1128/MCB.01505-09.

[293] Irene M Min *et al.* "The Transcription Factor EGR1 Controls Both the Proliferation and Localization of Hematopoietic Stem Cells". In: *Cell Stem Cell* 2.4 (Apr. 2008), pp. 380–391. URL: https://doi.org/10.1016/j.stem.2008.01.015.

[294] Nobuyuki Kakiuchi and Seishi Ogawa. "Clonal expansion in non-cancer tissues". In: *Nature Reviews Cancer* 21.4 (2021), pp. 239–256. URL: https://doi.org/10.1038/s41568-021-00335-3.

[295] L G Schuettpelz *et al.* "G-CSF regulates hematopoietic stem cell activity, in part, through activation of Toll-like receptor signaling". In: *Leukemia* 28.9 (2014), pp. 1851–1860. URL: https://doi.org/10.1038/leu.2014.68.

[296] Dagmar Walter *et al.* "Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells". In: *Nature* 520.7548 (Apr. 2015), pp. 549–552. URL: http://dx.doi.org/10.1038/nature14131.

[297] Corinna C Pedersen *et al.* "Changes in Gene Expression during G-CSF–Induced Emergency Granulopoiesis in Humans". In: *The Journal of Immunology* 197.5 (Sept. 2016), pp. 1989–1999. URL: http://www.jimmunol.org/content/197/5/1989.abstract.

[298] Jeffrey M Bernitz *et al.* "Granulocyte colony-stimulating factor mobilizes dormant hematopoietic stem cells without proliferation in mice". In: *Blood* 129.14 (Apr. 2017), pp. 1901–1912. URL: https://doi.org/10.1182/blood-2016-11-752923.

[299] C Craddock *et al.* "Azacitidine fails to eradicate leukemic stem/progenitor cell populations in patients with acute myeloid leukemia and myelodysplasia". In: *Leukemia* 27.5 (Apr. 2013), pp. 1028–1036. URL: http://dx.doi.org/10.1038/leu.2012.312%20http://10.0.4.14/leu.2012.312.

[300] Subhashisa Swain *et al.* "Comorbidities in Osteoarthritis : A Systematic Review and Analysis of Observational Studies". In: *Arthritis Care & Research* 72.7 (2020), pp. 991–1000. URL: https://doi.org/10.1002/acr.24008.

[301] Britta Will *et al.* "Stem and progenitor cells in myelodysplastic syndromes show aberrant stage-specific expansion and harbor genetic and epigenetic alterations". In: *Blood* 120.10 (Sept. 2012), pp. 2076–2086. URL: https://doi.org/10.1182/blood-2011-12-399683.

[302] Wendy W Pang *et al.* "Hematopoietic stem cell and progenitor cell mechanisms in myelodysplastic syndromes". In: *Proceedings of the National Academy of Sciences* 110.8 (Feb. 2013), pp. 3011–3016. URL: https://doi.org/10.1073/pnas.1222861110.

[303] Koen Van den Berge *et al.* "Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications". In: *Genome Biology* 19.1 (2018), p. 24. URL: https://doi.org/10.1186/s13059-018-1406-4.

[304] Shiwei Zheng *et al.* "Molecular transitions in early progenitors during human cord blood hematopoiesis". In: *Molecular Systems Biology* 14.3 (Mar. 2018), e8041. URL: https://doi.org/10.15252/msb.20178041.

[305] Johannes W Bagnoli *et al.* "Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq". In: *Nature Communications* 9.1 (2018), p. 2937. URL: https://doi.org/10.1038/s41467-018-05347-6.

[306] Alejo E Rodriguez-Fraticelli *et al.* "Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis". In: *Nature* 583.7817 (2020), pp. 585–589. URL: https://doi.org/10.1038/s41586-020-2503-6.

[307] Matthias Farlik *et al.* "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation". In: *Cell Stem Cell* 19.6 (Nov. 2016), pp. 808–822. URL: https://doi.org/10.1016/j.stem.2016.10.019.

[308] S. Haihua Chu *et al.* "FLT3-ITD Knockin Impairs Hematopoietic Stem Cell Quiescence/Homeostasis, Leading to Myeloproliferative Neoplasm". In: *Cell Stem Cell* 11.3 (2012), pp. 346–358. URL: https://doi.org/10.1016/j.stem.2012.05.027.

[309] James A Saba *et al.* "Translational control of stem cell function". In: *Nature Reviews Molecular Cell Biology* 22.10 (2021), pp. 671–690. URL: https://doi.org/10.1038/s41580-021-00386-2.

[310] Loïc Garçon *et al.* "Ribosomal and hematopoietic defects in induced pluripotent stem cells derived from Diamond Blackfan anemia patients". In: *Blood* 122.6 (Aug. 2013), pp. 912–921. URL: https://doi.org/10.1182/blood-2013-01-478321.

[311] Rajiv K Khajuria *et al.* "Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis". In: *Cell* 173.1 (2018), 90–103.e19. URL: https://doi.org/10.1016/j.cell.2018.02.036.

[312] Betsabeh Khoramian Tusi *et al.* "Population snapshots predict early haematopoietic and erythroid hierarchies". In: *Nature* 555.7694 (2018), pp. 54–60. URL: http://dx.doi.org/10.1038/nature25741.

[313] Fong-Ying Tsai *et al.* "An early haematopoietic defect in mice lacking the transcription factor GATA-2". In: *Nature* 371.6494 (1994), pp. 221–226. URL: https://doi.org/10.1038/371221a0.

[314] Jonathan W Snow *et al.* "A Single cis Element Maintains Repression of the Key Developmental Regulator Gata2". In: *PLOS Genetics* 6.9 (Sept. 2010), e1001103. URL: https://doi.org/10.1371/journal.pgen.1001103.

[315] Juan Bautista Menendez-Gonzalez *et al.* "Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia". In: *Stem Cell Reports* 13.2 (2019), pp. 291–306. URL: https://doi.org/10.1016/j.stemcr.2019.07.005.

[316] Adam J Mead and Ann Mullally. "Myeloproliferative neoplasm stem cells". In: *Blood* 129.12 (Mar. 2017), 1607 LP –1616. URL: http://www.bloodjournal.org/content/129/12/1607.abstract.

[317] Y-I Chang *et al.* "Loss of Dnmt3a and endogenous KrasG12D/+ cooperate to regulate hematopoietic stem and progenitor cell functions in leukemogenesis". In: *Leukemia* 29.9 (Sept. 2015), pp. 1847–1856. URL: http://dx.doi.org/10.1038/leu.2015.85.

[318] Jianming Jiang *et al.* "A core Klf circuitry regulates self-renewal of embryonic stem cells". In: *Nature Cell Biology* 10.3 (2008), pp. 353–360. URL: https://doi.org/10.1038/ncb1698.

[319] Anja Schuetz *et al.* "The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation". In: *Cellular and Molecular Life Sciences* 68.18 (2011), pp. 3121–3131. URL: https://doi.org/10.1007/s00018-010-0618-x.

[320] Emmalee R. Adelman *et al.* "Aging Human Hematopoietic Stem Cells Manifest Profound Epigenetic Reprogramming of Enhancers That May Predispose to Leukemia". In: *Cancer Discovery* 9.8 (Aug. 2019), pp. 1080–1101. URL: https://doi.org/10.1158/2159-8290.CD-18-1474.

[321] Atsuo Okamura *et al.* "Involvement of casein kinase Iϵ in cytokine-induced granulocytic differentiation". In: *Blood* 103.8 (Apr. 2004), pp. 2997–3004. URL: https://doi.org/10.1182/blood-2003-08-2768.

[322] Kim Jeong-Ho *et al.* "CREB coactivators CRTC2 and CRTC3 modulate bone marrow hematopoiesis". In: *Proceedings of the National Academy of Sciences* 114.44 (Oct. 2017), pp. 11739–11744. URL: https://doi.org/10.1073/pnas.1712616114.

[323] Hanae Sato *et al.* "DNMT3A and TET2 in the Pre-Leukemic Phase of Hematopoietic Disorders". In: *Frontiers in Oncology* 6 (2016), p. 187. URL: https://doi.org/10.3389/fonc.2016.00187.

[324] Jennifer M. SanMiguel, Kira Young, and Jennifer J. Trowbridge. "Hand in hand: intrinsic and extrinsic drivers of aging and clonal hematopoiesis". In: *Experimental Hematology* 91 (2020), pp. 1–9. URL: https://doi.org/10.1016/j.exphem.2020.09.197.