

GENETIC AND MOLECULAR BASIS OF MYELOID NEOPLASMS AT SINGLE CELL RESOLUTION

Alba Rodriguez Meira

A Thesis submitted for the Degree of Philosophy

Trinity Term 2019



Keble College

University of Oxford

ABSTRACT

Myeloproliferative neoplasms (MPN) originate in a single human haematopoietic stem cell which acquires a somatic mutation and undergoes clonal expansion. This initial clone usually accumulates further genetic lesions during disease progression, giving rise to a genetically heterogeneous population of tumour cells, which are also molecularly and functionally distinct. This makes it challenging to effectively target all cancer cells in a particular patient, and as a consequence, currently available drug therapies fail to completely eliminate clonal haematopoiesis.

Therefore, there is an urgent need to develop new therapies specifically targeting the mutant clone. To do so, it is essential to characterize clonal diversity in MPN at the genetic and molecular level. Single cell genomic techniques are ideally placed to resolve such heterogeneity. However, the lack of technologies integrating genetic and transcriptional readouts from the same single cell has limited their application to study intratumoral heterogeneity.

To overcome such limitation, I first developed TARGET-seq, a single cell multi-omic method for the high-sensitivity detection of mutations within single cells in parallel with whole transcriptome analysis and cell surface proteomics.

Following extensive validation in cell lines, I applied TARGET-seq to study haematopoietic stem and progenitor cells (HSPCs) from patients with myeloproliferative neoplasms. TARGET-seq analysis revealed a high degree of genetic heterogeneity, identifying both linear and branching patterns of clonal evolution. Different genetic subclones had distinct transcriptional signatures, with subclones carrying poor prognosis collaborating mutations showing increased transcriptional diversity as compared to clones carrying exclusively mutations in *JAK2*. Wild-type cells from MPN patients were enriched in inflammatory signatures as compared to cells from

normal donors, suggesting a cell-extrinsic effect disrupting gene expression in non-mutant cells. This could underly therapy resistance or influence the development of secondary haematological malignancies. Moreover, TARGET-seq analysis identified putative biomarkers of *JAK2*-mutant cells, which could be used for disease monitoring and are potential targets for immunotherapy.

I then analysed HSPCs from MPN patients undergoing disease transformation to secondary Acute Myeloid Leukaemia (sAML), a poor prognosis complication affecting up to 20 % of MPN patients. TARGET-seq analysis identified leukemic subclones emerging from haematopoietic stem cells rather than more mature progenitors, in contrast to findings in *de novo* AML. The acquisition of *TP53* mutations, associated with disease transformation, led to chromosomal aberrations and highly-complex patterns of clonal evolution. This supports a model in which acquisition of *TP53* mutations drive leukemogenesis by promoting increased genetic instability and molecular heterogeneity in HSPCs.

In summary, single-cell multi-omic analysis allowed me to identify distinct and biologically relevant molecular signatures of different genetic subclones in HSPCs from patients with myeloid neoplasms. Ultimately, this analysis could potentially guide the development of new therapeutic approaches to specifically target disease propagating cells.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Adam J Mead for the endless support and mentorship received during my DPhil. I am extremely grateful for all the guidance and opportunities I received during my time at his laboratory. It was a pleasure to have the opportunity to work in such an exciting scientific atmosphere, surrounded by inspiring scientists and where I had the freedom to put into practice many ideas that I thought would never be possible.

I would also like to thank Dr. Supat Thongjuea for giving me the opportunity to join his group and teaching me more bioinformatics than I ever thought I would be able to learn. I am extremely thankful to the many advisors and collaborators who provided crucial guidance and support for my project, specially my co-supervisor Professor Sten Eirik Jacobsen, and graduate advisors Professor Douglas Higgs and Professor Tatjana Sauka-Spengler.

I am also very grateful to my colleagues at the Weatherall Institute of Molecular Medicine, specially the Mead, Jacobsen, Thongjuea and Nerlov laboratories for advice and help with experiments or analysis, and their friendship over the years.

Last but not least, I would like to thank my fiancée, Miguel Cacho Soblechero, who very patiently stood by me throughout many stressful times (and still decided to marry me), my friend Aitor, who taught me all I know about how to make good scientific figures, and my family, without whom I would have never had the opportunity to do a DPhil at the University of Oxford.

DECLARATION

All experiments and analysis in this thesis were designed and performed by myself, with contributions of certain colleagues stated in the relevant sections. Part of the work in this thesis has previously been published as part of a research article (Rodriguez-Meira *et al*, *Molecular Cell*, 2019) (1) and therefore, certain figures presented throughout this thesis are adapted from this publication. When this is the case, it is indicated in the relevant figure legends.

TABLE OF CONTENTS

ABSTRACT	2
ACKNOWLEDGEMENTS	4
DECLARATION	5
LIST OF FIGURES	10
LIST OF TABLES	14
COMMONLY USED ABBREVIATIONS	16
CHAPTER 1: INTRODUCTION	20
1. THE HAEMATOPOIETIC SYSTEM AND HAEMATOPOIETIC STEM CELLS	20
2. MYELOPROLIFERATIVE NEOPLASMS	24
2.1. CLASSIFICATION OF MYELOPROLIFERATIVE NEOPLASMS	24
2.2. PHILADELPHIA-NEGATIVE MYELOPROLIFERATIVE NEOPLASMS	25
2.3. THE MUTATIONAL SPECTRUM OF MYELOPROLIFERATIVE NEOPLASMS	25
2.4. DISEASE-INITIATING CELLS IN MYELOPROLIFERATIVE NEOPLASMS .	35
2.5. TREATMENT OF MYELOPROLIFERATIVE NEOPLASMS	36
3. ACUTE MYELOID LEUKAEMIA	38
3.1. CLASSIFICATION OF ACUTE MYELOID LEUKAEMIA	38
3.2. GENETIC LANDSCAPE OF ACUTE MYELOID LEUKAEMIA	40
3.3. DISEASE-INITIATING CELLS AND CLONAL EVOLUTION IN ACUTE MYELOID LEUKAEMIA	42
3.4. TREATMENT OF ACUTE MYELOID LEUKAEMIA	45
3.5. SECONDARY ACUTE MYELOID LEUKAEMIA	47
3.6. <i>TP53</i> MUTANT SECONDARY ACUTE MYELOID LEUKAEMIA	51
4. SINGLE CELL TECHNOLOGIES AND THEIR APPLICATION TO STUDY NORMAL AND MALIGNANT HAEMATOPOIESIS	56
4.1. SINGLE CELL TRANSCRIPTOMIC TECHNIQUES	58
4.2. SINGLE CELL GENOMIC TECHNIQUES	66
4.3. SINGLE CELL MULTI-OMIC TECHNIQUES	68
4.4. APPLICATION OF SINGLE CELL TECHNOLOGIES TO STUDY THE HAEMATOPOIETIC SYSTEM	71
5. AIMS OF THIS WORK	73

CHAPTER 2: MATERIALS AND METHODS	75
1. BANKING AND PROCESSING OF HUMAN SAMPLES.....	75
2. CELL LINES	76
3. GENOMIC DNA EXTRACTION AND BULK SEQUENCING OF MONONUCLEAR CELLS.....	76
4. BULK KARYOTYPING	78
5. SAMPLE STAINING, FLUORESCENT ACTIVATED CELL SORTING (FACS) AND SINGLE-CELL ISOLATION.....	79
6. LYSIS BUFFER PREPARATION AND SINGLE CELL CDNA SYNTHESIS	83
7. TARGETED NGS SINGLE-CELL GENOTYPING	87
8. NEXTERA XT LIBRARY PREPARATION AND WHOLE TRANSCRIPTOME SEQUENCING	91
9. ASSAY FOR TRANSPOSASE-ACCESSIBLE CHROMATIN AND SEQUENCING (ATAC-SEQ).....	94
10. FULL-LENGTH SINGLE CELL RNA-SEQUENCING DATA PRE- PROCESSING	97
11. 3'-BIASED SINGLE CELL RNA-SEQUENCING DATA PRE- PROCESSING	98
12. MIXED HUMAN:MOUSE SINGLE CELL RNA-SEQUENCING DATA PRE- PROCESSING	99
13. WHOLE TRANSCRIPTOME VARIANT CALLING FROM SINGLE CELLS	99
14. DROPOUT FREQUENCY AND LIBRARY BIAS CALCULATION.....	100
15. TRANSCRIPT COVERAGE	100
16. COPY NUMBER VARIATION INFERENCE FROM SINGLE CELLS.....	100
17. CELL-TO-CELL CORRELATION ANALYSIS	101
18. COMPUTATIONAL RECONSTRUCTION OF CLONAL HIERARCHIES	101
19. DIFFERENTIAL EXPRESSION ANALYSIS	101
20. DIFFERENTIAL SPLICING ANALYSIS IN SINGLE CELLS	102
21. GENE-SET ENRICHMENT ANALYSIS	102
22. IDENTIFICATION OF HIGHLY VARIABLE GENES	102
23. BATCH CORRECTION	103
24. CELL CYCLE SCORE CALCULATION AND CORRECTION.....	103
25. SINGLE CELL CLUSTERING AND DIMENSIONALITY REDUCTION...	103

26. RANDOM FOREST ANALYSIS.....	104
27. FORCE-DIRECTED GRAPH ANALYSIS	105
28. CHROMATIN ACCESSIBILITY ANALYSIS.....	105
29. STATISTICAL ANALYSIS	106
CHAPTER 3: DEVELOPMENT OF TARGET-SEQ	107
1. INTRODUCTION.....	107
1.1. STATE OF THE ART SINGLE-CELL TECHNOLOGIES: OVERVIEW AND LIMITATIONS.....	108
1.2. HYPOTHESIS AND AIMS.....	110
2. RESULTS	111
2.1. OPTIMIZATION OF SINGLE-CELL RNA-SEQUENCING CONDITIONS....	111
2.2. DEVELOPMENT OF TARGET-SEQ.....	115
2.3. DEVELOPMENT OF A BIOINFORMATICS PIPELINE FOR SINGLE-CELL MUTATION ANALYSIS	121
2.4. VALIDATION OF TARGET-SEQ IN CELL LINES AND PRIMARY HAEMATOPOIETIC STEM CELLS.....	123
2.5. DEVELOPMENT OF A HIGH-THROUGHPUT, 3'-BIASED WHOLE TRANSCRIPTOME TARGET-SEQ METHOD	127
2.6. DEVELOPMENT OF A HIGH-THROUGHPUT SINGLE-CELL GENOTYPING METHOD.....	131
3. DISCUSSION	135
CHAPTER 4: SINGLE CELL ANALYSIS OF STEM AND PROGENITOR CELLS IN MYELOPROLIFERATIVE NEOPLASMS.....	139
1. INTRODUCTION	139
1.1. HYPOTHESIS AND AIMS.....	139
2. RESULTS	141
2.1. GENETIC AND TRANSCRIPTOMIC HETEROGENEITY OF <i>JAK2</i> -MUTANT MYELOPROLIFERATIVE NEOPLASMS	141
2.2. GENETIC AND TRANSCRIPTOMIC HETEROGENEITY IN <i>U2AF1</i> MUTANT MYELOPROLIFERATIVE NEOPLASMS	148
2.3. HIGH-THROUGHPUT TARGET-SEQ ANALYSIS OF HSPCs IN <i>JAK2</i> -MUTANT MYELOFIBROSIS	154
3. DISCUSSION	170
CHAPTER 5: SINGLE CELL CHARACTERIZATION OF LEUKEMIC STEM CELL POPULATIONS IN SECONDARY ACUTE MYELOID LEUKAEMIA ..	173
1. INTRODUCTION	173

1.1 HYPOTHESIS AND AIMS	174
2. RESULTS	175
2.1. GENETIC CHARACTERIZATION OF <i>TP53</i> MUTANT SECONDARY ACUTE MYELOID LEUKAEMIA.....	175
2.2. CLONAL ARCHITECTURE OF <i>TP53</i> MUTANT SECONDARY ACUTE MYELOID LEUKAEMIA.....	178
2.3. TRANSCRIPTOMIC AND EPIGENOMIC CHARACTERIZATION OF PRE-LEUKEMIC AND LEUKEMIC SUBCLONES	184
3. DISCUSSION	187
CHAPTER 6: CONCLUSIONS AND FUTURE WORK.....	190
6.1. IMPLICATIONS OF THIS WORK	190
6.2. LIMITATIONS OF THIS WORK	191
6.3. FUTURE WORK.....	192
BIBLIOGRAPHY	197
APPENDIX 1	221
BARCODED OLIGO-DT PRIMERS USED FOR 3'-TARGET-SEQ	221
APPENDIX 2	233
PRIMERS USED IN THE PRE-AMPLIFICATION OF GDNA, MRNA AND CDNA AMPLICONS DURING RT AND PCR STEPS	233
APPENDIX 3	237
TARGET-SPECIFIC BARCODING PRIMERS USED FOR FIRST BARCODING PCR, TO GENERATE SINGLE-CELL TARGETED GENOTYPING LIBRARIES	237

LIST OF FIGURES

Figure 1.1. Schematic representation of the haematopoietic hierarchy.

Figure 1.2. Schematic representation of the continuous differentiation model of haematopoiesis, based on Waddington's landscape model.

Figure 1.3. Schematic diagram of the JAK/STAT signalling pathway through homodimeric cytokine receptors (EPOR, MPL, GCSFR).

Figure 1.4. Schematic representation of the domain structure of JAK2, CALR and MPL.

Figure 1.5. *U2AF1* mutations affect the zinc finger domain of the protein, responsible for binding to canonical 3' splice sites.

Figure 1.6. Two-step model of leukemogenesis.

Figure 1.7. Schematic overview of the P53 pathway.

Figure 1.8. Schematic representation of p53 domains and recurrently mutated residues in myeloid malignancies.

Figure 1.9. Single-cell sequencing unravels masked heterogeneity not captured by bulk sequencing techniques.

Figure 1.10. Schematic workflow representation of the most widely used single-cell RNA-sequencing techniques.

Figure 1.11. Generation of full-length cDNA molecules using a Template Switching Oligonucleotide and a MMLV-derived RTase.

Figure 1.12. Sequencing library preparation methods and sequencing configurations for full-length and 3'-biased or 5'-biased approaches.

Figure 2.1. Gating strategy for HSPC isolation.

Figure 2.2. Schematic representation of targeted library preparation for single-cell genotyping.

Figure 2.3. Representative traces of tagmented and amplified ATAC-seq libraries.

Figure 3.1. SMART-seq plus improves the quality of cDNA libraries from single K562 cells.

Figure 3.2. Single-cell RNA-sequencing of primary HSCPs reveals high levels of allelic dropout.

Figure 3.3. Optimization of proteinase-based lysis conditions.

Figure 3.4. Schematic representation of full-length TARGET-seq method.

Figure 3.5. Bioinformatics pipeline for TARGET-seq single-cell mutation analysis.

Figure 3.6. TARGET-seq achieves high sensitivity mutation detection in single cells.

Figure 3.7. TARGET-seq produces high quality and unbiased whole-transcriptome libraries from single cells.

Figure 3.8. Schematic representation of 3'-TARGET-seq.

Figure 3.9. Technical performance of 3'-TARGET-seq.

Figure 3.10. Schematic representation of the barcoding strategy for high-throughput single-cell genotyping.

Figure 3.11. Technical performance of high-throughput single-cell genotyping.

Figure 4.1. TARGET-seq reveals genetic heterogeneity in the HSPC compartment of patients with myeloproliferative neoplasms.

Figure 4.2. TARGET-seq reveals transcriptional heterogeneity of *JAK2V617F* mutant subclones in MPN patients.

Figure 4.3. TARGET-seq identifies potential biomarkers of *JAK2V617F* mutant cells.

Figure 4.4. TARGET-seq reveals distinct transcriptional signatures associated with the presence of absence of somatic mutations in HSPCs from MPN patients.

Figure 4.5. Genetic architecture of *U2AF1*-mutant myeloproliferative neoplasms.

Figure 4.6. Transcriptional effects of *U2AF1* mutations in MPN patients.

Figure 4.7. Genetic subclones and patterns of clonal evolution from patients with myelofibrosis, identified using TARGET-seq single-cell genotyping.

Figure 4.8. Genetic subclones identified in patient IF0137 using mRNA targeting or TARGET-seq methods.

Figure 4.9. High-throughput 3'-biased TARGET-seq reveals molecular signatures of HSPCs from myelofibrosis patients.

Figure 4.10. Lineage differentiation trajectories in HSCPs from myelofibrosis patients and normal donors.

Figure 4.11. Wild-type cells from patients and normal donors are transcriptionally different.

Figure 4.12. *JAK2V617F* heterozygous and homozygous subclones present distinct molecular signatures.

Figure 4.13. Genetic subclones within individual patients present distinct molecular signatures.

Figure 4.14. Dimensionality reduction and unsupervised clustering fail to discriminate different genetic subclones in individual patients.

Figure 5.1. Summary of genetic alterations found in *TP53*-mutant secondary Acute Myeloid Leukaemia patients.

Figure 5.2. Immunophenotypic characterization of MPN-BP patient samples.

Figure 5.3. Clonal hierarchies of HSPCs from MPN-BP patients.

Figure 5.4. Detection of cytogenetic abnormalities in single cells using single-cell transcriptomics.

Figure 5.5. Molecular pathways dysregulated in HSPCs from MPN-BP patients.

Figure 5.6. Schematic representation of MPN development and transformation to secondary Acute Myeloid Leukaemia upon acquisition of *TP53* mutations.

LIST OF TABLES

Table 1.1. Immunophenotypic definition of different haematopoietic cell types in humans.

Table 1.2. Classification of the most common myeloproliferative neoplasms according to Philadelphia Chromosome status.

Table 1.3. Classification of Acute Myeloid Leukaemia according to the 2016 revision of the WHO guidelines.

Table 1.4. Classification of AML based on the 2017 European Leukaemia Net risk stratification.

Table 2.1. Genes and exons covered for targeted bulk sequencing of myeloid hotspots.

Table 2.2. Antibodies used for flow cytometry and fluorescent-activated cell sorting.

Table 2.3. Reagents used for lysis buffer preparation for SMART-seq2, SMART-seq plus and TARGET-seq methods.

Table 2.4. Reagents used for retrotranscription SMART-seq plus, mRNA targeting and TARGET-seq libraries.

Table 2.5. Thermocycler conditions used for retrotranscription of SMART-seq2, SMART-seq plus, mRNA targeting and TARGET-seq libraries.

Table 2.6. Reagents used for PCR amplification of SMART-seq2, SMART-seq plus, mRNA targeting and TARGET-seq libraries.

Table 2.7. Conditions used for PCR amplification of SMART-seq2, SMART-seq plus, mRNA targeting and TARGET-seq libraries.

Table 2.8. Reagents used for genotyping-PCR1 amplification.

Table 2.9. Conditions for target-specific amplification during genotyping-PCR1.

Table 2.10. Reagents used for genotyping-PCR2 amplification.

Table 2.11. Conditions used for genotyping-PCR2 amplification.

Table 2.12. Reagents used for PCR amplification of tagmented 3'-TARGET-seq libraries.

Table 2.13. Conditions used for PCR amplification of tagmented 3'-TARGET-seq libraries.

Table 2.14. Reagents used for lysis buffer preparation for ATAC-seq libraries.

Table 2.15. Primer sequences used to amplify ATAC-seq libraries during PCR.

Table 2.16. Reagents used for PCR amplification of transposed DNA to prepare ATAC-seq libraries.

Table 3.1. Summary of different Template Switching Oligonucleotides (TSO), retrotranscription and PCR conditions used for cDNA synthesis of single K562 cells.

Table 4.1. Summary of MPN patients used for full-length TARGET-seq analysis presented in chapter 4 section 2.1, including mutations identified in each patient and clinical characteristics.

Table 4.2. Summary of patients with myelofibrosis analysed in chapter 4 section 2.3, including mutations detected by bulk sequencing for each patient and clinical characteristics.

COMMONLY USED ABBREVIATIONS

2HG : (R)-2-hydroxyglutarate

5'AZA : 5' azacytidine

A3SS : Alternative 3' Splice Sites

A5SS : Alternative 5' Splice Sites

Abn: abnormal

ADC : antibody-drug conjugates

ADO : Allelic Dropout

ANOVA : Analysis of Variance

AML : Acute Myeloid Leukaemia

Amp : amplification

ASXL1: Additional Sex Combs-Like 1

ATAC-seq: Assay for Transposase-Accessible Chromatin and Sequencing

BAM : Binary Alignment Map

BCR-ABL1: Breakpoint Cluster Region Abelson tyrosine-protein kinase 1

BM: Bone Marrow

Bp : base pair

BWA : Burrows-Wheeler Aligner

CALR: calreticulin

CB : Cell Barcode

CBL: Casitas B-lineage Lymphoma

CHIP : Clonal Haematopoiesis of Indeterminate Potential

CML : Chronic Myeloid Leukaemia

CMML : Chronic Myelomonocytic Leukaemia

CMP : Common Myeloid Progenitor

CNV: Copy Number Variation

Decitabine: 5'aza-2'-deoxycytidine

Del: deletion

DNMT3A: DNA (cytosine-5)-methyltransferase 3A

EPOR : Erythropoietin Receptor

ER : Endoplasmic Reticulum

ERCC : External RNA Controls Consortium

ET : Essential Thrombocythemia

EZH2 : Enhancer of Zeste Homolog 2

FACS: Fluorescent Activated Cell Sorting
FLT3 : fms-like tyrosine kinase 3
FLT3 -ITD : fms-like tyrosine kinase 3 Internal Tandem Repeat
FDA : Food and Drug Administration
FERM : 4.1 protein ezrin radixin moesin
FMO : Fluorescence Minus One
GATK : Genome Analysis Toolkit
GCSFR : Granulocyte Colony Stimulating Factor Receptor
gDNA: genomic Deoxyribonucleic Acid
GMP : Granulocyte-Monocyte Progenitor
GSEA : Gene-set Enrichment Analysis
GTF : Gene Transfer Format
H3K27me3: Histone H3 lysine 27 tri-methylation
HISAT2: Hierarchical Indexing for Spliced Alignment of Transcripts
HSCs : Haematopoietic Stem Cells
HSPCs : Haematopoietic Stem and Progenitor Cells
HU : Hydroxyurea
IFN α : Interferon Alpha
IL-6: interleukin 6
IL-8: interleukin 8
IR : Intron Retention
HT1 : Hybridization Buffer 1
JAK : Janus Activated Kinases
JAK2: Janus Kinase 2
JH1 : JAK Homology 1
JH2 : JAK Homology 2
JMML : Juvenile Myelomonocytic Leukaemia
Knn : K-nearest Neighbourhood
LNA : Locked Nucleic Acid
LMPP : Lymphoid-Primed Multipotent Progenitors
LOH: Loss of Heterozygosity
LT-HSCs : Long Term Haematopoietic Stem Cells
MAF: Mutant Allele Frequency
MAPKs : Mitogen Activated Protein Kinases
MDA : Multiple Displacement Amplification
MDR : Minimal Residual Disease
MDS : Myelodysplastic Syndrome

MEP : Megakaryocyte-Erythroid Progenitor
MF : Myelofibrosis
MMLV : Moloney Murine Leukaemia Virus
MNCs : Mononuclear cells
MPL: myeloproliferative leukaemia protein
MPN : Myeloproliferative neoplasms
MPN-AP : Accelerated Phase Myeloproliferative neoplasm
MPN-BP : Blast Phase Myeloproliferative neoplasm
MPP : Multipotent Progenitors
mRNA: messenger Ribonucleic Acid
MSigDB : Molecular Signatures Database
MXE : Mutually Exclusive Exons
MYCN : v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog
NFE2 : NF-E2 45kDa protein
NGS : Next Generation Sequencing
NPM1 : Nucleophosmin
PB: peripheral blood
PBS : Phosphate Buffered Saline
PCA: Principal Component Analysis
PCR: Polymerase Chain Reaction
Ph : Philadelphia Chromosome
PI3K : Phosphoinositide 3-kinase
PPM1D : Protein Phosphatase Mg²⁺/Mn²⁺ Dependent 1D
PRC2 - Polycomb Repressive Complex 2
PSI: Percent Spliced In
PV : Polycythaemia Vera
QC : Quality Control
RNA-seq: Ribonucleic Acid Sequencing
ROS : Reactive Oxygen Species
RPKM : Reads per Kilobase per Million Mapped Reads
RT: Retrotranscription
RTase: Retrotranscriptase
RT-qPCR : Quantitative Reverse-Transcriptase Polymerase Chain Reaction
SAM : Sequence Alignment Map
sAML : Secondary Acute Myeloid Leukaemia
sc-RNA-seq: single cell Ribonucleic Acid Sequencing
SC3 : Single Cell Consensus Clustering

SD : Standard Deviation
SE : Exon-Skipping
SF3B1: Splicing factor 3B subunit 1
SH2 : Scr Homology 2
SMART : Switching Mechanism at the 5'-end of the RNA transcript
SNP: Single Nucleotide Polymorphism
snRNPs :Small Nuclear Ribonucleoproteins
SRSF2: Serine and Arginine Rich Splicing Factor 2
STAR: Spliced Transcripts Alignment to Reference
STAT : Signal Transducers and Activators of Transcription
ST-HSCs : Short Term Haematopoietic Stem Cells
SVC : Somatic Variant Caller
t-AML : Therapy-Related AML
TET2: Tet methylcytosine dioxygenase 2
TGF β : Transforming growth factor beta
TKI : Tyrosine Kinase Inhibitor
TN : Triple Negative
TNF α : Tumour Necrosis Factor Alpha
TP53: Tumour Protein 53
tSNE: T-distributed stochastic neighbourhood
TSO: Template Switching Oligonucleotides
U2AF1: U2 Small Nuclear RNA Auxiliary Factor 1
UMI : Unique Molecular Identifier
UPD : Uniparental Disomy
VAF: Variant Allele Frequency
WT : Wild-Type
WNT : Wingless-related integration site
ZF1 and ZF2 : Zinc Finger 1 & 2
ZIFA : Zero Inflated Factor Analysis

CHAPTER 1: INTRODUCTION

1. THE HAEMATOPOIETIC SYSTEM AND HAEMATOPOIETIC STEM CELLS

Haematopoiesis is the process by which all blood cells are formed. This process is tightly regulated and hierarchically organized (Figure 1.1), with haematopoietic stem cells (HSCs) at the top of the cellular hierarchy.

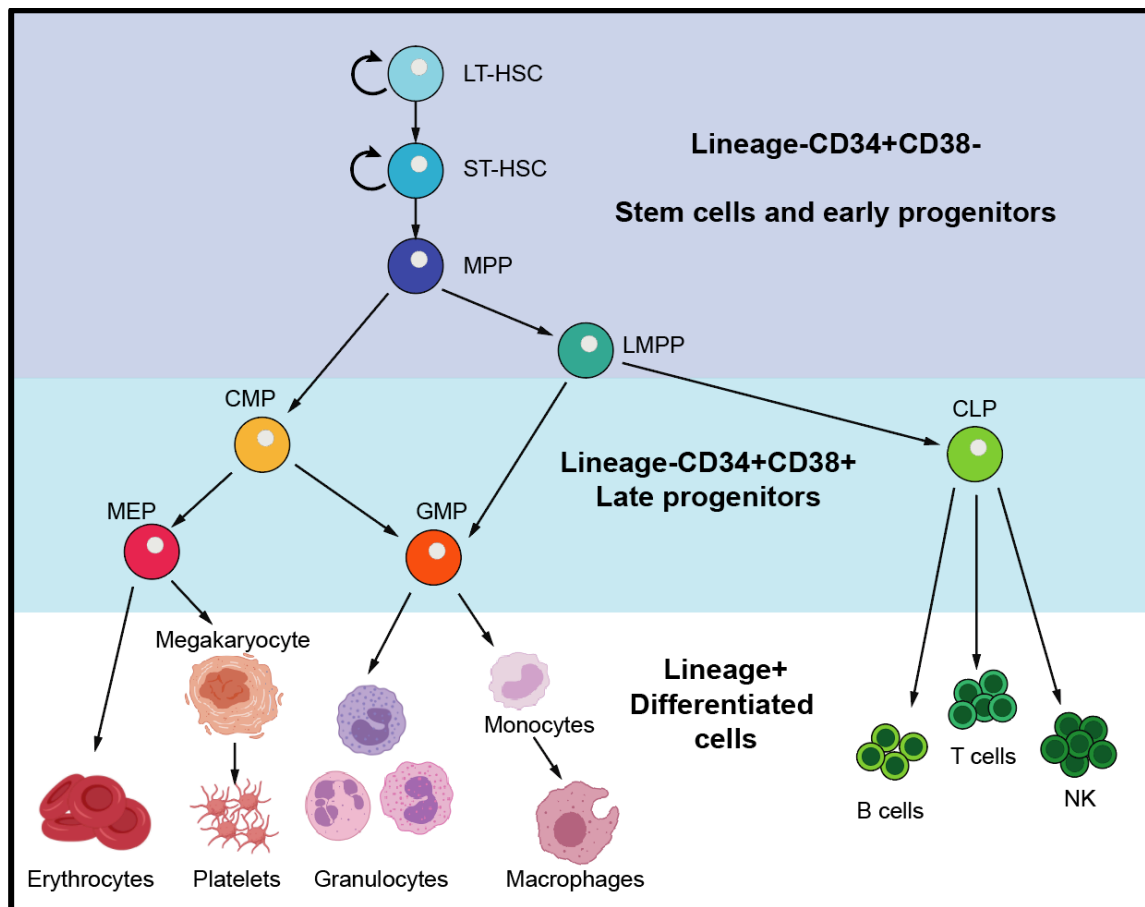


Figure 1.1. Schematic representation of the haematopoietic hierarchy. Haematopoietic stem cells give rise to all differentiated cell types through intermediate progenitors with restricted lineage potential. Lineage markers (marking mature cells; Table 1.1), CD34 and CD38 characterize early progenitors (lacking CD38 expression), late progenitors (acquiring CD38 expression) and fully differentiated cell types (expressing one or more lineage markers). Abbreviations used to label each cell type can be found in Table 1.1. Some elements of this figure were created with BioRender.

HSCs have two defining features. First, they are multipotent, i.e. capable of generating all of the rest of the blood cell types. Secondly, they possess self-renewal capacity, which maintains the stem cell pool throughout the life of an organism. HSCs are functionally defined as being able to reconstitute multilineage haematopoiesis in lethally irradiated recipient mice (2-4), and being serially transplantable (4). It is estimated that only 0.05% of the cells in adult mouse and human bone marrow are HSCs.

In the classical view of haematopoiesis (Figure 1.1), HSCs give rise to all types of differentiated blood and immune cells (granulocytes, monocytes, platelets, red blood cells and lymphoid cells) through a step-wise and progressive loss of multipotent capacity (5). This classical model of haematopoiesis illustrates homogeneous cellular differentiation states, in which each progenitor cell type gives rise to a subset of more differentiated progenitor cells with further restricted lineage potential, ultimately giving rise to terminally differentiated cell types.

All of the haematopoietic cell types have been traditionally defined according to the expression of specific combinations of cell surface markers, which allow the prospective isolation of each cell type and subsequent assessment of lineage potential (the capacity of a cell to differentiate towards a certain lineage) *in vitro* or *in vivo* (3, 4). In humans, the main haematopoietic cell types are usually defined according to the immunophenotypic definitions detailed in Table 1. However, some of these immunophenotypic definitions are not broadly accepted, and many other marker combinations are used to designate cell populations with similar lineage potential. For example, haematopoietic stem and progenitor cells (HSPCs) are commonly defined as Lineage-CD34⁺ cells; nevertheless, a subpopulation of CD34⁻ cells (lineage-CD34⁻CD38⁻CD93^{hi}) (6, 7) was shown to have stem cell properties and give rise to multilineage engraftment.

Cell type	Abbreviation	Immunophenotype
Long-term Haematopoietic Stem Cells	LT-HSC	Lineage-CD34+CD38-CD90+CD45RA-CD49f+
Short-term Haematopoietic Stem Cells	ST-HSC	Lineage-CD34+CD38-CD90+CD45RA-CD49f-
Multipotent Progenitors	MPP	Lineage-CD34+CD38-CD90-CD45RA-
Lymphoid-Primed Multipotent Progenitors	LMPP	Lineage-CD34+CD38-CD90-CD45RA+
Megakaryocyte-Erythroid Progenitor	MEP	Lineage-CD34+CD38+CD123-CD45RA-
Common Myeloid Progenitor	CMP	Lineage-CD34+CD38+CD123+CD45RA-
Granulocyte-Monocyte Progenitor	GMP	Lineage-CD34+CD38+CD123+CD45RA
Common Lymphoid Progenitor	CLP	Lineage-CD34+CD38+CD127+
B cells	B	CD34- CD3-CD19+
T cells	T	CD34-CD3+ (CD4+/CD8+)
Megakaryocytes	Meg	CD34-CD61+
Erythrocytes	Ery	CD235a+
NK cells	NK	CD3-CD56+
Monocytes	Mono	CD34-CD33+CD14+
Granulocytes	Gran	CD34-CD33+CD66b+
Platelets	Pla	CD41+
Macrophages	Macro	CD14+CD68+CD163+

Table 1.1. Immunophenotypic definition of different haematopoietic cell types in humans. Lineage refers to lineage markers of differentiated cell types such as CD2, CD3, CD4, CD7, CD8a, CD11b, CD14, CD16, CD56, CD66b and CD235a.

The definition and prospective isolation of these immunophenotypically defined cell populations and subsequent assessment of lineage potential has fundamental limitations. This strategy relies on the assumption that all cells within a particular population have homogeneous properties regarding lineage bias (which cell type will

preferentially give rise to) and lineage commitment (which cell types can give rise to) (8).

Over the past few years, technological advances made possible the characterization of molecular and functional properties of individual stem and progenitor cells. Since then, single-cell assays have challenged the classical model of the haematopoietic hierarchy. For instance, single cell transplantation (9, 10) and gene expression analysis (10, 11) resolved previously uncharacterized HSC heterogeneity, providing evidence of transcriptional lineage priming (this is, expression of genes associated with differentiation to a particular lineage in a multipotent or oligopotent cell type (12)). Single-cell RNA-sequencing also resolved transcriptional heterogeneity within CMPs and GMPs in murine bone marrow (13), identifying subpopulations of cells with distinct lineage priming within these immunophenotypically-defined compartments. Moreover, Paul and colleagues showed that mixed-lineage states were relatively rare in myeloid progenitors (13). Other studies identified different progenitor subsets with predominant erythroid or granulomonocytic potential (14). Single-cell RNA-sequencing studies of mouse and human HSPCs (15-17) consistently identified continuous differentiation trajectories, as opposed to discrete differentiation paths going through distinct progenitor populations. These lines of evidence have led to a revised model of the haematopoietic hierarchy. In this revised model, haematopoietic differentiation is represented as continuous of cell state transitions with different likelihoods towards progression to a particular lineage, usually visualized using Waddington's landscape (17) (Figure 1.2).

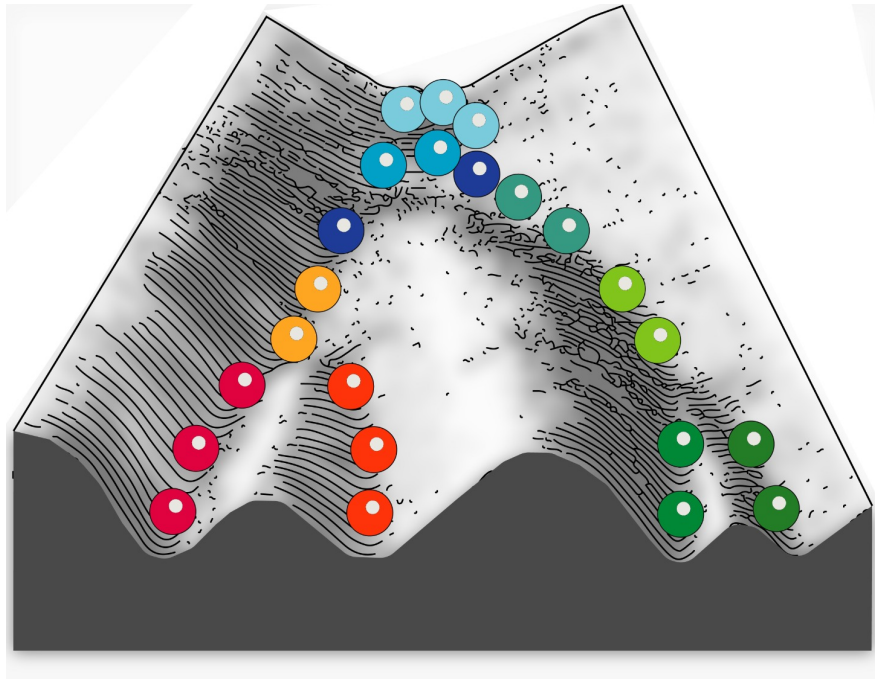


Figure 1.2. Schematic representation of the continuous differentiation model of haematopoiesis, based on Waddington's landscape model. Adapted from (17).

Understanding these transitions, as well as the cellular states and differentiation processes of native haematopoiesis has crucial implications to decipher how these are disrupted in haematopoietic malignancies.

2. MYELOPROLIFERATIVE NEOPLASMS

Myeloproliferative neoplasms (MPN) are bone marrow disorders characterized by hyperproliferation of one or more myeloid lineages (erythroid, megakaryocytic and/or granulocytic) (18), which results in increased risk of thrombosis, haemorrhages and increased risk of leukemic transformation (19). These disorders affect 5-6/100,000 individuals every year (19-22), typically over 60 years of age (22).

2.1. CLASSIFICATION OF MYELOPROLIFERATIVE NEOPLASMS

The most common MPN can be divided in two main categories, depending on the presence of the Philadelphia Chromosome $t(9;22)$ and different histopathological and clinical features (Table 1.2).

Other less frequent conditions (chronic eosinophilic leukaemia, chronic neutrophilic leukaemia and unclassifiable myeloproliferative neoplasms) are also considered myeloproliferative neoplasms under the WHO Classification of Tumours of the Haematopoietic and Lymphoid Tissues (23).

Philadelphia Chromosome Status	Positive	Chronic Myeloid Leukaemia (CML)
	Negative	Essential Thrombocythemia (ET)
		Polycythaemia Vera (PV)
		Myelofibrosis (MF)

Table 1.2. Classification of the most common myeloproliferative neoplasms according to Philadelphia Chromosome status.

2.2. PHILADELPHIA-NEGATIVE MYELOPROLIFERATIVE NEOPLASMS

Ph-negative myeloproliferative neoplasms represent a heterogeneous group of diseases characterized by distinct clinical features. PV patients present with erythrocytosis, low EPO levels in serum and trilineage proliferation (24). ET patients present with megakaryocyte expansion and high platelet counts (25). Primary MF patients typically present with enlarged spleen (splenomegaly), bone marrow fibrosis, anaemia and constitutional symptoms such as fatigue, night sweats, fever or weight loss (26). Approximately 16 % of PV patients (24) and 5% of ET patients (27) progress to myelofibrosis (post-PV MF or post-ET MF). All of these conditions are characterized by risk of thrombosis (obstruction of blood flow in a blood vessel), haemorrhage (uncontrolled bleeding) and progression to acute myeloid leukaemia (AML). Although they are considered different disease entities, the high degree of overlap of some clinical features suggests that they could be considered as different stages of a disease spectrum (28).

2.3. THE MUTATIONAL SPECTRUM OF MYELOPROLIFERATIVE NEOPLASMS

The main genetic driver of Philadelphia Chromosome positive myeloproliferative neoplasms is the t(9;22)(q34;q11) translocation, which generates the *BCR-ABL1* fusion

gene. *BCR-ABL1* encodes an oncogenic protein in which the ABL1 kinase is under the control of the *BCR* promoter, leading to constitutive tyrosine kinase activity (29, 30). CML represents a unique clinical entity within the myeloproliferative neoplasms, in which *BCR-ABL1* is necessary and sufficient to drive the disease.

On the other hand, over 85% Philadelphia Chromosome (Ph) negative MPN are driven by activating somatic mutations in genes belonging to the JAK/STAT signalling pathway (*JAK2*, *MPL* and *CALR*; (31-37)) mostly in a mutually-exclusive manner. Ph negative MPN are an excellent disease paradigm to understand how the same genetic events can lead to distinct disease phenotypes (ET, PV, MF).

For simplicity, Ph-negative MPN will be referred as MPN throughout the rest of this text.

2.3.1. Driver mutations in the JAK/STAT signalling pathway

JAKs (Janus Activated Kinases) are a family of tyrosine kinases associated with a variety of cytokine and growth factor receptors (38). They play a central role in the regulation of haematopoietic differentiation, survival and proliferation.

JAK2 proteins are bound to the cytosolic juxtamembrane domain of the three main homodimeric cytokine receptors: the erythropoietin receptor (EPOR); thrombopoietin receptor (TPOR, also known as MPL) and granulocyte colony stimulating factor receptor (GCSFR; Figure 1.3). Cytokine binding triggers receptor homodimerization, which causes a conformational change leading to JAK2 proteins coming in close proximity and *trans*-phosphorylating. JAK2 subsequently phosphorylates the signal transducers and activators of transcription (STATs) STAT3 and STAT5, and mitogen activated protein kinases (MAPKs), which enter the nucleus and activate target transcriptional targets (38-40), as well as Phosphoinositide 3-kinase (PI3K), which regulates cell proliferation (41).

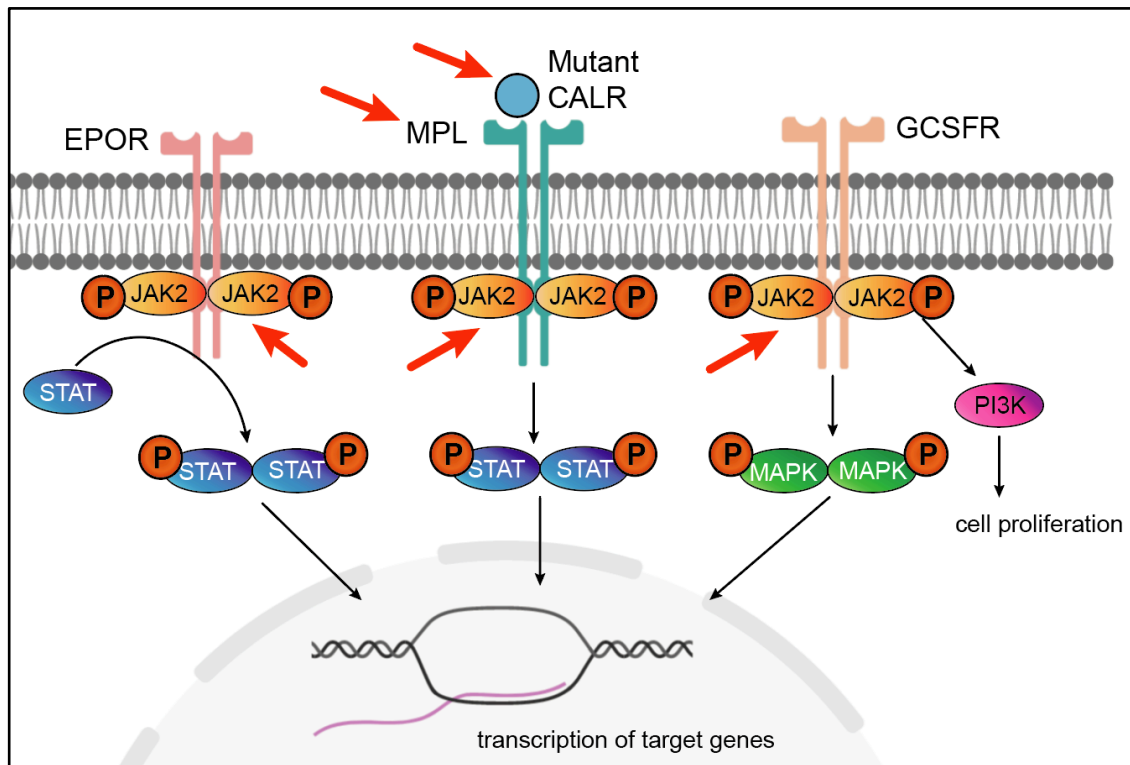


Figure 1.3. Schematic diagram of the JAK/STAT signalling pathway through homodimeric cytokine receptors (EPOR, MPL, GCSFR). Mutant CALR protein binds to MPL independently of cytokine signals and mutant MPL leads to constitutive JAK2 phosphorylation and signalling. Constitutively active JAK2V617F leads to phosphorylation of STAT proteins (STAT3 or STAT5), MAPK dimers and PI3K. Ultimately, phosphorylation of these downstream effectors results in transcription of targets genes and cell proliferation. Components of this pathway affected by somatic mutations in MPN patients are indicated by red arrows. Some of the elements in this figure were generated using BioRender.

A small subset of patients with Ph-negative MPN lack mutations in these three main driver genes and are referred as Triple Negative (TN) MPN. In some instances, TN patients have non-canonical mutations in *JAK2* or *MPL* (42), and in other cases, driver mutations are unknown. However, all MPN patients share aberrant JAK/STAT signalling transcriptional signatures, independently of their mutational status or clinical subtype (43). This illustrates the central role of JAK/STAT signalling in MPN pathogenesis.

Activating mutations in these three genes (*JAK2*, *MPL* and *CALR*) cluster in hotspot regions affecting domains essential for their signalling activity or subcellular location.

JAK2 mutations are the most common driver mutation in MPN (60-95 % depending on disease subtype), followed by mutations in *CALR* (18-25 % (44, 45)) and *MPL* (1-5 % (46)).

The *JAK2* protein has four main domains: a FERM domain (N-terminal band 4.1, ezrin, radixin, moesin), a SH2 (Scr Homology 2) domain, JH2 and JH1 domains (JAK Homology 2 and JAK Homology 1; Figure 1.4). FERM and SH2 domains mediated interactions with the cytokine receptors; the JH2 pseudokinase domain mainly exerts a regulatory function by inhibiting JH1 kinase domain activity, which is catalytically active (47-49). The most common *JAK2* (Janus Kinase 2) mutation (31-34) is a valine to phenylalanine substitution (p.V617F; c.1849 G>T). The *JAK2V617F* missense mutation destabilizes the interaction between JH1 and JH2 domains, resulting in loss of its autoinhibitory activity and increasing the chance of *trans*-phosphorylation, which consequently results in constitutive signalling (47, 49). *JAK2V617F* mutations are found in patients with ET, PV and MF.

A common feature affecting the *JAK2* gene is loss of heterozygosity in the short arm of chromosome 9 (9p-LOH), where *JAK2* is located, leading to *JAK2* homozygosity (50). This genetic event was identified before *JAK2* point mutations themselves. In fact, fine microsatellite mapping of this region led to the discovery of *JAK2V617F* activating mutations (31). 9p-LOH is usually a copy neutral loss of heterozygosity by acquired Uniparental Disomy (UPD), frequently caused by mitotic recombination. *JAK2V617F* dosage is associated with distinct clinical features: patient with *JAK2* homozygous mutations have higher haemoglobin levels, white blood cell counts and reduced platelet counts (51) compared to patients carrying *JAK2* heterozygous mutations. *JAK2* homozygous mutations are more frequently found in PV patients than ET patients (52), consistent with the observed erythrocytosis (53) and indicating that different *JAK2V617F* dosage influences disease phenotypes.

JAK2 mutations can also affect the exon 12 of the gene (spanning from residues 536 to 547; Figure 1.4), although this is more infrequent and is usually restricted to *JAK2V617F* negative PV (54). These mutations are mostly in-frame small deletions and insertions located in the linker region between the SH2 and JH2 pseudokinase domain, which also result in constitutive kinase activity (55).

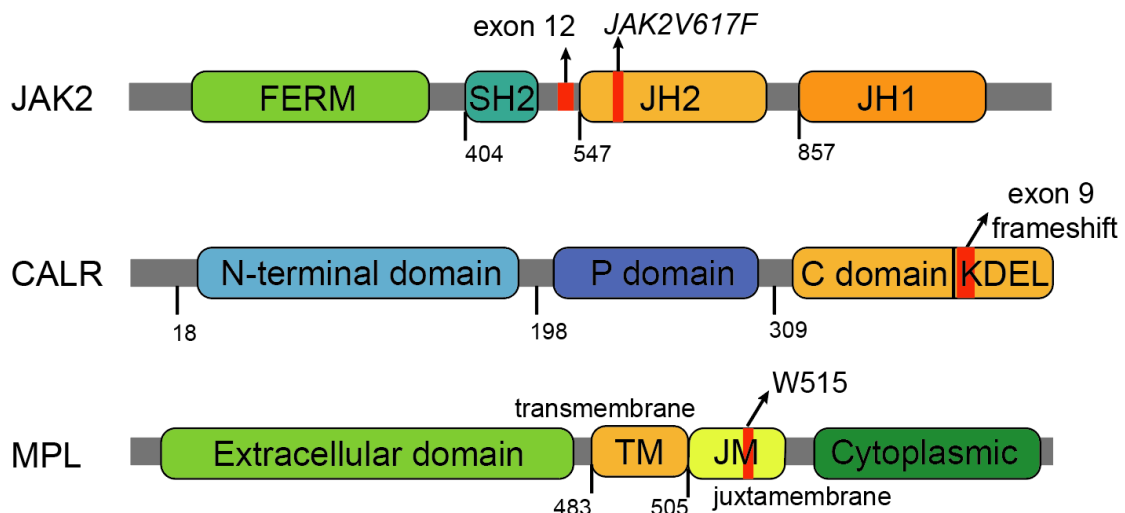


Figure 1.4. Schematic representation of the domain structure of JAK2, CALR and MPL. Commonly mutated residues and domains are labelled with arrows and indicated in red. Numbers below each schematic representation indicate the position of certain residues.

Mutations in *MPL* (Myeloproliferative Leukaemia Virus, or thrombopoietin receptor) mostly occur in exon 10 (35) and affect the transmembrane and juxta-membrane domains of the protein (Figure 1.4). These domains are responsible for receptor dimerization (transmembrane domain) and tyrosine kinase binding (juxtamembrane domain). Mutations in these regions cause autonomous receptor activation (56). The most commonly mutated residue is W515.

CALR (calreticulin) is a endoplasmic reticulum (ER) associated chaperone which under normal circumstances, regulates calcium homeostasis and protein folding (57, 58). *CALR* mutations (36, 37) in MPN affect the terminal exon of the gene (exon 9; Figure 1.4), and are frameshift heterozygous mutations generating a new C-terminal domain

of the protein. This new C-terminal domain lacks the endoplasmic reticulum retention signal (KDEL), which maintains CALR localized to the ER, and the mutant protein activates JAK/STAT signalling by directly binding to MPL localized in the cell membrane (59-61). The most common *CALR* mutations, which represent over 80 % of *CALR* pathogenic variants, are type 1 (52 bp deletion; p.L367fs*46) and type 2 (5 bp insertion; p.K385fs*47) (23, 45). Type 1 mutations are more commonly associated with myelofibrosis and type 2 mutations, with ET (23).

Mutations in *JAK2*, *CALR* and *MPL* are considered the major genetic events driving the myeloproliferative phenotype in MPN. However, most patients with myeloproliferative neoplasms (>55% (62)) also carry additional somatic mutations in other genes such as epigenetic regulators, spliceosome components, signalling molecules and metabolic enzymes.

The number of additional somatic mutations varies considerably according to disease subtype. Approximately 50 % of patients with ET and PV carry additional somatic mutations (63), as opposed to over 80 % of patients diagnosed with myelofibrosis (64). These mutations contribute to MPN pathogenesis and influence disease prognosis.

2.3.2. Somatic mutations in epigenetic modifiers

Mutations in epigenetic modifiers are the second most common type of mutations in MPN after canonical *JAK2*, *CALR* and *MPL* mutations. They mostly affect enzymes involved in DNA methylation (TET2, DNMT3A) and members of the Polycomb Repressive Complex (ASXL1, EZH2), and they often co-occur in the same patient (65). This illustrates the central role of epigenetic mechanisms in the regulation of self-renewal, differentiation, and oncogenic transformation of HSCs. Contrary to JAK/STAT signalling mutations described above, which are largely restricted to MPN, these epigenetic regulators are also widely mutated in other types of blood cancers and solid tumours.

The most common mutations in epigenetic modifiers in MPN are loss-of-function mutations in the tumour suppressor gene *TET2* (Ten Eleven Translocation 2) (62, 66), which can be identified in approximately 12 % of MPN patients. *TET2* demethylates DNA by hydroxylating 5-methylcytosine (5mC) residues to 5-hydroxymethylcytosine (5hmC). *TET2* loss-of-function increases HSC self-renewal *in vivo* (67, 68) and gives rise to skewed myeloid differentiation (68) by altering 5hmC levels at specific loci (69). *TET2* mutations can be acquired before or after *JAK2V617F* mutations, and the order of acquisition influences clinical features and response to targeted therapies (70).

DNMT3A (DNA Methyltransferase 3A) is a methyltransferase that catalyses the methylation of 5mC (71). The most common hotspot mutation *DNMT3A* is R882H, which exerts a dominant-negative effect by inhibiting wild-type *DNMT3A* enzymatic activity (72). Mutant *DNMT3A* enhances haematopoietic stem cell self-renewal, causing a dramatic clonal expansion (71, 73).

ASXL1 and *EZH2* are members of the Polycomb Repressive Complex 2 (PRC2) which are recurrently mutated in myeloproliferative neoplasms, in 13-23% and 7 % of patients respectively (74). Loss-of-function heterozygous mutations in these genes result in global loss of PRC2-mediated histone H3 lysine 27 tri-methylation (H3K27me3) (75) (76). This leads to transcriptional activation of genes regulating stem cell maintenance (*HOXA* cluster) (75) upon *ASXL1* knock-down and transcriptional activation of inflammatory pathways and tumour necrosis factor alpha (TNF α) signalling upon *EZH2* knock-out (76) in *in vitro* and *in vivo* models. *ASXL1* and *EZH2* mutations are associated with reduced overall survival and higher risk of fibrotic transformation (63, 77).

2.3.3 Somatic mutations in components of the spliceosome machinery

Recent whole-exome sequencing studies of patients with myeloid malignancies identified the components of the spliceosome machinery as a new functional group of

commonly mutated genes (78-80). These mutations represent up to 40 % of the genetic defects found in patients with myeloid disorders (79), and 9% of patients with myeloproliferative neoplasms, suggesting that alteration of mRNA splicing is an important feature driving the pathogenesis of myeloid neoplasms. Moreover, spliceosome mutations are usually detected in the founding clone, indicating a potential role in initial clonal expansion and establishment of clonal dominance (81).

The spliceosome is a complex molecular machine regulating alternative splicing in eukaryotic cells. It is formed by small nuclear ribonucleoproteins (snRNPs) and associated protein cofactors (82). The most commonly mutated spliceosome components in MPN are *SF3B1*, *SRSF2* and *U2AF1* (62), which affect the assembly of snRNPs, recognition of pre-mRNA splice sites or recruitment of splicing regulators. These mutations are mutually exclusive and heterozygous (78, 83), potentially indicating that cells can only tolerate splicing deregulation to a certain extent, or that they act via similar mechanisms and the acquisition of a second spliceosome mutation does not confer an additional competitive advantage.

Some of these mutations are mostly associated with certain disease phenotypes. For example, *SF3B1* mutations in myeloid malignancies are mostly associated with the presence of ring sideroblasts (78, 79) and *SRSF2* mutations are highly enriched in Chronic Myelomonocytic Leukaemia (CMML) (84). *U2AF1* mutations are the most common spliceosome mutation in primary MF (85) and are associated with more severe anaemia and poorer overall survival in this group of patients.

U2AF1 (U2 Auxiliary Factor 2, or U2AF35) is an accessory factor of the U2 pre-mRNA splicing complex snRNP (small nuclear RiboNucleoProteins) (86) which binds to a conserved intronic sequence (AG) in the 3' splice site of pre-mRNA molecules, and is required for splicing (Figure 1.5A) (87). *U2AF1* mutations in myeloid malignancies are localized in two hotspots within the zinc finger domains of the protein (79, 80) (S34F/Y

and Q157P/R, Figure 1.5B). These residues are responsible for sequence motif recognition in pre-mRNA molecules. S34 is more commonly mutated in MDS (Myelodysplastic Syndrome) and AML patients and Q157, in MPN. Both hotspot mutations lead to a defect in exon recognition and 3' splicing site specification *in vitro* (88, 89).

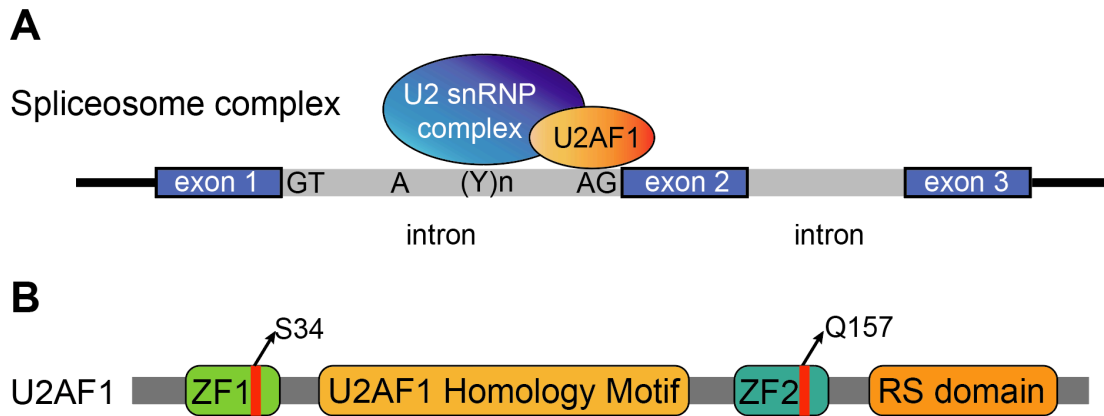


Figure 1.5. *U2AF1* mutations affect the zinc finger domain of the protein, responsible for binding to canonical 3' splice sites. Schematic representation of *U2AF1* interaction with the intronic 3' AG dinucleotide within the U2 snRNP spliceosome complex (A) and domain structure of *U2AF1* (B). “ZF1” and “ZF2” stand for Zinc Finger 1 and 2. Commonly mutated residues are labelled with arrows.

Splicing defects resulting from *U2AF1* mutations were initially identified *in vitro* (79), and suggested a widespread splicing failure. However, further functional studies *in vitro* identified changes in isoform usage and isoform patterns in specific subsets of genes (80, 88), resulting from different binding affinities of *U2AF1* mutant proteins to 3' splice sites and preferential recognition of alternative 3' splice site sequences (88, 89). This was also confirmed in transgenic mouse models (90, 91) and primary human HSPCs (89, 92), suggesting that *U2AF1* mutations are gain-of-function mutations, and demonstrating the contribution of mRNA splicing defects to MPN pathogenesis.

2.3.4 Other recurrent genetic abnormalities in myeloproliferative neoplasms

Excluding the most frequent JAK/STAT signalling, epigenetic and spliceosome mutations, recurrent somatic mutations in MPN mostly affect other signalling components (such as *CBL*), transcription factors (such as *NFE2*) (62, 93) and members of the P53 pathway (*TP53* and *PPM1D*).

CBL (Casitas B-lineage lymphoma) is an E3 ubiquitin ligase responsible for the ubiquitination of tyrosine kinase receptors, including *MPL* (94). Missense mutations result in loss of ubiquitin ligase activity, preventing homeostatic receptor internalization and subsequent degradation. *CBL* mutations are most commonly found in MF patients and are associated with progression to Acute Myeloid Leukaemia (40).

NFE2 (NF-E2 45kDa protein) is a transcription factor involved in megakarythroid lineage differentiation. *NFE2* mutations are small indels which result in a truncated protein enhancing wild-type *NFE2* function (93). They are mostly subclonal and are associated with disease progression (62).

TP53 (Tumour Protein 53) is a transcription factor with strong tumour suppressor activity, and it is the most commonly mutated gene in cancer (95). Missense mutations in MPN often co-occur with 17p aberrations and are subclonal (62, 96). They are strongly associated with progression to MF, leukemic transformation and markedly reduced overall survival (62).

PPM1D (protein phosphatase Mn^{2+}/Mg^{2+} -dependent 1D) is a component of the p53-pathway; it inhibits p53-mediated transcription and is recurrently mutated in many types of cancer. Mutations in *PPM1D* are often truncating, localized in the terminal exon of the gene. These mutations are often subclonal in MPN and expand after cytotoxic chemotherapy (62, 97).

2.4. DISEASE-INITIATING CELLS IN MYELOPROLIFERATIVE NEOPLASMS

MPNs are clonal HSC disorders; this is, they originate through the acquisition of a somatic driver mutation in a single HSC which undergoes clonal expansion (98). This driver mutation confers a competitive proliferative advantage to mutant haematopoietic stem cells and promotes myeloid differentiation, giving rise to a myeloproliferative phenotype. The hyperproliferation of myeloid lineages is a consequence of constitutive tyrosine kinase signalling, which promotes cytokine-independent growth and leads to defective regulation of proliferation, apoptosis and differentiation of progenitor populations.

MPN-mutant haematopoietic stem cells, which reside at the top of the haematopoietic cellular hierarchy, are thought to be the cells responsible for the initiation and propagation the disease, constituting disease reservoirs (98-101). This has been demonstrated *in vivo* using a *JAK2V617F* mouse model, in which transplantation of a single LT-HSC into a lethally irradiated recipient mouse led to MPN development (99). Loss-of function *TET2* and *EZH2* mutations, which can be acquired before *JAK2V617F*, enhance the MPN phenotype in conditional *JAK2V617F* knock-in mouse models in which those mutations are restricted to haematopoietic stem cells (102-105), suggesting that other recurrent somatic mutations in MPN are also acquired within HSCs.

In patient samples, *JAK2V617F* mutations can be traced back to the Lineage-CD34+CD38-CD90+ stem cell compartment (106), and *JAK2V617F* can be detected in both myeloid and lymphoid cell lineages (107, 108), indicating that it is acquired in a multipotent cell in MPN patients. Transplantation of Lineage-CD34+ cells in a humanized ossicle xenograft model, which allows the efficient engraftment of MPN patients' cells (109), showed that only the Lineage-CD34+CD38-CD90+CD45RA- stem cell fraction from *JAK2V617F* and *CALR* mutant MPN patients successfully engrafted.

Lineage-CD34+CD38+ progenitors and Lineage-CD34+CD90-CD45RA- MPPs did not engraft from any of the patients, demonstrating that MPN-initiating cells reside within the HSC compartment.

Characterizing the cells that initiate and sustain the disease is crucial for the development of successful treatments. In principle, targeting and eliminating MPN stem cells is not only essential but also potentially sufficient to completely eradicate the disease.

2.5. TREATMENT OF MYELOPROLIFERATIVE NEOPLASMS

MPN patients have been conventionally treated with cytoreductive compounds such as Hydroxyurea (HU), interferon alpha (IFN α) or anagrelide. Low doses of aspirin are also used in combination with some of these agents to reduce the risk of thrombosis, given aspirin's inhibitory effect on platelet aggregation.

Hydroxyurea (also known as hydroxycarbamide) is a ribonucleotide reductase inhibitor. It reduces the production of deoxyribonucleotides and inhibits replication. It is mostly used to reduce red blood cell counts (PV patients), platelet counts (ET and PV) and splenomegaly (MF patients). One of the concerns of HU treatment is the development of therapy-related leukaemia (110), although its leukemogenicity has not been definitely proved (27). Moreover, up to 24% patients develop resistance or show HU intolerance (111-113).

IFN α was introduced over 35 years ago for the treatment of MPN, and it is also used to treat other myeloproliferative conditions such as CML (114, 115). IFN α induces clinical, haematological and molecular responses, normalizing or lowering down platelet, erythrocyte and leukocyte counts through a wide range of biological effects. It exerts an immunomodulatory effect (116), pro-apoptotic effect in myeloid progenitors (117) and targets disease-initiating stem cells by inducing cell cycle entry and exhaustion (118). However, its use was initially associated with many short term adverse reactions and

toxicity, some of which can now be avoided by using different forms of pegylated IFN α (119). IFN α has been shown to induce molecular responses in both *JAK2* and *CALR* mutant patients (119, 120), although there is a rapid molecular relapse after IFN α discontinuation (121), indicating that IFN α does not completely eliminate the mutant clone.

Anagrelide blocks megakaryocyte differentiation and proliferation (122), which reduces platelet production and subsequently lowers platelet counts in PV and ET patients. However, it can change disease course increasing the risk of fibrotic transformation (123). All of these therapies (HU, IFN α and anagrelide) manage symptoms, but none of them have been shown to improve overall survival of MPN patients.

The discovery of the crucial role of JAK/STAT signalling in the pathogenesis of the disease led to the development of compounds specifically inhibiting constitutive JAK/STAT activity. The first compound introduced in the clinic was ruxolitinib, a JAK1/JAK2 inhibitor (124) used to treat myelofibrosis (primary MF, post-PV MF and post-ET MF) and polycythaemia vera resistant to HU. Ruxolitinib markedly reduces spleen size and constitutional symptoms (125), leading to a significant improvement of life quality and overall survival (126). This effect is partly due to its anti-inflammatory activity (reduction of IL-6 and TNF α levels) rather than specifically related to mutant JAK2 inhibition (127). However, ruxolitinib does not induce clonal remissions (reduction of mutant allele burden) (128, 129), indicating that it does not preferentially target MPN cells over their normal counterparts. Besides, it does not have a marked cytoreductive activity and does not improve bone marrow fibrosis (126).

Despite ameliorating symptoms and in some instances, overall survival, none of the treatments presented above completely eradicate disease-initiating cells. Therefore, the only curative option for MPN patients is haematopoietic stem cell transplantation. Unfortunately, stem cell transplantation is only indicated for younger patients (mostly

below 65 years old) due to its high toxicity and treatment related mortality in the elderly (130). Given that MPN patients are usually over 60 years old, most patients are not eligible for stem cell transplantation. Moreover, approximately 20% of patients relapse (131) and median survival post-transplantation is 6 years (130, 131), indicating the urgent need for the development of new therapeutic strategies in MPN.

3. ACUTE MYELOID LEUKAEMIA

AML is an aggressive and rapidly growing blood cancer affecting 3.7/100,000 people (132), with this incidence rising to 17.9/100,000 in people over 65 years of age (133). It is the second most common subtype of leukaemia diagnosed in adults, and it is often associated with extremely poor outcome; the 5-year survival rate is lower than 25 %.

AML originates by the acquisition of somatic mutations in HSPCs. These mutations give rise to uncontrolled growth of clonal immature myeloid progenitors which lack the ability to differentiate to fully mature cell types (133). This results in an accumulation of such immature progenitors and subsequently, multilineage cytopenia (anaemia, thrombocytopenia and leukopenia), leading to bone marrow failure, which is the main cause of mortality in AML (133). Poorly differentiated progenitors infiltrate the bone marrow, blood, and ultimately other tissues such as spleen, liver, lungs and brain, the consequences of which are often fatal. AML is diagnosed when the blast count reaches 20 % of the bone marrow or peripheral blood (134).

3.1. CLASSIFICATION OF ACUTE MYELOID LEUKAEMIA

The 2016 revision of the WHO classification of Acute Myeloid Leukaemia and myeloid neoplasms integrates genetic, molecular, morphological and clinical features (135) to classify AML in four main categories: AML with recurrent genetic abnormalities, AML with myelodysplasia-related changes, therapy related AML and AML not otherwise specified. Details of the subgroups included within each category can be found in Table 1.3.

AML category	
AML with recurrent genetic abnormalities	AML with t(8;21)(q22;q22.1); <i>RUNX1-RUNX1T1</i>, also known as <i>AML1-ETO</i> fusion AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); <i>CBFB-MYH1</i> AML with <i>PML-RARA</i> AML with t(9;11)(p21.3;q23.3); <i>MLLT3-KMT2A</i> AML with t(6;9)(p23;q34.1); <i>DEK-NUP214</i> AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); <i>GATA2, MECOM</i> AML (megakaryoblastic) with t(1;22)(p13.3;q13.3); <i>RBM15-MKL1</i> AML with <i>BCR-ABL1</i> AML with mutated <i>NPM1</i> AML with biallelic <i>CEBPA</i> mutations AML with mutated <i>RUNX1</i>
AML with myelodysplasia-related changes	AML from prior MDS/MPN AML with MDS-related cytogenetic abnormality AML with morphologic multilineage dysplasia
Therapy-related AML	
AML not otherwise specified	AML with minimal differentiation AML without maturation AML with maturation Acute myelomonocytic leukaemia Acute monoblastic/monocytic leukaemia Pure erythroid leukaemia Acute megakaryoblastic leukaemia Acute basophilic leukaemia Acute panmyelosis with myelofibrosis

Table 1.3. Classification of Acute Myeloid Leukaemia according to the 2016 revision of the WHO guidelines. Adapted from (135).

The first category, AML with recurrent genetic abnormalities, is defined by the presence of recurrent cytogenetic and genetic abnormalities. AML with myelodysplasia-related (MDS-related) changes is characterized by the presence of

multilineage dysplasia or an antecedent haematological malignancy (AHD-AML) such as MDS or MPN. Therapy-related AML includes patients who develop AML after cytotoxic therapy and AML not otherwise specified is mainly based on morphological features.

AML can also be classified as *de novo* AML or secondary AML (sAML). The first refers to AML which develops in patients without any previous history of haematological malignancies, whereas the latest includes AML developed from an antecedent haematological disease (MDS, MPN) or therapy-related AML. Secondary AML patients are usually older and have a poorer prognosis when compared to *de novo* AML (136).

3.2. GENETIC LANDSCAPE OF ACUTE MYELOID LEUKAEMIA

The landscape of genetic abnormalities in AML is markedly different depending on the disease subtype (*de novo* AML or secondary AML), which imply that these are biologically and clinically distinct disease entities.

Somatic mutations in *de novo* AML affect nine main functional categories (137): tumour suppressor genes, DNA methylation, chromatin modifiers, spliceosome components, activated signalling, transcription factor fusions, myeloid transcription factors, the cohesin complex and nucleophosmin (*NPM1*). Over 98 % of patients carry a known somatic alteration in one of these main functional categories, and 86% of patients carry two or more known *driver* mutations (138).

Mutation in genes affecting signalling pathways, which can be detected in 59% of patients, are the most frequently mutated group in *de novo* AML. This includes mutations in the tyrosine kinase FLT3 (missense or internal tandem repeats; FLT3-ITD) and mutations in RAS pathway genes (*NRAS*, *KRAS*, *HRAS*, *KIT* and *PTPN11*). Mutations in the JAK/STAT signalling pathway such as *JAK2*, *CALR* or *MPL* are rarely seen in *de novo* AML; however, they are commonly mutated in sAML from an antecedent MPN.

Genes regulating DNA methylation are the second most commonly mutated category in *de novo* AML (44% of patients). This group includes *DNMT3A*, *DNMT1*, *TET2*, *IDH1* and *IDH2*. *NPM1* mutations can be detected in up to 27% of patients; mutations in chromatin-modifying genes (*EZH2*, *ASXL1*) are detected in 30 % of patients; 22 % of cases have mutations in transcription factors (*RUNX1*, *CEBPA*), 16% in tumour suppressor genes (*TP53*, *WT1*, *PHF6*), 14% in spliceosome components (*U2AF1*, *SF3B1*, *SRSF2*, *ZRSR2*) and 13 % in members of the cohesin complex (*STAG2*, *RAD21*)(137).

Fusion genes are also a recurrent genetic abnormality in *de novo* AML. They can be detected in 45% of patients (137), and they mostly affect transcription factors. Recurrent events include *PML-RARA*, *CBFB-MYH11*, *AML1-ETO* and multiple fusions involving *MLL* (also known as *KMT2A*). Fusion genes are however rarely detected in sAML.

Moreover, approximately 50 % of AML patients also carry other chromosomal abnormalities such as copy number gain, loss, or loss of heterozygosity (LOH). The most common are loss of chromosome 7 (-7), 7q (-7q) or del(7q); loss of chromosome 5 (-5), 5q (-5q) or del(5q), loss of chromosome Y (-Y), trisomy 8 (+8), abnormal 12p (abn(12p)), trisomy 21 (+21) and abnormal 17 p (abn(17p)). Some patients present a complex karyotype with three or more (10.7%) or five or more (8.8%) karyotypic abnormalities (139).

Cytogenetic abnormalities and mutations are often used to stratify patients in different risk categories with prognostic value: favourable, intermediate and adverse (134) (Table 1.4).

Patients belonging to favourable risk categories have relatively good outcomes when treated with induction chemotherapy whereas patients belonging to adverse risk categories mostly require allogeneic stem cell transplantation. This classification, as

well as other new prognostic systems based on patterns of mutation co-occurrence, are beginning to inform clinical practice (138).

Risk category	Genetic abnormality
Favourable	<i>RUNX1-RUNX1T1</i> <i>MYH11-CBFB</i> biallelic <i>CEBPA</i> mutation <i>NPM1</i> mutation without <i>FLT3</i> -ITD
Intermediate	<i>MLLT3-KMT2A</i> co-mutated <i>NPM1</i> and <i>FLT3</i> -ITD ^{high} (allelic ratio higher than 0.5) wild-type <i>NPM1</i> (without <i>FLT3</i> -ITD or with <i>FLT3</i> -ITD allelic ratio lower than 0.5) Cytogenetic abnormalities not classified as favourable or adverse
Adverse	<i>RUNX1</i> , <i>ASXL1</i> and/or <i>TP53</i> mutations <i>FLT3</i> -ITD with wild-type <i>NPM1</i> <i>MLL (KMT2A)</i> rearrangement <i>BCR-ABL1</i> <i>DEK-NUP214</i> <i>GATA2/MECOM</i> -5, del(5q), -7, -17, abn(17p) Complex or monosomal karyotype

Table 1.4. Classification of AML based on the 2017 European Leukaemia Net risk stratification. Adapted from (134).

3.3. DISEASE-INITIATING CELLS AND CLONAL EVOLUTION IN ACUTE MYELOID LEUKAEMIA

Acute Myeloid Leukaemia arises through the sequential acquisition of somatic mutations and cytogenetic abnormalities in HSPCs. The vast majority of patients have two or more known *driver* mutations, indicating that AML is not driven by a single genetic alteration and that AML is genetically more complex than MPN (138).

Many studies have characterized the order and stage in which these mutations are acquired, to elucidate the genetic architecture of the disease and its functional consequences. Traditionally, variant allele frequencies obtained from bulk sequencing studies were used to infer the phylogenetic tree of each tumour (137). Bulk sequencing studies performed in patient samples and patient-derived xenografts identified higher variant allele frequencies in genes involved in epigenetic regulation (*DNMT3A*, *ASXL1*, *TET2*) in AML (140, 141). These cells are capable of multilineage engraftment, demonstrated by their presence in cells from both myeloid and lymphoid lineages. *FLT3*-ITD mutations and other activating signalling mutations such as *KRAS* and *NRAS* are usually present at lower variant allele frequencies (140), indicating that they are acquired as a second genetic hit. This indicates that signalling mutations in *de novo* AML do not promote initial clonal expansion, as opposed to the essential role of initiating *JAK2* or *CALR* mutations promoting clonal dominance in MPN.

Targeted-sequencing of fractionated cell populations was used to determine the precise cell type in which these mutations were first acquired. Mutations in epigenetic modifiers or transcription factor fusion genes (*DNMT3A*, *TET2*, *MLL* fusions, *AML1-ETO*) could be traced back to HSCs (142, 143), but these lacked other cooperating mutations. Secondary mutations could only be detected in more mature progenitors. Interestingly, mutations in *DNMT3A* and *TET2* detected in pre-leukemic clones are also the most frequent genetic defects found in individuals with Clonal Haematopoiesis of Indeterminate Potential (CHIP). These individuals carry somatic mutations associated with the development of a haematological malignancy in their peripheral blood or bone marrow (144, 145). However, they do not fulfil the criteria to be diagnosed with a haematological malignancy (blood counts are normal and they are asymptomatic), but they are at increased risk of developing leukaemia (146).

These lines of evidence led to the proposal of a two-step model of leukemogenesis (Figure 1.6), in which mutations in epigenetic modifiers are first acquired in HSCs

(Lineage-CD34+CD38-CD90+CD45RA-; see Table 1.1) and second-hit mutations in signalling components are acquired in more mature progenitors. In this hierarchical model, the first set of cells, which precedes the development of leukaemia, is usually termed “preleukemic”, whereas the second are usually known as leukemic stem cells, or leukemic clones.

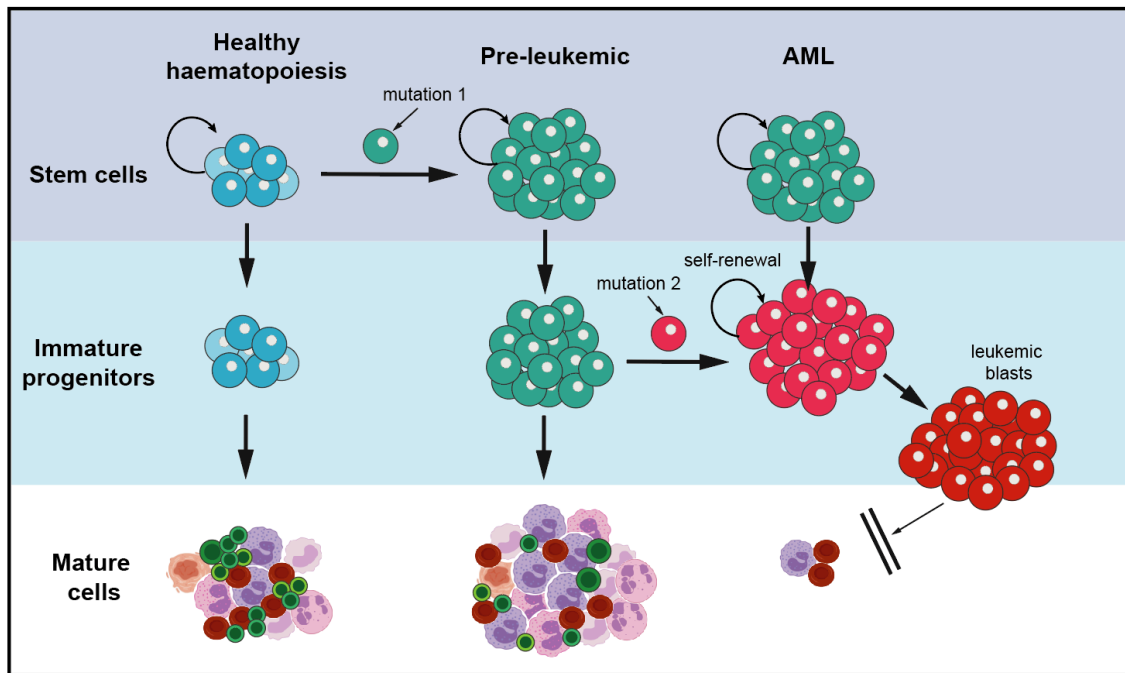


Figure 1.6. Two-step model of leukemogenesis. In healthy haematopoiesis, stem cells give rise to mature blood cells through intermediate immature progenitor states. When a preleukemic mutation is acquired in a stem cell, this gives rise to clonal expansion and subsequent increased production of progenitor cells, which can still differentiate. When a second-hit mutation is acquired in progenitor cells, these acquire self-renewal capacity and proliferate abnormally to produce leukemic blasts, which overcrowd the bone marrow but are unable to differentiate. Some elements of this figure were generated using BioRender.

Under this model, mutations acquired in pre-leukemic stem cells confer a competitive advantage which gives rise to clonal expansion. Subsequently, the acquisition of a secondary mutation activating constitutive signalling in a more mature progenitor cell, confers *de novo* self-renewal ability to those cells, which will expand and populate the bone marrow (Figure 1.6) (140, 147, 148). These leukemic stem cells can initiate and

maintain the disease when transplanted in immunodeficient mice and can also give rise in more differentiated leukemic blasts, which lack the ability to self-renew (148).

Preleukemic cells lacking second hit mutations are often difficult to study because they are outcompeted by highly proliferative leukemic clones. However, these cells are recurrently resistant to chemotherapy, present during remission (140, 141) and are responsible for disease relapse (149, 150), indicating the importance of preleukemic stem cell monitoring as a diagnostic tool during clinical remission.

3.4. TREATMENT OF ACUTE MYELOID LEUKAEMIA

Cytotoxic therapy is the first line of treatment in AML (151). Generally, this involves a first regime of intensive induction chemotherapy, aimed at quickly reducing blood counts and achieving remission, and a second regime of consolidation therapy.

Induction chemotherapy usually involves the use of an anthracycline in combination with cytarabine. Anthracycline is an intercalating DNA agent that inhibits topoisomerase II, leading to DNA damage. Cytarabine is a nucleoside analogue impeding replication by blocking DNA polymerase. 60-85% of adults <60 years achieve a complete response, and 40-60% >60 years old. Induction therapy is considered to be successful if patients achieve remission, defined as hypocellular bone marrow and less than 5% blasts. After that, consolidation therapy usually involves several cycles of high-dose cytarabine, allogeneic stem cell transplant (younger patients) or non-myeloablative stem cell transplant (older patients). Treatment is more effective in younger adults, and 35-40% of patients <60 years old are cured. This percentage is reduced to 5-15% percent in patients who are older than 60 (151), partially explained by the higher rates of treatment-related mortality.

Given that approximately 60 % of patients are positive for a well-characterized molecular marker, minimal-residual-disease (MDR) monitoring can be achieved by quantitative reverse-transcriptase polymerase-chain-reaction (RT-qPCR). These

assays are much more sensitive than standard cytogenetic techniques or morphological and histopathological evaluation, achieving a sensitivity of 10^{-3} to 10^{-7} and accurately predicting relapse (134, 152).

Patients carrying specific genetic abnormalities or blasts expressing a particular immunophenotypic marker might be eligible for targeted therapies, which are likely to overcome some of the disadvantages of cytotoxic chemotherapy and ultimately, improve the outcomes of patients.

The FLT3 inhibitor Midostautin was the first tyrosine kinase inhibitor approved by the Food and Drug Administration (FDA) for the treatment of *FLT3* mutant AML. FLT3 point mutations or internal tandem duplications (FLT3-ITD), which result in constitutive tyrosine kinase activity, occur in approximately 25% of AML patients. Midostautin showed a significant improvement in overall survival (7% improvement) in this subset of patients (153).

IDH2 (isocitrate dehydrogenase 2) inhibitors were developed for the treatment of *IDH2*-mutant AML (representing 10 % of all AMLs). *IDH2* mutations lead to the accumulation of (*R*)-2-hydroxyglutarate (2HG). 2HG is an oncometabolite which inhibits proteins of the TET family, and therefore causes an epigenetic dysregulation which results in a differentiation block. The FDA-approved IDH2 inhibitor Enasidenib (AG-221) reduces 2HG levels (154) and mutant IDH allele frequency (153).

Another targeted strategy for AML treatment relies on targeted delivery of cytotoxic drugs using antibody-drug conjugates (ADC). Gemtuzumab ozogamicin is a FDA-approved ADC for CD33 positive AML. Gemtuzumab ozogamicin is composed of a CD33 monoclonal antibody covalently attached to the DNA intercalant calicheamicin. This drug showed an improvement in event-free and overall survival (153).

Ultimately, patients in which these therapies are ineffective will undergo allogeneic stem cell transplantation (134), with the associated treatment-related mortality. Moreover, similarly to MPN patients, AML often occurs in patients >65 years old, making most of them ineligible for transplant and highlighting the need for new therapeutic approaches.

3.5. SECONDARY ACUTE MYELOID LEUKAEMIA

Secondary Acute Myeloid Leukaemia (sAML) is a group of diseases characterized by a previous history of haematological disorder (MDS or MPN) or which originate after treatment with a cytotoxic drug. Approximately 30 % of patients with MDS (155) and 10 % of patients with MPN (156) will develop sAML, which is associated with a dismal prognosis.

All types of MPN (ET, PV and MF) can progress to secondary AML. However, there are major differences in the risk of progression, which is much higher in MF (10-20% in the first 10 years) as compared to PV (3-8%) and ET (1%). Based on the stage of progression, three different phases are differentiated in these patients: chronic phase MPN refers to the myeloproliferative phase (PV, ET or MF), when there is less than 10 % blasts in PB or BM; accelerated phase (MPN-AP) refers to an intermediate stage with 10-19% blasts and blast phase (MPN-BP) or secondary AML, with >19% blasts in PB or BM.

Transformation to acute myeloid leukaemia in MPN patients is associated with certain clinical and genetic factors that are distinct from those seen in most cases of *de novo* AML. Risk factors predicting transformation include advanced age, anaemia, excess blasts, leucocytosis and importantly, mutations in certain genes (*TP53*, *EZH2*, *SRSF2*, *RUNX1*, *IDH2* and *ASXL1*) as well as cytogenetic abnormalities. The presence of four or more somatic mutations per patient and three or more cytogenetic abnormalities is also an strong predictor of leukemic transformation (157). Treatment with cytotoxic

agents such as radiophosphorous, busulpham or pipobroman (158, 159) have also been associated with leukemic progression in MPN.

Genetic aberrations associated with MPN-BP development are mostly distinct from those found in patients with *de novo* AML described before. For example, recurrently mutated genes in *de novo* AML such as *NPM1*, *FLT3-ITD*, *DNMT3A* or *CEBPA* are rarely mutated in MPN-BP. On the other hand, MPN-BP patients frequently carry a *JAK2V617F*, *CALR* or *MPL* mutation, which are very rarely found in *de novo* AML. Moreover, signalling mutations are usually acquired as a second hit in *de novo* AML, whereas mutations in the JAK/STAT signalling pathway are usually initiating mutations in MPN. Equally, epigenetic modifiers are usually acquired early in *de novo* AML, whereas these mutations are associated with progression from the chronic to the blast phase of the disease in MPN. This suggests different mechanisms of leukemic transformation in both types of malignancies, and that epigenetic mechanisms might be major drivers of disease progression as secondary events in MPN-BP.

Another question regarding MPN-BP evolution is whether mutations associated with progression are acquired *de novo* or are present at the chronic phase of the disease. Studies analysing paired chronic phase MPN and MPN-BP samples showed that *TET2* (160) and *TP53* (161) mutations were acquired at the time of leukemic transformation. Moreover, *TP53* loss was shown to cause leukemic transformation in a *JAK2V617F* mutant setting *in vivo* (161). In some cases, *TP53* mutations could be detected at low levels in the chronic phase of the disease (161), indicating that they expanded rather than emerged during leukemic transformation. On the contrary, *SRSF2* (162) and *ASXL1* (160) mutations were usually detected at high levels prior leukemic transformation, despite been associated with leukemic progression in MPN.

In some instances, MPN patients carrying a *JAK2V617F* mutation develop *JAK2V617F*-negative sAML (163), suggesting sAML development from a clonally

unrelated subclone different to the one giving rise to MPN. A similar phenomenon has been recently shown in MDS (164).

The poor sensitivity of the assays used to detect genetic alterations in the majority of studies presented above (Sanger sequencing or low-depth targeted next generation sequencing) limit the conclusive determination of drivers of transformation in MPN-BP. The use of highly sensitive next generation sequencing or single cell assays are required to determine if these mutations are present prior sAML development in the majority of patients, as well as the exact order of mutation acquisition during progression to the blast phase of the disease. The specific causative role of the genetic events leading to transformation has not been characterized yet.

This poor characterization of the genetic events driving transformation results in limited therapeutic options for these patients and consequently, extremely poor outcomes. MPN patients progressing to sAML have a median overall survival of 1.5-2.5 months when they only receive palliative care and 3.9-9.4 months when they receive treatment (165), and the five year survival rate is 4% (159). This is in contrast with a >10 year median overall survival for those patients who do not progress to sAML (166).

Induction chemotherapy is one of the first lines of treatment in MPN-BP. However, despite the fact that >40% patients achieve remission to a chronic myelofibrotic state, these remissions are short lived (167, 168). Ultimately, the median survival of patients treated with induction chemotherapy is only 4-5 months, indicating that new therapeutic approaches are urgently needed to this group of patients.

In keeping with the relevance of epigenetic dysregulation in disease progression, transformation to blast phase MPN is also associated with global DNA hypermethylation (169). Specific hypermethylation of p15INK4b and p16INK4a, which regulate cell cycle progression, has been described in MPN-BP patients (170). These genes are also commonly hypermethylated in other types of myeloid tumours

(MDS, AML) and other types of cancers (170), suggesting an oncogenic role. These lines of evidence led to the introduction of hypomethylating agents to treat MPN-BP, which also showed promising results in post-MDS sAML (171).

The most widely used hypomethylating agents in MPN-BP are 5'azacytidine (5'AZA) and 5'aza-2'-deoxycytidine (Decitabine). These are nucleoside analogues that inhibit the activity of DNA methyltransferases irreversibly (172). Treatment with 5'-AZA and Decitabine reduced global patterns of DNA methylation (173, 174). However, responses to these agents were only observed in approximately 25 % of the patients enrolled in clinical trials, and they did not correlate with reduction in DNA methylation (173). The duration of the response was short, most patients did not achieve complete remission and the median overall survival was only improved from 3-4 to 10-11 months (173, 175). Recently, a dual targeting approach combining Decitabine and ruxolitinib treatment showed synergistic *in vitro* activity in a mouse model of MPN-BP (161), suggesting that dual JAK/STAT and DNA methylation inhibition might be a promising therapeutic strategy. This led to a Phase 1 clinical trial, which unfortunately did not show an improvement in overall survival or response rates (176). Importantly, despite the constant development of novel therapies, there has not been an improvement in survival trends for the past 30 years (177).

Given the poor results achieved with current therapies, allogeneic stem cell transplantation remains the only curative option for MPN-BP patients, and it is the only treatment able to prolong survival (73 % of patients were alive after 31 months of follow-up in a recent study (167)). However, the high incidence of the disease in patients >70 years old (177) makes most of them ineligible for transplant, and 5-year survival rates are still extremely poor (10 %) (159). Therefore, the characterization of the genetic, molecular and epigenetic drivers of transformation is an active area of research, which will be used to inform the development of new therapies to improve the dismal outcome of MPN-BP patients.

3.6. *TP53* MUTANT SECONDARY ACUTE MYELOID LEUKAEMIA

Mutations or cytogenetic abnormalities in *TP53* are the most common genetic abnormality in cancer (95), and they are also recurrently present in myeloid neoplasms. Approximately 2-5 % of MPN patients (40, 62, 178), 5-10% of MDS patients (179) and 8% of *de novo* AML patients (151) carry somatic mutations in *TP53*. The presence of a *TP53* mutation is invariably associated with poorer prognosis and complex karyotype (3 or more cytogenetic abnormalities) (180, 181).

In patients with MPN, the presence of *TP53* mutations is strongly associated with transformation to sAML or MPN-BP (40). In particular, patients with homozygous somatic mutations in *TP53*, consequence of loss of heterozygosity, are at high risk of transformation (182). *TP53* mutations in the chronic phase of the disease are associated with older age (96), and treatment with cytotoxic therapies has been suggested to increase the risk of *TP53*-mediated transformation (183). However, it seems that in the specific case of hydroxyurea treatment, this effect might be partially confounded by the more frequent use of this therapy in older patients (96, 184).

TP53 (tumour protein 53) is a tumour suppressor gene which regulates a variety of cellular processes such as DNA repair, apoptosis or oncogenic proliferation (185). The high frequency of *TP53* mutations across all major cancer types (95) highlights its essential role in tumour suppression.

The P53 pathway is activated by a wide range of cellular stresses: oncogene expression, DNA damage, hypoxia, metabolic dysfunction, replicative stress, etc (185). *TP53* functions mainly as a transcription factor (185). P53 basal levels are regulated by MDM2 (Figure 1.7), an E3-ubiquitin ligase which ubiquitinates p53 for proteasomal degradation (186). MDM2 activity is also regulated by MDM4, which stabilizes MDM2. Cellular stresses activate p53, which forms an homotetramer, by inhibiting MDM2 or phosphorylating p53; once activated, addition of protein modifications (acetylation,

methylation) increase p53 stability and site-specific DNA binding. P53 recruits transcriptional cofactors and binds to p53 response elements, resulting in the transcriptional activation of target genes (187). Phosphatases such as PPM1D and deacetylases such as HDAC2 also regulate p53 transcriptional activity.

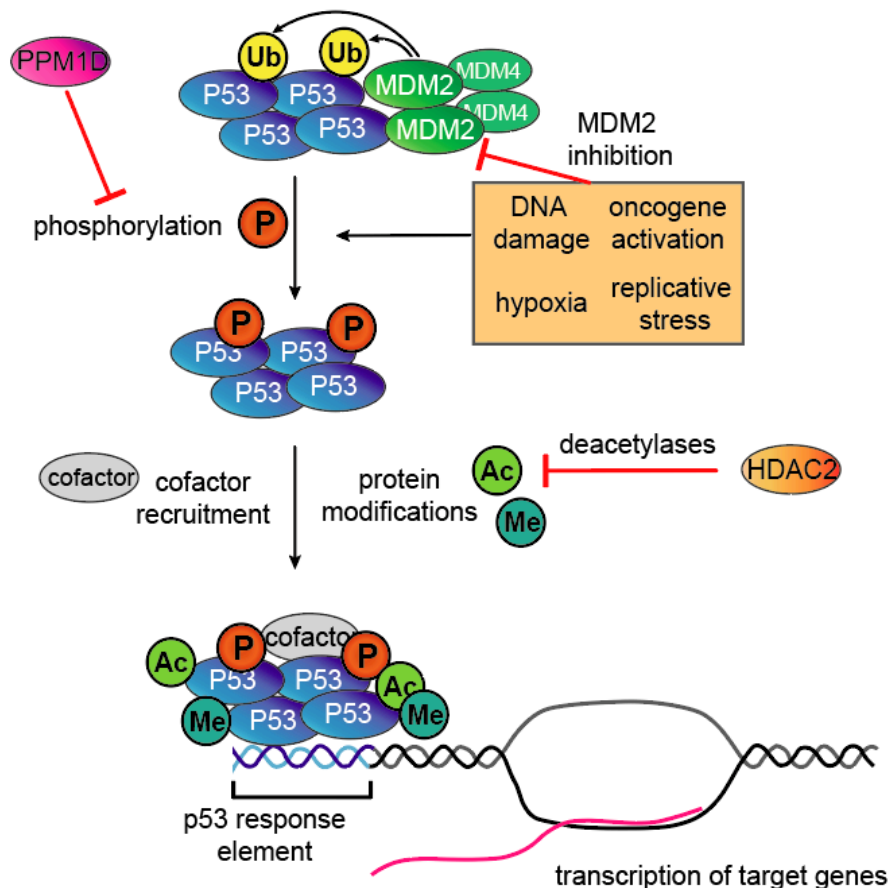


Figure 1.7. Schematic overview of the P53 pathway. Ubiquitin ligase MDM2 regulates the level of p53 protein in the cell. When a cellular stress is detected, p53 is phosphorylated and hence activated; subsequently, further protein modifications such as acetylation and methylation increase p53 stability, leading to cofactor recruitment and binding to p53 response elements in the genome, ultimately activating transcriptional targets.

In patients with myeloid malignancies, *TP53* mutations are usually missense mutations affecting the DNA binding domains of the protein (188), and up 70% of patients harbour two or more *TP53* mutations. Small deletions, insertions and premature stop codons are also frequent. *TP53* mutations are often homozygous: approximately 30% of patients with *TP53* mutations also carry 17p loss and 47 % of patients carry

hemizygous mutations. The most commonly mutated residues are R248 and R273 (188) (Figure 1.8), in the DNA binding domain of the protein, but mutations affecting the transactivation and tetramer formation domains are also frequent in AML (189).

TP53 mutations have been shown to contribute to acquisition of self-renewal and abnormal myeloid differentiation *in vivo* (190), as well as give rise to large genomic rearrangements (181). A MPN-BP mouse model in which *TP53* null mice were transduced with a retroviral *JAK2V617F* construct developed AML, suggesting that *TP53* loss and *JAK2V617F* cooperate to give rise to AML (161), although the specific mechanisms remain elusive.

Despite the vast number of studies characterizing the role of *TP53* genetic defects in the haematopoietic system, the functional role of *TP53* mutations and *TP53* loss (-17/-17q) in myeloid tumours is still subject of controversy. Whilst some studies have described a neomorphic gain-of-function activity of the most common *TP53* missense mutations in transgenic mouse models, human and murine cell lines (191), others could not find evidence of a gain-of-function effect in both isogenic human cell lines *in vitro* as well as in mouse models *in vivo* (192). This latest study proposed that *TP53* missense mutations in AML drive clonal expansion through a dominant-negative-effect rather than gain-of-function.

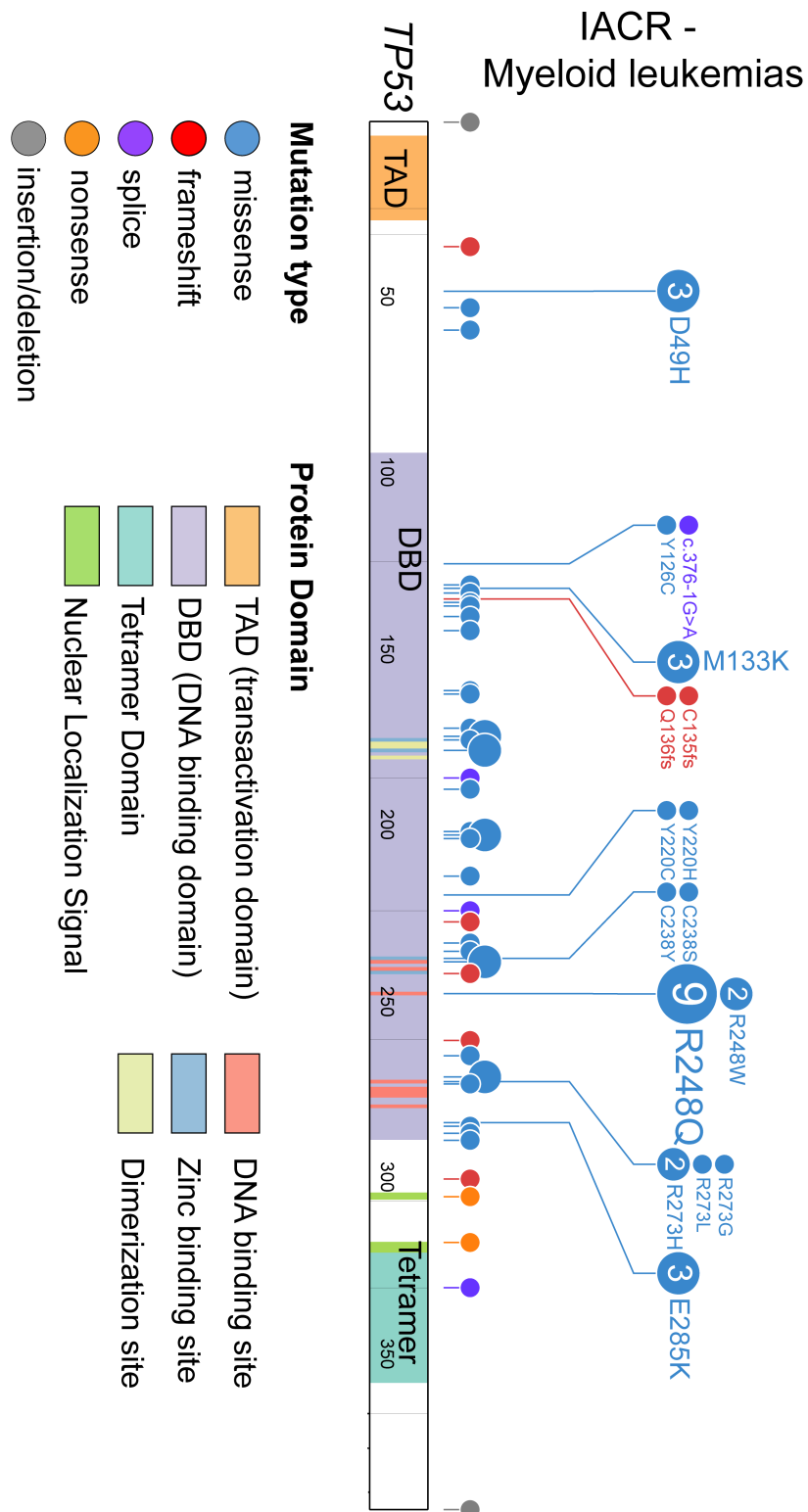


Figure 1.8. Schematic representation of p53 domains and recurrently mutated residues in myeloid malignancies. Residues mutated in myeloid neoplasms (acute myeloid leukaemia, chronic myeloid leukaemia, essential thrombocythemia, myelodysplastic syndrome, therapy-related leukaemia) were extracted from the IARC TP53 database (p53.iarc.fr) (193) and plotted using PeCan Data Portal (<https://pecan.stjude.cloud>). The colours of each circle indicate the

type of mutation identified in each residue, and the numbers inside each circle indicate the number of patients in which a mutation was identified. The numbers inside the rectangle represent the location of each residue in the schematic representation of the protein domains, which are coloured according to the legend provided.

The precise mechanisms of transformation in MPN-BP patients carrying *TP53* mutations remain unknown. *TP53* mutations can often be detected at very low levels (<5%) using NGS (Next Generation Sequencing) during chronic phase MPN, many years prior sAML diagnosis (96, 182, 194). Furthermore, patients with *TP53* mutations do not always transform to the blast phase of the disease (182). This indicates that the role of *TP53* mutations in MPN transformation might be context dependent and affected by other regulatory mechanisms currently unidentified. Given that *TP53* mutations are strongly associated with loss of chromosome 17p and complex karyotype, it raises the possibility that additional collaborating genetic events also cooperate in disease development.

The presence of *TP53* mutations in the chronic phase of the disease correlates with markedly reduced overall survival (182), and AML patients carrying *TP53* mutations are classified as unfavourable risk (Table 1.4), reflecting the dismal prognosis of the disease. This poor prognosis is partly explained by the fact that *TP53* mutations confer resistance to cytotoxic therapies (188), which restricts the therapeutic options for this subset of patients. Recently, a clinical trial using Decitabine in MDS and AML patients (195) showed higher responses rates in patients with *TP53* mutations and/or complex karyotype (100% and 67% response rates, respectively). Despite the short duration of the responses, Decitabine improved overall survival and the variant allele frequency of *TP53* mutations was markedly reduced. However, *TP53* mutant AML still remains a poorly understood disease with an extremely poor prognosis, stressing the importance of characterizing the mechanisms of transformation to develop better therapies.

4. SINGLE CELL TECHNOLOGIES AND THEIR APPLICATION TO STUDY NORMAL AND MALIGNANT HAEMATOPOIESIS

Traditional approaches to study haematopoiesis usually involved the use of small subset of cell surface markers (Table 1.1) to purify cell populations and characterize their molecular and functional properties. However, this population-based analysis masks any underlying heterogeneity within immunophenotypically-defined cell types, and therefore limits the conclusions that can be reached with these types of assays.

Moreover, during the development of myeloid malignancies (as well as many types of cancers), the serial acquisition of mutations in different stem and progenitor cells gives rise to genetically-heterogeneous populations. This is commonly known as intratumoral heterogeneity (196-198), and occurs not only at the genetic level, but also at the transcriptomic, epigenomic and microenvironmental levels (199, 200).

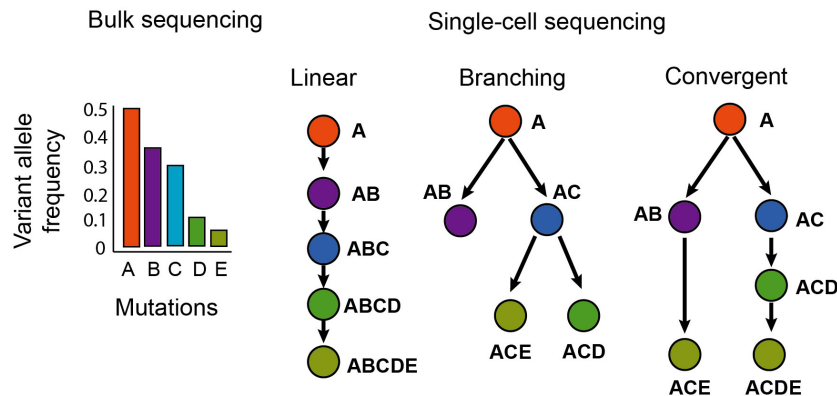
Tumours are highly dynamic and complex ecosystems under constant selective pressure to evolve and adapt to adverse conditions such as treatment. Characterizing each of the layers of intratumoral heterogeneity is essential to understand the biological mechanisms of tumour initiation, development, progression, treatment resistance and relapse. Ultimately, unravelling this multi-layered heterogeneity will lead to a more comprehensive understanding of leukemogenesis, with huge potential to guide the design of more effective therapeutic strategies. To resolve this heterogeneity, it is essential to study the fundamental biological unit: a single cell.

During the past decade, advances in molecular biology and next generation sequencing led to the development of new technologies enabling the analysis of the genome, transcriptome, proteome and epigenome from individual cells (201). These techniques now allow to study the composition of cellular systems with much higher resolution. For example, bulk sequencing techniques were traditionally used to study the clonal evolution of tumours and combinational patterns of mutation, by inferring

mutation order from the variant allele frequency (Figure 1.9A). However, these techniques do not allow researchers to confidently distinguish between linear, branching and convergent patterns of clonal evolution, or to confidently identify mutation co-occurrence (201). The isolation and subsequent molecular profiling of distinct pre-defined cell populations at the bulk level also has major limitations (Figure 1.9B). Bulk gene expression analysis techniques, such as RNA-sequencing, generate average expression values of all of the cells present in a particular cell population; however, this type of analysis precludes the identification of subpopulations with distinct molecular signatures within these pre-defined cell types. This molecular heterogeneity also has crucial implications during cancer progression and treatment; particular subpopulations of cells might be selectively resistant to therapies and might drive disease relapse (Figure 1.9C). The molecular characterization of those persistent subpopulations can provide crucial insights into the mechanisms of drug resistance.

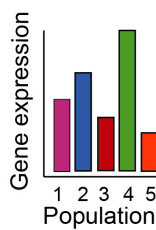
Therefore, the continuous development and application of single-cell technologies to study myeloid malignancies has the potential to provide mechanistic insights into the development and progression of these diseases, as well as the development of better targeted treatments.

A Tumour clonal evolution

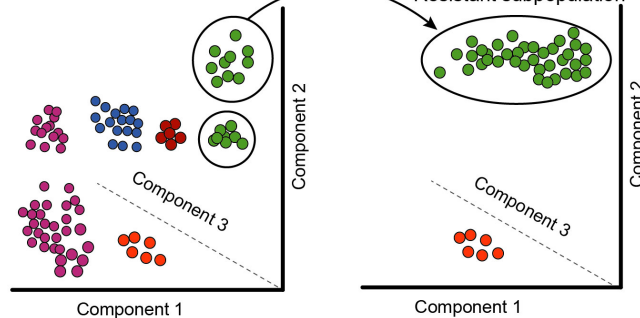


B Transcriptional heterogeneity

Bulk sequencing



Single-cell sequencing



C Therapy resistant subpopulations

Single-cell sequencing

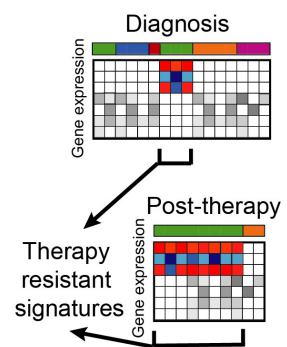


Figure 1.9. Single-cell sequencing unravels masked heterogeneity not captured by bulk sequencing techniques. (A) Tumour phylogeny inference from variant allele frequencies obtained from bulk sequencing assays do not distinguish between different patterns of clonal evolution (linear, branching, convergent), which can only be definitely resolved using single-cell sequencing techniques. (B,C) Gene expression analysis of cell populations does not distinguish between molecular subpopulations within pre-defined cell types (B, left panel). Single cell gene expression analysis allows unbiased clustering of single cells based on their molecular profile (B, right panel) and discovery of new subpopulations of cells which might be resistant to drugs (C, left panel). This ultimately allows to characterize molecular signatures of therapy resistance (C, right panel). Adapted from (201).

4.1. SINGLE CELL TRANSCRIPTOMIC TECHNIQUES

Single cell transcriptomic techniques are undoubtedly the most commonly used single cell technology to explore the heterogeneity of cellular systems. Their high scalability

and the constant technological improvements in the field have made them widely established.

Single-cell qPCR was the first technique used to obtain gene expression readouts from single cells (202, 203), usually allowing the detection of up to 96 genes per cell. Single-cell qPCR provides high-sensitivity amplification of targets, making it the technique of choice to confidently identify gene expression differences in lowly expressed transcripts (204). However, single-cell qPCR is limited to the amplification of a preselected panel of genes, which might not fully capture the molecular heterogeneity of the populations studied. Moreover, qPCR is a relatively low throughput technique despite been implemented in microfluidic platforms which increased scalability (204, 205) .

Subsequently, the development of whole-transcriptome single-cell methods allowed the unbiased discovery of gene expression differences between single cells. The first single-cell whole transcriptome technique was developed in 2009 (206), rapidly followed by a rapid expansion in the number of new single-cell RNA-sequencing (scRNA-seq) technologies per year (207), with more than a hundred different methods available to date.

Single cell transcriptomic techniques rely on the efficient retrotranscription and amplification of minute amounts of RNA from individual cells. This is associated with a high degree of technical noise, which should be taken into account when designing experiments and comparing different scRNA-seq technologies. Common metrics used to assess the efficiency of amplification are: number of genes detected per single cell, percentage of reads mapping to exonic regions, dropout probability and accuracy. The number of genes detected per cell is a measurement on how good a particular technique is at amplifying a high percentage of the total transcripts present in a particular cell (sensitivity of the technique); the percentage of exonic reads assesses

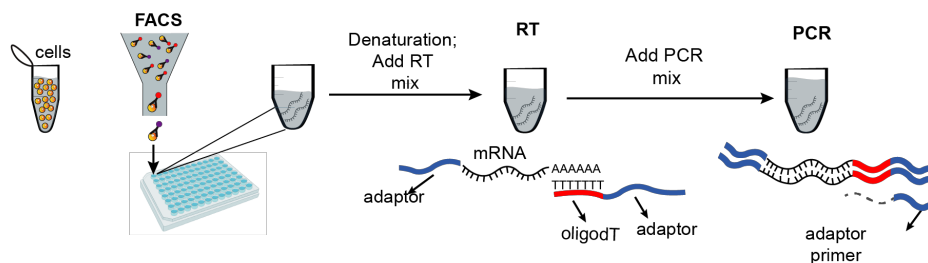
how specific is the technique at amplifying transcripts versus other regions of the genome (which would represent unspecific amplification or noise). The dropout probability (which is a measurement of the precision of the technique) is the probability that a certain mRNA molecule present in the mix is not detected in a particular cell. Finally, the accuracy of the technique refers to how well the quantification of a particular mRNA species reflects the true concentration of that particular mRNA in the mix. This is usually measured using mRNA control molecules (External RNA Controls Consortium spike-ins or ERCCs) added at known concentrations in the mix. Ultimately, the performance of each technique for these metrics informs the selection of a particular method for each type of application (208-210).

scRNA-seq techniques can be classified according to the platform used to isolate and generate cDNA libraries (plate-based or droplet-based; Figure 1.10) and the type of library preparation and sequencing configuration (full-length, 5'-biased or 3'-biased; Figure 1.12). The most common plate-based technique is SMART-seq 2 (211) and the most commonly used droplet-based technique, the Chromium Single Cell Expression Solution by 10X Genomics (more commonly referred to as Chromium) (212). Plate-based techniques were the first single-cell techniques developed, but they have limited throughput. The complexity of cellular systems soon required the development of higher throughput techniques allowing analysis of tens of thousands of cells, such as droplet-based technologies.

Plate-based scRNA-seq techniques rely on isolation of single cells usually by a FACS (Fluorescent Activated Cell Sorting) instrument, which encapsulates each single cell into individual nanoliter-droplets and deposits them into one well of a PCR plate containing lysis buffer (Figure 1.10A). This lysis buffer consists of a detergent (i.e., Triton X-100, NP40), dNTPs, an oligodT primer tagged to a PCR adaptor, and optionally, ERCC spike-in synthetic mRNA molecules (used as control mRNA). Cells

are directly lysed once deposited on the lysis buffer and subsequent reactions are performed in the same PCR plate.

A PLATE-BASED TECHNIQUES



B DROPLET-BASED TECHNIQUES

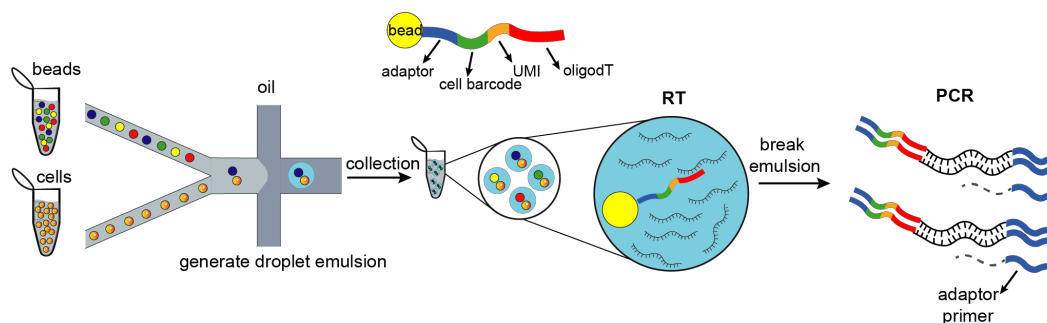


Figure 1.10. Schematic workflow representation of the most widely used single-cell RNA-sequencing techniques. (A) Plate-based techniques rely on FACS to perform single-cell isolation, and reagents are consecutively added to individual wells of a PCR plate, each of them containing a single cell. (B) Schematic representation of droplet-based techniques. In this type of techniques, cells are encapsulated with beads in an emulsion. Each bead contains a primer with a unique cellular barcode, to allow prospective identification of mRNA molecules from each single cell. Cells are collected in a single microfuge tube to perform retrotranscription. After that, the emulsion is broken and cDNA molecules are amplified by PCR. Some elements of this figure were generated using BioRender.

In droplet-based techniques, single cells are introduced into a microfluidic device which flows two aqueous solutions (one containing beads and the other one containing cells and reagents) into an oil channel (Figure 1.10B). This encapsulates a bead and a single cell into a single droplet, generating an emulsion of hundreds of thousands of droplets (212, 213). Cells are introduced in the microfluidic device at limiting dilution, which minimizes the chances of capturing two single cells within the same bead. The beads are functionalized with oligonucleotides comprising an adaptor sequence, a cell

barcode, a Unique Molecular Identifier (UMI) and oligodT sequence. Lysis occurs immediately after cell capture.

After lysis, the retrotranscription mix (RT) is added to the same PCR plate in plate-based techniques. The RT mix contains a Moloney Murine Leukaemia Virus derived (MMLV-derived) retrotranscriptase (RTase), RNase inhibitor and a template-switching oligonucleotide (TSO). In this step, the oligodT primer added during the lysis step is used to prime polyadenylated mRNA molecules, which are subsequently retrotranscribed by the RTase. This oligodT primer is usually tagged to a known adaptor sequence (SMART-seq, SMART-seq2) and in some scRNA-seq techniques (CEL-Seq2, MARS-seq), it also contains a unique molecular barcode (UMI) and unique cell barcode (CB). The addition of this adaptor sequence allows subsequent amplification of cDNA molecules during PCR.

In droplet-based techniques, droplets are collected into a single microfuge tube. The beads release the oligonucleotides (containing adaptor sequence, cell barcode, UMI and oligodT) and retrotranscription occurs within each droplet. Droplet-based techniques also use MMLV-derived RTases and a template-switching oligonucleotide (TSO) during RT reaction.

MMLV-derived retrotranscriptases (RTases) add a few extra bases (mostly deoxycytidine) at the 3'-end of each cDNA molecule during first strand synthesis, thanks to their terminal transferase activity. This can be exploited to design an oligonucleotide complementary to the deoxycytidines added by the RTase, which is known as template-switching oligonucleotide. Once the RTase reaches the 5'-end of the RNA template, it switches strand (from the RNA molecule to the TSO) and continues replication until the end of the cDNA molecule. This strategy is known as Switching Mechanism at the 5'-end of the RNA transcript (SMART) (214, 215), and it

generates full-length molecules surrounded by specific adaptor sequences tagged to the oligodT and TSO molecules (Figure 1.11).

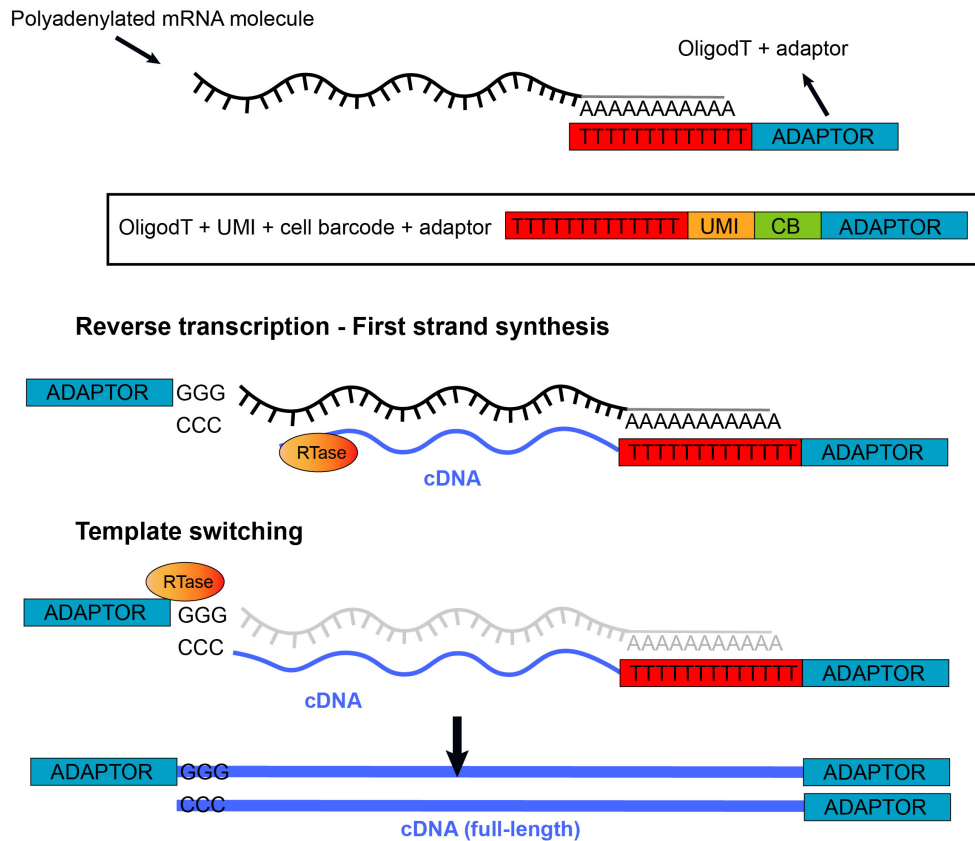


Figure 1.11. Generation of full-length cDNA molecules using a Template Switching Oligonucleotide and a MMLV-derived RTase. An oligodT primer tagged with an adaptor sequence (and alternatively, a cell barcode and UMI) is used to prime cDNA molecules. RTase will retrotranscribe the mRNA until it reaches then end of the molecule, and will add extra cytosines thanks to its terminal transferase activity. The TSO, complementary to the extra cytosines added on the 3'-end of the primer, will bind to the end of the newly synthesized cDNA molecule and the RTase will switch strand, continuing replication until it reaches the 3'-end of the cDNA molecule, generating full-length cDNA.

Template switching mechanism is based on 3 guanoses added to the 3'-end of the TSO. To increase the efficiency of this pairing, several studies modified of the 3'-end of TSO substituting deoxyguanoses by riboguanosines (214, 216), isomeric nucleotides (217) or locked nucleic acids (211, 216). The later enhances thermostability of the DNA/LNA hybrid and results in increased yield.

In a subsequent PCR step, a primer binding to the adaptor sequences added during the RT is used to amplify cDNA libraries from each single cell (Figure 1.10). In plate-based techniques such as SMART-seq2, PCR buffer is added to each well containing a single-cell derived cDNA library. In droplet-based techniques, the droplet emulsion is broken and uniquely-barcoded cDNA molecules are amplified in a single reaction, using a primer complementary to the adaptor sequence added during the RT step. The purpose of this PCR step is to generate enough material for subsequent library preparation or downstream applications.

Library preparation methods can generate full-length, 3'-biased or 5'-biased libraries for sequencing, depending on the part of the cDNA molecule which will be sequenced in each case (the entire molecule, or just the 3' or 5' ends, respectively). All of them rely on random DNA fragmentation (also known as tagmentation) using an hyperactive transposase optimized for low-input DNA amounts (such as Tn5 (218); Figure 1.12). This transposase is loaded with overhang adaptor sequences, which are tagged to each end of the tagmented DNA and are subsequently used during PCR to amplify these fragments and at the same time, add Illumina-compatible sequencing adaptors for flow-cell binding as well as a unique index sequence.

To generate full-length libraries (Figure 1.12A), a transposition enzyme is performed independently for each single-cell derived cDNA library, making this protocol labour intensive and quite costly. However, this type of sequencing configuration allows the detection of splice sites and is associated with higher sensitivity and gene-capture rates (208), making it suitable for differential isoform analysis or the study of rare cell populations with high depth. Given that each single-cell derived library is independently tagmented and barcoded with a unique Illumina-compatible adaptor containing an index sequence, only plate-based techniques are suitable for full-length library preparation protocols.

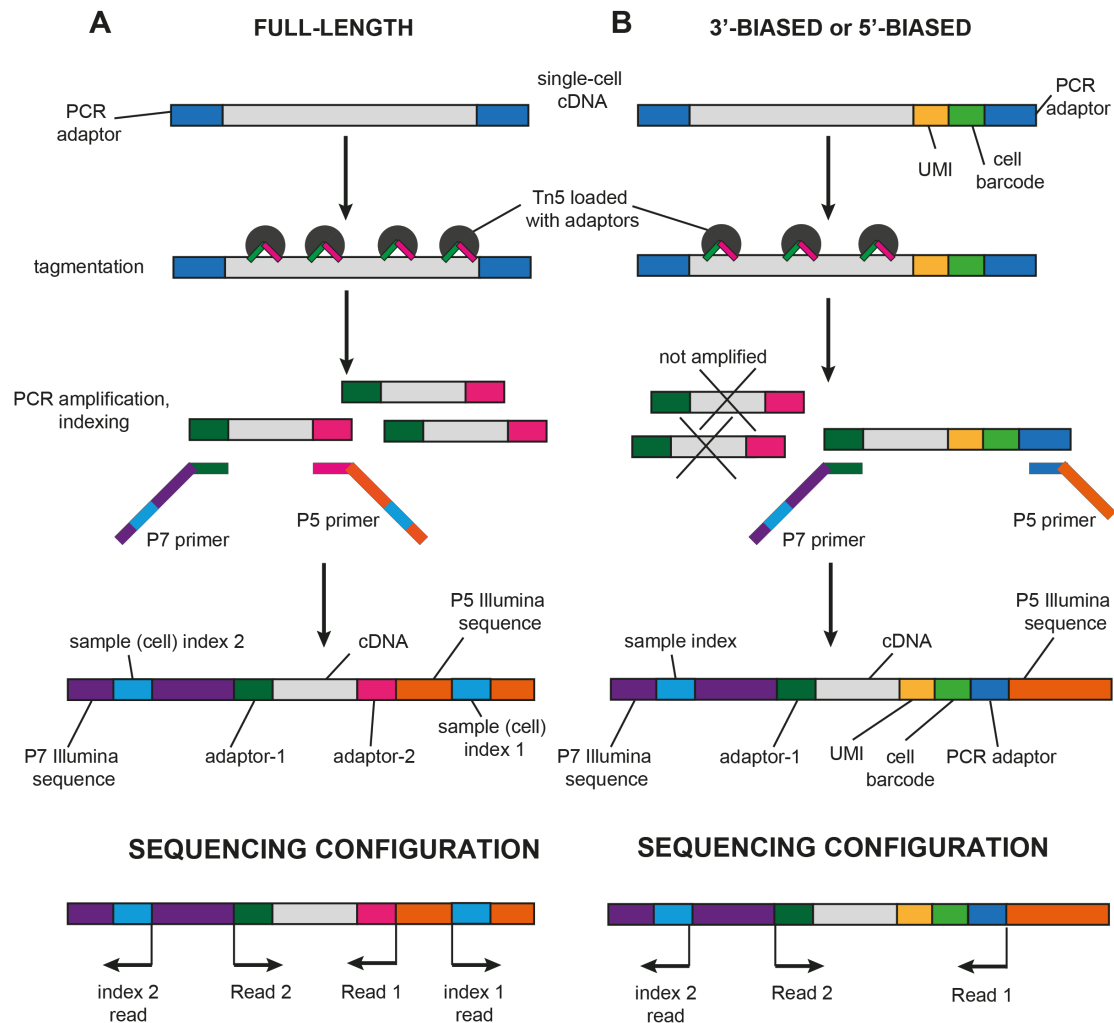


Figure 1.12. Sequencing library preparation methods and sequencing configurations for full-length (A) and 3'-biased or 5'-biased approaches (B). Single-cell cDNA is tagmented and PCR amplified using adaptor sequences introduced by the transposase (Tn5). In full-length methods, the forward and reverse adaptor sequences are used for PCR amplification; in 3'-biased or 5'-biased methods, the adaptor sequence previously added during PCR and one of the adaptor sequences added during tagmentation are used to amplify the libraries, which preserve the cell barcodes.

In droplet-based techniques, cDNA amplification is performed in a pool of cells, each of them tagged to a unique cell barcode sequence in the 3'-end of the molecules (in some cases, the cell barcode can also be introduced on the 5'-end of each molecule (219)). Tagmentation and addition of adaptor sequences is also performed in this amplified cDNA pool, but to retain the information from each cell barcode, only the fragments containing the cell barcode (3'-end or 5'-end) are amplified for sequencing (Figure

1.12B). To achieve this, a primer binding to the 3'-end or 5'-end of the molecules is used during PCR amplification of the tagmented libraries, to specifically enrich for these fragments (Figure 1.12B). This strategy allows demultiplexing each cDNA molecule according to the original cell barcode it was tagged to *a posteriori*. Thanks to the one-reaction setup, this type of library preparation is high-throughput and extremely economical. However, it is associated with much lower sensitivity (208, 210).

This type of sequencing configuration is also usually termed 3' or 5' molecule counting, because it allows accurate counting of the number of molecules mapped to each gene in each single cell. Thanks to the presence of the UMI sequence, which uniquely barcodes each mRNA molecule before amplification, duplicate counts generated from the same initial mRNA molecule during PCR amplification can be removed. However, the limitation of this sequencing configuration is that it is restricted to gene-level counting, and isoform detection is not possible using Illumina short-read sequencing. Recently, new protocols have incorporated long-read sequencing of pre-amplified cDNA libraries to obtain isoform-level information from droplet-based techniques (220, 221). However, the high error rates of these technologies (222) does not allow confident identification of cell barcodes.

4.2. SINGLE CELL GENOMIC TECHNIQUES

Single cell genomic techniques allow profiling the genetic landscape of single cells to detected germline genetic variation, somatic mutations, gene fusions or copy number variations. Single cell genomic techniques can be mainly classified in discovery-type, providing genome-wide or exome-wide genomic DNA coverage, or targeted techniques, which rely on the amplification of a small panel of pre-selected genes. Each of them has their own benefits and limitations, which mostly restrict their use for certain types of applications.

The most widely used single-cell whole genome amplification method is based on multiple displacement amplification (MDA). This method relies on random hexamers (6 base pair oligonucleotides) annealing to previously denatured DNA, followed by strand displacement DNA synthesis by a Φ 29 polymerase at constant temperature (223). This technique allows confident identification of CNVs, but suffers from uneven genome coverage, high allelic dropout rates (>25%), and a high percentage of false positive results (>5%) associated with the low fidelity of the Φ 29 polymerase (224, 225). This makes the method unsuitable to identify point mutations in cancer samples. To help address this problem, a wide range of bioinformatic tools have been developed to correctly resolve the branches of clonal trees when applied to the study of tumour evolution (226-228).

Single-cell whole exome sequencing relies on prior single-cell whole genome amplification and subsequent hybridization capture targeting regions of interest. This approach has been used to identify mutations in breast cancer genomes (225), clonal evolution in myeloproliferative neoplasms (229) and kidney tumours (230). However, it has not been widely implemented due to the low sensitivity to detect mutations, false positive rates and high cost (owing to the library preparation method and sequencing depth required).

Targeted mutation analysis approaches have a much higher sensitivity of mutation detection and much lower allelic dropout (approximately 5%) and false positive rates. These techniques rely on cell sorting, subsequent genomic DNA lysis from each single cell and amplification of target regions using target-specific primers (231). They can confidently identify point mutations, small indels, copy number variations and gene fusions (231-233), and their throughput is usually restricted to hundreds to few thousands of cells.

More recently, the throughput of targeted mutation analysis has been massively increased by implementing this method in a droplet-based platform (234). This platform (Tapestri®) now allows targeted mutational analysis of up to 300 different amplicons in >10,000 cells per sample, providing a 10-100 fold throughput compared to previous studies. Moreover, it retains the sensitivity of single-cell targeted mutational analysis approaches, with an allelic dropout rate estimated at 5 %.

Despite the relatively limited number of regions (tens to hundreds) that can be profiled using targeted-methods, they provide a much higher sensitivity of mutation detection, and they are the preferred choice for resolving clonal evolution in haematopoietic tumours (231-233) due to their relatively low mutational burden.

4.3. SINGLE CELL MULTI-OMIC TECHNIQUES

The integration of several genomic, proteomic and epigenetic readouts from the same single cell is an extremely powerful approach to decode the relationships between different cellular modalities (235-237). This will allow dissecting the regulatory mechanisms governing cellular hierarchies both in healthy and disease settings, and will ultimately contribute to our understanding their coordinated regulation at unprecedented sensitivity.

For example, one of the major limitations to investigate intratumoral heterogeneity using single-cell analysis when this study first started was the integration of mutational and transcriptomic readouts from the same single cell. Tumours are genetically and molecularly heterogeneous, hence characterizing the molecular (and functional) consequences of the serial acquisition of somatic mutations in single cells is of paramount importance to understand cancer evolution and response to therapies. Linking genetic and whole transcriptome readouts would allow genotype-phenotype correlations (inferred from the transcriptome) to be determined unambiguously. However, the high dropout rate of single-cell transcriptomic techniques prevented the

correlation of transcriptional and genetic readouts from the same cell (1, 238); likewise, single cell genomic techniques were not compatible with parallel whole transcriptome amplification from the same single cell.

A number of studies (238-240) tried to identify point mutations from full-length scRNA-seq data, but failed to do so in the vast majority of cells (0.3-9%) (238). Other studies developed methods to physically separate genomic DNA and mRNA from each single cell, and perform targeted mutation and whole transcriptome amplification from each cell fraction respectively (G&T-seq, SDR, scTRIO-seq) (241-243) or parallel amplification of both types of molecules and subsequent masking of coding regions (244). Again, these approaches had high allelic dropout rates which prevented the reliable detection of point mutations. Parallel mutation and transcript analysis could only be achieved with enough sensitivity using targeted gene expression assays (245, 246) in a limited number of genes and transcripts.

The first method confidently identifying mutations in parallel with whole transcriptome analysis of the same single cell was pioneered by the Mead and Jacobsen groups in the context of Chronic Myeloid Leukaemia (247), driven by the *BCR-ABL1* fusion gene. In this study, we adapted the widely used SMART-seq2 technique to allow parallel targeted amplification of *BCR-ABL1*, which provided single-molecule sensitivity to detect this transcript in single cells. However, this approach was ineffective for other point mutations due to the high allelic dropout rate at the transcript level (1, 247).

Techniques integrating mutational and whole transcriptomic readouts have since been developed, mainly based on two different approaches: building on the parallel targeted amplification of mutant transcripts developed by Giustacchini and colleagues (248-250), or by incorporating parallel amplification of mutations from genomic DNA of the same cell (1, 251). These techniques, which now allow parallel interrogation of

mutations and whole transcriptomes from the same single cell, will be discussed later in the context of the results of this work.

Another powerful approach for the multimodal analysis of single cells is parallel epigenomic and transcriptomic analysis of the same single cell. This was initially achieved by physically separating genomic DNA (using G&T-seq and scTRIO-seq methods) and performing bisulphite conversion on the purified DNA fraction to obtain single-cell methylomes. This could also be achieved using a targeted assay for DNA methylation, transcripts and mutations (245), which allows the analysis of tens of loci per cell and revealed methylation patterns associated with the presence of *EGFR* mutations in lung cancer. Further improvements of these methods incorporated parallel chromatic accessibility (252, 253) in parallel with whole transcriptome analysis, and recently developed methods integrated DNA-protein interactions (through single-cell DNA adenine methyltransferase identification; DamID) and RNA-sequencing of the same cell (254, 255). This latest method, named scDam&T-seq, revealed how genome-lamina contacts correlated with gene expression in individual cells.

Additionally, the development of new computational tools for the analysis and integration of several layers of genomic information from the same sample made feasible multi-omics single-cell profiling. These computational methods (256, 257) integrate several datasets from the same sample obtained by independent genomic readouts (for example, by performing scRNA-seq and single-cell ATAC-seq independently from the same group of cells). They rely on the assumption that different cell types are equally represented in each method, and that certain features within the different datasets are correlated, allowing the cells to be jointly clustered.

In summary, the parallel interrogation of several genomic layers from individual cells has the potential to significantly improve our understanding of cellular decision making

in normal and cancerous tissues, building a much more comprehensive molecular snapshot of each cell.

4.4. APPLICATION OF SINGLE CELL TECHNOLOGIES TO STUDY THE HAEMATOPOIETIC SYSTEM

Many of the single cell techniques described above were pioneered on the haematopoietic system. Haematopoiesis is one of the best characterized systems of cellular differentiation, and how this is disrupted in disease, making it an ideal model to apply single cell technologies (201, 258). Indeed, the use of single cell assays led paradigm shifts in our understanding of lineage specification during haematopoietic differentiation (16, 259) and clonal evolution during tumour development and relapse.

One of the most prominent applications of single cell analysis in the haematopoietic system has been to characterize lineage specification, lineage priming and developmental routes leading to specific differentiation pathways. Single cell transplantations (10), single-cell whole transcriptome (260, 261) and single cell targeted gene expression analysis (11, 262) identified a distinct route for megakaryocyte differentiation emerging directly from single haematopoietic stem cells in both mouse and human. High-throughput single-cell analysis has also been later used to identify abnormal pathways of megakaryocyte differentiation in patients with myelofibrosis (263). Moreover, integration of single cell transcriptomics, index sorting and single-cell functional assays (264) was used to refine the purification strategy of murine HSCs with durable self-renewal, highlighting the power of single cell technologies to identify haematopoietic cell subtypes.

Single cell analysis also offers an unprecedented opportunity to study intratumoral heterogeneity in the haematopoietic system, which is one of the major causes of therapy resistance. The presence, in the same tumour, of many different genetic and molecularly distinct subclones makes it difficult to target all subpopulations of tumour

cells. Therefore, the identification of specific subpopulations of cells remaining after therapy or emerging during disease relapse could lead to the development of better treatments.

Single-cell gene expression analysis was used to identify therapy resistant leukemic stem cells in patients with CML treated with tyrosine kinase inhibitors (TKI) using a targeted (246) or scRNA-seq approach (247). Using single cell transcriptomics, we demonstrated the emergence of distinct *BCR-ABL1*⁺ subpopulations of tumour cells after TKI therapy, and how different molecular signatures of *BCR-ABL1*⁻ cells were associated with molecular response to TKI treatment.

Single-cell targeted mutation profiling in AML patients (265) revealed marker clonal complexity after treatment with the FLT3 inhibitor Quizartinib. Another recent study using high-throughput single-cell targeted sequencing of myeloid hotspots also identified the emergence of subclonal RAS pathway mutations in AML patients treated with the FLT3 inhibitor Gilteritinib. Single-cell analysis revealed that RAS mutations were acquired in FLT3-ITD mutant clones, and that they were present at low levels prior to Gilteritinib treatment, suggesting that they expanded in the presence of the drug (266).

Single cell epigenomic techniques were also pioneered on the haematopoietic system (267-269). A groundbreaking study applying single-cell ATAC-seq to HSPCs assessed the degree to which epigenetic heterogeneity contributes to the development of AML (267). Single leukemic stem cells were heterogeneous and showed mixed “regulome” profiles corresponding to divergent cellular states, indicating the aberrant differentiation stage of AML cells. Interestingly, many studies have described now that single-cell epigenomic readouts can better resolve cellular states as compared to single-cell transcriptomic techniques (267, 270-272). Given the high frequency of mutations affecting epigenetic modifiers in myeloid tumours (40, 151), these techniques are likely

to be extremely powerful to characterize epigenetic and transcriptional mechanisms driving the development of myeloid malignancies.

Single-cell genomic techniques have revolutionized our understanding of physiological and malignant haematopoiesis. The development and application of novel techniques to resolve previously uncharacterized heterogeneity has the potential to conceptually change many paradigms in the field. This will hopefully lead to a better characterization of mechanisms of leukemogenesis, ultimately leading to the development of better therapies.

5. AIMS OF THIS WORK

The high degree of genetic and molecular heterogeneity acquired during the development of myeloid neoplasms leads to diverse tumour populations with different genetic and molecular properties. Therefore, targeting all the different subpopulations of tumour cells is challenging, and as a result, currently available therapies are often unable to completely eradicate the disease.

As MPN stem cells are the unit of selection driving cancer development, I hypothesized that resolving the heterogeneity of MPN HSPCs using high sensitivity single cell analysis will lead to a better understanding of the cellular and molecular mechanisms driving disease initiation, progression and ultimately, to the development of biomarkers of disease and better treatments.

However, studying MPN HSPCs at the single cell level has a number of associated challenges:

- The acquisition of somatic mutations in MPN gives rise to genetically heterogeneous subpopulations. Therefore, to understand the molecular consequences of this genetic heterogeneity, it is essential to correlate genetic and transcriptomic readouts from the same single cell.

- To confidently identify tumour clonal hierarchies, it is essential to perform high resolution mutation detection with low allelic dropout rates. This includes the confident identification of loss of heterozygosity events (which are of high relevance in the context of *JAK2V617F* mutations in MPN).
- Parallel interrogation of cell surface markers would be essential to trace back mutations and molecular readouts to canonically defined haematopoietic cell populations.

To solve these limitations, the specific aims of this project were to:

- (1) develop a technology for the parallel interrogation of genetic, molecular and cell surface proteomic readouts from the same single cell.
- (2) apply this technology to the study of HSPCs of patients with myeloid neoplasms (MPN and sAML) to characterize patterns of clonal evolution and molecular signatures of the disease.
- (3) identify mutant-specific biomarkers and molecular pathways amenable of therapeutic targeting, to ultimately be able to specifically eliminate mutant cells.

CHAPTER 2: MATERIALS AND METHODS

1. BANKING AND PROCESSING OF HUMAN SAMPLES

Samples were banked under the ethical approval of the INForMeD Study (REC:199833, University of Oxford). All patients and normal donors provided written consent in accordance with the Declaration of Helsinki for sample collection and use in research.

Peripheral blood and bone marrow samples were first diluted 1:1 with Phosphate Buffered Saline (PBS) and filtered through 70 μ m cell strainers to remove any solid particles. Then, 20 mL of sample were deposited in a 50 mL Falcon tube containing 10 mL of Ficoll-Plaque (GE Healthcare; Catalog# 17-5442-03) to perform mononuclear cell isolation by gradient centrifugation (350 g during 30 minutes at room temperature). During centrifugation, erythrocytes and granulocytes sediment in the bottom layer and mononuclear cells, at the interface between Ficoll-Plaque and plasma. Mononuclear cells were aspirated from this interface and washed with Roswell Park Memorial Institute (RPMI; Gibco, Cat# 21056-023) medium supplemented with 10 % Foetal Bovine Serum (FBS; Sigma, Cat# F7524) and 1% DNase I (1mg/mL; Merck). Then, cells were counted, resuspended in FBS supplemented with 10 % Dimethyl Sulfoxide (DMSO) at a concentration of 5-25 million cells per mL, and aliquoted into cryovials. All samples were processed and cryopreserved within 48 hours of collection.

Granulocytes were isolated from the gradient centrifugation pellet. 15 mL of Red Blood Cell Lysis solution (Qiagen, Catalog No. 158904) was added to 5 mL of pellet, incubated for 10 minutes at room temperature and centrifuged at 2000 g for 5 minutes. The supernatant, containing erythrocyte lysate, was discarded, and granulocytes were washed and stored in microfuge tubes at -80°C for subsequent DNA extraction.

Samples were thawed by introducing cryovials in a waterbath set at 37°C. Each sample was diluted with 1 mL of FBS and transferred dropwise to a 15 mL Falcon tube. Then, each sample was further diluted with 8 mL of RPMI supplemented with 10 % FBS + 11% DNase I ("Thaw Media"). Samples were centrifuged at 350 g for 10 minutes, washed with 10 mL of Thaw Media, and cells were counted with an haematocytometer. Samples were washed once more by adding 10 mL of RPMI supplemented with 10 % FBS ("FACS Buffer"), supernatant was removed and a maximum of 25 million cells per sample were resuspended in 105 µL of FACS Buffer for subsequent staining by fluorochrome-conjugated antibodies.

2. CELL LINES

MOLT4, JURKAT and K562 cell lines were obtained from the American Type Culture Collection. NALM6 cell line was obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ). SET2 cell line was kindly provided by Dr. Jacqueline Boulwood (Radcliffe Department of Medicine, University of Oxford). All cell lines were cultured in RPMI supplemented with 10 % FCS in a humidified incubator. Cell line authentication was performed by targeted sequencing of known mutations and Single Nucleotide Polymorphism Array karyotyping.

3. GENOMIC DNA EXTRACTION AND BULK SEQUENCING OF MONONUCLEAR CELLS

Genomic DNA extraction from mononuclear cells or granulocytes was performed with DNeasy Blood and Tissue DNA extraction kit (Qiagen; Catalog#69504) following manufacturer's instructions.

50-250 ng of genomic DNA were used to perform library preparation for targeted next generation sequencing. Libraries were prepared using a TruSeq Custom Amplicon panel (Illumina), containing 341 amplicons designed around 32 exons of frequently mutated genes in myeloid malignancies (MDS, AML, MPN; Table 2.1) (1, 273) and

were sequenced on a MiSeq instrument (Illumina). Amplicons covered were chosen on the basis of been frequently mutated in myeloid malignancies or likely to alter clinical practice; this is, with diagnostic, prognostic, predictive or monitoring value.

Gene	Exons Targeted	Gene	Exons Targeted
<i>ASXL1</i>	12	<i>KRAS</i>	2, 3
<i>ATRX</i>	8, 9, 10, 17-31	<i>MPL</i>	10
<i>CBL</i>	8-9	<i>NPM1</i>	11
<i>CBLB</i>	9-10	<i>NRAS</i>	2,3
<i>CBLC</i>	9-10	<i>PDGFRA</i>	12, 14, 18
<i>CEBPA</i>	1	<i>PHF6</i>	2-10
<i>CSF3R</i>	14-17	<i>PTEN</i>	5-7
<i>DNMT3A</i>	23	<i>RUNX1</i>	3-8
<i>ETV6</i>	1-8	<i>SETBP1</i>	4
<i>EZH2</i>	2-20	<i>SF3B1</i>	14,15
<i>FLT3</i>	14, 15, 20	<i>SRSF2</i>	1
<i>HRAS</i>	2,3	<i>TET2</i>	3-11
<i>IDH1</i>	4	<i>TP53</i>	4-9
<i>IDH2</i>	4	<i>U2AF1</i>	2, 6
<i>JAK2</i>	12,14	<i>WT1</i>	7, 9
<i>KIT</i>	2, 8-11, 13, 17	<i>ZRSR2</i>	1-11

Table 2.1. Genes and exons covered for targeted bulk sequencing of myeloid hotspots. Adapted from (1).

Read alignment and variant calling were performed in BaseSpace (Illumina) using BWA (Burrows-Wheeler Aligner) and GATK (Genome Analysis Toolkit) or SVC (Somatic Variant Caller), as per standard practice in the Oxford Molecular Diagnostics Laboratory. Variant annotation and filtering were performed using Variant Studio (Illumina). Both alignment and variant calling were performed by Dr. Angela Hamblin.

Coverage of each targeted amplicon was assessed using CoverMi software (<https://github.com/eawilson/CoverMi>); at least 100 reads per amplicon were required.

Variants were filtered by quality and consequence (missense, frameshift, stop gained, in-frame insertion, in-frame deletion, splice region). Variants identified as potential somatic mutations were further assessed for germline/somatic frequency and pathogenicity against public databases such as COSMIC, dbSNP or gnomAD as well as published literature. Those variants with >1% population frequency, or less than 1% but with ethnicity bias and variant allele frequency close to 50%, were considered polymorphisms (1). Variants retained after these steps with at least 1 % variant allele frequency were considered somatic mutations.

A subset of patients diagnosed with post-MPN secondary Acute Myeloid Leukaemia were analysed using an extended myeloid hotspot panel spanning 29 additional genes (*ABL1*, *ANKRD26*, *ASXL2*, *ATM*, *BCOR*, *BCORL1*, *BRAF*, *CHEK2*, *CUX1*, *DDX41*, *ETNK1*, *GATA2*, *IKZF1*, *KDM6A*, *MECOM*, *NF1*, *NFE2*, *NIPBL*, *PPM1D*, *PTPN11*, *RAD21*, *RIT1*, *SH2B3*, *SMC1A*, *SMC3*, *SRP72*, *STAG2*, *TERC* and *TERT*). *CALR* mutational status was separately assessed using a targeted-sequencing assay spanning exon 9 of the gene.

4. BULK KARYOTYPING

Karyotyping of bulk mononuclear cells was performed using a Single Nucleotide Polymorphism (SNP) Array (Infinium HTS assay on Human Omniexpress-24 Beadchips v1.3 , Illumina). This array contains 714,238 common markers with a mean spacing between probes of 4 kb, and allows to capture common variants, small insertions and deletions (indels) and Copy Number Variations (CNVs).

200 ng of genomic DNA from 96 samples (MPN-BP patient samples, normal donor controls and cell lines) were used for hybridization in four Human Omniexpress-24

arrays. Samples were processed by the Oxford Genomics Centre at the Wellcome Centre for Human Genetics. Copy Number Variants and Loss of Heterozygosity regions were analysed using cnvPartition Module within GenomeStudio Software (Illumina) using default parameters.

5. SAMPLE STAINING, FLUORESCENT ACTIVATED CELL SORTING (FACS) AND SINGLE-CELL ISOLATION

Peripheral blood and/or bone marrow HSPC staining from patient samples was performed using two different panels of fluorochrome conjugated antibodies (Table 2.2). Antibodies were added to each sample in a 1.5 mL microfuge tube and incubated for 20 minutes on ice. After that, samples were washed once with 1 mL of FACS buffer, resuspended in 500 μ L of FACS Buffer and 7-AAD or DAPI were added for dead cell exclusion.

Antibody-Fluorochrome	Dilution	Panel	Company	Identifier
CD8-FITC (Lineage)	1:100	A	BioLegend	Clone: RPA-T8; Cat#: 301006; RRID: AB_314124
CD20-FITC (Lineage)	1:150	A	BioLegend	Clone: 2H7; Cat#: 302304; RRID: AB_314252
CD66b-FITC (Lineage)	1:15	A	BioLegend	Clone: G10F5; Cat#: 305104; RRID: AB_314496
CD10-FITC (Lineage)	1:30	A	BioLegend	Clone: HI10a; Cat#: 312208; RRID: AB_314919
CD127-FITC (Lineage)	1:30	A	eBioscience	Clone eBioRDR5; Cat#: 11-1278-42; RRID: AB_1907343
Human Hematopoietic Lineage Cocktail – FITC (Lineage)	1:15	A	eBioscience	Cat# 22-7778-72; RRID: AB_1311229
CD123-PECy7	1:60	A and B	BioLegend	Clone: 6H6; Cat#: 306010; RRID: AB_493576

CD38-PETxRed	1:21	A and B	Invitrogen	Clone: HIT2; Cat#: MHCD3817; RRID: AB_10392545
CD90-BV421	1:10	A	BioLegend	Clone: 5E10; Cat#: 328122; RRID: AB_2561420
CD45RA-PE	1:150	A	eBioscience	Clone: HI100; Cat#: 12-0458-41; RRID: AB_10717397
CD34-APC-eF780	1:150	A	eBioscience	Clone: 4H11; Cat#: 47-0349-42; RRID: AB_2573956
CD34-PerCP/Cy5.5	1:100	B	BioLegend	Clone: 562; Cat# 343611, RRID:AB_2566787
CD90-PE	1:25	B	BioLegend	Clone: 5E10; Cat# 328109; RRID: AB_893442
CD45RA-FITC	1:200	B	Invitrogen	Clone: MEM56; Cat# MHCD45RA01; RRID: AB_10373858
CD2-PE/Cy5 (Lineage)	1:150	B	BioLegend	Clone: RPA-2.10; Cat# 300209; RRID:AB_314033
CD3-PE/Cy5 (Lineage)	1:150	B	BioLegend	Clone: HIT3a; Cat# 300310; RRID: AB_314046
CD4-PE/Cy5 (Lineage)	1:80	B	BioLegend	Clone: RPA-T4; Cat# 300510; RRID: AB_314078
CD7-PE/Cy5 (Lineage)	1:20	B	BioLegend	Clone: 6B7; Cat# 343110; RRID: AB_2075096
CD8-PE/Cy5 (Lineage)	1:150	B	BioLegend	Clone: RPA-T8; Cat# 301010; RRID: AB_314128
CD10-PE/Cy5 (Lineage)	1:40	B	BioLegend	Clone: HI10a; Cat# 312206; RRID: AB_314917
CD11b-PE/Cy5 (Lineage)	1:80	B	BioLegend	Clone: ICRF44; Cat# 301308; RRID: AB_314160
CD14-PE/Cy5 (Lineage)	1:80	B	Invitrogen	Clone: 61D3; Cat# 15-0149-41; RRID: AB_2573057

CD19-PE/Cy5 (Lineage)	1:150	B	BioLegend	Clone: HIB19; Cat# AB_314240; RRID: 302210
CD20-PE/Cy5 (Lineage)	1:200	B	BioLegend	Clone: 2H7; Cat# AB_314256; RRID: 302308
CD56-PE/Cy5 (Lineage)	1:40	B	BD Biosciences	Clone: B159; Cat# 555517; RRID: AB_395907
CD235a,b-PE/Cy5 (Lineage)	1:150	B	BioLegend	Clone: HIR2; Cat# 306606; RRID: AB_314624

Table 2.2. Antibodies used for flow cytometry and fluorescent-activated cell sorting.

All experiments included Fluorescence Minus One controls (FMOs), in which at least 300,000 cells were stained with all minus one antibody to control for spreading of other fluorochromes into the channel of interest. Gates for each fluorochrome were set based on FMOs, except in the case of CD38 antibody, in which the percentage of cells in CD38⁺ and CD38⁻ gates (90:10) was also taken into account as per standard practice (274). Single colour control compensation beads (CompBeads, BD Biosciences) were used to initially set laser voltages and compensations in each instrument for each fluorochrome and antibody. Purity of sorted cell populations was assessed by sorting 100 cells from each population in a 1.5 microfuge tube containing 100 μ L of FACS Buffer and re-analyzing their immunophenotypic profile. A purity higher than 95% was required for all experiments.

FACS was performed using BD Aria III, BD Fusion I, BD Aria II or BD Aria III instruments (Becton Dickinson) for bulk and 96-well plate experiments. A SH800S (SONY) instrument was used for 384-well plate experiments. For all experiments involving single-cell isolation in 96 or 384-well plates I chose a sorting purity mode in which the 50 % of the previous and next droplet to the target one is empty, and the target event is within the 75 % central part of the target droplet. This was essential to maximize cell recovery whilst making sure only one cell per well was sorted.

To prepare the sorter for single-cell sorting, I first performed an alignment test sort using a 96 or 384 well tissue culture plate (Corning® 96 or 384 well microplate, Catalog# CLS3925-100EA and CLS3702-100EA) and fluorescent beads. I sorted one fluorescent bead per well and checked under the microscope that only one bead was deposited in each well of the 96 or 384 well plate. I then performed the same test using a 96 or 384-well PCR plate (ThermoFisher Scientific, Catalog No. AB0900 or FrameStar, 4titude, Catalog No. 4ti-0384/C) covered with a PCR film and sorted 50 droplets in the corners of the plate. The sorter was calibrated until the drops fell exactly in the centre of each well. After that, the PCR cover was removed and 20 droplets were sorted in each well. When the drops were deposited on the bottom of each well and there was no liquid splashing on the sides of the wells, alignment check was considered complete. Immediately after, single cells were sorted in 96-well (for full-length TARGET-seq) or 384-well plates (for high-throughput TARGET-seq) containing lysis buffer.

For experiments involving isolation of human HSPCs, viable lineage-CD34⁺ cells were sorted (unless otherwise specified in the figure legends). Cells were indexed for CD38, CD90, CD45RA and CD123 markers, which allowed the mean fluorescence intensity of each of these markers for each single cell to be recorded. The gating strategy for FACS is outlined in Figure 2.1. Analysis of flow cytometry data was performed using Flowjo Software version 10.0.8 (Flowjo LCC, Oregon, USA).

Most FACS experiments were performed with the assistance of Dr. Sally-Ann Clark at the Flow Cytometry Facility in the Weatherall Institute of Molecular Medicine.

unique oligodT-barcoded primer, was aliquoted in each well of a 384 well-plate. After lysis buffer aliquoting, plates were span down briefly and kept on ice until they were used for single-cell sorting. Immediately after sorting, plates were briefly vortexed, span down and stored at -80°C for up to three months.

Lysis	SMART-seq+/ SMART-seq2/ mRNA targeting	Full-length TARGET-seq (96-well)	3'- TARGET- seq (384-well)	Company, Catalog Number
Triton X-100 0.4 %	1.9 µL	1.9 µL	0.95 µL	Sigma-Aldrich, Cat# T8787
RNase Inhibitor (40 U/µL)	0.1 µL	0.1 µL	0.05 µL	TAKARA Clontech, Cat# 2313A
dNTPs (10 mM)	1 µL	1 µL	0.5 µL	Life Technologies, Cat#R0192
OligodT ₃₀ (10 µM)	1 µL	1 µL	0.5 µL (barcoded OligodT)	Biomers, custom HPLC oligo
Protease (1.07 AU/mL)	-	0.1 µL	0.05 µL	Qiagen, Cat#19155
ERCC spike-in (*)	0.1 µL	0.1 µL	0.02 µL	Invitrogen, Cat# 4456740
TOTAL	4.1 µL	4.2 µL	2.07 µL	

Table 2.3. Reagents used for lysis buffer preparation for SMART-seq2, SMART-seq plus and TARGET-seq methods. (*) 1:400,000 ERCC dilution for 3'TARGET-seq; 1:2,000,000 ERCC dilution for SMART-seq and full length TARGET-seq.

After that, plates were transferred to a PCR-clean area to perform heat inactivation. During heat inactivation, mRNA is denatured and for TARGET-seq method, the protease contained within the lysis buffer is heat-inactivated. Heat inactivation was performed at 72°C for 3 minutes for SMART-seq2 and SMART-seq plus, and at 72°C for 15 minutes for TARGET-seq. After that, retrotranscription mix (Table 2.4) was immediately added to the plates, and plates were incubated on a thermocycler using the program outlined in Table 2.5.

RT	SMART-seq+	mRNA targeting	TARGET-seq (96-well)	3'-TARGET-seq (384-well)	Company, Catalog Number
5X Buffer	2 μ L	2 μ L	2 μ L	1 μ L	Clontech, 639537
SMARTScribe (100 U/ μ L)	0.5 μ L	0.5 μ L	1 μ L	0.5 μ L	
RNase Inhibitor (40 U/ μ L)	0.25 μ L	0.25 μ L	0.25 μ L	0.125 μ L	TAKARA Clontech, 2313A
*TSO-LNA (100 μ M)	0.1 μ L	0.1 μ L	0.1 μ L	0.05 μ L	Custom oligonucleotide, Qiagen
mRNA primers (200 μ M stock)	0	Variable (0.035 μ L per primer)	Variable (0.035 μ L per primer)	Variable (0.0175 μ L per primer)	Custom oligonucleotide, biomers.net
RT-PCR Grade Water	2.25 μ L	Variable	Variable	Variable	Life Technologies, AM9935
TOTAL	5.6 μ L	5.6 μ L	5.6 μ L	2.8 μ L	NA
TOTAL volume (cumulative)	9.7 μ L	9.8 μ L	9.8 μ L	4.87 μ L	NA

Table 2.4. Reagents used for retrotranscription SMART-seq plus, mRNA targeting and TARGET-seq libraries. SMART-seq2 was performed as previously described (211). For the optimization experiments described in Chapter 3 “optimization of single-cell RNA-sequencing conditions”, SMART-seq2 protocol was modified by using the different enzymes described and keeping the rest of reaction parameters the same.

*TSO-LNA: AAGCAGTGGTATCAACGCAGAGTACATrGrG+G, HPLC purified.

RT PROGRAM		
Temperature	Time (min:sec)	Cycles
42 °C	90:00	1
50 °C	02:00	10
42 °C	02:00	
72 °C	15:00	1
4 °C	HOLD	1

Table 2.5. Thermocycler conditions used for retrotranscription of SMART-seq2, SMART-seq plus, mRNA targeting and TARGET-seq libraries.

Subsequently, PCR mix was added to the plates (Table 2.6) and PCR was performed as outlined in Table 2.7, using 20 cycles of amplification for cell lines, 22 cycles of amplification for HSPC libraries synthesized using full-length TARGET-seq and 24 cycles of amplification for HSPC libraries synthesized using 3'-TARGET-seq.

PCR	SMART-seq+	mRNA targeting	TARGET-seq	3'-TARGET-seq	Company, Catalog Number
2X Buffer	12.5 μ L	12.5 μ L	12.5 μ L	6.25 μ L	Clontech, 638509
SeqAMP (100 U/ μ L)	0.5 μ L	0.5 μ L	0.5 μ L	0.25 μ L	
*ISPCR primer (10 μ M)	0.125 μ L	0.125 μ L	0.125 μ L	0.0625 μ L	Custom oligonucleotide, biomers.net
gDNA primers (200 μ M stock)	0	Variable (0.1 μ L per primer)	Variable (0.1 μ L per primer)	Variable (0.05 μ L per primer)	
cDNA primers (20 μ M stock)	0	Variable (0.035 μ L per primer)	Variable (0.035 μ L per primer)	Variable (0.0175 μ L per primer)	
RT-PCR Grade Water	1.875 μ L	Variable	Variable	Variable	Life Technologies, AM9935
TOTAL	15 μ L	15 μ L	15 μ L	7.5 μ L	NA
TOTAL volume (cumulative)	24.7 μ L	24.8 μ L	24.8 μ L	12.37 μ L	NA

Table 2.6. Reagents used for PCR amplification of SMART-seq2, SMART-seq plus, mRNA targeting and TARGET-seq libraries. *ISPCR primer: AAGCAGTGGTATCAACGCAGAGT.

PCR PROGRAM		
Temperature	Time (min:sec)	Cycles
98 °C	03:00	1
98 °C	00:20	20 (cell lines); 22 (full-length HSPC); 24 (3'-TARGET-seq HSPC)
67 °C	00:15	
72 °C	06:00	
72 °C	05:00	1
4 °C	HOLD	1

Table 2.7. Conditions used for PCR amplification of SMART-seq2, SMART-seq plus, mRNA targeting and TARGET-seq libraries.

After PCR, SMART-seq2, SMART-seq plus, mRNA targeting and full-length TARGET-seq libraries were individually purified with Ampure XP beads, following manufacturer's instructions and using a 0.6:1 beads to cDNA ratio (16 µL of beads and 25 µL of cDNA library). Libraries were resuspended in 8 µL of EB (Qiagen, Catalog# 19086). 3'-TARGET-seq cDNA libraries were pooled using a Biomek FxP Automated Liquid Handler platform and each pool was purified twice with Ampure XP, following manufacturer's instructions and using a 0.6:1 beads to cDNA ratio (66 µL of beads and 110 µL of cDNA library). Libraries were resuspended in 30 µL of EB (Qiagen, Catalog# 19086). cDNA quality was assessed on a microcapillary array (Agilent Bioanalyzer; High Sensitivity DNA Kit, Agilent, Cat# 5067-4626) and libraries were quantified using Qubit dsDNA HS Assay Kit (Invitrogen, Cat# Q32854). Target-specific primers used for retrotranscription and PCR steps can be found in Appendix 2.

7. TARGETED NGS SINGLE-CELL GENOTYPING

To prepare libraries for single cell genotyping, cDNA+amplicon mix obtained from the PCR step described above was first diluted 1:1 with water in a Biomek FxP Liquid Handling Platform. Then, two PCR steps were performed to generate single-cell Illumina compatible libraries ready for sequencing (Figure 2.2).

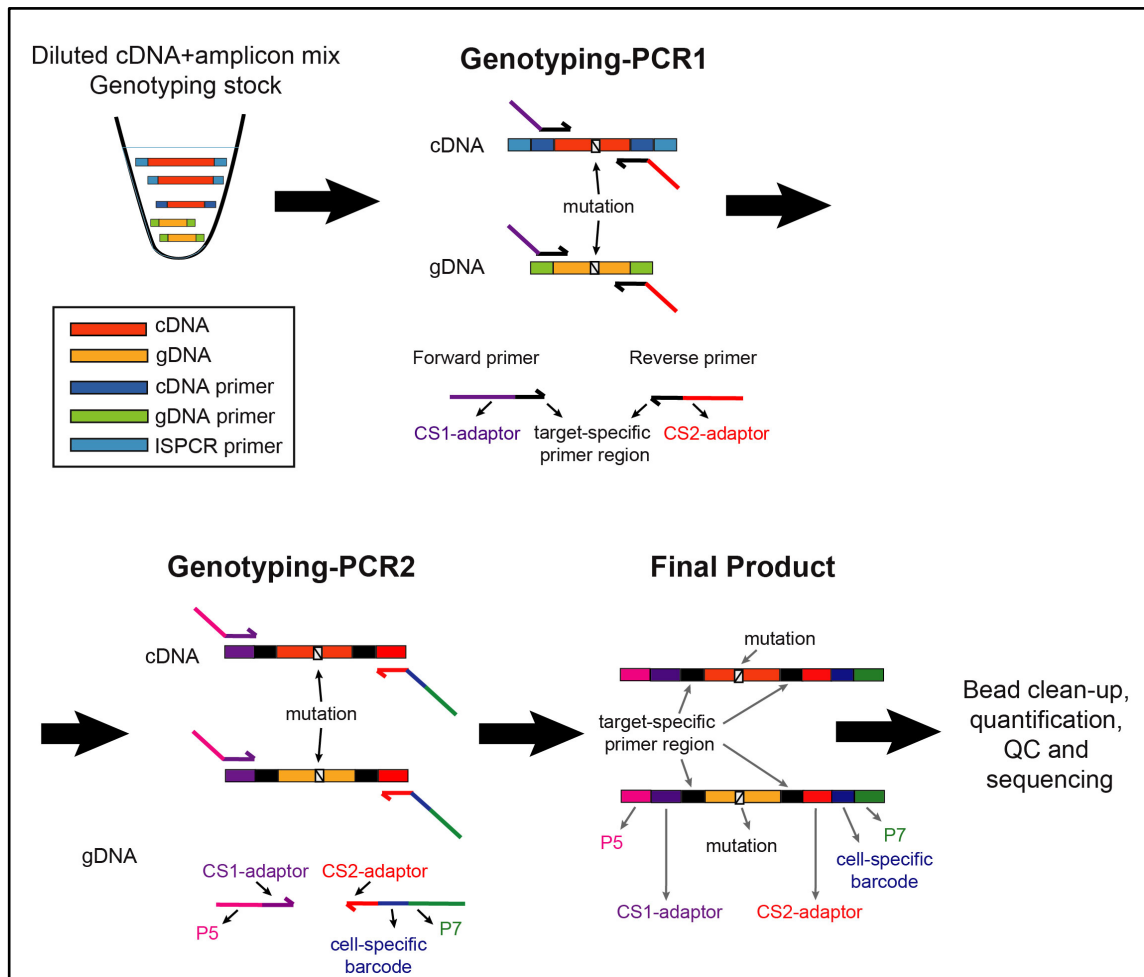


Figure 2.2. Schematic representation of targeted library preparation for single-cell genotyping.

During the first PCR step (genotyping-PCR1), target-specific primers containing CS1 (forward primer) or CS2 (reverse primer) adaptor sequences were used to amplify specific gDNA and cDNA targets, which had been pre-amplified in the previous RT and PCR step (see single-cell cDNA synthesis). Genotyping-PCR1 was performed in 384 well plates using a Biomek FxP Liquid Handling Platform using the reagents specified in Table 2.8. Up to four different PCR1 reactions were performed per single cell, each reaction containing different target-specific primers. Targeted-libraries were split due to the dissimilar amplification efficiencies of some amplicons, which would result in disparate amplification of target regions. Target-specific primers used for genotyping-PCR1 can be found in Appendix 3.

Genotyping-PCR1 Master Mix	1 cell	Catalog Number	Company
KAPA2G Robust HS Ready Mix	3.125 μ L	KK5702	Sigma (Roche)
Primer pair 1 (F+R; 20 μM)	0.375 μ L	Custom desalted oligos; resuspend in RNase Free TE	Invitrogen
Primer pair 2 (F+R; 20 μM)	0.375 μ L		
Primer pair 3 (F+R; 20 μM)	0.375 μ L		
Water (DNase/RNase free)	0.5 μ L	10977035	Life Technologies
TOTAL	4.75 μ L		

Table 2.8. Reagents used for genotyping-PCR1 amplification.

Once aliquoted, plates were incubated in a thermocycler using the program specific in Table 2.9.

Temperature	Time (min:sec)	Cycles
95 °C	3 min	1 cycle
95 °C	00:15	20 cycles
60 °C	00:20	
72 °C	1 min	
72 °C	5 min	1 cycle
4 °C	HOLD	1 cycle

Table 2.9. Conditions for target-specific amplification during genotyping-PCR1.

During the subsequent PCR step (genotyping-PCR2), Illumina compatible adaptors containing a cell-specific barcode are attached to the genotyping-PCR1 products through the CS1 and CS2 adaptor regions (Figure 2.2). Genotyping-PCR2 was performed in 384 well plates using a Biomek FxP Liquid Handling Platform using the reagents specified in Table 2.10.

Genotyping-PCR2 Master Mix	1 reaction	Catalog Number	Company
Buffer 10 X	1 μ L	FastStart High Fidelity PCR System (Cat# 04738292001)	Roche
MgCl ₂	1.8 μ L		
DMSO	0.5 μ L		
Nucleotide Mix	0.2 μ L		
Enzyme	0.1 μ L		
RNAse/DNAse free Water	1.9 μ L	10977035	Life Technologies
TOTAL	5.5 μL	NA	NA
cDNA (aliquoted from genotyping-PCR1 plate)	2.5 μ L	NA	NA
Access Array™ Barcode Library for Illumina® Sequencers-384	2 μ L	100-4876	Fluidigm
TOTAL (cumulative)	10 μL	NA	NA

Table 2.10. Reagents used for genotyping-PCR2 amplification.

Once aliquoted, plates were incubated in a thermocycler with the program specified in Table 2.11.

Genotyping-PCR2 PROGRAM		
Temperature	Time (min:sec)	Cycles
98 °C	10:00	1
98 °C	00:20	10
60 °C	00:15	
72 °C	01:00	
72 °C	03:00	1
4 °C	HOLD	1

Table 2.11. Conditions used for genotyping-PCR2 amplification.

After that, Illumina-compatible barcoded libraries were pooled using a Mosquito TTP Labtech Liquid Handling Platform and each pool was purified using Ampure XP beads, following manufacturer's instructions and using 0.8:1 beads to cDNA ratio (40 μ L of beads and 50 μ L of cDNA library). Libraries were resuspended in 30 μ L of EB (Qiagen,

Catalog# 19086), specific target amplification was checked using a TapeStation instrument (High Sensitivity D1000 ScreenTape and Reagents, Agilent, Cat# 5067-5583 and Cat# 5067-5584). Libraries were quantified using Qubit dsDNA HS Assay Kit (Invitrogen, Cat# Q32854), pooled in equimolar concentrations and sequenced on a MiSeq instrument (Illumina) with 151 cycles for read 1, 10 cycles for index read and 151 cycles for read 2. Custom primers binding to CS1 and CS2 tags were used for sequencing (CS1-seq: A+CA+CTG+ACGACATGGTTCTACA; CS2-seq: T+AC+GGT+AGCAGAGACTTGGTCT; CS1rc-seq: T+GT+AG+AACCATGTCGTCAGTGT; CS2rc-seq: A+GAC+CA+AGTCTCTGCTACCGTA; plus sign denotes LNA-modified bases). 500 nM of CS1-seq and CS2-seq in a total volume of 700 µL of Hybridization Buffer 1 (HT1) were used for R1 and R2; 500 nM of CS1rc-seq and CS2rc-seq primers in a total volume of 700 µL of HT1 were used for Index Read when using the MiSeq platform, Illumina.

For high-throughput single-cell genotyping (described in Chapter 3), genotyping PCR1 reaction was performed using primers containing a 6 bp barcode sequence between the CS1/CS2 tag region and the target-specific primer sequence. After genotyping-PCR1, 1 µL from each genotyping-PCR1 reaction was pooled using a Mosquito TTP Labtech liquid handling platform. Pooled genotyping-PCR1 libraries were purified with Ampure XP beads using different beads to cDNA ratios (0.6:1, 0.8:1) and resuspended in 50 µL of EB buffer, and then pooled libraries were used directly for genotyping-PCR2. After PCR2, libraries were pooled, quantified and sequenced as described in the paragraph above.

8. NEXTERA XT LIBRARY PREPARATION AND WHOLE TRANSCRIPTOME SEQUENCING

Library preparation for whole transcriptome sequencing for full-length TARGET-seq was performed with the commercially available Nextera XT Kit (Illumina; Catalog FC-131-1096, FC-131-2001 and FC-131-2003), following manufacturer's instructions but

using one-fourth of the recommended volume (total reaction volume 12.5 μ L). After PCR, libraries were diluted 1:1 with RNase/DNase free water (up to a total volume of 25 μ L) purified with Ampure XP beads (Beckman Coulter; Catalog#A63881) using 0.6:1 beads to cDNA ratio (16 μ L of beads and 25 μ L of tagmented library) and resuspended in 20 μ L of EB Buffer (Qiagen, Catalog#19086).

Library preparation for whole transcriptome sequencing for 3'-biased high-throughput TARGET-seq was performed with the commercially-available Nextera XT kit with modifications on the PCR-amplification step. DNA tagmentation was performed following manufacturer's instructions: 1 ng of pooled cDNA was mixed with 10 μ L of Tagmentation Buffer (Illumina) and 5 μ L of Amplicon Tagment Mix (containing Tn5 enzyme), and incubated at 55 °C for 6 minutes. Then 5 μ L of Neutralization Buffer (Illumina) were added to stop the tagmentation reaction. In the PCR step, a custom P5 PCR primer specifically binding to the ISPCR and 3'-PCR handle regions of the TARGET-seq libraries was used for the amplification (P5_index: AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCA GAGT*T*G) as indicated in Table 2.12. Conditions used for the amplification of tagmented libraries are outlined in Table 2.13.

Reagents	1 reaction	Company, Catalog Number
P7 primer + i7 index (2 μ M)	5 μ L	Nextera XT index Kit, Illumina, FC-131-2001/2003
P5_index (10 μ M)	1 μ L	Custom HPLC purified primer, biomers. net
NPM (PCR master mix)	15 μ L	Nextera XT Kit; Illumina, FC-131-1096
Water	4 μ L	Life Technologies, 10977035
TOTAL	25 μ L	NA
TOTAL (cumulative)	50 μ L	NA

Table 2.12. Reagents used for PCR amplification of tagmented 3'-TARGET-seq libraries.

Temperature	Time	Cycles
72 C	3 minutes	1
95 C	30 seconds	1
95 C	10 seconds	14 cycles
55 C	30 seconds	
72 C	30 seconds	
72 C	5 minutes	1
4 C	HOLD	1

Table 2.13. Conditions used for PCR amplification of tagged 3'-TARGET-seq libraries.

This strategy specifically enriched for tagged fragments containing the 3'-end of the cDNA molecules. A P7 primer containing an 8 bp index sequence was also attached to each library (Nextera XT Index Kit; Illumina, Catalog FC-131-2001). After PCR, libraries were purified twice using Ampure XP beads (Beckman Coulter; Catalog#A63881) using 0.7:1 beads to cDNA ratio (34 μ L of beads and 50 μ L of library). Libraries were resuspended in 20 μ L of EB Buffer (Qiagen, Catalog#19086).

Size distributions of full-length and 3'-TARGET-seq tagged and amplified whole transcriptome libraries were checked on a TapeStation instrument (High Sensitivity D5000 ScreenTape and Reagents, Agilent, Cat# 5067-5592 and Cat# 5067-5593). Libraries were quantified using Qubit (ThermoFisher, Catalog#32854) and pooled at equimolar concentrations.

Full-length whole transcriptome libraries were sequenced on a NextSeq, HiSeq or NovaSeq instrument with the following read configuration: for libraries sequenced on the NextSeq, 75 cycles were used for read1 and 8 cycles for index read; for libraries sequenced on the HiSeq or NovaSeq, 151 cycles were used for read1, 151 cycles for read 2 and 8 cycles for index read.

3'-TARGET-seq whole transcriptome libraries were sequenced on a NextSeq, using a custom primer for read1 (P5_seq: GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGA

GTTGC*T; PAGE purified, * denotes a phosphorothioate bond) which allows the cell barcode contained within the oligodT-ISPCR primer for each cell to be read. Libraries were loaded at a final concentration of 3 pM diluted in 1.3 mL of HT1 Buffer (Illumina); 900 nM of P5_seq primer in a total volume of 3 mL of HT1 buffer was added to the cartridge. Libraries were sequenced with the following read configuration: 20 bp read1; 8 bp index read; 64 cycles R2. Read 1 contains the cell-specific barcode added during retrotranscription and identifies each single cell.

9. ASSAY FOR TRANSPOSASE-ACCESSIBLE CHROMATIN AND SEQUENCING (ATAC-SEQ)

ATAC-seq was performed using a protocol optimized for low cell number experiments, based on the Omni-ATAC-seq protocol (275). 500 cells from each population (HSC, Lineage-CD34+CD38+ cells) were directly sorted in 0.5 mL PCR strips containing 23.5 µL of lysis buffer as per the table below (Table 2.14):

Reagent	Amount per sample	Company	Catalog Number
2x tagmentation Buffer	12.5 µL	Nextera DNA Library Prep Kit; Illumina	FC-121-1030
1% digitonin	0.25 µL	Promega	G9441
10% Tween-20	0.25 µL	Sigma Aldrich	P9416
PBS	8.25 µL	Life Technologies	10010056
H2O	2.5 µL	Life Technologies	10977049
Total	23.75 µL		

Table 2.14. Reagents used for lysis buffer preparation for ATAC-seq libraries.

Cells were kept chilled during cell sorting. 1.25 µL of transposase enzyme (Nextera DNA Library Prep Kit; FC-121-1030; Illumina) were added to each tube and samples were incubated at 37°C for 30 minutes with agitation (300 rpm). Immediately after transposition, DNA was purified using Qiagen MinElute Kit (Qiagen; Catalog#28004)

following manufacturer's instructions. DNA was eluted in 11 µL of Elution Buffer (Qiagen, Catalog# 19086), incubated at room temperature for 5 minutes, spun down at 13000 rpm and subsequently used for PCR.

PCR amplification was performed with Illumina-compatible primers binding to Nextera adaptor regions added to each DNA molecule during transposition (Nextera P1: AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG; P2: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTCTCGTGGGCTCGGAGATGT). Nextera primer 2 contains a variable 8 bp index (labelled with N) sequence to uniquely barcode each sample; the index sequences used in this study can be found in Table 2.15.

Nextera P2 primer	Primer sequence
P2-index1	CAAGCAGAAGACGGCATACGAGATTGCGCTTAGTCTCGTGGGCTCGGAGATGT
P2-index2	CAAGCAGAAGACGGCATACGAGATCTAGTACGTCTCGTGGGCTCGGAGATGT
P2-index3	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
P2-index4	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
P2-index6	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
P2-index7	CAAGCAGAAGACGGCATACGAGATGTAGAGAGTCTCGTGGGCTCGGAGATGT
P2-index8	CAAGCAGAAGACGGCATACGAGATCCTCTCTGTCTCGTGGGCTCGGAGATGT
P2-index9	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
P2-index10	CAAGCAGAAGACGGCATACGAGATCAGCCTCGTCTCGTGGGCTCGGAGATGT
P2-index11	CAAGCAGAAGACGGCATACGAGATTGCTCTTGTCTCGTGGGCTCGGAGATGT
P2-index12	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT
P2-index13	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
P2-index14	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
P2-index15	CAAGCAGAAGACGGCATACGAGATCAGATCCAAGTCTCGTGGGCTCGGAGATGT
P2-index16	CAAGCAGAAGACGGCATACGAGATACAAACGGTCTCGTGGGCTCGGAGATGT
P2-index17	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT

P2-index18	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
P2-index19	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
P2-index20	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
P2-index21	CAAGCAGAAGACGGCATACGAGATGAAACCCA GTCTCGTGGGCTCGGAGATGT
P2-index22	CAAGCAGAAGACGGCATACGAGATTGTGACCA GTCTCGTGGGCTCGGAGATGT
P2-index23	CAAGCAGAAGACGGCATACGAGATAGGGTCAA GTCTCGTGGGCTCGGAGATGT
P2-index24	CAAGCAGAAGACGGCATACGAGATAGGAGTGG GTCTCGTGGGCTCGGAGATGT

Table 2.15. Primer sequences used to amplify ATAC-seq libraries during PCR. Sequences labelled in blue correspond to Illumina's P7 PCR primer and sequences labelled in green, partial sequences binding to the tags added by the transposase enzyme. Sequences labelled in black are unique sample indexes (8 bp).

PCR reaction was performed as per the table below (Table 2.16) and incubated in a thermal cycler with the following conditions: 72°C, 5 min, 98°C, 30 sec; 14 cycles of 98°C, 10 sec, 63°C, 30 sec, 72°C, 1 min; hold at 4°C.

Reagent	Amount per sample	Company	Catalog Number
Purified transposed DNA	10 µL	NA	NA
Nextera P1 (25 µM)	2.5 µL	IDT	Custom primer
Nextera P2 (25 µM)	2.5 µL	IDT	Custom primer
NEBNext High-Fidelity 2x PCR master mix	25 µL	NEB	M0541S
H2O	10 µL	Life Technologies	10977035
Total	50 µL		

Table 2.16. Reagents used for PCR amplification of transposed DNA to prepare ATAC-seq libraries.

Each PCR reaction was purified with 90 µL of Ampure XP beads (Beckman Coulter; Catalog# A63881; 1.8:1 beads to cDNA ratio) following manufacturer's instructions. Library size distributions were checked on a TapeStation instrument (High Sensitivity D1000 ScreenTape and Reagents, Agilent, Cat# 5067-5583 and Cat# 5067-5584). Libraries showing the expected nucleosome periodicity (Figure 2.3) were considered to pass quality control and were subsequently prepared for sequencing.

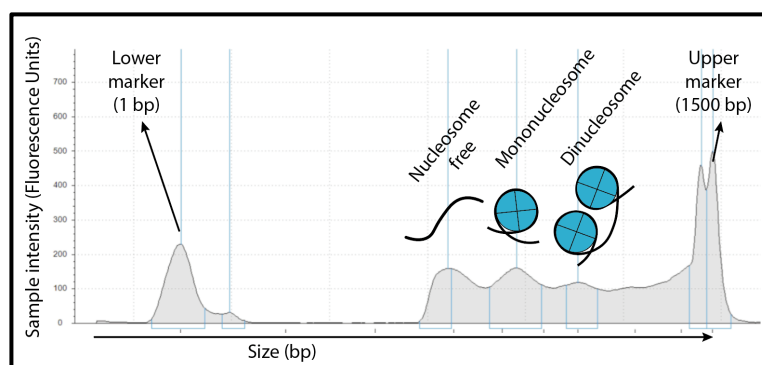


Figure 2.3. Representative traces of tagged and amplified ATAC-seq libraries. Distinctive peaks corresponding to nucleosome free, mononucleosome and dinucleosome regions are labelled.

Libraries were quantified using NEBNext Library Quant Kit for Illumina (NEB, Catalog# E7630S), pooled at equimolar concentrations and sequenced in a NextSeq instrument (Illumina), using the following sequencing configuration: 40 cycles read 1, 8 cycles index read and 40 cycles read 2.

10. FULL-LENGTH SINGLE CELL RNA-SEQUENCING DATA PRE-PROCESSING

RNA-sequencing reads were first trimmed for Nextera adaptor sequences using TrimGalore (version 0.4.1) and were then aligned to the human genome (hg19) using STAR with default settings (version 2.4.2a) (276). Counts mapping to each gene were obtained with featureCounts (version 1.4.5-p1; options `-primary`) (277) and were normalized to reads per kilobase per million mapped reads (RPKM) for subsequent analysis. Those genes with RPKM values lower than 1 were considered as not detected, and their gene expression values were transformed to zero. Quality control (QC) filtering was performed with the following parameters: reads mapping in exons should be >50 % of the mapping reads; number of genes detected per cell (with $\text{RPKM} \geq 1$) should be >6000 for JURKAT cells, >5000 for K562 and SET2 cells and >1500 for HSPCs, and the library size (number of reads mapping in genes) should be higher than 100,000. The gene expression matrix obtained after QC filtering was used for all subsequent analysis.

For the analysis of *U2AF1*-mutant MPN patient samples described in Chapter 4, section 2.2, RNA-sequencing reads were first trimmed for Nextera adaptor sequences using TrimGalore (version 0.4.1). For each cell, high-speed splice-aware aligner HISAT2 (278) was used to align raw RNA-sequencing reads to the human reference genome (GRCh38). Resulting sequence alignment map (SAM) files were subsequently compressed, sorted, and indexed using Samtools. A gene transfer format (GTF) file was generated using the transcript assembly StringTie (279). This in-house generated GTF file and the publicly available GENCODE GTF file contain a catalogue of exons were used as reference files to quantify alternatively spliced exons for downstream analysis. Gene expression counts were computed using Salmon (280).

11. 3'-BIASED SINGLE CELL RNA-SEQUENCING DATA PRE-PROCESSING

Libraries were demultiplexed to fastq files with bcl2fastq (version 2.20, Illumina), matching the first 12 or 14 bp of the read1 to the barcode sequences introduced during retrotranscription (which identify each single cell, Appendix 1) and the 8 bp from the index read to a known Nextera XT i7-index sequence (Illumina) introduced during whole transcriptome library preparation (read configuration Y12N*,I8,Y64N* or Y14N*,I8,Y64N*). RNA-sequencing reads were then trimmed for polyA sequences using TrimGalore (version 0.4.1) and were aligned to the human genome (hg19) using STAR with default settings (version 2.4.2a). Counts mapping to each gene were obtained with featureCounts (version 1.4.5-p1; options –primary) and were normalized by dividing the counts of each gene in each cell by the library size of that cell, multiplied by the mean library size of all cells processed. Those genes with normalized count values lower than 1 were considered as not detected, and their gene expression values were transformed to zero. I further normalized counts into the log2 scale. Quality control (QC) filtering was performed with the following parameters: library size >2000 reads; number of genes detected per cell (with normalized counts ≥ 1) > 500;

percentage of reads mapping to the mitochondrial chromosome < 10 % and percentage of reads mapping to ERCC spike-in molecules < 50%. The gene expression matrix obtained after QC filtering was used for all subsequent analysis.

12. MIXED HUMAN:MOUSE SINGLE CELL RNA-SEQUENCING DATA PRE-PROCESSING

For experiments involving human and mouse cell mixing using 3'-biased TARGET-seq method (Figure 3.9), RNA-sequencing reads were aligned using STAR (version 2.4.2.a) to an ensemble human and mouse genome with default parameters and `outFilterMultimapNmax=1`, to exclude reads mapping to more than one region in the ensemble genome. Counts mapping to each gene were obtained with `featureCounts` (version 1.4.5-p1; options `-primary`), using an ensemble GTF file containing both human and mouse genes. For each cell, the proportion of reads mapping to mouse or human genomes were identified by the Ensemble code of each gene ("ENSMUSG" identified genes from the mouse genome). Cells in which the number of counts mapping to the mouse genome were higher than 5000 were labelled as "mouse cells" and cells in which the number of counts in the human genome was higher than 5000 were labelled as "human cells".

13. WHOLE TRANSCRIPTOME VARIANT CALLING FROM SINGLE CELLS

Whole transcriptome variant calling from single cells was performed using LoFreq software (281) to generate the results presented in Figures 3.1 and 3.2. First, bam files from 48 K562 cells or 36 HSPCs were merged to create single-cell ensembles. Then LoFreq software was used to identify heterozygous SNPs in each ensemble (defined as variant allele frequency >30 % and <70%) requiring a minimum coverage of 10 reads. These heterozygous regions were used to perform variant calling in each single cell, requiring a minimum of 10 reads and a minimum base quality of 30. A particular variant was considered heterozygous in a particular single cell if the allele frequency

was >0.05 and <0.95 . Variants with allele frequencies <0.05 or >0.95 were considered homozygous.

14. DROPOUT FREQUENCY AND LIBRARY BIAS CALCULATION

Dropout frequency for each single cell was calculated as the percentage of K562 cells from a particular condition (SMART-seq2 and SMART-seq plus) in which a gene is not detected ($RPKM < 1$), as compared to bulk sequencing of the same K562 cells (100 cells per condition, $n=3$ independent technical replicates per chemistry). For global dropout frequency calculations, all of the genes detected by bulk sequencing were used for the analysis ($n=9096$ genes expressed at $RPKM > 1$); for lowly expressed genes dropout frequency calculation, only genes with $RPKM < 2$ and $RPKM > 1$ were used for the analysis ($n=1347$ genes).

Library bias was calculated as the mean RPKM of the top 10% highly expressed transcripts divided by the mean RPKM values of all genes. This metric measures the bias of a particular method towards amplifying highly expressed genes.

15. TRANSCRIPT COVERAGE

Transcript coverage for each single cell library was computed using “geneBody_coverage.py” script from the RSeQC package, using a list of 4040 housekeeping genes downloaded from <http://rseqc.sourceforge.net/>. This script divides each transcript into 100 bins, extracts the coverage across those bins for each transcript and normalizes the coverage across all transcripts to produce a normalized transcript score per library.

16. COPY NUMBER VARIATION INFERENCE FROM SINGLE CELLS

Copy number variation from single cells was performed using inferCNV package in R (<https://github.com/broadinstitute/inferCNV/>) using “cut-off-1, denoise=TRUE,

HMM=TRUE". Cells from normal donors were used to normalize expression intensities across each chromosome.

17. CELL-TO-CELL CORRELATION ANALYSIS

Spearman's correlation between RPKM values of highly variable genes for each single cell was performed using "cor" function in R (method='spearman', use='pairwise.complete.obs'). Pearson's correlations between gene expression values for each genetic subgroup of cells was performed using the log2(normalized counts) gene matrix with "rcorr" function in R (type='pearson').

18. COMPUTATIONAL RECONSTRUCTION OF CLONAL HIERARCHIES

The phylogenetic tree of each tumour was computed from the targeted single-cell genotyping results obtained from TARGET-seq using SCITE (226), with default parameters and "-r 1 -l 900000 -fd 0.001 -ad 0.01 0.01 -cc 0". Each *JAK2* allele was introduced independently in the mutation matrix to account for loss of heterozygosity in this region.

19. DIFFERENTIAL EXPRESSION ANALYSIS

Differential gene expression analysis was performed using a combination of Fisher's exact test and non-parametric Wilcoxon test. Fisher's test was used to compare the frequency of expression of each gene and Wilcoxon test, to compare the mean expression of each gene in each group. Results from each test were combined using Fisher's method and Benjamini & Hochberg procedure was used to derive adjusted p-values. I used log2(RPKM) or log2(normalized counts) gene matrices to perform differential expression analysis, using genes expressed in at least two cells when analysing less than 200 cells and expressed in at least 5 cells when analysing more than 200 cells. Significant genes were selected on the basis of adjusted p-value<0.1 and absolute log2(fold-change)>0.5.

Differentially expressed genes between multiple genetic subclones from the same patient were identified using analysis of variance (ANOVA) implemented within the “genefilter” package in R. Significant genes were selected on the basis of $p\text{-value} < 0.05$.

20. DIFFERENTIAL SPLICING ANALYSIS IN SINGLE CELLS

I used a Bayesian approach (BRIE) (282) to quantify alternatively spliced exons in single cells, where both sequence features and coverage are used as informative priors to predict percent spliced-in (PSI) for alternatively spliced exons. The five categories of alternative splicing events analysed are exon-skipping (SE), mutually exclusive exons (MXE), intron retention (IR), and alternative 5'(A5SS) and 3' splice sites (A3SS).

21. GENE-SET ENRICHMENT ANALYSIS

Gene-set enrichment analysis (GSEA) was performed using GSEA software downloaded from <http://software.broadinstitute.org/gsea> with default parameters and 1000 permutations on the phenotype (i.e., each single cell sample). The genesets used for the analysis were downloaded from the Molecular Signatures Database (MSigDB) or relevant literature (17, 247, 283-285). Single sample GSEA was performed with ssGSEA Projection Module, freely available at <https://genepattern.broadinstitute.org>, with default parameters and combine mode ‘combine.off’.

22. IDENTIFICATION OF HIGHLY VARIABLE GENES

Highly variable genes were identified by fitting a lowess model to the $\log_2(\text{mean expression})$ values against the coefficient of variation for each gene. Genes with a coefficient of variation above the model and $\log_2(\text{mean expression value}) \geq 0$ were considered above technical noise and were selected for downstream analysis.

23. BATCH CORRECTION

Batch correction was performed using “removeBatchEffect” function within the limma package in R (286) unless specified in the figure legends. This package uses an algorithm that fits a linear model to the data and then removes the components corresponding to the batch effect (or other cofounder variables), whilst preserving those specified in an experimental design matrix. The parameters used for counts matrix correction are specified in each figure legend for each analysis. For the analysis presented in Figure 4.14, batch effect was removed using “glm” function of the PAGODA2 package (downloaded from <https://github.com/hms-dbmi/pagoda2>).

24. CELL CYCLE SCORE CALCULATION AND CORRECTION

Cell cycle is a major confounder of scRNA-seq analysis (287). To remove the effect of cell cycle on the gene expression matrix, I first computed a cell cycle score for each cell by averaging the expression values of S-phase and G2M-phase cell cycle associated genes (obtained from (283)). Where cell cycle effect has been removed, this is specified on the figure legends.

25. SINGLE CELL CLUSTERING AND DIMENSIONALITY REDUCTION

T-distributed stochastic neighbourhood (tSNE) was used to create a low-dimensional representation of the gene expression data, using highly variable genes identified as described above or differentially expressed genes between different genetic subclones. tSNE was performed using “Rtsne” package in R. Perplexity was adjusted for each analysis based on the number of cells considered and genes included. Perplexity parameters and number of genes used for each analysis are specified in each figure legend.

K-nearest neighbourhood (knn) analysis was performed using “nn2” function from the RANN package in R, using the embedded matrix derived from the first 30 PCA

components and 15 nearest neighbours. Then, a weighted graph was calculated from the normalized shared number of nearest neighbours and the force-directed layout algorithm from the igraph package was used to generate three dimensional coordinates from the KNN weighted graph. Then, k-means clustering (k=8) was performed on the three-dimensional matrix of the graph layout. The identified clusters were superimposed on the graph layout presented in Figure 4.10A.

ZIFA (288) (Zero Inflated Factor Analysis) was used to analyse the transcriptional heterogeneity of patients with three or more genetic subclones (IF0137, IF0138 and IF0101) using highly variable genes identified in each patient, with default parameters. SC3 (289) (Single Cell Consensus Clustering) was also used to assess the transcriptional heterogeneity of patients IF0137 (k=4; as this patient had four major subclones), IF0138 and IF0101 (k=3; as these patients had three genetic subclones). K-nearest neighbourhood (knn), implemented within the PAGOGA2 package was used to analyse the same patients: first, I calculated a PCA reduction of the batch corrected gene expression matrix and selected the top 50 components and 3000 overdispersed genes to compute nearest neighbours using “cosine” distance (k=15), and identified distinct clusters using “multilevel community” method.

26. RANDOM FOREST ANALYSIS

Random forest analysis was performed in R using “randomForest” package, trained on the genotype of each single cell and using ntree=2000. The input gene expression matrix was batch and donor corrected, whilst genotypes were preserved, as described in section “batch correction”. Genes were ranked based on the Gini index and the top 2000 genes were used to generate the tSNE representations presented in Figure 4.4C and Figure 4.12A.

27. FORCE-DIRECTED GRAPH ANALYSIS

ForceAtlas2 software (290) was used to visualize the force-directed graph layout represented in Figure 4.10, using the knn weighted graph described in the section “single cell clustering and dimensionality reduction” as an input, and using the following parameters: “iterations =1,000, edgeWeightInfluence=1, barnesHutTheta=2, scalingRatio=1, gravity=0.05, jitterTolerance=1”. The output was plotted using ggplot2 in R and the normalized mean gene expression values of megakaryocytic (*C6orf25*, *PF4*, *GP9*, *ITGA2B*, *SELP*), erythroid (*EPOR*, *KLF1*, *TFR2*, *TFR1*, *CSF2RB*, *APOE*, *APOC1*, *CNRIP1*) and myeloid (*EGR1*, *ID2*, *KLF2*, *TNFAIP2*, *B3GNT7*, *SERPINB9*; selected based on (17)) genes were superimposed to the graph.

28. CHROMATIN ACCESSIBILITY ANALYSIS

Fastq files were demultiplexed using bcl2fastq (version 2.20) with default parameters, trimmed for Nextera adaptors using TrimGalore (version 0.4.1) and aligned to the human genome (hg19) using bowtie2 with the following parameters: “--very-sensitive -p 4 -k 1 -X 2000”. PCR duplicates were removed using PicardTools version 2.3.0 and read start sites were adjusted by +4 bp on the positive strand and -5 base pairs on the negative strand using “shiftGAlignmentsList” function implemented in the ATACseqQC package in R. Peak calling was performed on read ensemble from all samples analysed, using MACS2 version 2.1.2 with the following parameters: “-g hs -B -q 0.05 --nomodel --extsize 200 --shift -100 --keep-dup=1 -f BAM --call-summits”. Blacklisted peaks (identified as technical artefacts) were removed with bedtools version 2.26.0 “intersect” function and option “-v”, using a catalogue of hg19 blacklisted regions downloaded from ENCODE (291). This generated a list of good quality consensus peaks, which was subsequently used to count the number of reads mapping to these genomic regions in each sample using featureCounts. Differential peak accessibility was performed with the DESeq2 package in R. Motif identification was performed using

HOMER, using a 200 bp window from each peak centre and specifically selecting peaks differentially expressed in each group.

29. STATISTICAL ANALYSIS

Unpaired Student t-test with Welch's correction was used for the comparisons presented in Figure 3.1A, 3.1B, 3.1C, 3.1E; Figure 3.7C, 3.7D and Figure 5.2C, 5.2D. For the comparison of Pearson's correlations in Figure 4.9C and 4.9D, Kolmogorov-Smirnov test was used. The statistical test performed for each type of analysis is specified in the figure legends.

CHAPTER 3: DEVELOPMENT OF TARGET-SEQ

1. INTRODUCTION

Tumours are composed of heterogeneous cell populations with different genetic and molecular properties. This diversity poses a major therapeutic challenge in cancer: the presence, in a single tumour, of many different genetic subclones, which might also be molecularly and functionally distinct, makes it difficult to effectively target all cancer cells in a particular patient, and this is a major cause of treatment failure. Therefore, it is important to characterize the molecular and functional consequences of intratumoral heterogeneity in order to develop better cancer treatments.

Genetic heterogeneity is by far the best characterized source of diversity in cancer. The advances in next generation sequencing over the past decades have contributed to the comprehensive genetic characterization of most tumour types (292), and the recognition of the crucial contribution of genetic defects to cancer. This genetic variability has been used to stratify patients (62, 138) and predict responses to treatment (197, 293). More recently, with the advent of whole-transcriptome analysis techniques such as RNA-sequencing, the molecular diversity of cancer has also been widely characterized (294). Crucially, although transcriptional diversity is partly driven by genetic heterogeneity, many other factors also contribute to transcriptional heterogeneity independently of genetic events (43). For example, cell cycle is a major source of transcriptional heterogeneity in tumours (239). Thus, linking genetic and transcriptional readouts is important to gain mechanistic insights into the molecular pathways driving tumour development.

For many years, these sources of heterogeneity were inferred from bulk sequencing techniques. However, they do not provide the necessary resolution to confidently resolve the phylogenetic tree of each tumour or to identify molecularly distinct subpopulations of cells. The development of single cell genomic techniques enabled the study of intratumoral heterogeneity at unprecedented resolution, resolving tumoral cellular ecosystems and assessing how they respond to current therapies. Nevertheless, genetic and molecular sources of heterogeneity at the single cell level have been frequently studied separately for a particular sample. This was mainly because of the lack of multi-omic single cell techniques allowing parallel interrogation of genetic and molecular (transcriptomic) readouts from the same single cell. Yet to understand the molecular mechanisms driving tumour initiation, progression and relapse, it is essential to effectively correlate mutational and transcriptomic readouts at the single cell level. Only this would allow to characterize the mechanisms driving clonal evolution in cancer, and how different genetic subclones respond to treatment.

1.1. STATE OF THE ART SINGLE-CELL TECHNOLOGIES: OVERVIEW AND LIMITATIONS

When this project started, the application of single cell technologies to cancerous tissues was restricted to the study of transcriptional heterogeneity using single-cell RNA-seq (239), or genetic clonal evolution using single-cell whole genome sequencing (225) or targeted mutational analysis of a few mutations (232, 295). However, technologies to incorporate both types of readouts from the same single cell did not exist at the time.

Mutations in coding regions could potentially be identified from RNA-sequencing reads spanning the variants of interest in single-cell transcriptomic protocols. However, the coverage across key mutations hotspots in commonly used scRNA-seq techniques is extremely poor, which prevents the detection of point mutations from single cells. This lack of coverage relates to the use of 3' and 5'-biased techniques, which preferentially

sequence the 3' or 5'-end of the transcripts, but also to the high dropout rate of full-length single-cell RNA-sequencing techniques. For instance, pioneering studies using full-length single-cell RNA-sequencing (SMART-seq2) to study intratumoral heterogeneity in glioblastoma could only identify point mutations from scRNA-seq reads in 0.3-9% of the cells. Distinct genetic subclones could be inferred from CNV profiles inferred from the scRNA-seq reads (238-240). However, point mutations, fusion genes and small indels are the most common genetic alterations identified in cancer genomes (296), and it is therefore essential to confidently detect these genetic defects in single cells to correlate transcriptomic readouts with genetic tumour heterogeneity.

This limitation of scRNA-seq techniques was first addressed by the Mead and Jacobsen laboratories in the context of CML, by incorporating parallel targeted amplification of transcripts and scRNA-seq of the same single cell. We modified the commonly used plate-based SMART-seq2 protocol by adding primers targeting the *BCR-ABL1* fusion gene, the major genetic driver of the disease, in the retrotranscription and PCR steps. Then, single-cell qPCR was used to quantify the presence of *BCR-ABL1* transcripts in each cell. This method achieved single molecule resolution, allowing to genotype single cells with high sensitivity and to characterize the molecular signatures associated with each genetic subclone during TKI resistance.

However, this method is dependent on the expression of the mutated gene/allele in every cell. For many autosomal genes, expression is undetectable or highly allelic biased (this is, one of the alleles is preferentially expressed at the transcript level; Figure 3.1) making this method unsuitable for profiling the majority of mutations found in cancer.

To circumvent this limitation, the detection of mutations from the gDNA of single cells offers a suitable alternative, in parallel with whole transcriptome amplification. Several

methods have been described to amplify gDNA and mRNA from the same single cell. These methods either rely on the physical separation of each type of molecule (G&T-seq, SIDR, scTRIO-seq) (241, 243), which leads to loss of genetic material, or perform parallel amplification of total gDNA and mRNA using *in vitro* transcription, with subsequent masking of coding regions (244). These techniques lack the sensitivity to detect point mutations with enough resolution due to their high allelic dropout rate (estimated at >30%), and its use was restricted to the analysis of large Copy Number Variations (297). Moreover, these techniques mostly rely on whole-genome amplification of genomic DNA, with the subsequent increase in the number of false negative and false positive results and a significant increase in cost.

Targeted-sequencing techniques amplifying mutations from genomic DNA and concomitant gene expression analysis showed high sensitivity mutation detection (245, 298), but only allowed profiling of a restricted set of pre-selected transcripts per cell. As such, these methods were not suitable for a transcriptome-wide discovery type analysis.

1.2. HYPOTHESIS AND AIMS

To address the technological limitations in the field described above, the aim of the work described in this chapter was to develop a method that would allow the high sensitivity detection of mutations from single cells and at the same time, parallel whole transcriptome analysis. This technique would be ideally placed to study genetic and molecular heterogeneity in myeloid neoplasms, as well as other types of cancer.

Such technique should fulfil the following characteristics:

- High-sensitivity mutation detection at the single cell level (low allelic dropout), allowing detection of mutations in all cells, as well as confident detection of both alleles. This is essential in order to differentiate heterozygous versus homozygous statuses and to minimize the number of false negative results.

- High mRNA/cDNA molecular capture rate, minimizing transcript dropout, to obtain reliable transcriptional readouts from each single cell.
- The technique should be easily scalable, to be able to profile thousands of cells from each tumour and detect minor subpopulations of cancer cells such as cancer stem cells or therapy-resistant subclones.

Due to the high allelic dropout rate associated with whole-genome single-cell sequencing approaches, as well as significant cost, this approach would not be suitable for high-sensitivity mutation detection and would not be easily scalable. Therefore, I hypothesized that a targeted approach for mutation detection, in which only previously-identified known *driver* mutations will be analysed, would provide high resolution single-cell genotyping. I also hypothesized that plate-based techniques, associated with much higher molecular capture rates than droplet-based techniques, would be more suitable to achieve high sensitivity mutation detection. A plate-based approach is also much more easily customizable, and allows concomitant analysis of cell surface proteomics using index sorting. For this reason, all of the experiments described below refer to a plate-based implementation.

2. RESULTS

2.1. OPTIMIZATION OF SINGLE-CELL RNA-SEQUENCING CONDITIONS

Full-length single-cell RNA-sequencing techniques provide extremely poor coverage across key mutation hotspots, preventing the confident detection of point mutations from RNA-sequencing reads (1, 239). It was unclear whether this poor sensitivity was due to inefficient lysis or mRNA/cDNA amplification, or whether it related to true biological heterogeneity in the expression of mutant transcripts in single cells. Therefore, I first set out to optimize the retrotranscription and PCR conditions of SMART-seq2, the most widely used full-length plate-based scRNA-seq method, to try to improve the detection of coding mutations in single cells.

I first compared four different combinations of MMLV-derived retrotranscriptases, TSO oligonucleotides and PCR enzymes to SMART-seq2, which was used as a control (Table 3.1).

Condition	SMART-seq2	Cond#1	Cond#2	Cond#3	Cond#4
RTase	SuperScriptII (ThermoFisher)	ProtoScript (NEB)	SMART Scribe (Clontech)	SMART Scribe (Clontech)	SMART Scribe (Clontech)
PCR enzyme	KAPA 2G (Roche)	KAPA 2G (Roche)	KAPA 2G (Roche)	SeqAMP (Clontech)	SeqAMP (Clontech)
TSO	SMART-seq2 TSO-LNA	SMART-seq2 TSO-LNA	SMART-seq2 TSO-LNA	SMART-seq2 TSO-LNA	Proprietary Clontech TSO (SMART-seq v4)
Price per cell	≈2 GBP	≈2 GBP	≈2 GBP	≈2 GBP	≈60 GBP

Table 3.1. Summary of different Template Switching Oligonucleotides (TSO), retrotranscription and PCR conditions used for cDNA synthesis of single K562 cells.

All of the conditions tested showed a significant increase in the number of genes per cell compared to SMART-seq2 (Figure 3.1A), with the highest number of genes detected in Condition#3 and Condition#4. However, Condition#4 was associated with a much higher cost per cell, and therefore I decided to select Condition#3 for all subsequent experiments and analysis, referred to as SMART-seq plus (SMART-seq+) throughout the rest of this text.

I then assessed several metrics in SMART-seq plus libraries and compared them to SMART-seq2. Dropout probability was lower in SMART-seq plus (Figure 3.1B), including for lowly expressed genes (Figure 3.1C). Pearson correlation of global gene expression patterns confirmed the higher detection of lowly expressed genes in SMART-seq plus (Figure 3.1D); library bias was also significantly lower as compared with SMART-seq2 (Figure 3.1E), indicating that SMART-seq plus improved the molecular capture rate of cDNA molecules in single cells.

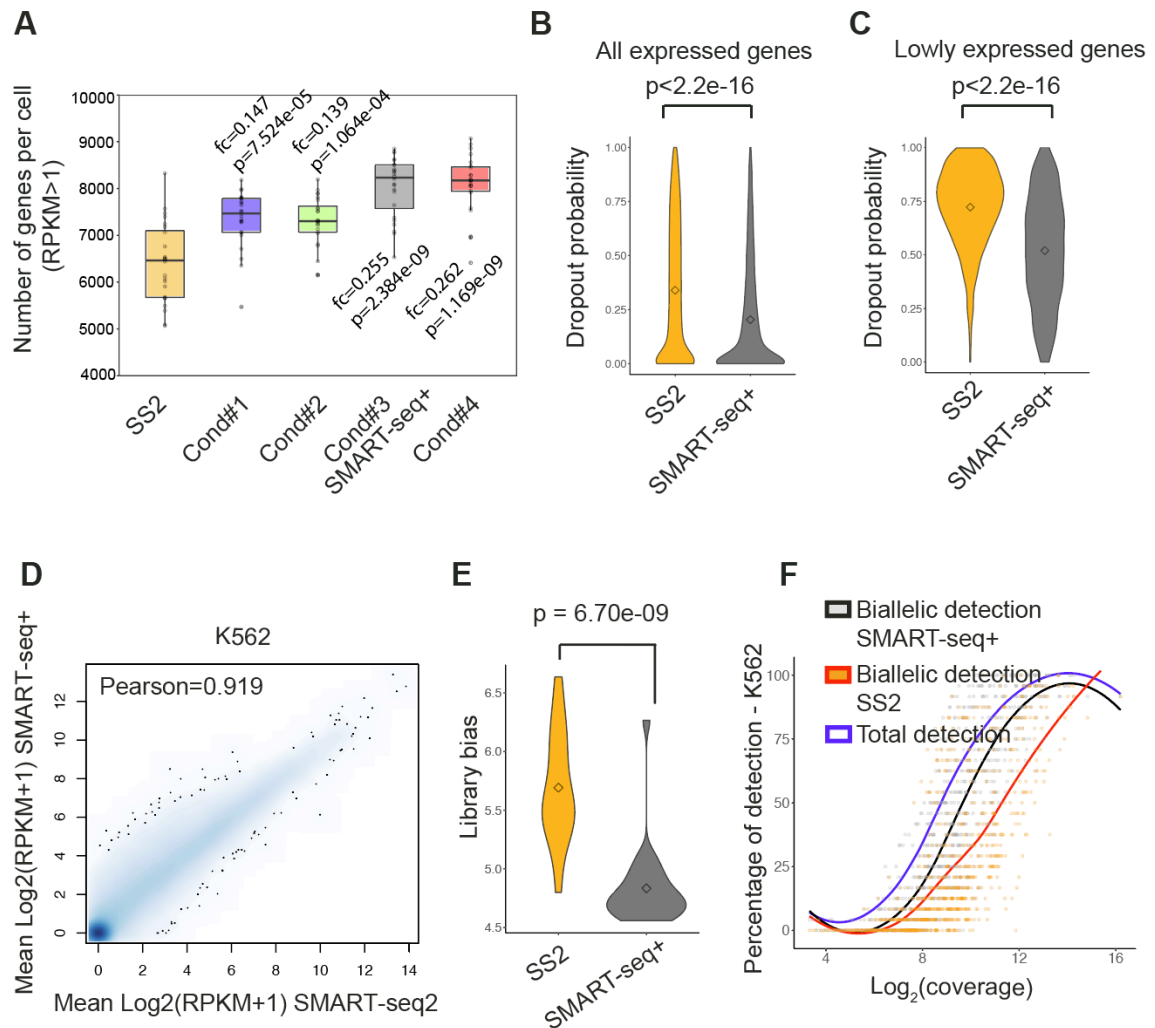


Figure 3.1. SMART-seq plus improves the quality of cDNA libraries from single K562 cells. (A) Number of genes detected per cell in (RPKM>1) in K562 cells profiled with each of the five different methods outlined in Table 21. Boxes represent median and quartiles and each point represents the value for each single cell. P indicates the p-value from Student's t-test for each method compared to SMART-seq2 (SS2), which is shown on the top of each bar for each comparison "fc" indicates fold-change of the mean value for each method compared to SMART-seq2. (B-C). Dropout probability for all expressed genes (B) or lowly expressed genes (C) for SMART-seq2 and SMART-seq plus methods. (D) Pearson's correlation between SMART-seq 2 and SMART-seq plus methods using all expressed genes (RPKM>1). (E) Library bias for single cell from each method. Library bias is defined as the number of reads mapping to the top 10% expressed transcripts divided by the total number of reads. (F) Total percentage of detection (blue line; y-axis) or bi-allelic detection (red line for SMART-seq plus and black line for SMART-seq2 or SS2; y-axis) of heterozygous single nucleotide variants from the same cells as in (B-E) with respect to the number of reads spanning each position (coverage; x-axis). Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

I then performed whole-transcriptome analysis of allele specific gene expression, to determine the allelic dropout (ADO) rate associated with each method; this is, how many times both alleles from a known heterozygous single nucleotide polymorphism (SNP) are detected by scRNA-seq, versus detection of a single allele (homozygous; false negative result). This analysis revealed a marked improvement in allelic dropout rates using SMART-seq plus (Figure 3.1F) as compared to SMART-seq2. However, ADO rates were still markedly high, especially for lowly expressed genes, which would make the false negative rate of this approach extremely high.

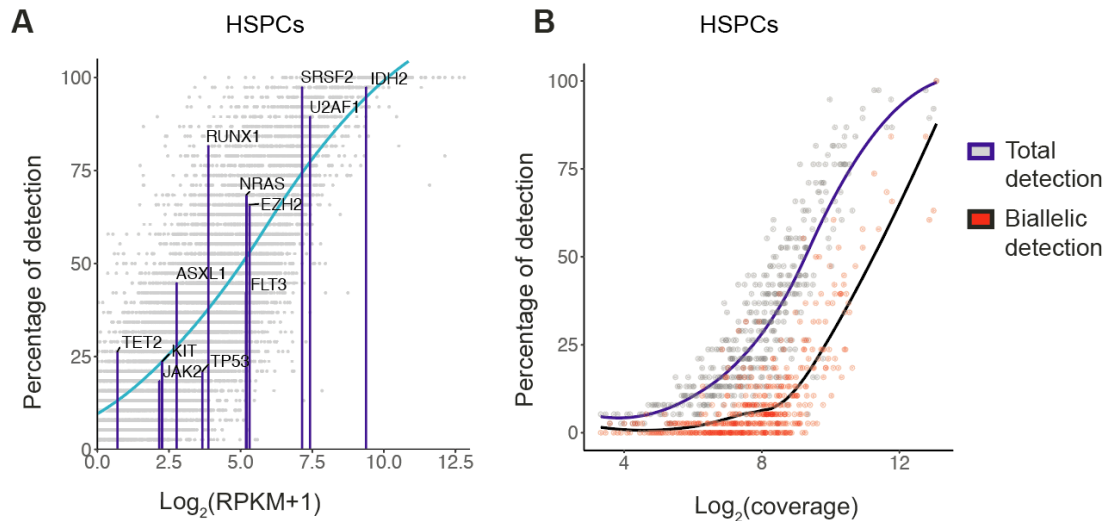


Figure 3.2. Single-cell RNA-sequencing of primary HSCPs reveals high levels of allelic dropout. (A) Percentage of detection of each transcript in Lineage-CD34+CD38- cells (n=38; y-axis) with respect to the expression level of each transcript (x-axis). Blue bars represent transcript recurrently mutated in myeloid malignancies. (B) Total percentage of detection (blue line; y-axis) or bi-allelic detection (red line; y-axis) of heterozygous single nucleotide variants from the same cells as in (A) with respect to the number of reads spanning each position (coverage; x-axis). Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

A similar analysis performed in Lineage-CD34+ HSPCs confirmed the low percentage of detection of transcripts from genes recurrently mutated in myeloid malignancies such as *TET2*, *JAK2*, *ASXL1*, *TP53* or *EZH2* (Figure 3.2A), and the extremely high allelic dropout rate in primary cells (Figure 3.2B). These results suggest that

improvements in cDNA synthesis conditions alone are highly unlikely to provide enough sensitivity to detect cancer-associated mutations from scRNA-seq.

2.2. DEVELOPMENT OF TARGET-SEQ

The high allelic dropout rate of single-cell RNA-sequencing methods led me to develop a technique that would allow high-sensitivity mutation detection and whole transcriptome analysis of the same single cell. For the reasons outlined above, I decided to implement a technique which would allow parallel detection of genomic DNA and mRNA from the same single cell, as this approach was likely to give higher sensitivity of mutation detection as compared to mRNA analysis alone.

To efficiently amplify genomic DNA from single cells, it would be necessary to first digest the proteins bound to it, in order to release genomic DNA and make it accessible for targeted amplification. To do this, a protease treatment to digest DNA-bound proteins is usually the method of choice in common genomic DNA extraction protocols (299). However, protease treatment interferes with subsequent enzymatic steps, as the protease will also digest any enzymes present in the reaction, and therefore it would not be compatible with parallel retrotranscription and PCR amplification in a one-step reaction unless this protease could be efficiently inactivated.

Firstly, I tested several proteinases and lysis conditions to determine the optimal combination generating good quality cDNA libraries. I tested two different proteases for the efficient retrotranscription and amplification of cDNA: proteinase K from New England Biolabs (Catalog# P8107S) and Qiagen Protease (Catalog# 19155). Two different conditions were also tested: addition of the proteinase in the lysis buffer, and subsequent heat inactivation before performing retrotranscription, as well as addition of the proteinase after the retrotranscription step, with subsequent inactivation prior performing PCR amplification. The cDNA yield obtained from single cells was used as

a measurement of efficient retrotranscription and cDNA amplification, compared to SMART-seq plus (control condition).

In the first instance, addition of the proteinase in the lysis buffer was tested (Figure 3.3A). Proteinase K can only be efficiently inactivated by heat at 95 °C, at which temperature the mRNA is degraded. Therefore, Proteinase K was inactivated by the addition of a proteinase inhibitor (Proteinase K inhibitor, Calbiochem, Catalog# 539470-10MG). Qiagen Protease is a thermolabile protease which can be heat inactivated at 72 °C, at which temperature mRNA is stable, and therefore Qiagen Protease was heat-inactivated. Digestion with Proteinase K and proteinase inhibitor resulted in a complete inhibition of cDNA amplification as compared to control conditions (98% reduction in cDNA yield; Figure 3.3A). Digestion with Qiagen Protease (Catalog# 19155) and subsequent heat inactivation at 72 °C did not reduce cDNA yield, indicating that protease was efficiently inactivated by heat (Figure 3.3A).

The addition of proteinase after the retrotranscription and prior performing a PCR amplification step was also tested (Figure 3.3B). Proteinase K (New England Biolabs, Catalog# P8107S) and Qiagen Protease (Qiagen, Catalog# 19155) were tested. Both enzymes were added to each library and were incubated at 56 °C for 10 minutes and heat-inactivated at 95°C for 20 minutes. At this stage, mRNA has been retrotranscribed to cDNA, which is stable at high temperatures, and therefore this heat inactivation step can be performed at higher temperatures. However, the addition of proteinase in this step resulted in a 35 % (Proteinase K) and 47 % (Qiagen Protease) reduction in cDNA yield (Figure 3.3B).

Addition of Qiagen Protease in the lysis buffer was the only condition preserving cDNA yield from single cells and therefore was selected for further validation in primary HSPCs. Two different amounts (2.7×10^{-5} and 5.4×10^{-5} Anson Units per microliter) and lengths of heat inactivation (10 and 15 minutes) at 72°C were tested to determine the

optimal concentration and duration of heat inactivation of the Qiagen Protease (Figure 3.3C-E).

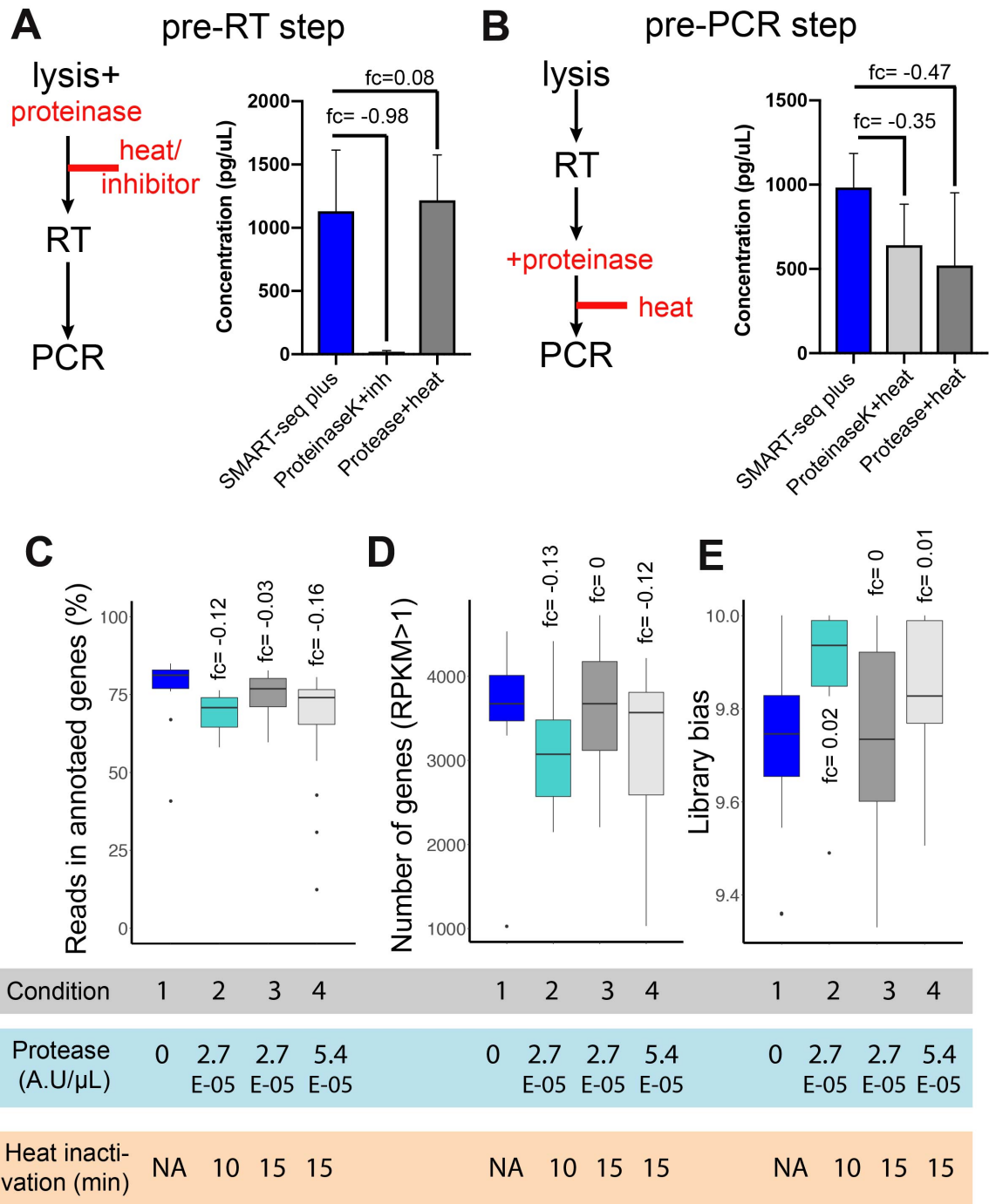


Figure 3.3. Optimization of proteinase-based lysis conditions. (A) Comparison of different cDNA yields obtained from single cells when adding Proteinase K or Qiagen Protease to the lysis buffer, and inactivating then with Proteinase K inhibitor or heat prior retrotranscription. cDNA yield was compared to SMART-seq plus libraries, which were used as a control. Bar plot indicates the mean value for each group and error bars indicate standard deviation from the

mean; mean fold-change (“fc”) difference compared to SMART-seq plus libraries is indicated on the top of each bar. (B) Comparison of different cDNA yields obtained from single cells when adding Proteinase K or Qiagen Protease to single-cell libraries after the retrotranscription step, and subsequently heat-inactivating them prior PCR amplification of cDNA. cDNA yield was compared to SMART-seq plus libraries, which were used as a control. Bar plot indicates the mean value for each group and error bars indicate standard deviation from the mean; mean fold-change (“fc”) difference compared to SMART-seq plus libraries is indicated on the top of each bar. (C-E) Comparison of sequencing statistics between libraries generated with different lengths of heat-inactivation and concentrations of Qiagen Protease, as compared to SMART-seq plus, which was used as a control condition. Qiagen Protease was added in the lysis buffer at different concentrations and heat inactivated prior retrotranscription for 10 or 15 minutes at 72 °C. Boxplot indicates median and quartiles for each group; mean fold-change (“fc”) difference compared to SMART-seq plus libraries is indicated on the top of each bar. Outliers (values lower or higher than 1.5 times the interquartile range) are indicated by points.

After protease based lysis, samples were sequenced, and sequencing metrics of test conditions were compared to SMART-seq plus (condition 1), which was used as a control (Figure 3.3C-E). Inactivating Qiagen Protease for 10 minutes using 2.7E-05 AU/μL led to a 12% reduction in reads mapping in known transcripts, 13% reduction in genes detected per cell and an increase in library bias (Figure 3.3C-E; condition 2). The addition of 5.4E-05 Anson Units per microliter and 15 minutes heat inactivation led to a 16% reduction in the reads mapping in transcripts, a 12% reduction in the number of genes detected per cell and an increase in library bias (Figure 3.3C-D; condition 4). The addition of 2.7E-05 Anson Units per microliter and heat inactivation for 15 minutes achieved sequencing metrics comparable to SMART-seq plus control condition (condition 1; 3% decrease in reads mapping in known transcripts, and no differences in number of genes detected per cell or library bias; Figure 3.3C-E). This indicated that applying 72 degrees Celsius for 15 minutes efficiently inactivated Qiagen Protease and led to generation of good quality cDNA libraries from single cells; moreover, this condition showed a remarkable increase in amplification efficiency of genomic DNA amplicons at the single cell level (Figure 3.6B).

This led me to choose an incubation step of 15 minutes using $2.7\text{E-}05$ AU/ μL of Qiagen Protease for all subsequent steps. Additionally, to efficiently detect mutations from the mRNA and genomic DNA of single cells, target-specific primers were added during retrotranscription (primers targeting mutant transcripts) and PCR amplification (primers targeting mutant transcripts and genomic DNA amplicons), which constitute the TARGET-seq workflow outlined in Figure 3.4. This TARGET-seq workflow generates full-length cDNA libraries, and has a throughput of approximately 300 cells per week.

METHOD LAYOUT

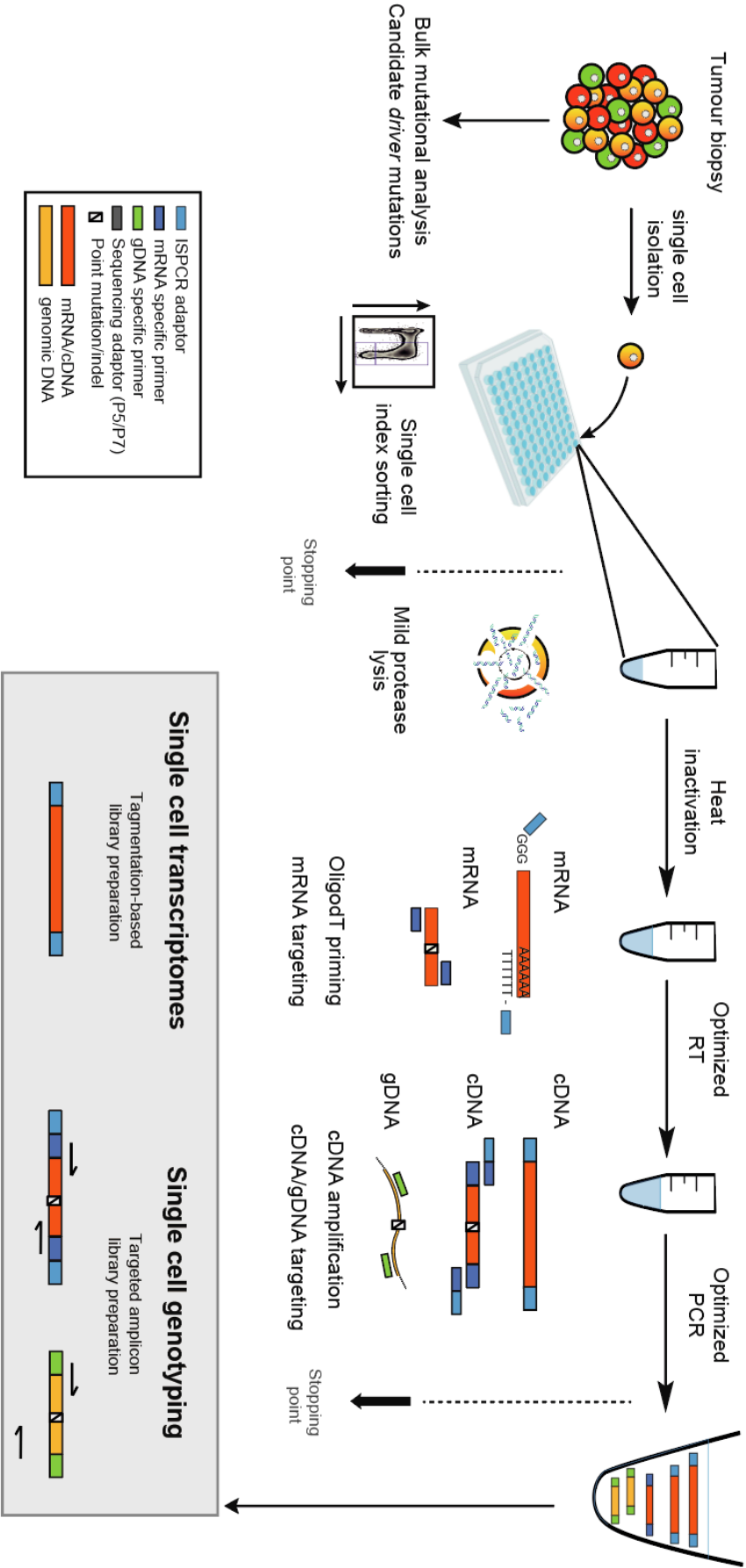


Figure 3.4. Schematic representation of full-length TARGET-seq method. Each tumour is analysed using a bulk-sequencing method to identify candidate *driver* mutations for each sample, which will be subsequently used for targeted amplification with TARGET-seq. The tumour is dissociated into a single cell suspension and single cells are directly sorted into plates using FACS. The mean fluorescence intensity of each marker is recorded for each single cell. Once cells are deposited into the PCR plate containing lysis buffer, cell lysis occurs immediately. At this point, plates can be stored in -80 C. Then, the protease contained within the lysis buffer is heat inactivated so it does not interfere in subsequent enzymatic steps. After that, the retrotranscription mix is added to each plate. Total polyadenylated mRNA molecules are primed using an oligodT tagged to adaptor sequences, and retrotranscribed with a MMLV-derived enzyme and TSO to generate full-length transcripts. Mutated transcripts (identified through bulk sequencing) are retrotranscribed using target-specific primers. Subsequently, cDNA is amplified during PCR using a primer binding to the ISPCR adaptor sequence added during the RT step, and genomic DNA primers are added to amplify amplicons spanning the mutations of interest. After that, an aliquot of the resulting cDNA+amplicon mix is used for whole transcriptome library preparation and another aliquot, for single-cell targeted genotyping. Figure adapted from Rodriguez-Meira *et al*, 2019 (1). Some of the elements in this figure were generated using BioRender.

2.3. DEVELOPMENT OF A BIOINFORMATICS PIPELINE FOR SINGLE-CELL MUTATION ANALYSIS

TARGET-seq generates unique targeted sequencing data, in which cDNA and gDNA molecules are analysed from each single cell, but should be analysed independently to obtain separate mutational readouts from each type of molecule. Moreover, TARGET-seq generates targeted-sequencing data from thousands of cells, which cannot be efficiently processed using common variant calling pipelines. Therefore, this type of data requires the development of an efficient bioinformatics pipeline to genotype each single cell.

To efficiently process mutations readouts from single-cell genotyping data generated with TARGET-seq, I developed a pipeline in which fastq files are automatically processed to generate a matrix of single-cell genotypes (Figure 3.5A). The pipeline's code is available at: <https://github.com/albarmeira/TARGET-seq>. First, each single-cell derived fastq file is aligned to the human genome using a splice-aware aligner (STAR); then, amplicons corresponding to cDNA or gDNA were separated in individual files for

each single cell. This was achieved by matching the chromosomal coordinates of the primers used for target-specific genotyping PCR1 amplification with the chromosomal coordinates of each read. Then, variant calling is performed individually for each cell and each type of molecule (cDNA/gDNA) using samtools mpileup. Blank controls are used to set detection cut-offs (Figure 3.5B) for each variant and wild-type (WT) controls (cells from normal donors) are used to set mutation detection cut-offs (Figure 3.5C).

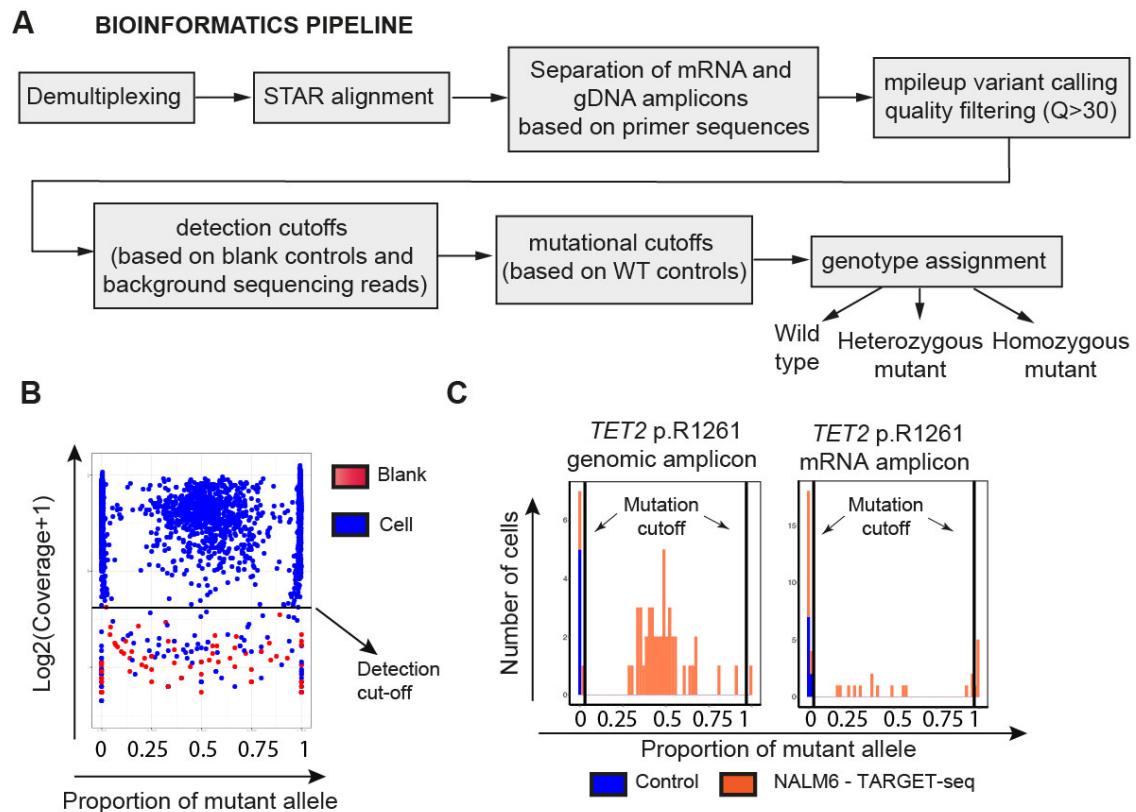


Figure 3.5. Bioinformatics pipeline for TARGET-seq single-cell mutation analysis. (A) Schematic overview of each step of the pipeline. (B) Representative example of detection cut-offs based on blank controls and (C) mutation cut-offs used to assign genotypes based on WT controls. Cells falling below 3% of mutant allele frequency were assigned as WT; cells between 3-97% were assigned as heterozygous genotype and those falling above 97%, as homozygous. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

Finally, gDNA and mRNA readouts for each variant were merged to generate a consensus genotype for each cell. Cells were considered heterozygous if at least one of the molecules (gDNA or mRNA) gave a heterozygous readout for an specific variant,

or if one of the molecules gave a WT readout and another one an homozygous mutant readout. Cells were considered homozygous if only the mutant allele was detected and wild-type if only the wild-type allele was detected.

2.4. VALIDATION OF TARGET-SEQ IN CELL LINES AND PRIMARY HAEMATOPOIETIC STEM CELLS

Using the pipeline described above, I first validated TARGET-seq in cell lines with known heterozygous point mutations or small indels in regions recurrently mutated in myeloid malignancies. I compared TARGET-seq performance to detect these mutations with SMART-seq plus (not including target-specific amplification or protease-based lysis) and mRNA targeting (247) (including mRNA target-specific amplification but not protease-based lysis) methods. In all cases, single-cell mutational analysis and variant calling was performed from targeted-sequencing reads. TARGET-seq provided better sensitivity of mutation detection across all of the coding and non-coding regions profiled in cell lines (Figure 3.6A). Importantly, TARGET-seq achieved >98 % mean bi-allelic detection for coding regions, indicating that the allelic dropout rate associated with the method was extremely low (<2%). Analysis of genomic DNA amplicons alone (Figure 3.6B) revealed that the low allelic dropout rate in TARGET-seq condition was largely due to the high amplification efficiency of genomic DNA amplicons, which provided 93% mean bi-allelic detection (Figure 3.6B).

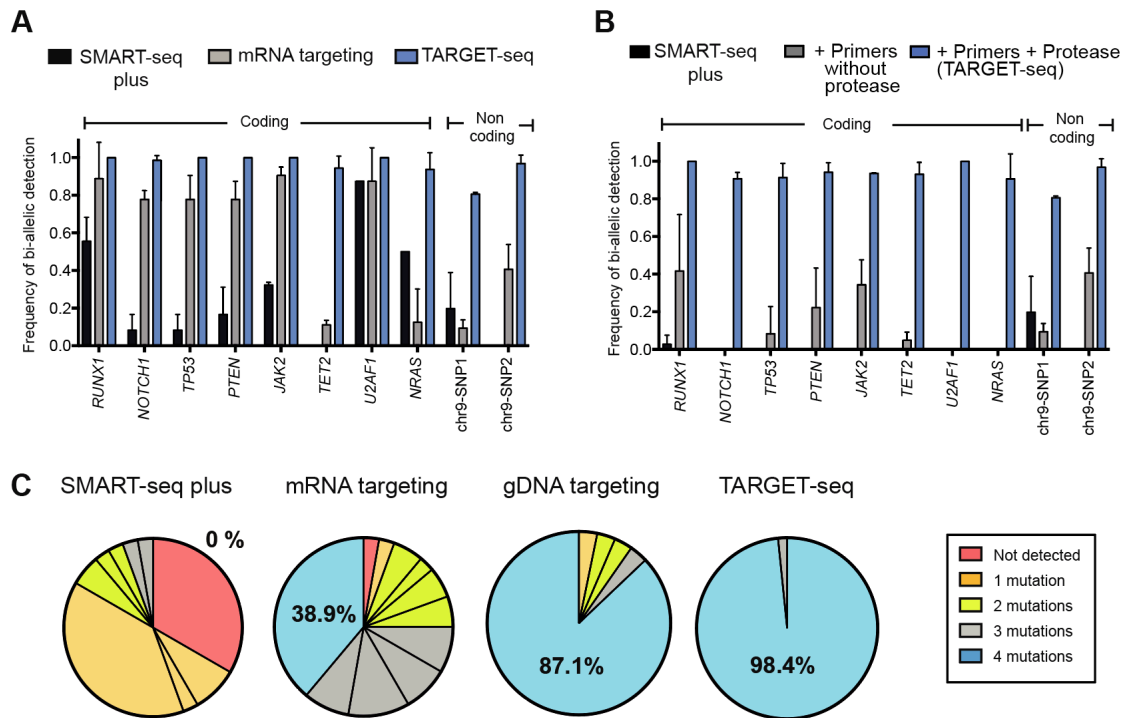


Figure 3.6. TARGET-seq achieves high sensitivity mutation detection in single cells. (A) Frequency of bi-allelic detection of heterozygous mutations in ten coding (n=8) and non-coding (n=2) regions recurrently mutated in myeloid malignancies in MOLT4, JURKAT and SET2 cells lines (n=376 single cells, 2-3 independent experiments per amplicon). Bars represent mean $\pm$ standard deviation. SMART-seq plus, mRNA targeting and TARGET-seq approaches are compared. (B) Frequency of bi-allelic detection of heterozygous mutations in the same single cells as in A, displaying results from gDNA detection alone for each condition. Bars represent mean $\pm$ standard deviation. (C) Cumulative frequency of heterozygous mutation detection in JURKAT cells when profiling four different mutations in the same single cell (*RUNX1*, *TP53*, *PTEN*, *NOTCH1*), using SMART-seq plus (n=36 cells), mRNA targeting (n=36), genomic DNA targeting (n=62) and TARGET-seq (n=62) methods in three independent experiments. Each colour represents the number of heterozygous mutations detected per single cell and each slice of the pie, the number of cells with a particular combination of mutations. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

To reconstruct tumour clonal hierarchies, it is essential to detect several combinations of mutations in single cells with high sensitivity. To test TARGET-seq performance when more than one mutation was profiled per single cell, I profiled four heterozygous mutations in JURKAT cell line (*RUNX1*, *PTEN*, *NOTCH1*, *TP53*) and compared it to SMART-seq plus, mRNA targeting, gDNA targeting or TARGET-seq methods

(combined mRNA and gDNA targeting; Figure 3.6C). In gDNA targeting condition, protease based lysis was performed and gDNA-specific primers were added, but only reads from genomic DNA amplicons were considered for the analysis; for TARGET-seq, both mRNA and gDNA derived reads were considered.

This analysis confirmed the low allelic dropout rate achieved with TARGET-seq (combined mRNA+gDNA targeting), which detected the four mutations in 98.4% of the single cells profiled, in contrast with 38.9% and 87.1% of the cells when using mRNA or gDNA targeting (Figure 3.6C). SMART-seq plus did not detect the four mutations in any of the cells analysed (Figure 3.6C).

I then evaluated the quality of single-cell transcriptome libraries generated by TARGET-seq in two cell lines (JURAT, SET2) and primary HSPCs. Cells clustered by cell type rather than method (Figure 3.7A, 3.7B), and the number of genes detected per cell was not significantly different (Figure 3.7C). Equally, sequencing statistics (Figure 3.7D) and the number of cells passing quality controls (QC) were comparable between methods (Figure 3.7E). TARGET-seq also generated full-length libraries comparable to SMART-seq plus (Figure 3.7F). These results demonstrate that TARGET-seq generates good quality whole transcriptome sequencing libraries, and at the same time, provides high sensitivity mutation analysis from single cells.

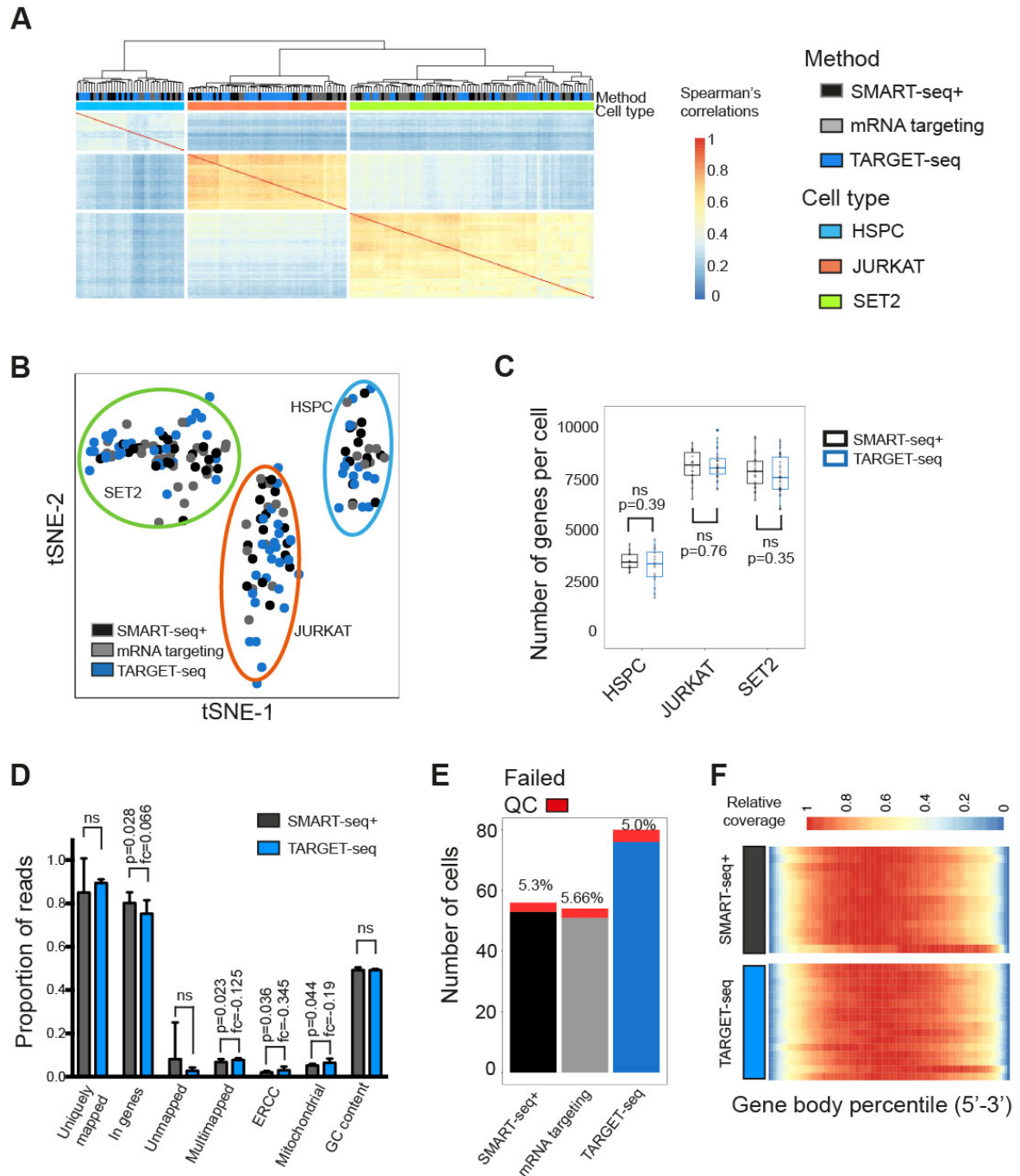


Figure 3.7. TARGET-seq produces high quality and unbiased whole-transcriptome libraries from single cells. (A) Unsupervised hierarchical clustering of Spearman's correlation from JURKAT (n=56), SET2 (n=86) and HSCPs (n=38) using SMART-seq plus, mRNA targeting or TARGET-seq methods. 4,088 highly variable genes were used for the analysis. (B) tSNE representation of the same cells as in (A) using 4,088 highly variable genes. The color of each dot indicates the method used for cDNA synthesis and the circles indicated each cell type (green, SET2 cells; blue, HSCPs; orange, JURKAT cells). (C) Number of genes detected per cell in JURKAT, SET and HSCPs using SMART-seq plus and TARGET-seq methods. The boxes represent median and quartiles, and each point represents the value of each single cell. P-value from Student's t-test is shown on the top of each bar for each comparison. "ns" indicates

non-significance. (D) Comparison of sequencing statistics between SMART-seq plus and TARGET-seq libraries. Bars indicate the mean value of the proportion of reads for each sequencing statistic and condition, and error bars indicate standard deviation of the mean. P-value from Student's t-test is shown on the top of each bar for each comparison. "ns" indicates non-significance. (E) Number of cells passing or failing quality control (QC) for SMART-seq plus, mRNA targeting and TARGET-seq methods for the same cells as in (A-B). The percentage of cells failing for each method is indicated on the top of each bar. (F) Normalized transcript coverage across 4040 housekeeping genes in HSPCs using SMART-seq plus or TARGET-seq methods. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

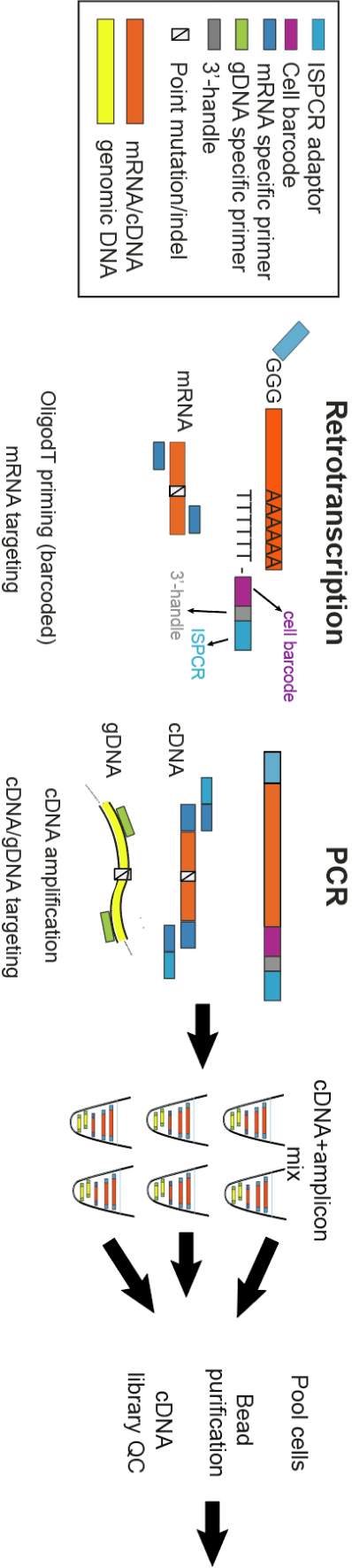
2.5. DEVELOPMENT OF A HIGH-THROUGHPUT, 3'-BIASED WHOLE TRANSCRIPTOME TARGET-SEQ METHOD

The complexity of cellular systems sometimes requires the use of higher throughput single-cell techniques which allow profiling thousands of cells. Full-length scRNA-sequencing techniques, such as the full-length implementation of TARGET-seq, are usually restricted to the analysis of a few hundreds of cells. This is due to the time-consuming and expensive library preparation protocols, which require an individual library preparation from each single cell to be generated, and the high cost associated with full-length single-cell RNA-sequencing, in which at least one million reads are required per cell on average to reach saturation of gene detection (247). 3'-biased or 5'-biased techniques do not require such high sequencing depth, because only a small part of each transcript is sequenced, and libraries are barcoded during retrotranscription, which means that library preparation can be performed on pools of single cells, reducing library preparation cost and hands-on processing time.

Therefore, I set out to adapt TARGET-seq to allow for higher throughput processing of single cells by introducing cellular barcodes in the oligodT-adaptor primers used to prime total polyadenylated mRNA molecules (Figure 3.8), to allow for early pooling of mRNA libraries and to adapt the method to a 3'-biased sequencing configuration. During retrotranscription, a barcoded oligodT tagged to a cell barcode, 3'-specific PCR handle and a known adaptor sequence (ISPCR adaptor; Figure 3.8) is used to retrotranscribe mRNA molecules from each single cell. A uniquely barcoded oligodT

primer is used in each well of a PCR plate in which each single cell is sorted. After that, cDNA is amplified using the ISPCR adaptor sequence added during the RT step to generate single-cell barcoded and amplified libraries. Then, libraries from each plate are pooled in a unique reaction, bead purified and subsequently used for tagmentation. During the tagmentation step, a custom primer binding to the 3'-specific PCR handle and ISPCR adaptor sequence is used to specifically enriched for the 3'-end of each molecule, and this is used during sequencing to retrieve the cell barcode.

I first evaluated the ability of the method to identify single cells based on each cellular barcode introduced during the retrotranscription step. To do so, I performed a cell mixing experiment in which K562 (human) and BaF3 (mouse) cells were mixed in the same tube and sorted into individual plates containing uniquely-barcoded oligodT primers. If the transcriptome of an individual cell was correctly assigned using cellular barcodes, each cell should only contain mouse or human transcripts, but not both. Mixed transcriptomes were not detected in any of the cells profiled (n=177; Figure 3.9A) and neither mouse or human transcripts could be detected in blank controls (n=5; Figure 3.9A). I then validated the detection of cellular barcodes result in primary HSPCs (Figure 3.9B). The 16 cell barcodes used for cDNA synthesis of HSPCs could be reliably demultiplexed and comprised 95.9% of the reads, whereas none of the barcodes that were not included in the experiment (n=80) could be detected (Figure 3.9B).



Single cell whole transcriptome library preparation

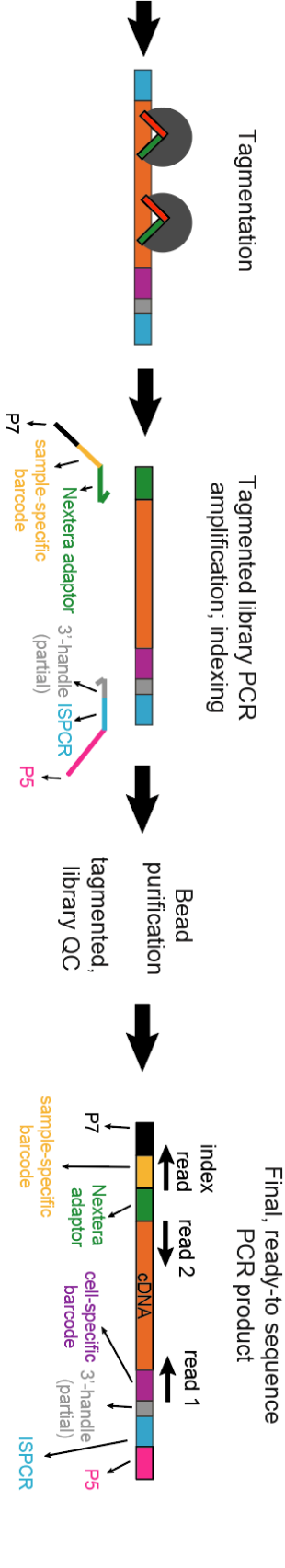


Figure 3.8. Schematic representation of 3'-TARGET-seq. Each sequence is represented by a rectangle labelled according to the colour codes provided in the text boxes. QC: quality control. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

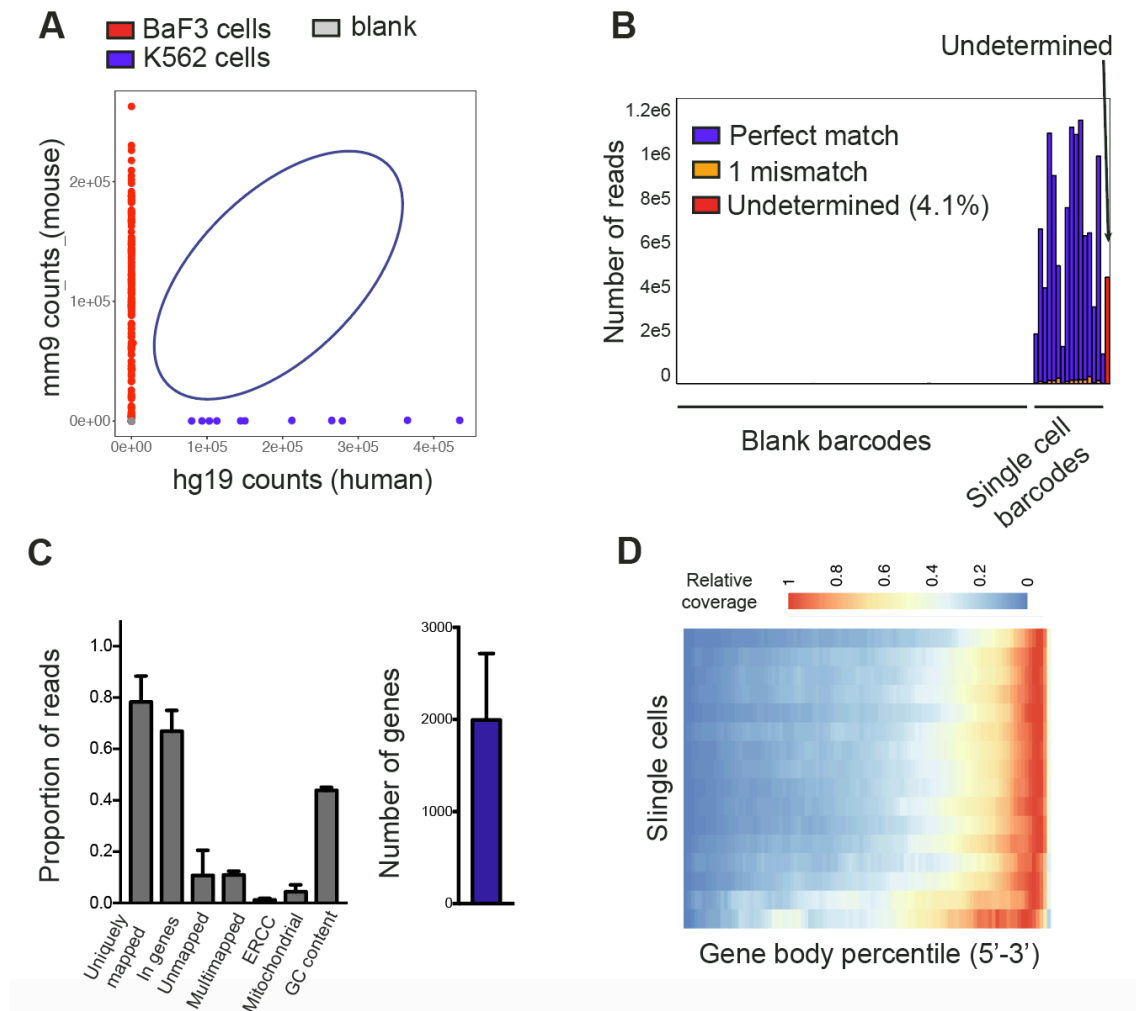


Figure 3.9. Technical performance of 3'-TARGET-seq. (A) Number of reads mapping to the mouse or human genome detected in a mixture of K562 and BaF3 cells (n=177). The black circle represents the region in which mixed human:mouse transcriptomes should be detected. Blank wells (coloured in grey, n=15) in which cells were not sorted are also included. (B) Detection of cell barcodes in HSPCs (n=16). Blue bars represent the 16 barcodes included in the reaction and the red bar represents reads that could not be demultiplexed to any of the 96 barcodes available; blank barcodes (n=80) represent those barcodes not used for cDNA synthesis. (C) Sequencing statistics of the same cells as in (B) (left panel) and number of genes detected per cell (right panel). (D) Normalized transcript coverage across 4040 housekeeping genes in the same 16 HSPCs as in (B,C), demonstrating coverage bias towards the 3'-end of the transcript. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

Sequencing statistics (Figure 3.9C) were similar to those obtained with full-length SMART-seq plus (Figure 3.9D), were in line with those obtained from other 3'-biased single-cell RNA-sequencing methods (13, 212) and the transcriptome libraries showed the expected 3'-bias (Figure 3.9D). This indicates that 3'-TARGET-seq method provides good quality single-cell transcriptomes whilst allowing for much higher throughput processing of single cells.

2.6. DEVELOPMENT OF A HIGH-THROUGHPUT SINGLE-CELL GENOTYPING METHOD

The previously described single-cell genotyping strategy (see Chapter 2, section "Targeted NGS single-cell genotyping") relies on two consecutive PCR steps in which first, targets are amplified using target-specific primers and secondly, cell-specific Illumina-compatible commercial barcodes are added to each pre-amplified single-cell library. This strategy is relatively low throughput, as only 384 cells can be processed per day even when using liquid handling robotic platforms, and it is expensive because two PCR steps need to be performed for each single cell before these libraries can be pooled in a single reaction. Moreover, the use of commercially available Illumina-compatible barcodes restricts the multiplexing capacity of the method to 384 cells per sequencing run. This precludes the use of high-throughput sequencing platforms such as NextSeq, HiSeq or NovaSeq sequencers (Illumina), which would require a much higher multiplexing capacity in a single sequencing run (i.e., allow for pooling of thousands of single cells with non-overlapping barcodes), making the protocol expensive as well as time consuming.

To circumvent this limitation, I designed a barcoding strategy to allow pooled amplification of targeted amplicons and to achieve higher multiplexing capacity. This approach incorporates a unique 6 bp cellular barcode in each forward and reverse target-specific primer used for genotyping PCR1. After PCR1 amplification, single-cell targeted libraries are pooled, but the original cell-of-origin from each molecule is

recorded in the cellular barcode added during PCR1. This pool is then cleaned up and used for genotyping PCR2, in which a pool-specific barcode tagged to Illumina-compatible adaptors is attached to each library (Figure 3.10).

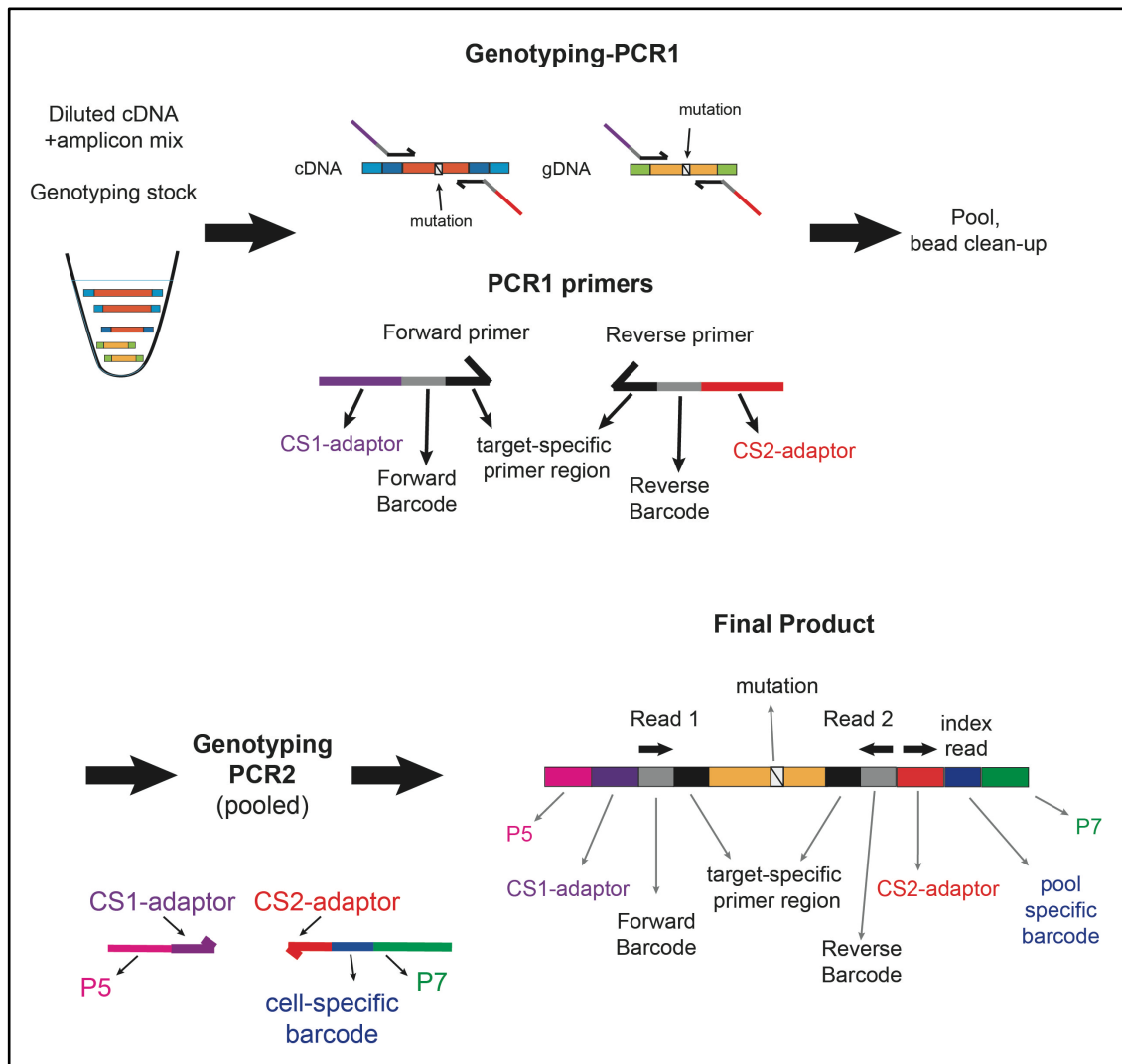
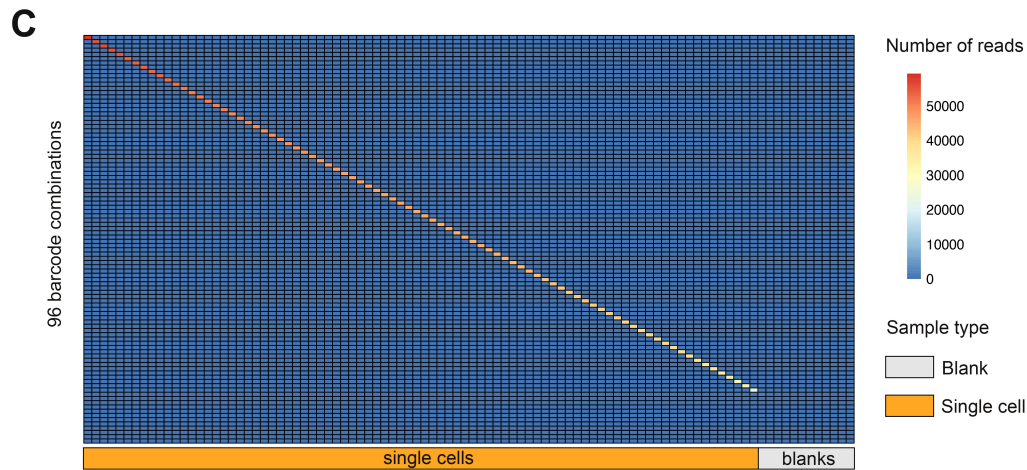
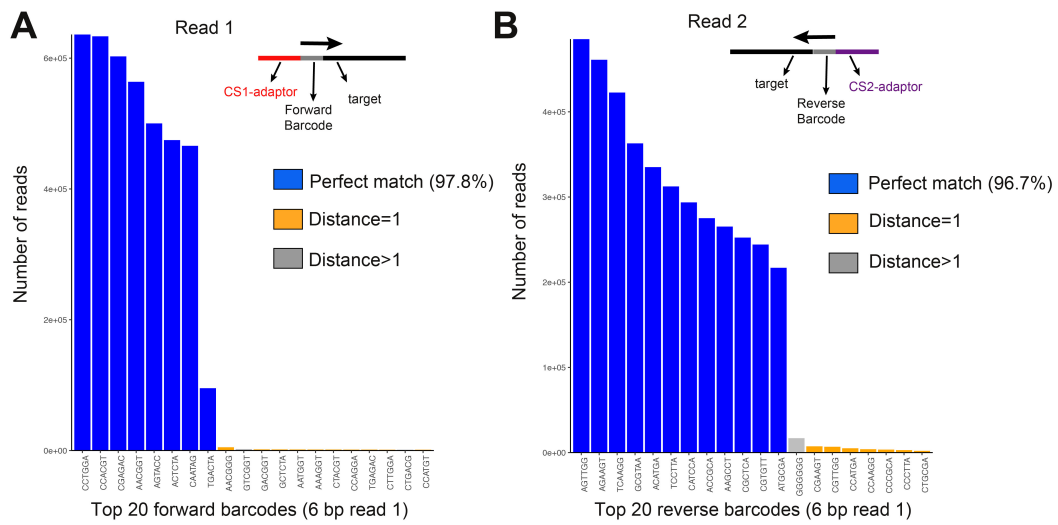


Figure 3.10. Schematic representation of the barcoding strategy for high-throughput single-cell genotyping. During PCR1, a 6 bp barcode is introduced in each side of the molecule in between the universal CS1/CS2 adaptor and the target-specific primer sequence. Each single-cell derived targeted library is pooled, and during PCR2, Illumina compatible adaptors containing a unique index sequence for each pool are attached to the libraries. During sequencing, the first 6 bp of read1 and read2 will contain the cellular barcode, which will be used to demultiplex each library according to the original cell it came from.

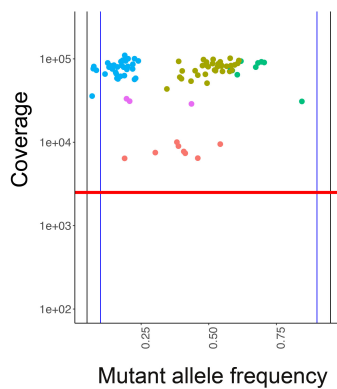
Using eight forward primers and twelve reverse target-specific primers during PCR1, 96 unique combinations of forward and reverse barcodes are generated for a specific amplicon. Then, up to 384 different barcodes can be used during PCR2 to amplify each

pool, resulting in a theoretical multiplexing capacity of 96 times 384 (36864 unique barcode combinations).

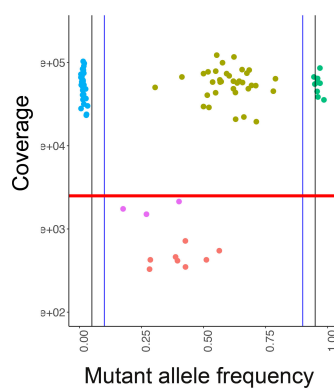
Cellular barcodes tagged to target-specific primers could be reliably detected in single cells (Figure 3.11A, 3.11B) and there was no barcode cross-contamination in reactions performed in individual PCR wells (Figure 3.11C). However, barcode cross-contamination was observed when genotyping-PCR1 reactions were pooled and barcoded in a single-reaction (Figure 3.11D), indicating that the remainder target-specific barcoded primers in the pooled genotyping-PCR2 were overriding the pre-attached cellular barcodes during PCR1. To overcome this issue, I performed progressively more stringent bead purification conditions and reduced the primer concentrations in genotyping-PCR1 pools prior performing genotyping PCR2. This showed a gradual decrease in barcode cross-contamination, achieving an optimal result when using 0.6:1 beads to cDNA ratio (Figure 3.11E) or reducing the primer concentration to 600 nM (Figure 3.11F).



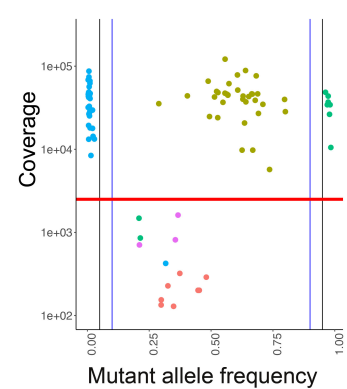
D 0.8:1 beads to cDNA ratio
1200 nM primer
pooled genotyping-PCR2



E 0.6:1 beads to cDNA ratio
1200 nM primer
pooled genotyping-PCR2



F 0.8:1 beads to cDNA ratio
600 nM primer
pooled genotyping-PCR2



Reference sample type

blank failed wild-type heterozygous homozygous

detection cut-off 0.05/0.95 MAF 0.10/0.90MAF

Figure 3.11. Technical performance of high-throughput single-cell genotyping. (A-B) Top 20 cell barcodes identified in the first 6 base pairs of the forward (A) and reverse (B) reads. Barcodes are coloured according to the Levenshtein distance from the expected “ground truth” barcodes. Eight forward barcodes and twelve reverse barcodes were used for the analysis. (C) Number of reads mapping to each cellular barcode combination (forward and reverse) in each single cell. Cells were individually barcoded during genotyping-PCR1 and genotyping-PCR2, adding a sample-specific barcode for each single cell. Sample-specific barcodes added during PCR were used for demultiplexing, and the reads mapping to each combination of cell-specific barcodes used in the experiment were computed for each sample. (D-F) Correlation between genotyping-PCR2 performed in individual wells (as described in Chapter 2, “Targeted NGS single-cell genotyping”) and pooled genotyping-PCR2 with different bead purification ratios and primer concentrations for genomic *JAK2* amplicon in 84 single cells and 12 blanks. The coverage of *JAK2V617F* genomic amplicon is plotted against the mutant allele frequency in each single cell; cells are coloured according to the genotype of the control condition (genotyping-PCR2 performed in individual wells). (D) 0.8:1 bead to cDNA ratio and 1200 nM primers, (E) 0.6:1 bead to cDNA ratio and 1200 nM primers (F) 0.8:1 bead to cDNA ratio and 600 nM primers. MAF: Mutant Allele Frequency.

3. DISCUSSION

Single-cell multi-omic techniques are ideally placed to resolve intratumoral heterogeneity, which poses many challenges to develop effective cancer treatments. However, when this study started, technological limitations prevented the parallel interrogation of genetic heterogeneity (mutations) and molecular diversity (transcriptome) from the same single cell.

To tackle such limitation, I developed TARGET-seq (1), a plate-based single-cell RNA-sequencing technique which detects mutations from single cells with extremely high sensitivity and provides whole transcriptome readouts from the same single cell. I implemented TARGET-seq for low throughput and high depth full-length sequencing, as well as higher throughput 3'-biased transcriptome profiling. Full-length TARGET-seq provides higher molecular capture rates, with more genes detected per cell on average, and it allows the interrogation of alternative splicing at the single cell level; however, its throughput is limited to processing approximately 300 cells per week and

the cost per cell is high (>15 GBP). 3'-biased TARGET-seq has lower molecular capture rates, and due to the 3'-biased nature of the libraries generated, does not allow detection of alternative splicing within the entire body of the transcript. However, it allows for much higher throughput processing of single cells (more than 1000 cells per week) at a much lower cost ($\approx$ 5 GBP) than full-length TARGET-seq.

Other methods have been simultaneously developed to try to tackle this same limitation. An adaptation of Seq-Well (300), a 3'-biased technology in which single cells are deposited in array nanowells and subsequent reactions are performed in the same array, was used to characterize cellular hierarchies in AML (250). In this study, van Galen and colleagues used biotinylated primers to capture mutant transcripts after whole transcriptome amplification of pooled cells. This targeted library was then used for targeted-sequencing of common myeloid hotspots, and the cellular barcode of the 3'-end of the mutant transcript was used to match the genotype of each single cell to its whole transcriptome library. However, this technique has an extremely low detection rate (1-5%) for most mutations, especially for those far away from the 3'-end of the molecule.

Another technology inspired by this same principle adapted the mRNA targeting approach described by Giustacchini and colleagues (247) to the Chromium droplet-based platform from 10X Genomics (248). This technique, named GoT-seq (Genotyping of Transcriptomes) also relies on the cellular barcode at the 3'-end of each molecule to assign each mutant transcript to its cell of origin. Nam and colleagues incorporated a ligation approach to resolve the low detection of mutations in the 5'-end of the transcripts, in which by a series of targeted amplification steps, molecules are circularized to bring closer the site of the point mutation and the cellular barcode. These targeted libraries can then be sequenced on a standard Illumina platform, and this strategy can efficiently assign mutant transcripts to its corresponding cellular barcodes independently of the position of the mutation across the body of the

transcript. However, despite the substantial improvement in cell barcode assignment, the capture rate of most transcripts is extremely low, and only a few tens of cells could be genotyped for lowly expressed genes such as *JAK2*, illustrating the limitations of mRNA targeting to efficiently capture mutated transcripts in single cells.

Despite the higher throughput provided by other methods implemented in nanowell or droplet-based microfluidic platforms, TARGET-seq provides a much higher sensitivity of mutation detection and is currently the only method allowing interrogation of combinatorial patterns of mutations as well as confident reconstruction of tumour clonal hierarchies. However, the characterization of complex and highly heterogeneous cellular systems such as some solid tumours might require the analysis of tens of thousands of cells per sample, which could only be currently achieved with droplet-based or nanowell based approaches.

The incorporation of robotic platforms to fully automate the TARGET-seq protocol and to miniaturize reaction volumes will increase the scalability of the method, and it will reduce its cost. The implementation of the high-throughput single-cell genotyping strategy described in this chapter in section 2.6 will also result in much higher throughput single-cell genotyping, achieving a comparable throughput to that of droplet-based techniques.

In summary, the low allelic dropout and high-sensitivity mutation detection achieved with TARGET-seq is ideally placed to identify clonal hierarchies in cancerous tissues, and at the same time, obtain single-cell transcriptomic and cell surface proteomic readouts. Compared to other single cell multi-omic methods, TARGET-seq provides the highest sensitivity of mutation detection, allowing to confidently distinguish between heterozygous and homozygous cells. This is essential to identify molecular signatures associated with the acquisition of *JAK2V617F* mutation in HSPCs from MPN patients, which often undergo loss of heterozygosity of chromosome 9. Moreover, TARGET-seq

integrates cell surface proteomics, which can be used to trace back each single-cell to canonical immunophenotypically defined haematopoietic cell types.

The versatility of the method, which can be used for high-depth, low throughput or low-depth, higher throughput analysis would allow researchers to implement a different type of approach for different biological questions and sample sizes.

Ultimately, the integration of high sensitivity mutation detection, transcriptomics and cell surface proteomics using TARGET-seq would be a powerful tool to characterize intratumoral heterogeneity, and to identify rare subpopulations of tumour cells.

CHAPTER 4: SINGLE CELL ANALYSIS OF STEM AND PROGENITOR CELLS IN MYELOPROLIFERATIVE NEOPLASMS

1. INTRODUCTION

Myeloproliferative neoplasms are clonal haematopoietic disorders driven by mutations in the JAK/STAT pathway (*JAK2*, *CALR* and *MPL*) leading to constitutive signalling. These mutations, as well as other collaborating mutations acquired in haematopoietic stem cells, confer a competitive advantage to HSPCs, which undergo clonal expansion. The mechanisms by which these cells become clonally dominant and eventually give rise to the disease are poorly understood.

Whilst there are therapies specifically targeting aberrant JAK/STAT signalling activation (*JAK2* inhibitors), none of them induce durable clonal remissions (this is, eliminate mutant cells). If they do, these are short-lived, and patients frequently relapse. This is partly explained by the unspecific mechanism of action of these therapies, where clinical responses are thought to be in large part due to their anti-inflammatory activity, rather than specific targeting of the mutant MPN clone. Therefore, there is an urgent need to identify the mechanisms of clonal expansion and molecular pathways dysregulated in mutant cells, to ultimately develop new therapeutic strategies for MPN patients.

1.1. HYPOTHESIS AND AIMS

Single-cell multi-omic technologies offer a unique opportunity to identify dysregulated molecular pathways in MPN HSCPs, and generate testable hypothesis relating to the development of new therapeutic targets. Therefore, I hypothesized that applying single-

cell technologies to characterize the genetic and transcriptomic heterogeneity of MPN stem cells would lead to a better understanding of the cellular and molecular mechanisms of disease evolution and progression, as well as the identification of molecular signatures amenable of therapeutic targeting.

My specific aims were to apply TARGET-seq to:

- (1) Comprehensively characterize the clonal evolution of each tumour and determine combinatorial patterns of mutations that might not be apparent through bulk sequencing approaches
- (2) Identify molecular signatures of different genetic subclones and molecularly distinct subpopulations of tumour cells
- (3) Identify biomarkers of mutant cells that would allow the prospective isolation and targeting of the malignant MPN clone

In this chapter, I used TARGET-seq to characterize the genetic and transcriptomic heterogeneity of patients with myeloproliferative neoplasms carrying mutations in *JAK2* independently of disease phenotype, and cooperating mutations in the spliceosome component *U2AF1* using a full-length, high depth sequencing approach. Given that splicing is another source of cellular heterogeneity independently from transcript expression (301), this later approach would serve as a proof of principle validating the use of TARGET-seq to identify alternative splicing driven by spliceosome mutations in single cells.

Then, I specifically focused on the MPN subtype with poorer prognosis, and performed a comprehensive characterization of *JAK2*-mutant myelofibrosis using the higher throughput, 3'-biased version of the technique.

2. RESULTS

2.1. GENETIC AND TRANSCRIPTOMIC HETEROGENEITY OF *JAK2*-MUTANT MYELOPROLIFERATIVE NEOPLASMS

To determine the clonal hierarchy and transcriptional effects of MPN-associated mutations, I first selected a cohort of two normal donors and five myeloproliferative neoplasm patient samples with different disease phenotypes (MF, PV, ET and MDS/MPN overlap; Table 4.1) to perform TARGET-seq analysis. All of these patients carried *JAK2V617F* mutations, and two of them also carried collaborating loss of function mutations in epigenetic modifiers (*TET2*, *EZH2*).

Analysis of the three patients carrying exclusively *JAK2V617F* mutations showed marked differences in the clonal structure of the Lineage-CD34+CD38- compartment. *JAK2V617F* mutation was detected at comparable variant allele frequencies in patient IF0602 (Figure 4.1A, Table 4.1; patient with myelofibrosis receiving JAK1/2 inhibitor treatment) and patient IF0111 (Figure 4.1B, Table 4.1; patient with PV receiving IFN treatment) at the bulk level, as determined by bulk sequencing of total mononuclear cells.

However, single-cell analysis exposed remarkable differences in the clonal composition of the Lineage-CD34+CD38- compartment: patient IF0602 presented a predominant *JAK2V617F* heterozygous mutant clone with a normal immunophenotypic distribution of different stem and progenitor fractions, whereas patient IF0111 had a minor *JAK2V617F* homozygous subclone. Interestingly, *JAK2V617F* homozygous cells in patient IF0111 were restricted to the Lineage-CD34+CD38-CD90+ compartment, including an aberrant CD90+CD45RA+ population previously described in JMML (302) and high-risk MDS (303). This population was also present in an additional donor (OX4739; Figure 4.1C; Table 1) carrying a *JAK2V617F* homozygous mutation, highlighting the power of single-cell analysis to identify novel immunophenotypic populations associated with the presence of somatic mutations.

Sample code	Mutation(s) and VAF(s)	Donor type	Diagnosis	Treatment	Source	Age	Gender	PB Blasts >1%	Risk Group (DIPSS Score)
IF0602	JAK2 p.Val617Phe (25.67%)	Patient	Myelofibrosis (PMF)	Mometotinib (JAK1 and JAK2 inhibitor)	PB	68	Female	No	Interme- date-1
OX2123	JAK2 p.Val617Phe (37.41%) + EZH2 p.Glu249AsnfsTer16 (44.03%) + TET2 (96.34%)	Patient	CMML 1/ Myelofibrosis overlap	None	Bone marrow	76	Male	Yes (3%)	Interme- date-2
OX4739	JAK2 p.Val617Phe (93.49%)	Patient	Myelofibrosis (PMF)	Ruxolitinib (JAK1 and JAK2 inhibitor)	PB	65	Male	No	Interme- date-1
IF0111	JAK2 p.Val617Phe (15.77%)	Patient	PV	Pegylated IFN alpha-2a	Bone marrow	67	Male	No	N/A
SMD32316	JAK2 p.Val617Phe (22.18%) + TET2 p.Gln958Ter (7.02%)	Patient	ET	Aspirin	PB	68	Male	No	N/A
HD7650	-	Normal	-	NA	PB	44	Male	N/A	N/A
HD7643	-	Normal	-	NA	PB CD34 enriched	42	Male	N/A	N/A

Table 4.1. Summary of MPN patients used for full-length TARGET-seq analysis presented in chapter 4 section 2.1, including mutations identified in each patient and clinical characteristics. PB: peripheral blood; BM (Bone Marrow). Table adapted from Rodriguez-Meira *et al*, 2019 (1).

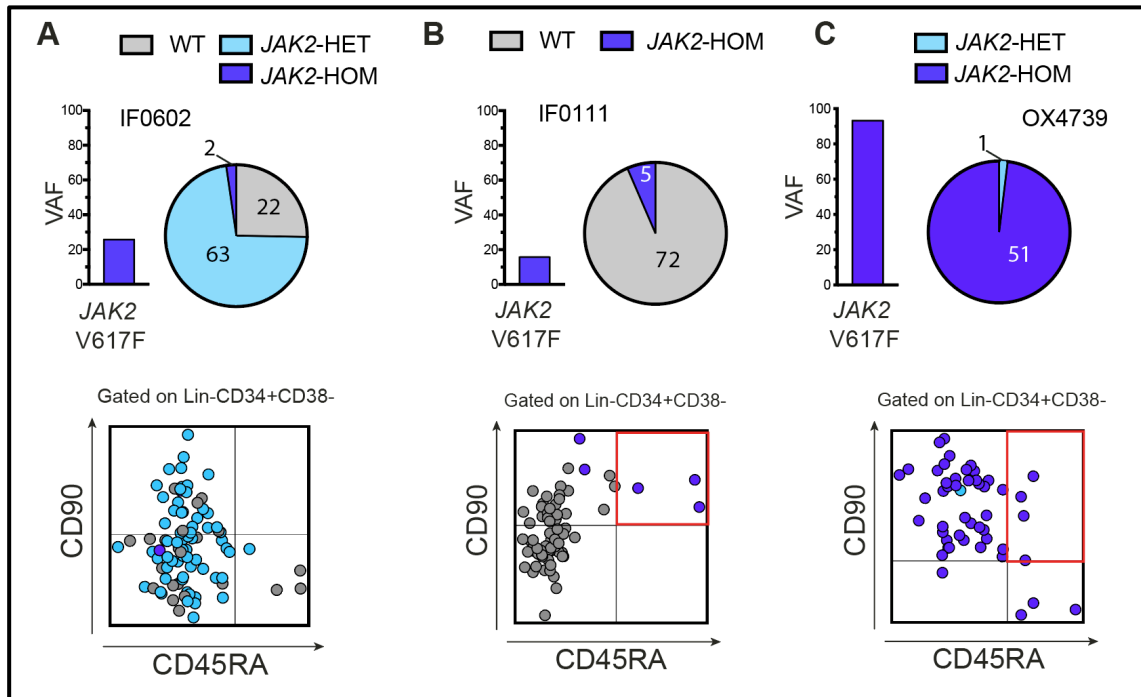


Figure 4.1. TARGET-seq reveals genetic heterogeneity in the HSPC compartment of patients with myeloproliferative neoplasms. (A-C) Variant allele frequency (VAF) of the *JAK2V617F* mutation identified by bulk sequencing of total mononuclear cells (MNCs) and number of cells that carry the mutation for each patient in the Lineage-CD34+CD38- compartment (upper panel), and integration of mutation and index sorting data (lower panel) for patient IF0602 (A), IF0111(B) and OX4739 (C). Each slice of the pie represents the number of cells from each genotype group. Patients are labelled according to the code provided in Table 4.1. Lineage-CD34+CD38-CD90+CD45RA+ subpopulation is indicated with a red box in patients IF0111 and OX4739. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

Differential expression between *JAK2V617F* mutant and wild-type cells from the same patient (Figure 4.2) is uniquely enabled by TARGET-seq. This identified dysregulation of biologically-relevant genes such as *LEPR* (304), a regulator of JAK/STAT signalling, and oncogenes such as *MYCN* (305), *TP53* (185) or *PPP2R5A* (306) in both heterozygous (Figure 4.2A) and homozygous (Figure 4.2C) mutant cells. *JAK2V617F* heterozygous mutant cells were enriched in P53, cell-cycle and hedgehog signalling associated-transcription, as well as genes upregulated in platelets ("COAGULATION" geneset, Figure 4.2.B) whereas *JAK2V617F* homozygous mutant cells were enriched in WNT β -catenin signalling, oxidative phosphorylation and genes involved in DNA repair (Figure 4.2D). This indicates that the acquisition of somatic mutations in

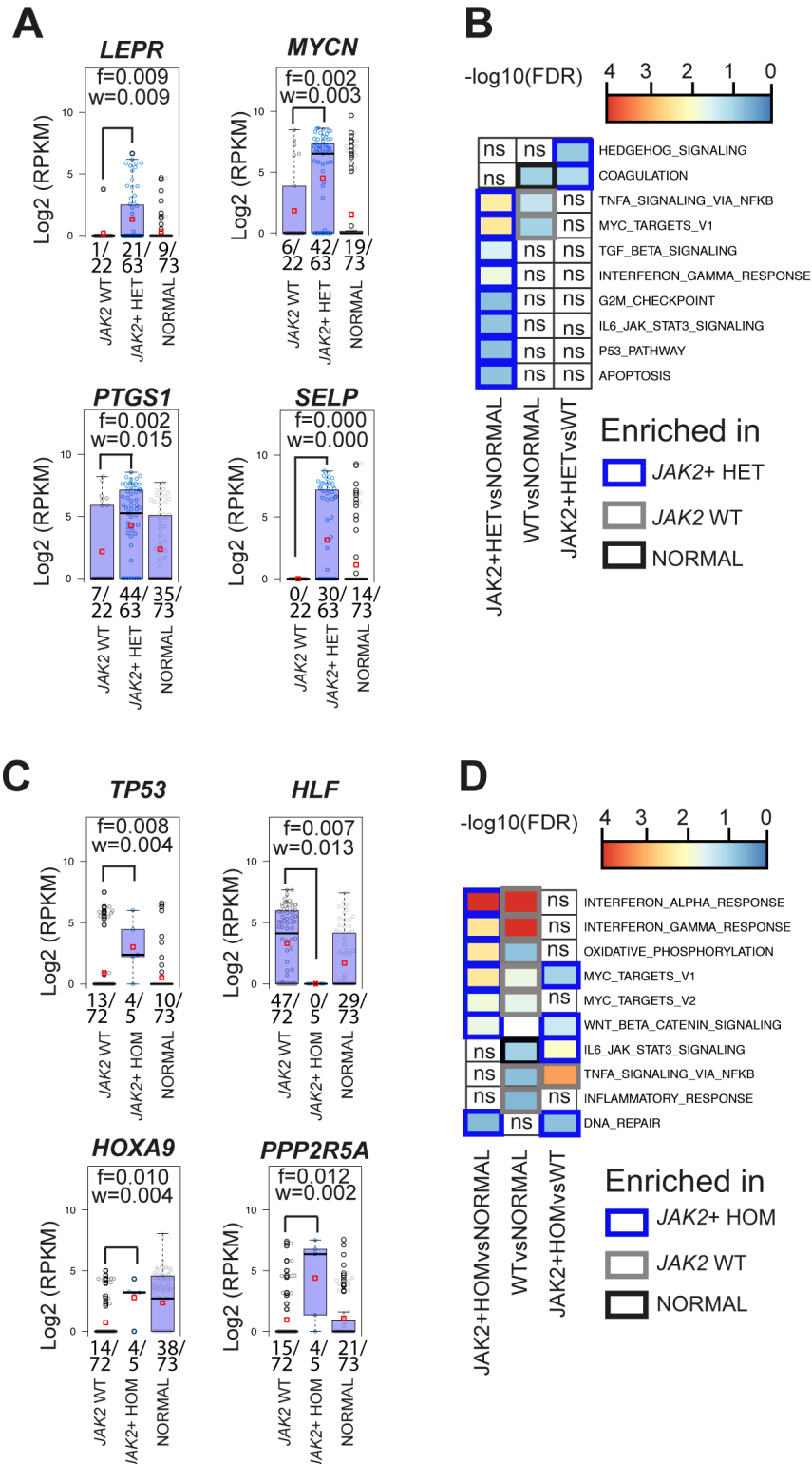


Figure 4.2. TARGET-seq reveals transcriptional heterogeneity of *JAK2V617F* mutant subclones in MPN patients. (A-B) Selected differentially expressed genes (A) and differentially enriched pathways (B) between *JAK2V617F* heterozygous mutant and WT cells from patient IF0602, as well as cells from two normal donors (HD7650, HD7643). (C-D) Selected differentially expressed genes (C) and differentially enriched pathways (D) between *JAK2V617F*

homozygous mutant and WT cells from patient IF0111, as well as cells from two normal donors (HD7650, HD7643). The p-value from Fisher's test ("fisher") and Wilcoxon ("wilcoxon") test is indicated on the top of each box for each comparison. FDR: False Discovery Rate; RPKM: Reads per Kilobase per Million; ns indicates non-significance. Each dot indicates the expression value of each single cell, red squares indicate the mean for each group and the boxes indicate median and quartiles. Frequency of expression is indicated below each box for each group. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

JAK2V617F leads to transcriptional activation of inflammatory and oncogenic signatures in HSCPs.

TARGET-seq analysis also identified genes specifically expressed in *JAK2V617F* mutant cells (Figure 4.3A) such as *RXFP1*, *GAS2* and *WDR86*. *VWF*, a marker of platelet-biased HSCs (10, 307) was specifically overexpressed in *JAK2V617F* mutant cells from patients with myelofibrosis (IF0602 and OX4739) but not in *JAK2V617F* mutant cells from patient IF0111, diagnosed with Polycythaemia Vera (Figure 4.3B). Myelofibrosis patients present increased megakaryocyte production, whereas polycythaemia vera patients usually present erythrocytosis. This observation indicates that the acquisition of *JAK2V617F* mutation in stem cells with different lineage biases might influence the disease phenotype.

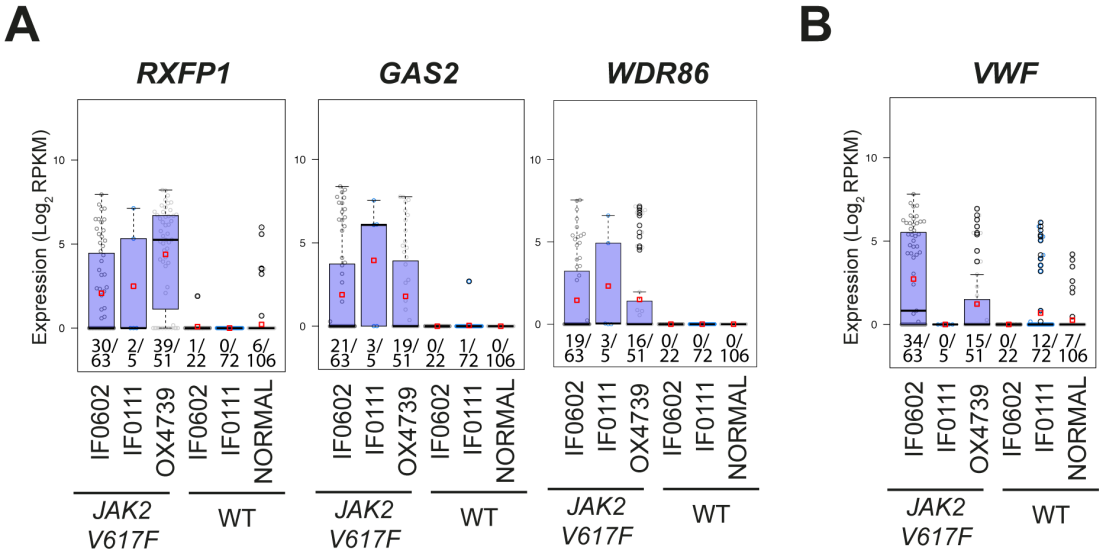


Figure 4.3. TARGET-seq identifies potential biomarkers of *JAK2V617F* mutant cells. (A-B) Gene expression values of *RXFP1*, *GAS2* and *WDR86* (A) or *VWF* (B) in *JAK2V617F* mutant or

WT cells from patient samples and normal donors. Each dot indicates the expression value of each single cell, red squares indicate the mean for each group and the boxes indicate median and quartiles. Frequency of expression is indicated below each box for each group. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

Specific analysis of global gene expression patterns in wild-type cells from patient samples and normal donors revealed markedly distinct transcriptional signatures (Figure 4.4A), potentially indicating effects of the tumour microenvironment on wild-type cells from patients or cell-intrinsic non-genetic differences in these cells. Moreover, wild-type cells from patient IF0111, undergoing treatment with pegylated IFN, showed enrichment in interferon alpha and gamma associated-transcription (Figure 4.4B), validating the relevance of the gene signatures obtained using TARGET-seq.

Random forest analysis of all genetic subclones from patients and normal donors showed specific clustering of mutant cells based on the presence of collaborating mutations (*EZH2*, *TET2*) and *JAK2V617F* zygosity (Figure 4C). Cells carrying mutations in *EZH2* and *TET2* were enriched in pathways such as apoptosis, P53 signalling, hypoxia and cell cycle, all of which had been previously associated with loss of PRC2 function (284) (Figure 4D), were depleted in genes downregulated upon *EZH2* knock-down (308) (Figure 4D). These cells also showed enrichment in HSC genes (Figure 4D), associated with *TET2* loss of function, and were depleted in genes downregulated upon *TET2* knockout (285). *JAK2V617F* mutant cells also showed expected dysregulation of JAK/STAT signalling (Figure 4E). These results indicate that TARGET-seq captures biologically relevant gene signatures in HSPCs from MPN patients, independently of their disease phenotype.

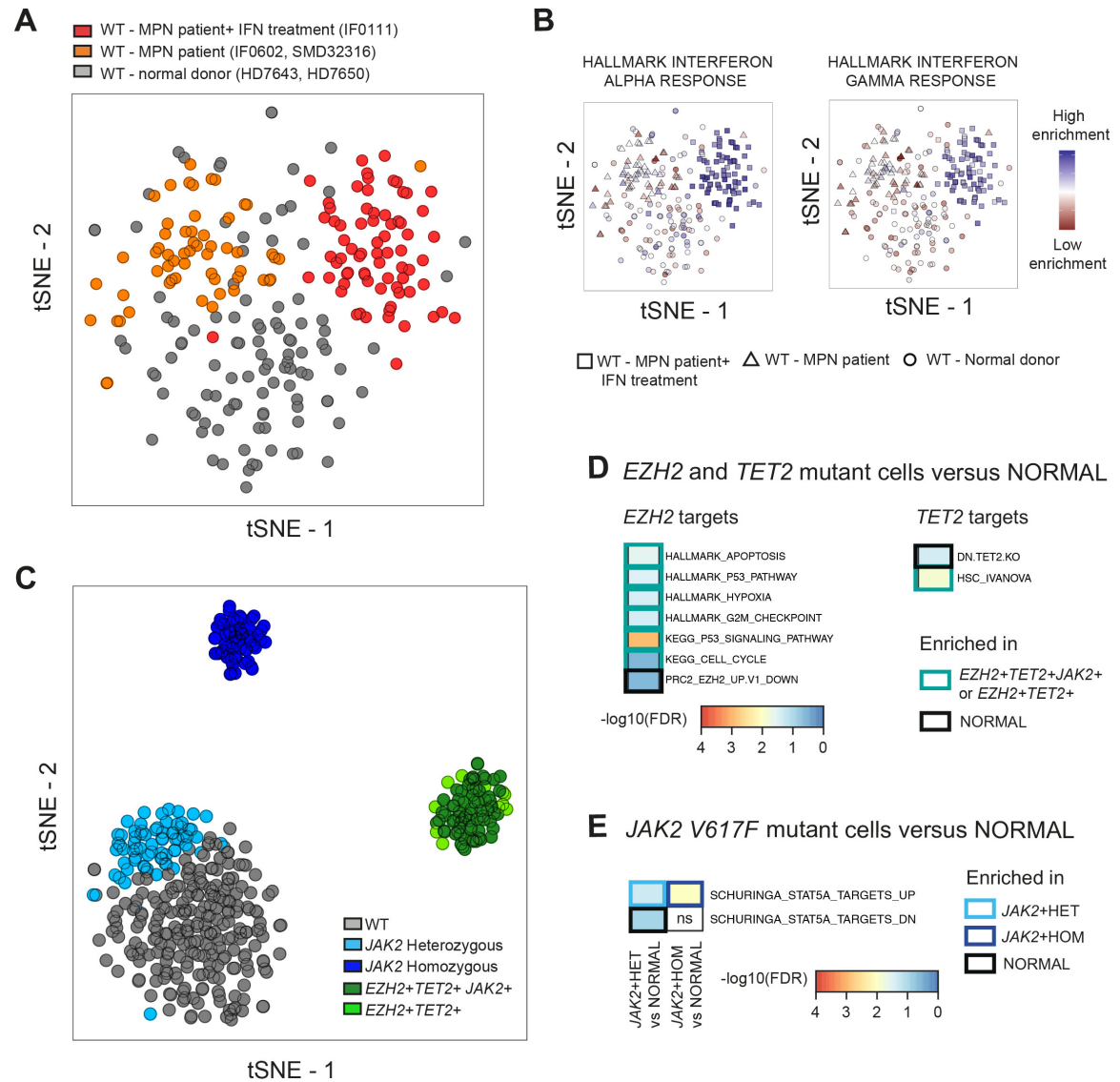


Figure 4.4. TARGET-seq reveals distinct transcriptional signatures associated with the presence of absence of somatic mutations in HSPCs from MPN patients. (A) tSNE representation of 236 wild-type cells from patient samples (n=3; IF0602, SMD32616 and IF0111) and normal donors (n=2) using 5365 highly-variable genes and perplexity=15. Wild-type cells from patient samples are coloured in orange (IF0602, SMD32616) or red (IF0111; patient been treated with pegylated interferon). Wild-type cells from patients are coloured in grey. (B) Projection of the enrichment of interferon alpha and interferon gamma associated gene signatures using the same cells and tSNE coordinates as in (A). Each shape represents a group of donors, following the same categories as in (A). (C) tSNE representation of 448 HSPCs from patients and normal donors using the top 2000 genes (ranked by the Gini index) identified from random forest analysis and perplexity=20. Cells are coloured according to their genotype; only genotypes present in at least five cells were used for the analysis. The gene expression matrix was batch and donor corrected, whilst preserving the genotype of each single cell. Figure adapted from Rodriguez-Meira *et al*, 2019 (1). (D-E) Pathway enrichment analysis of

EZH2, *TET2* (D) or JAK/STAT related (E) genesets in cells carrying mutations in these genes as compared to cells from two normal donors. The heatmap represents $-\log_{10}(\text{FDR})$ values for each comparison, using a FDR value cut-off < 0.25 . The colour of the border in each square of the heatmap represents the particular group in which a pathway is enriched. White colour with “ns” represents non-significance.

2.2. GENETIC AND TRANSCRIPTOMIC HETEROGENEITY IN *U2AF1* MUTANT MYELOPROLIFERATIVE NEOPLASMS

Full-length TARGET-seq can identify molecular signatures of different genetic subclones in *JAK2*-mutant MPN; however, it was unclear whether the transcriptional readouts obtained with TARGET-seq could identify alternative splicing patterns in single cells. Splicing is a major source of cellular heterogeneity in cancer, and approximately 9% of patients with myeloproliferative neoplasms and 30 % of patients with primary myelofibrosis carry collaborating mutations in components of the spliceosome machinery (85). Therefore, establishing whether full-length TARGET-seq can resolve heterogeneity of tumour cell populations driven by alternative splicing would be essential to apply the technique to study spliceosome mutant myeloid malignancies. Moreover, given that spliceosome components are recurrently mutated in other human cancers, this proof of principle study would be of broad relevance for many other types of tumours.

U2AF1 mutation is the most frequency somatic mutation in primary myelofibrosis, present in up to 16% of the patients (85), and it is associated with poorer overall survival (85), higher risk of transformation to secondary acute myeloid leukaemia and transfusion-dependency (309).

The mechanism through which *U2AF1* mutations contribute to neoplastic disease remains poorly understood. These mutations are usually detected at high variant allele frequencies (80), indicating that they are initiating mutations in MPN potentially contributing to clonal dominance. However, the low resolution provided by bulk

sequencing techniques does not allow to precisely determine the clonal hierarchies of *U2AF1*-mutant MPN.

Several transcriptomic studies have identified specific splicing defects associated with *U2AF1* mutations *in vitro* (88, 92) and *in vivo* (90, 91). However, no studies to date have investigated the transcriptional effect of *U2AF1* mutations in single Lineage-CD34+ cells in MPN, which would be essential to determine specific patterns of differential splicing in disease propagating cells.

Therefore, I selected six MPN patient samples to perform full-length TARGET-seq analysis and investigate the clonal hierarchy and transcriptional effects of *U2AF1* mutations in MPN.

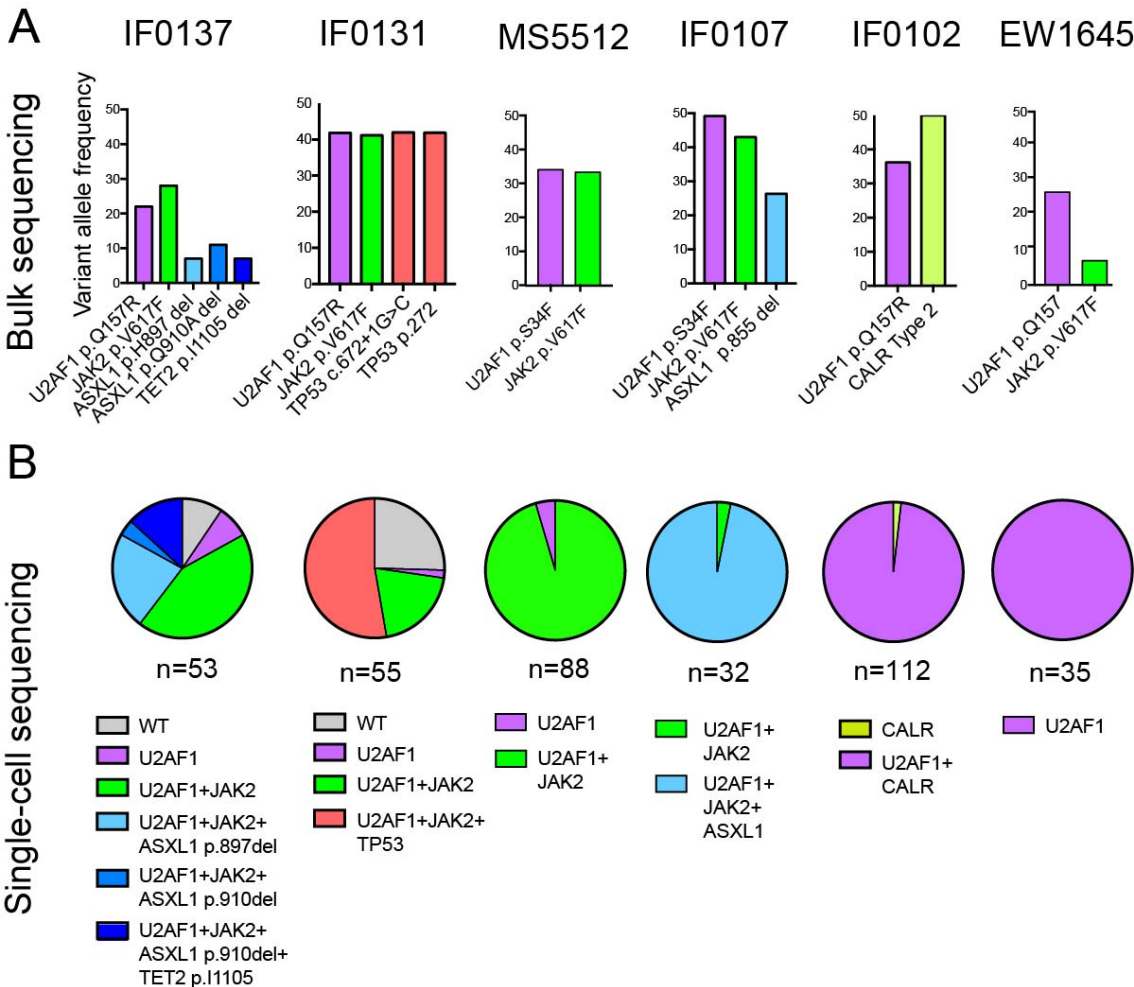


Figure 4.5. Genetic architecture of *U2AF1*-mutant myeloproliferative neoplasms. (A) Variant allele frequency of myeloid hotspot mutations detected by bulk sequencing of total mononuclear cells in six patients carrying *U2AF1* mutations. Each bar represents the variant allele frequency of each mutation. (B) Combinatorial patterns of mutations detected by TARGET-seq single-cell sequencing in the same patients as in (A). Each slice of the pie represents a different combination of somatic mutations, and is labelled according to the legend provided below each pie chart. The number of cells analysed for each patient is indicated below each graph.

To determine the clonal evolution of *U2AF1* mutant MPNs, I first performed single-cell genotyping and analysed the combinations of mutations detected in single cells to determine the clonal hierarchy of each tumour. *U2AF1* mutations were initiating genetic events in three out of the six patients analysed (Figure 4.5; IF0137, IF0131 and MS5512), which was also confirmed by higher throughput analysis of IF0137 patient (Figure 4.7). This is not readily apparent from bulk sequencing VAF; for example, *JAK2V617F* VAF is higher than *U2AF1* VAF in patient IF0137 (Figure 4.5). *U2AF1* mutations completely clonal in two of the remaining patients (Figure 4.5; IF0107, EW1645). In patient IF0102, *CALR* mutation was detected prior acquisition of *U2AF1* mutation. However, only 2 out of 112 cells profiled were *U2AF1* WT but *CALR* mutant. The sensitivity of the technique does not allow confident discrimination of subclones lower than 1 %, and therefore *U2AF1* mutation in this patient might have been acquired concomitantly with *CALR* mutation. Nevertheless, this analysis indicates that *U2AF1* mutation is usually acquired as a first genetic hit in MPN.

Then, I performed whole transcriptome analysis of single cells using full-length TARGET-seq. Full-length TARGET-seq uniquely enables to detect splicing events in single cells and correlate them to the mutational status of each cell. Four samples from *U2AF1*-mutant patients were analysed, as well as two normal donors. A patient carrying only a *JAK2V617F* mutation was also included in the analysis to account for the transcriptional effect of *JAK2V617F* mutations in this cohort, given the co-occurrence of *U2AF1* and *JAK2V617F* mutations in most of the patients (Figure 4.5).

tSNE representation using highly variable genes revealed distinct molecular signatures of *U2AF1* mutant cells as compared to cells from normal donors. Different *U2AF1* point mutations (S34 and Q157) were also associated with distinct transcriptional signatures (Figure 4.6A), suggesting that both types of mutations have different transcriptional effects. These results are in agreement of the different clinical phenotypes associated with *U2AF1* S34 and Q157 mutations in MPN (310). tSNE representation of PSI (Percent Splice In) values from all the alternative splicing events detected in the dataset also showed specific clustering based on the genotype of each single cell, and different splicing patterns in *U2AF1* S34 and Q157 mutant cells (Figure 4.6B). These results suggest that *U2AF1* mutations disrupt transcription both at the gene and isoform level in Lineage-CD34+ cells in a variant-specific manner.

Differentially spliced events between *U2AF1* mutant and normal donor cells were enriched in exon skipping and alternative 3'-splice sites (Figure 4.6C, Figure 4.6D), in keeping with the previously described transcriptional effect of *U2AF1* mutations (92) and *U2AF1*'s role on 3'-splice site recognition (87). Differential pathway analysis (Figure 4.6E) consistently identified antigen processing and presentation genesets (indicated in green) enriched in *U2AF1* mutant cells, which is of interest as *U2AF1* mutations might lead to the generation of novel peptides through aberrant splicing, which can be presented on the surface of cancer cells. The identification of aberrant splicing patterns generating neoantigens could be a promising therapeutic strategy for *U2AF1* mutant leukaemias.

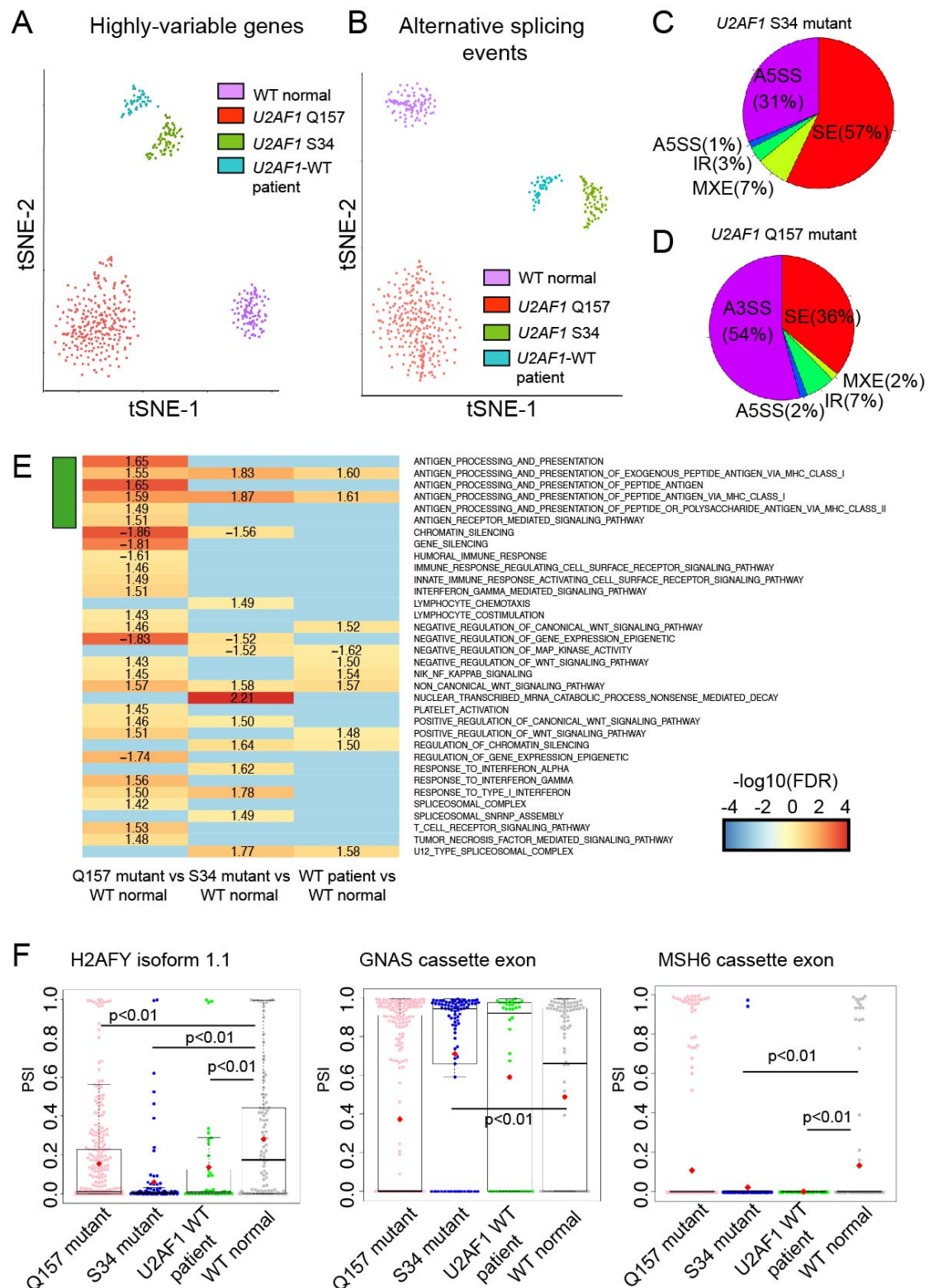


Figure 4.6. Transcriptional effects of *U2AF1* mutations in MPN patients. (A) tSNE representation of 513 cells using 8784 highly variable genes. Cells are coloured according to their mutational status and donor. Only *U2AF1* mutational status was considered for this analysis. Counts matrix was batch and donor corrected, and the genotype of each single cell was preserved. (B) tSNE representation of the same cells as in (A) using 23320 alternative

splicing events identified in the entire dataset. PSI matrix was batch and donor corrected, and the genotype of each single cell was preserved. (C-D) Frequency of alternative splicing events differentially enriched in S34 (C) and Q157 (D) *U2AF1* mutant cells compared to cells from normal donors. Each slice of the pie represents the frequency of each alternative splicing event (SE: skipping exon; 3ASS: 3' alternative splice site; 5AAS: 5AAS alternative splice site; IR: intron retention; MXE: mutually-exclusive exons). (E) Pathway enrichment analysis of *U2AF1* Q157 mutant, *U2AF1* S34 mutant, WT cells from patient and WT cells from normal donors. Heatmap is coloured according to the $-\log_{10}(\text{FDR})$ value for each comparison. Blue colour indicates non-significance, using a FDR value threshold of <0.25 . The numbers in each square of the heatmap indicate the Net Enrichment Score. (F) Percent Spliced In (PSI) distributions of alternatively spliced exons in *H2AFY*, *GNAS* and *MSH6* genes stratified by *U2AF1* genotype. Boxplot represents median and quartiles; the red point represents the mean value for each group and each point represents the PSI value for each single cell. The p-value ("p") for each pairwise comparison is shown on the top of each bar. Non-significant comparisons are not indicated. Bioinformatic analysis presented in this figure was performed with Wei Wen, PhD Student in the Mead and Thongjuea Laboratories.

Differential splicing analysis also identified cassette exons specifically included or excluded in *U2AF1* mutant cells, some of which had been previously identified in bulk sequencing studies (*H2AFY*, *GNAS*) (90, 92), validating the results of this analysis (Figure 4.6F) and further highlighting the differences in splicing patterns between *U2AF1* Q157 and S34 mutant cells. Moreover, novel splicing events in genes such as *MSH6* (Figure 4.6F), implicated in DNA repair, could also be identified.

These proof-of-concept results demonstrate that TARGET-seq analysis can identify differentially spliced events in single HSPCs associated with the acquisition of spliceosome mutations in MPN patients. Analysis of higher numbers of cells and donors would likely provide crucial insights into the mechanisms of clonal expansion driven by mutations in *U2AF1*.

2.3. HIGH-THROUGHPUT TARGET-SEQ ANALYSIS OF HSPCs IN *JAK2*-MUTANT MYELOFIBROSIS

Full-length TARGET-seq analysis of HSPCs from patients with myeloproliferative neoplasms revealed molecular signatures of distinct genetic subclones, highlighting the power of the technique to identify mutation-specific transcriptional signatures independently of disease phenotypes. However, the high degree of genetic and transcriptional heterogeneity identified in HSPCs from MPN patients in the previous sections did not provide enough statistical power to identify phenotype-specific gene signatures or complex combinational patterns of mutations.

Therefore, I then decided to apply 3'-biased high-throughput TARGET-seq to analyse HSPCs from a cohort of patients with defined genetic lesions and disease phenotype. Myelofibrosis is the MPN subtype associated with poorer outcomes and higher risk of transformation to secondary acute myeloid leukaemia (77); however, the molecular mechanisms driving disease progression are poorly understood.

I selected eight patients with myelofibrosis (primary myelofibrosis or secondary myelofibrosis to ET or PV) carrying *JAK2V617F* mutations, as well as two aged-matched normal donors (Table 4.2) for TARGET-seq analysis. Given the higher clonal complexity of patients with myelofibrosis (64), I decided to use the higher throughput, 3'-biased version of the TARGET-seq method described in Chapter 3.

Sample code	Mutation(s) and VAF(s)	Donor type	Diagnosis	Treatment	Source	Age	Gender	PB Blasts >1%	Risk Group (DIPSS Score)
IF0602	JAK2 p.Val617Phe (25.67%)	Patient	Myelofibrosis (PMF)	Momelotinib	PB	68	Female	No	Intermediate-1
IF0137	JAK2 p.Val617Phe (47.26%), U2AF1 p.Gln157Arg (45.8%), TET2 p.Ile1105MetfsTer8 (14.43%), ASXL1 p.Gln910AlafsTer13 (16.87%), ASXL1 p.Trp898ArgfsTer5 (2.53%)	Patient	Myelofibrosis (PMF)	None	PB	59	Female	Yes (3%)	Intermediate-1
IF0138	JAK2 p.Val617Phe (63.91%), ASXL1 p.Gly646TrpfsTer12 (19.89%), ASXL1 p.Gly644TrpfsTer12 (2.5%)	Patient	Myelofibrosis (post-PV)	Hydroxycarbamide	PB	61	Male	No	Intermediate-1
IF0123	JAK2 p.Val617Phe (34.87%), SF3B1p.Lys666Asn (16.69%)	Patient	Myelofibrosis (PMF)	Ruxolitinib 5mg BD	PB	86	Male	Yes (3%)	High
IF0101	JAK2 p.Val617Phe (87.2%), CBL p.Cys404Tyr (22.5%), SRSF2 p.Pro95His (47.9%)	Patient	Myelofibrosis (PMF)	Ruxolitinib 10mg BD	PB	71	Male	Yes (<5%)	Intermediate-2
IF0140	JAK2 p.Val617Phe (93.0%), TET2 p.Ser1612LeufsTer4 (47.0%)	Patient	Myelofibrosis (post-PV)	Ruxolitinib 20mg BD	PB	73	Female	Yes (4%)	Intermediate-2
IF0157	JAK2 p.Val617Phe (28.34%)	Patient	Myelofibrosis (post-PV)	Ruxolitinib 10mg BD	PB	68	Female	No	Intermediate-1
IF0155	JAK2 p.Val617Phe (79.63%)	Patient	Myelofibrosis (post-ET)	Anagrelide	PB	59	Female	No	Intermediate-1
HD85	-	Normal	-	NA	PB	69	Male	N/A	N/A
Aph1	-	Normal	-	NA	PB	63	Female	N/A	N/A

Table 4.2. Summary of patients with myelofibrosis analysed in section 2.3, including mutations detected by bulk sequencing for each patient and clinical characteristics. PB: peripheral blood; BD: twice a day; VAF: Variant Allele Frequency. Table adapted from Rodriguez-Meira *et al*, 2019 (1).

Single-cell genotyping of these patients using TARGET-seq revealed linear and branching patterns of clonal evolution (Figure 4.7), with up to seven different subclones identified per patient (Figure 4.7; patient IF0137). Mutations in spliceosome components (*U2AF1*, *SF3B1*, *SRSF2*) were acquired early, whereas mutations in epigenetic modifiers (*TET2*, *ASXL1*) or other signalling components (*CBL*) were acquired late (Figure 4.7). *JAK2V617F* heterozygous clones frequently underwent loss of heterozygosity, giving rise to *JAK2V617F* homozygous clones (Figure 4.7). Interestingly, two different loss-of-function *ASXL1* mutations were independently acquired in the same tumour (Figure 4.7), suggesting a strong effect of *ASXL1* loss of function driving clonal evolution in this patient.

Importantly, this analysis identified clonal hierarchies with over 95% resolution even in patients with complex clonal hierarchies (Figure 4.8; patient IF0137), which could not have been determined using bulk sequencing approaches or mRNA-based mutational analysis alone (Figure 4.8).

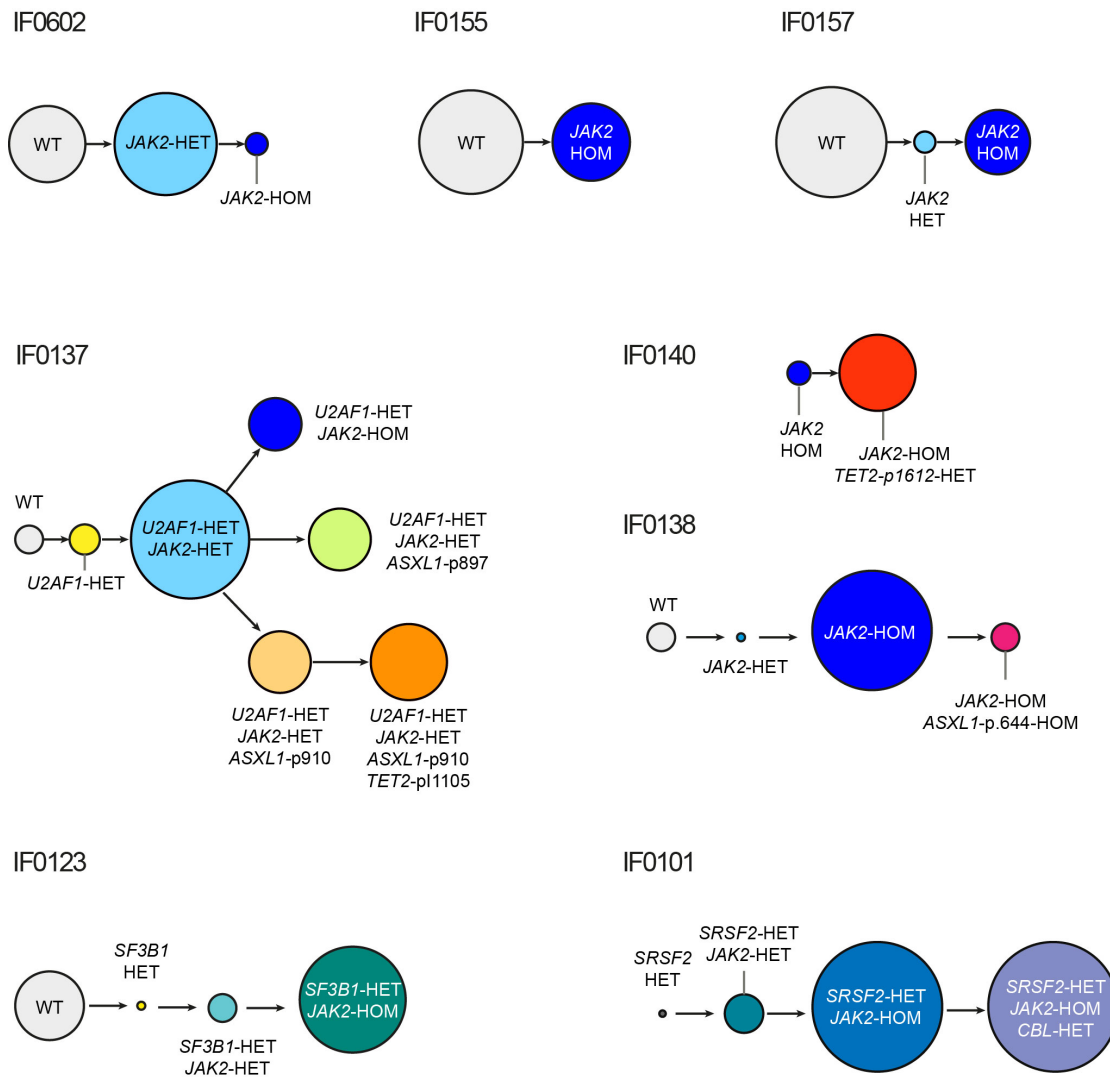


Figure 4.7. Genetic subclones and patterns of clonal evolution from patients with myelofibrosis, identified using TARGET-seq single-cell genotyping. Each circle represents a different genetic subclone and arrows indicate the evolution of each tumour; the size of the circles is proportional to the number of cells from each genetic subclone. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

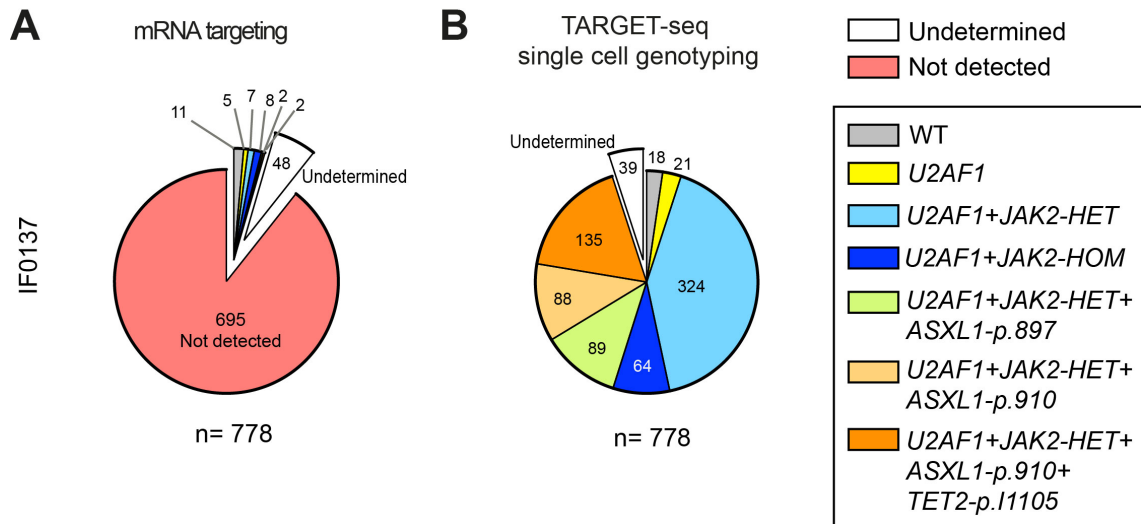


Figure 4.8. Genetic subclones identified in patient IF0137 using mRNA targeting or TARGET-seq methods. Each slice of the pie represents the number of cells carrying each combination of mutations using each method, which are coloured according to the legend on the right; the total number of cells analysed is indicated below each pie. Undetermined indicates cells which could not be assigned to the most likely clonal hierarchy computationally determined using SCITE algorithm (226), and “not detected” indicates cells in which at least one of the mutations targeted was not detected. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

Whole transcriptome analysis of 2734 HSPCs from myelofibrosis patients and normal donors showed distinct clustering according to the genotype of each single cell (Figure 4.9) and donor type. Cells from normal donors (Figure 4.9A; labelled in light grey) formed a discrete cluster as compared to patient-derived cells (Figure 4.9A). HSPCs carrying exclusively *JAK2V617F* mutations clustered with WT cells from patient samples, whereas cells carrying cooperating mutations in spliceosome components or epigenetic modifiers showed highly-distinct transcriptional signatures (Figure 4.9A). Differential expression analysis identified genes specifically upregulated in each of these genetic subclones (*B4GALT1*, *PITX1*, *ZFP36*, *PHB*; Figure 4.9B).

Cells from patient samples showed higher transcriptional diversity than cells from normal donors (Figure 4.9C), including within different genotype subgroups (Figure 4.9D). Interesting, wild-type cells from patients also showed increased transcriptional diversity (Figure 4.9D), indicating that transcriptional heterogeneity is not driven by genetic mechanisms alone.

Acquisition of somatic mutations in HSPCs was associated with enrichment in angiogenesis, hypoxia and *TP53*-associated transcription, inflammatory signatures and JAK/STAT signalling (Figure 4.9E). Mutant cells were depleted in oxidative-phosphorylation, fatty acid, heme and glucose metabolism, protein secretion and DNA repair, indicating defects in the metabolic activity of mutant MPN cells, concomitantly with aberrant inflammation-associated gene expression.

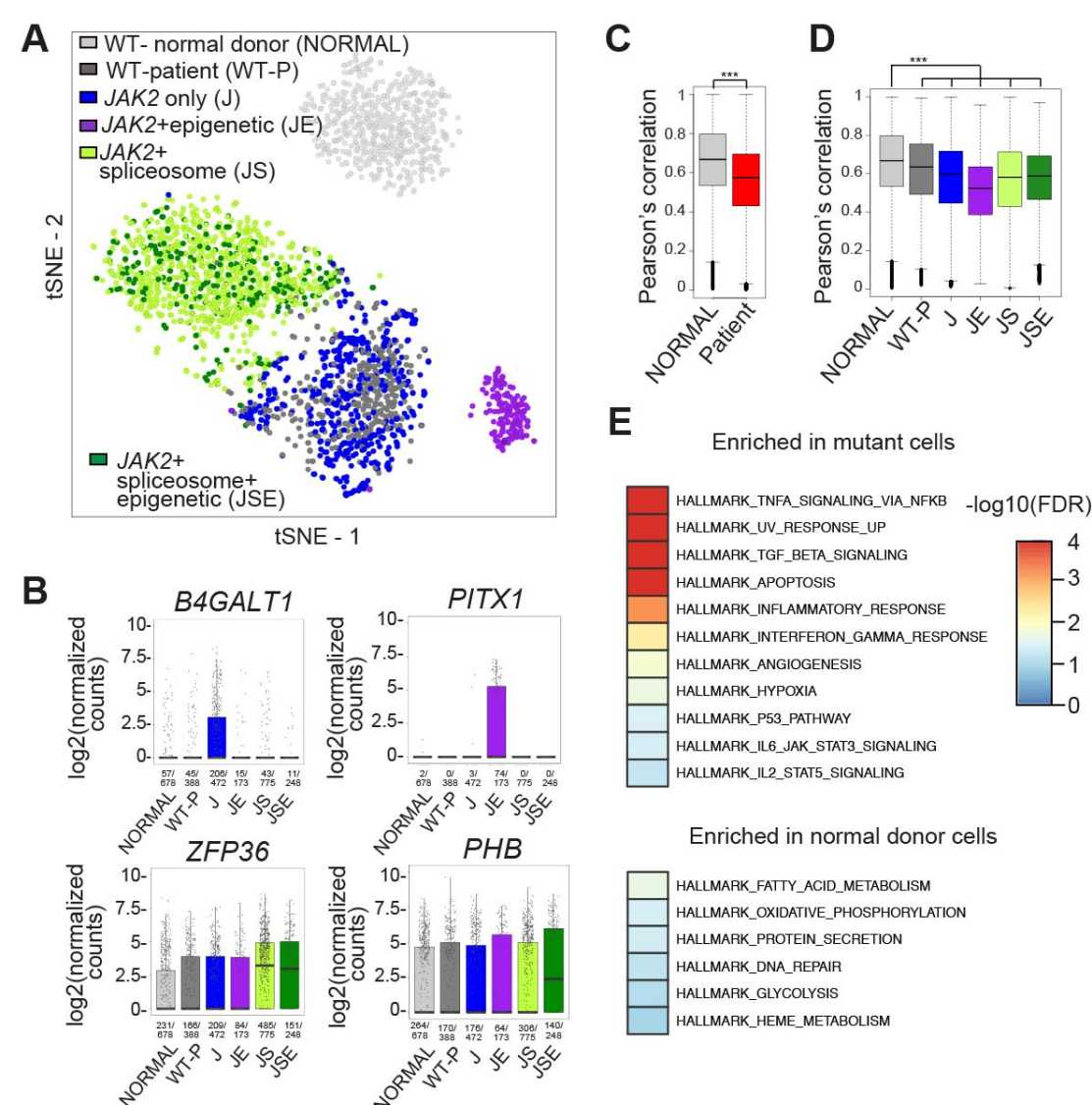


Figure 4.9. High-throughput 3'-biased TARGET-seq reveals molecular signatures of HSPCs from myelofibrosis patients. (A) tSNE representation of 2734 cells from myelofibrosis patients (n=8; Table 4.2) and two aged-matched normal donors (n=2), using 3286 highly variable genes and perplexity=30. Cells from normal donors ("NORMAL") are coloured in light grey; WT cells from patients are coloured in dark grey ("WT-P"); cells carrying exclusively mutations in *JAK2*

(“J”) are coloured in blue; cells carrying mutations in epigenetic modifiers (“JE”; *ASXL1*, *TET2*) are coloured in purple; cells carrying mutations in spliceosome components (“JS”; *U2AF1*, *SF3B1* and *SRSF2*) are coloured in light green and cells carrying mutations in spliceosome components and epigenetic modifiers (“JSE”) are coloured in dark green. (B) Boxplots of selected differentially expressed genes between different genetic subclones. Boxplots represent median and quartiles; each dot represents the gene expression value for each cell and expression frequencies are shown below each bar. (C-D) Boxplot of overall Pearson’s correlation between the transcriptomes of cells from normal donors (“NORMAL”) and myelofibrosis patients (“patients”) (C) and cells from normal donors and cells from patients divided per genotype group (D). Boxplots represent median and quartiles, and points represent outlier values (outside of 1.5 times the interquartile range). Kolmogorov-Smirnov test was used to assess statistical significance (***; p-value<0.001). (E) Selected differentially enriched pathways between mutant cells from patients and WT cells from normal donors. Heatmap is coloured according to $-\log_{10}(\text{FDR})$ value. A cut-off of $\text{FDR}<0.25$ was used. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

To identify transcriptomic differences driven by differentiation trajectories rather than mutational status, I performed a completely unbiased analysis in which the genotype of each single cell was not accounted for. Cells were ordered in gene expression space using forced directed graphs (Figure 4.10) and k-means clustering and lineage specific signatures were used to identify molecularly distinct subpopulations of cells.

K-means clustering identified eight distinct HSPC clusters (Figure 4.10A). Superimposition of haematopoietic lineage signatures on the force-directed graph layout identified cluster 3 as a myeloid cluster and cluster 8, as a megaerythroid cluster (Figure 4.10B). Mutant cells were enriched in myeloid and megaerythroid clusters, in keeping with the leucocytosis characteristic of MPN patients and the dramatic increase in megakaryocytes. Normal donor cells were depleted in cluster 3 and 8, indicating that HSPCs from MPN patients present transcriptional lineage biases not present in normal donor cells. Interestingly, WT cells from patients were also enriched in cluster 3, indicating that these lineage biases might not be exclusively mediated by the acquisition of somatic mutations.

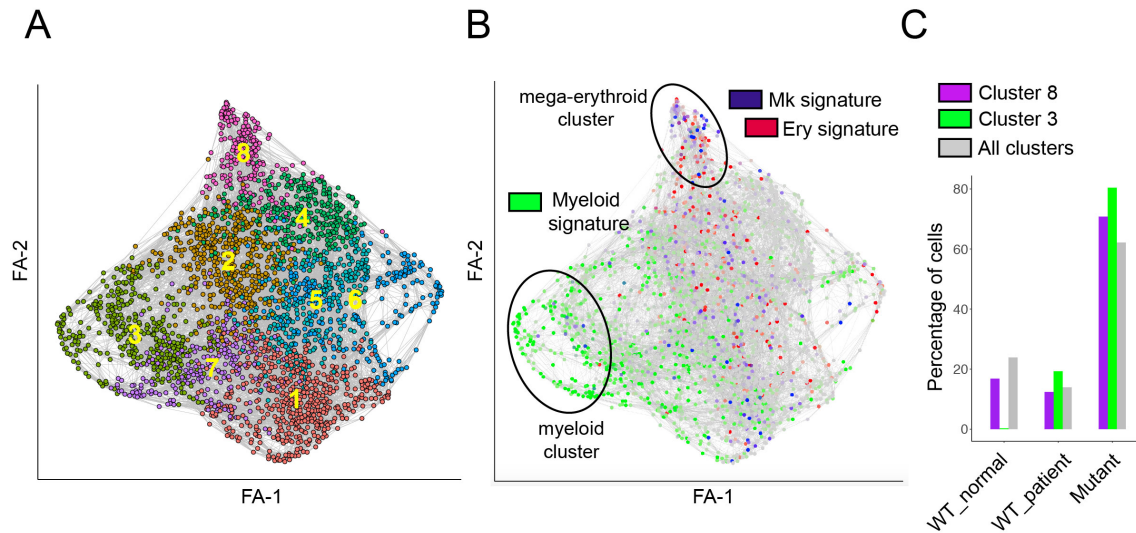


Figure 4.10. Lineage differentiation trajectories in HSCPs from myelofibrosis patients and normal donors. (A) Force-directed graph representation of 2734 cells from myelofibrosis patients (n=8) and normal donors (n=2). Cells are coloured according to each cluster identified by k-means clustering method, and the number assigned to each cluster is indicated by the yellow numbers displayed on the graph. (B) Enrichment of myeloid, erythroid and megakaryocyte gene signatures in the same cells and graph coordinates as in (A). (C) Percentage of wild-type cells from normal donors, wild-type cells from patients and mutant cells in myeloid (cluster 3) and megaerythroid (cluster 8) clusters identified in (A-B) as compared to the percentage of cells from each genotype in all clusters. Gene expression matrix was batch corrected.

To further characterize transcriptional effects not driven by genetic defects in single cells, mutant cells were then excluded from the analysis. tSNE analysis of wild-type cells from patients and normal donors revealed separate clustering based on each type of donor (Figure 4.11A), similarly to the observation made using full-length TARGET-seq. WT cells from patients with myelofibrosis were enriched in inflammatory signatures (Figure 4.11B) and genes associated with inflammation (*DDIT3*, *IRF1*, *NFKBIA*; Figure 4.11C), potentially indicating cell extrinsic effects mediated by the tumour microenvironment or by factors secreted by mutant cells. Importantly, the inflammatory signatures identified in WT cells have predictive prognostic value in CML (247), suggesting that the identification of inflammatory pathways in non-mutant cells could help stratify patients with MPN. These molecular signatures can only be derived

from the combined mutational and transcriptomic readouts made possible through TARGET-seq.

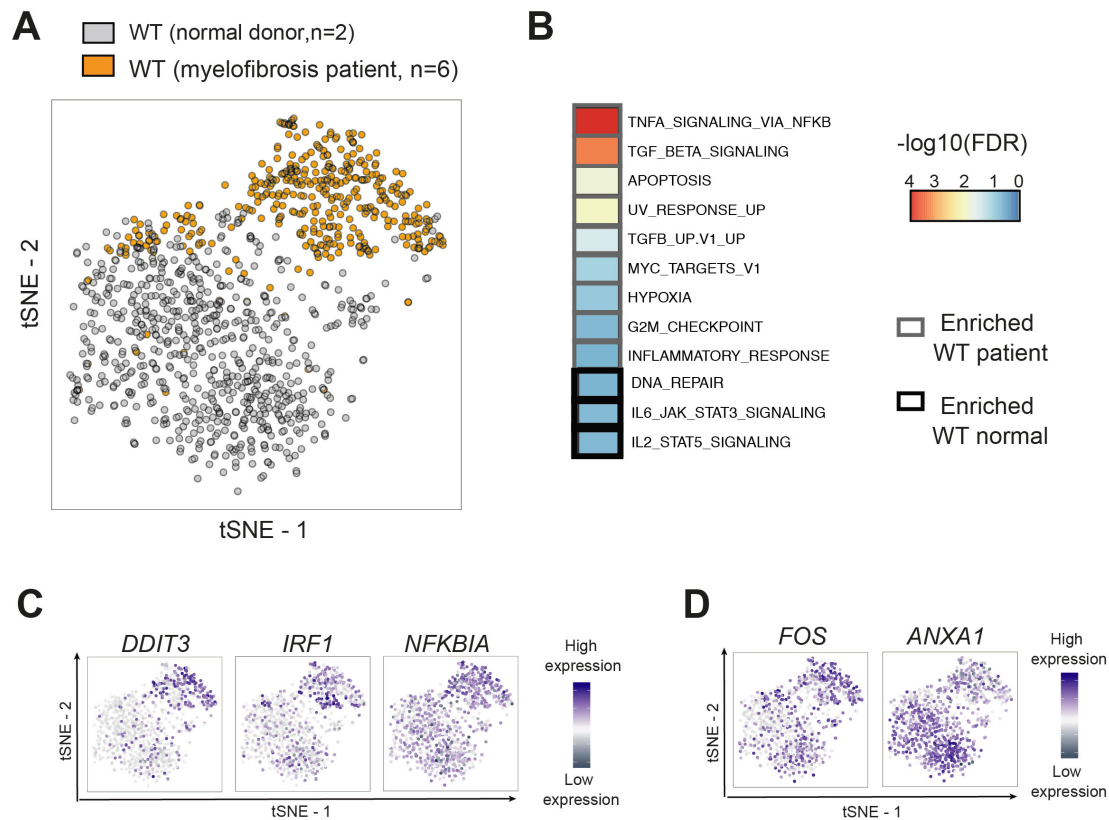


Figure 4.11. Wild-type cells from patients and normal donors are transcriptionally different. (A) tSNE representation of 1066 wild-type cells from myelofibrosis patients (n=4; IF0138, IF0602, IF0155, IF0157; Table 4.2) and aged-matched normal donors (n=2; Aph1 and HD85; Table 4.2) using 3436 highly-variable genes and perplexity=30. (B) Differential pathway analysis between wild-type cells from patients and aged-matched normal donors using the same cells as in (A). Signatures enriched in WT cells from patients are indicated by grey squares and those enriched in WT cells from normal donors, by black squares. Heatmap is coloured according to the -log₁₀(FDR) value for each comparison. Only significant genesets (FDR<0.25) are shown. (C-D) Projection of the normalized gene expression values for genes involved in inflammation (C; *DDIT3*, *IRF1*, *NFKBIA*) or genes involved in oncogenesis and apoptosis (D; *FOS*, *ANXA1*) using the same cells and tSNE coordinates as in (A). Adapted from Rodriguez-Meira *et al*, 2019 (1).

WT HSCPs from patients specifically upregulated oncogenes (*FOS*; Figure 4.11D) and were enriched in MYC targets (Figure 4.11B), and they were depleted in genes mediating apoptosis (*ANXA1*; Figure 4.11D) as well as DNA-repair associated transcription (Figure 4.11B). These raises the possibility that WT HSCPs from patients might be more susceptible to malignant transformation and progression to secondary

acute myeloid leukaemia, which is illustrated by the development of *JAK2*-negative haematopoietic malignancies following a *JAK2*-positive MPN (163).

To study the effect of *JAK2V617F* mutations independently of other collaborating mutations in myelofibrosis, I selected clones carrying exclusively *JAK2V617F* mutations. Random forest analysis revealed marked transcriptomic differences between *JAK2V617F* heterozygous and homozygous subclones, which were also different to cells from normal donors (Figure 4.12A). Differential pathway analysis identified consistently dysregulated pathways in *JAK2V617F* mutant cells such as STAT5A target genes and STAT3/STAT5 signalling (Figure 4.12B), in keeping with the aberrant JAK/STAT signalling activation in MPN.

Interesting, despite the majority of patients receiving JAK2 inhibitor treatment (Table 4.2), *JAK2V617F* cells still showed transcriptional evidence of JAK/STAT signalling pathway hyperactivation, suggesting that the effect of JAK2 inhibitor treatment is limited to a certain extent in Lineage-CD34⁺ mutant HSPCs. This might partly explain why *JAK2* inhibitor treatment does not induce clonal remissions in myelofibrosis patients.

Inflammatory pathways, IFN- α and IFN- γ signalling, TNF- α and TGF- β signalling were also enriched in *JAK2V617F* independently of zygosity status (Figure 4.12), indicating a strong inflammatory signature in mutant cells, in keeping with the elevated levels of proinflammatory cytokines characteristic of MPN patients (311). Interestingly, *JAK2V617F* homozygous cells were specifically enriched in WNT β -catenin and hedgehog signalling, but not *JAK2V617F* heterozygous cells (Figure 4.12B). WNT- β catenin regulates stem cell maintenance, is implicated in bone marrow fibrosis, and pharmacological inhibition of β -catenin targets therapy-resistant leukemic stem cells in CML (312). Hedgehog inhibitors in combination with JAK2 inhibitors have shown promising results for high-risk myelofibrosis patients (313). The differential enrichment

of these pathways in *JAK2V617F* homozygous cells indicates that *JAK2V617F* zygosity status could identify patients who might benefit from targeted therapies specifically inhibiting these pathways.

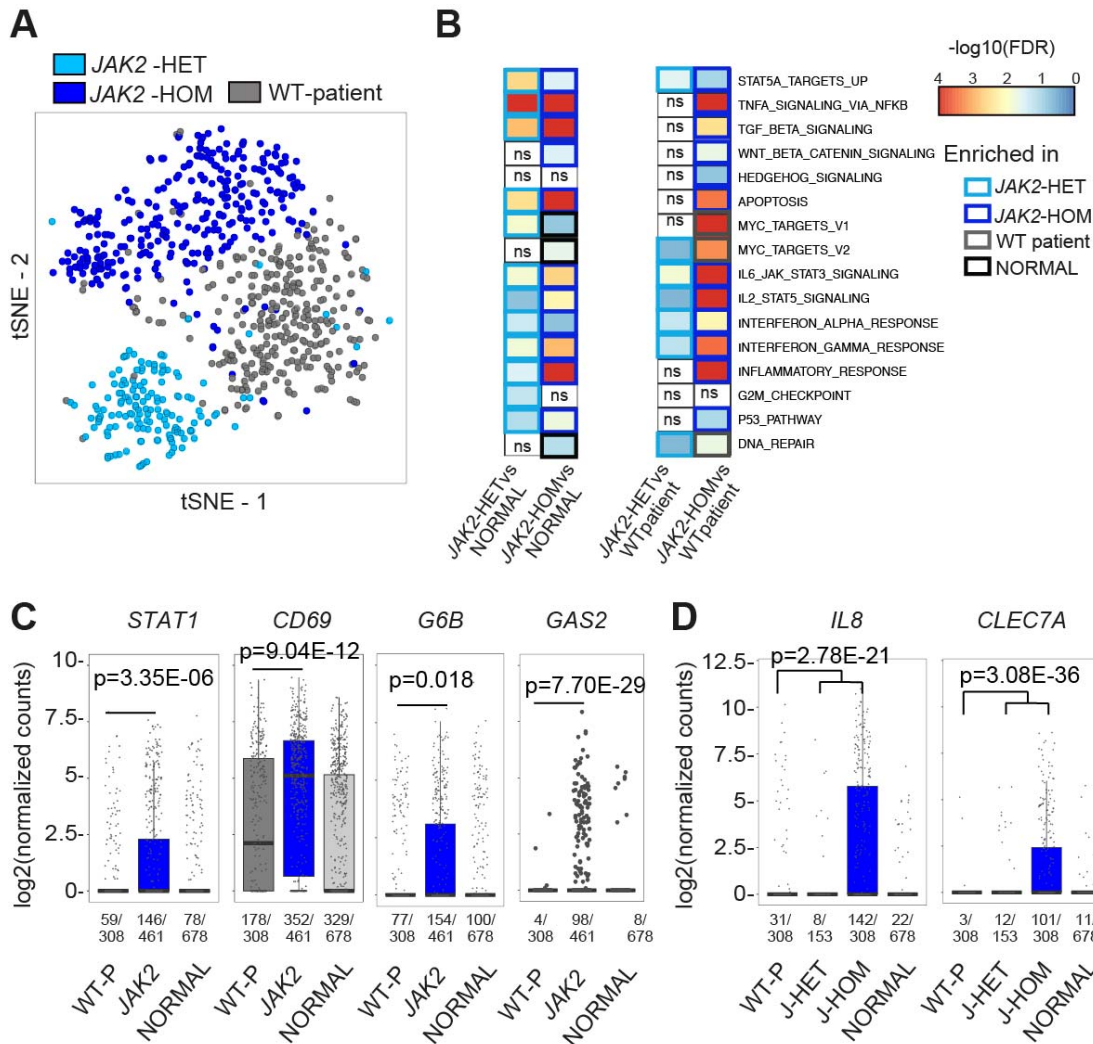


Figure 4.12. *JAK2V617F* heterozygous and homozygous subclones present distinct molecular signatures. (A) tSNE representation of 769 cells from four patient samples (IF0602, IF0155 IF0157 and IF0138) using the top 2000 genes selected from a random forest classification trained on the genotype of each single cell and perplexity=20. (B) Gene set enrichment analysis of differentially expressed pathways between *JAK2V617F* heterozygous and homozygous clones, compared to cells from normal donor (left panel) and WT cells from patient samples (right panel). Heatmaps are coloured according to the $-\log_{10}(\text{FDR})$ value for each comparison. Only significant genesets ($\text{FDR} < 0.25$) are shown. The borders of each square of the heatmap indicate the genotype group in which a particular pathway is enriched. “ns” represents non-significance. (C-D) Expression of selected genes differentially expressed in *JAK2V617F* mutant (“JAK2”) and WT cells from patient samples (“WT-P”) (C) or *JAK2V617F* homozygous mutant (“J-HOM”) cells but not *JAK2V617F* heterozygous mutant (“J-HET”) cells. (D). Gene expression

values from normal donors ("NORMAL") are also shown for comparison. Adjusted p-value from combined Fisher's and Wilcoxon test is shown on the top of each comparison. Each dot indicates the gene expression value for each single cell; bars indicate median and quartiles and expression frequencies are shown below each bar for each genotype group. Adapted from Rodriguez-Meira *et al*, 2019 (1).

Differential expression analysis between *JAK2V617F* mutant and WT cells from patients identified genes specifically upregulated in *JAK2V617F* subclones such as *STAT1*, *CD69*, *G6B* and *GAS2* (Figure 4.12C). *STAT1* is a IFN signalling effector and promotes megakaryocytic differentiation (314), in keeping with the aberrant megakaryopoiesis characteristic of myelofibrosis patients. *CD69* is a marker of *BCR-ABL1* tyrosine kinase activity in CML (315), and might serve as a biomarker of aberrant JAK/STAT activation in myelofibrosis. *G6B* (also known as *MPIG6B*; Megakaryocyte and Platelet Inhibitory Receptor G6b) is a critical regulator of haematopoietic differentiation as well as megakaryocyte and platelet function (316, 317). *G6B* deficient mice show reduced megakaryocyte and platelet production, suggesting that *G6B* overexpression in mutant cells could lead to the increase in megakaryocytes and thrombocytosis commonly observed in myelofibrosis (316). Besides, *CD69* and *G6B* are cell surface markers that could be used to prospectively isolate mutant cells using flow cytometry, and could be potentially implemented as a diagnostic and prognostic tool in the clinical setting to monitor the presence of mutant cells.

Moreover, certain genes such as *IL8* and *CLEC7A* (Figure 4.12D) were specifically upregulated in *JAK2V617F* homozygous but not *JAK2V617F* heterozygous cells. Interestingly, high *IL8* levels in the serum of primary myelofibrosis patients are associated with inferior overall survival and constitutional symptoms, and have predictive prognostic value (318).

These results highlight the differences at the molecular level between *JAK2V617F* heterozygous and homozygous mutant clones. These distinct molecular signatures might partly explain the different distribution of *JAK2V617F* homozygous and

heterozygous mutations in different disease subtype (*JAK2V617F* homozygous are more frequently found in PV and *JAK2V617F* heterozygous cells, in ET).

I then analysed different genetic subclones within individual patients with complex clonal hierarchies (at least 3 genetic subclones, patients IF0137, IF0138 and IF0101; Figure 4.13 and Table 4.2) to try to identify molecular signatures associated with clonal evolution in these patients. Dimensionality reduction using differentially expressed genes in between subclones identified distinct transcriptional signatures associated with each combination of somatic mutations. These transcriptional differences were driven by pro-apoptotic genes (*MCL1*; Figure 4.13B), genes implicated in JAK/STAT signalling (*STAT2*; Figure 4.13D), chemokines (*CXCL2*; Figure 4.13D), genes previously implicated in leukemogenesis (*PHB*, *STAG2*; Figure 4.13B, 4.13F), stem cell maintenance (*BCL11A*; Figure 4.13F) (319) or drug resistance (*GSTK1*; Figure 4.13F).

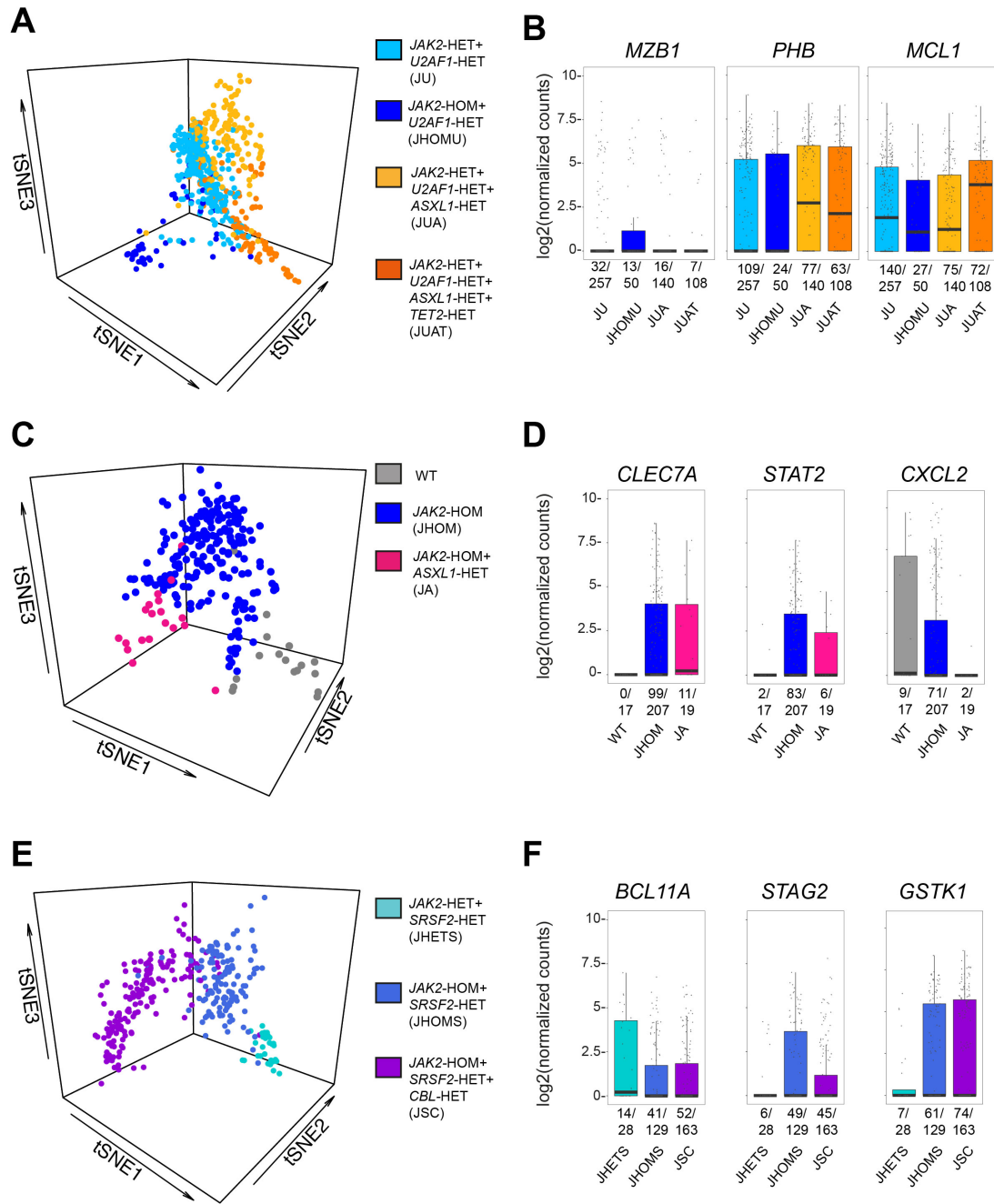


Figure 4.13. Genetic subclones within individual patients present distinct molecular signatures. (A-B) Transcriptional signatures of genetic subclones in patient IF0137. tSNE representation of 555 cells using 633 differentially expressed genes identified using ANOVA (A) and selected differentially expressed genes in each genetic subclone (B). Only genetic subclones with at least 20 cells were used for analysis. The two subclones carrying *JAKV617F* heterozygous mutation, *U2AF1* mutation and *ASXL1* p.897 or p.910 mutation are labelled as “JAK2-HET+U2AF1-HET+ASXL1-HET” and were analysed together, due to the same predicted functional consequences of each mutation. (C-D) Transcriptional signatures of genetic subclones in patient IF0138. tSNE representation of 243 cells using 418 differentially expressed

genes identified using ANOVA (C) and selected differentially expressed genes (D). (E-F) Transcriptional signatures of genetic subclones in patient IF0101. tSNE representation of 320 cells using 500 differentially expressed genes identified using ANOVA (E) and selected differentially expressed genes (F). Boxplots represent median and quartiles, and each dot represents the expression value of each individual cell. Perplexity was set at 25 for all tSNE representations. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

Importantly, commonly-used single-cell clustering and dimensionality reduction methods did not identify genetically-distinct subclones. tSNE analysis using highly-variable genes (Figure 4.14A-C) could not identify specific clustering based on genetic subclones, neither when regressing out the effect of cell cycle (Figure 4.14D-F) or when specifically modelling zero-inflation (Figure 4.14G-I) using ZIFA (288). SC3 K-means clustering (289), previously reported to distinguish genetically-distinct subclones of *JAK2V617F* mutant cells, could not identify genetic subclones in myelofibrosis patients (Figure 4.14J-L), and KNN-clustering implemented within the PAGODA2 package was also unable to identify subclone-specific gene signatures (Figure 4.14M-O). However, molecular signatures of different genetic subclones could be identified by dimensionality reduction when using differential expressed genes between these subclones. This indicates that TARGET-seq uniquely allows to identify subclone-specific gene signatures, which could not have been identified using unsupervised clustering methods. This also suggests that gene expression differences detected by unsupervised methods (Figure 4.14) are not driven by genetic alterations found in each single cell, and might be driven by other transcriptional, microenvironmental or epigenetic mechanisms.

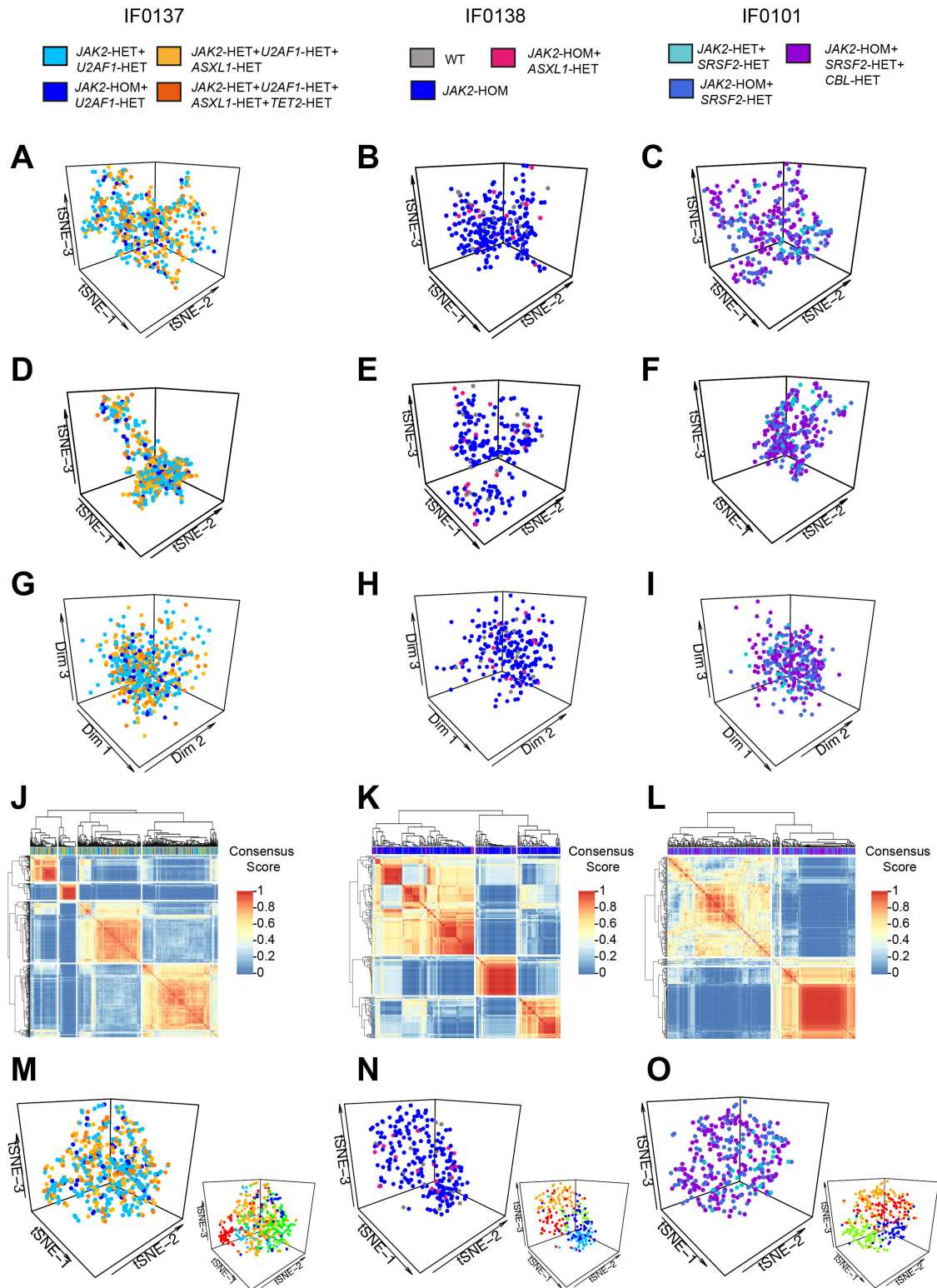


Figure 4.14. Dimensionality reduction and unsupervised clustering fail to discriminate different genetic subclones in individual patients. (A-C) tSNE representation of 555 cells from patient IF0137 (A), 243 cells from patient IF0138 (B) and 320 cells from patient IF0101 (C) using 3031 (A), 2605 (B) and 3023 (C) highly variable genes. Gene expression matrices were batch

corrected and genotypes of each single cell were preserved. Cells are coloured according to the genetic subclones found in each patient. (D-F) tSNE representation of the same patients and cells as in (A-C) using highly variable genes and removing the effect of cell cycle using a linear regression model implemented within the *limma* package. Gene expression matrices were batch and genotype corrected. (G-I) Dimensionality reduction using ZIFA from the same single cells and patients as in (A-C). (J-L) SC3 K-means clustering from the same single cells and patients as in (A-C) using k=4 for patient IF0137 (J), k=3 for patient IF0138 (K) and k=3 for patient IF0101 (L). Heatmaps are coloured according to the consensus score computed by SC3. (M-O) tSNE representation of clusters identified by PAGODA2 from the same single cells and patients as in (A-C), coloured by genotype (left panel) or by clusters identified by PAGODA2 (right panel). We identified 7 clusters in patient IF0137 (M), 5 clusters in patient IF0138 (N) and 4 clusters in patient IF0101 (O). Gene expression matrices were batch corrected using the batchNorm function (method='glm'). Perplexity was set at 25 for all tSNE representations. Figure adapted from Rodriguez-Meira *et al*, 2019 (1).

3. DISCUSSION

Integrated multi-omic single-cell analysis using TARGET-seq revealed marked genetic and transcriptomic heterogeneity in the stem and progenitor cell compartment of MPN patients. High-sensitivity mutational analysis resolved clonal hierarchies with over 95 % resolution, identifying both linear and branching patterns of clonal evolution, and up to seven different genetic subclones per patient. Moreover, the high sensitivity mutational analysis provided by TARGET-seq enabled the confident identification of *JAK2V617F* heterozygous and homozygous subclones, and in many patients, clonal evolution was driven by 9p-LOH.

Integration of whole transcriptome and mutational analysis revealed distinct molecular signatures associated with the presence or absence of somatic mutations in single cells. Cells carrying exclusively *JAK2V617F* mutations were transcriptionally highly similar to their wild-type counterparts, but acquisition of collaborating mutations in spliceosome components or epigenetic modifiers, which are associated with poorer prognosis, increased transcriptional diversity. *JAK2V617F* heterozygous and homozygous clones (which can only be confidently identified using single-cell analysis)

were also transcriptionally distinct, indicating molecular differences which could potentially be exploited therapeutically.

Moreover, full-length TARGET-seq uniquely enabled the analysis of alternative splicing in single cells. This proof of principle study identified molecular heterogeneity in *U2AF1* mutant HSPCs from MPN patients driven by alternative splicing. In the future, the analysis of larger cohorts of patients would potentially allow the identification of transcriptional signatures and splicing events driving clonal expansion, as well as the identification of neoantigens which could be targeted therapeutically. The implementation of long-read sequencing technologies (Nanopore, PacBio) (320, 321) to identify alternative splicing patterns in single cells could also provide a more comprehensive picture of the coordinated regulation of splicing events at the single cell level.

The high confidence mutational readouts obtained through TARGET-seq analysis also enabled the identification molecular signatures of non-mutant cells. Wild-type cells from MPN patients showed enriched inflammatory signatures compared to wild-type cells from normal donors, suggesting a cell-extrinsic effect of the tumour microenvironment disrupting gene expression in non-mutant cells. This might also underly therapy resistance or development of secondary haematological malignancy. Using mouse models to assess the impact of MPN mutant cells over their wild-type counterparts in a competitive transplantation setting could provide crucial insights into the functional mechanisms of these effects. This will allow to determine whether the presence of MPN mutant cells can induce leukemogenesis in wild-type cells residing in the same microenvironment.

Importantly, TARGET-seq analysis identified genes specifically expressed in mutant cells, including cell surface markers such as *RXFP1*, *CD69* or *G6B*, which might allow the prospective isolation and characterization of these genetic subclones. Further

characterization of these markers at the protein level would potentially validate their use as biomarkers of disease progression, or their implementation for disease monitoring and response to treatment in the clinical setting.

Furthermore, candidate surface markers specifically expressed in mutant cells identified through TARGET-seq analysis would also pave the way for their implementation as immunotherapy targets, to ultimately be able to eliminate mutant cells and specifically target the malignant clone in MPN.

In summary, TARGET-seq analysis revealed previously uncharacterized genetic and molecular heterogeneity in both mutant and non-mutant HSPCs in MPN, which could not have been identified with bulk sequencing techniques or other less sensitive single cell approaches. Applying TARGET-seq to understand the progression to other more aggressive and genetically complex forms of the disease, such as sAML, has the potential to unravel new clonal relationships and molecular signatures of disease progression, which could ultimately guide the development of new therapies for a currently untreatable myeloid neoplasm.

CHAPTER 5: SINGLE CELL CHARACTERIZATION OF LEUKEMIC STEM CELL POPULATIONS IN SECONDARY ACUTE MYELOID LEUKAEMIA

1. INTRODUCTION

Up to 20 % of patients with MPN will eventually transform to secondary acute myeloid leukaemia, also known as MPN-BP (MPN blast-phase). Whilst MPN patients receiving targeted therapies can live for many years, transformation is associated with extremely poor clinical outcomes, due to the poor response to standard AML induction chemotherapy and the lack of targeted therapies for this stage of the disease. As a consequence, the median overall survival of MPN-BP patients is less than four months (159), a prognosis which has not changed in the past decades.

Progression to acute myeloid leukaemia is associated with the acquisition of *TP53* missense mutations or alterations of the p53 pathway (*PPM1D* mutations, *MDM2/MDM4* amplification, -17, -17p) in 50% of cases (161, 178), suggesting that *TP53* has an essential role driving leukemogenesis. This is further supported by the dramatic increased in *TP53* variant allele frequency during the acute phase of the disease, and the collaboration between *JAK2V617F* mutation and *TP53* to give rise to acute myeloid leukaemia *in vivo* (161).

Patients with *TP53* mutations usually have high risk cytogenetic profiles (157) and particularly poor clinical outcomes. Therefore, there is an urgent need to develop new therapeutic approaches to treat this poor prognosis disease. However, the role of *TP53*

in development of sAML is poorly understood, making it essential to characterize the genetic and molecular mechanisms by which *TP53* might promote leukemogenesis.

1.1 HYPOTHESIS AND AIMS

The high frequency of *TP53* in MPN-BP suggests a major role of *TP53* as driver of transformation. In addition, acquisition of missense mutations in *TP53* are highly correlated with complex karyotypes, suggesting that increased genetic and molecular heterogeneity might drive leukemogenesis in this context.

I hypothesized that acquisition of *TP53* mutations might promote genetic instability and consequently, complex patterns of clonal evolution. This high degree of heterogeneity acquired during transformation would make extremely challenging to effectively target all subpopulations of tumour cells. Ultimately, this is likely to explain the mostly universal treatment failure in MPN-BP.

Moreover, the cellular hierarchies of *TP53* mediated leukemic transformation are likely to differ from *de novo* AML. In *de novo* AML, signalling mutations are often secondary hits and are acquired in committed progenitors, whereas *JAK2*, *CALR* and *MPL* are often initiating events acquired in HSCs in MPN. On the other hand, mutations in epigenetic regulators are usually detected in the founding clone in *de novo* AML, whereas these mutations are usually associated with disease progression in MPN. This suggests different patterns of clonal evolution in MPN-BP compared to *de novo* AML. This distinct molecular hierarchies might also explain the poor responses of MPN-BP patients to commonly used cytotoxic therapies for *de novo* AML patients, and it is therefore essential to characterize the precise genetic and molecular mechanisms of disease progression.

To investigate the role of *TP53* mutations in transformation to MPN-BP and cellular hierarchies of the disease, I decided to perform a comprehensive genetic characterization at the bulk level as well as high-resolution genetic and molecular

analysis of lineage-CD34⁺ cells in MPN-BP patients using TARGET-seq and ATAC-seq. This analysis would identify patterns of clonal evolution in MPN-BP patients and identify in which cell type *TP53* mutations are acquired. Moreover, the majority of evidence suggests that the tumour suppressor function of *TP53* is mediated by its transcriptional effect (322-324), which would be captured by the transcriptomic readouts obtained using TARGET-seq. Additionally, epigenomic profiling using ATAC-seq will uniquely allow the identification of regulatory networks of transcription factors driving disease progression. Given the strong association of mutations in epigenetic modifiers with transformation to MPN-BP, epigenomic dysregulation is likely to be a major driver of disease progression. Moreover, *TP53* is a transcription factor and therefore, the epigenomic readouts obtained from ATAC-seq would provide an additional layer of genomic information to characterize the effect of *TP53* mutations in HSPCs. This comprehensive genetic and molecular characterization of *TP53* mutant MPN-BP samples would ultimately identify drivers of leukemogenesis.

In order to address the questions outlined above, my specific aims were to:

- (1) Comprehensively characterize the genetic landscape of *TP53*-mutant MPN-BP
- (2) Characterize the cellular hierarchies and patterns of clonal evolution in *TP53*-mutant MPN-BP
- (3) Identify the molecular mechanisms leading to transformation to *TP53*-mutant secondary acute myeloid leukaemia.

2. RESULTS

2.1. GENETIC CHARACTERIZATION OF *TP53* MUTANT SECONDARY ACUTE MYELOID LEUKAEMIA

To characterize the genetic landscape of *TP53* mutant transformed myeloproliferative neoplasms, I first analysed driver mutations and chromosomal abnormalities at the

acute phase of the disease in 33 patient samples carrying missense driver mutations in *TP53* (Figure 5.1). At the bulk level, 31 out of 33 patients also carried mutations in *JAK2*, *CALR* and/or *MPL*, whereas the two remaining patients were *JAK2/CALR/MPL* wild-type, suggesting that secondary Acute Myeloid Leukaemia emerged from a different clone to the one driving the chronic phase of the disease.

TP53 mutation had a median variant allele frequency of 58.9 (± 28.15 SD), and 45% of the patients had VAF > 45%. Higher variant allele frequencies were universally associated with loss of heterozygosity or loss of chromosome 17 (Figure 5.1). 79% of patients (22/28) for which karyotype was available had chromosomal abnormalities in chromosome 17, in keeping with the previously reported association of *TP53* mutations with cytogenetics alterations in chromosome 17 (188). The most common co-occurrent mutations were *ASXL1*, *TET2*, *U2AF1* and *RUNX1*, highlighting the role of epigenetic modifiers in progression to MPN-BP (Figure 5.1).

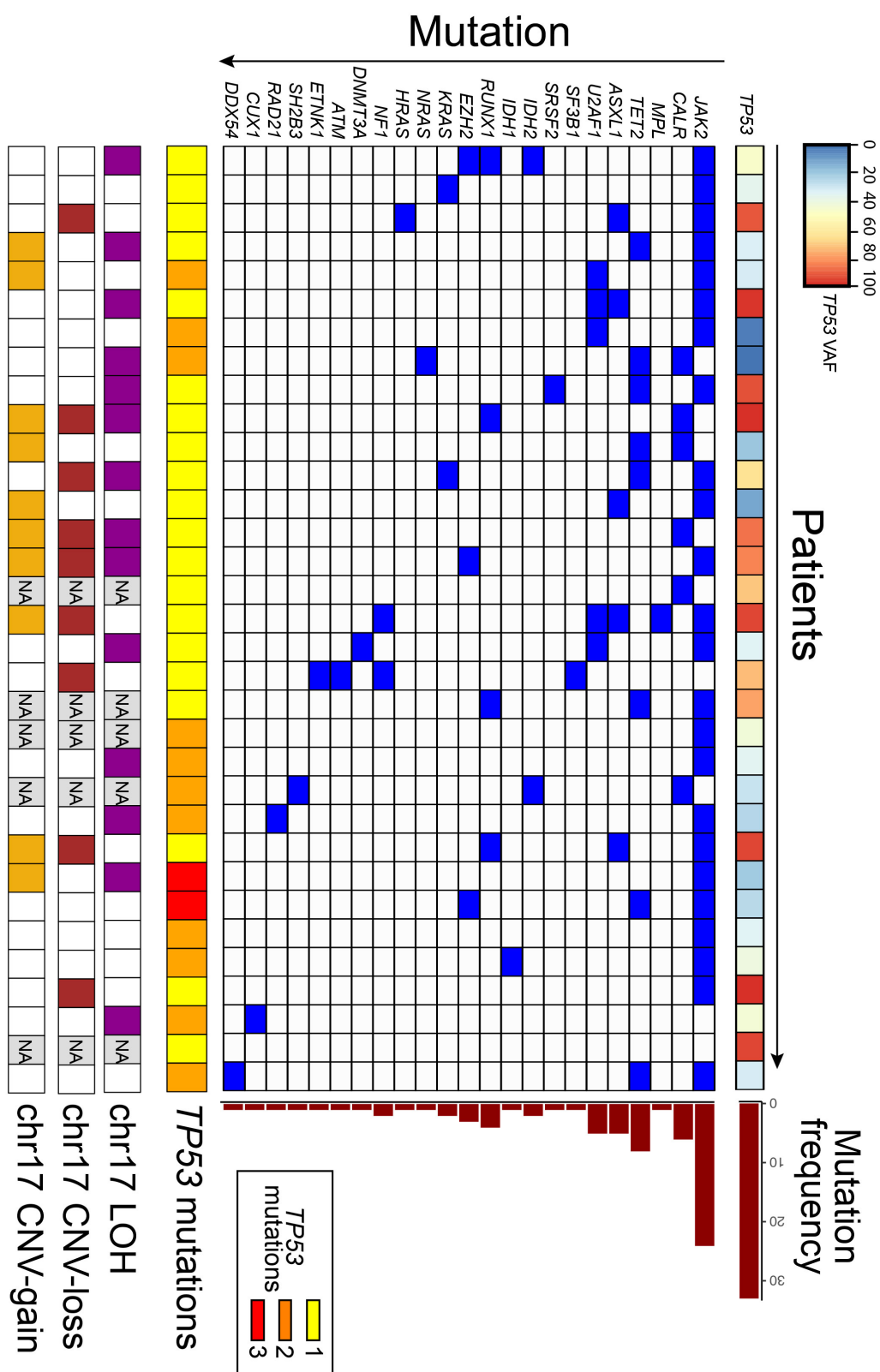


Figure 5.1. Summary of genetic alterations found in *TP53*-mutant secondary Acute Myeloid Leukaemia patients. The upper panel shows the mutations (rows) detected in each patient (columns) using a bulk sequencing assay for genes commonly mutated in myeloid malignancies (described in Chapter 2). The frequency of each mutation is indicated in the bar plot in the right

panel. The number of *TP53* mutations detected in each patient is shown on the lower panel for each patient, as well as the chromosomal abnormalities in chromosome 17 (LOH: Loss of Heterozygosity, CNV: Copy Number Variation).

2.2. CLONAL ARCHITECTURE OF TP53 MUTANT SECONDARY ACUTE MYELOID LEUKAEMIA

In order to characterize the cellular hierarchies in the acute phase of the disease, I first assessed the immunophenotypic distribution of lineage-CD34⁺ cells in eight MPN-BP patient samples. Flow cytometry profiles revealed a disrupted haematopoietic hierarchy in most patients analysed, with an expansion of MPP/CMP or LMPP/GMP populations (Figure 5.2A, 5.2B). This expansion is also characteristic of *de novo* Acute Myeloid Leukaemia patients (325), indicating that MPN-BP patient samples have similar disrupted cellular populations at the immunophenotypic level. Overall, MPN-BP patients had comparable proportions of lineage-CD34⁺CD38⁺ (Figure 5.2C) and lineage-CD34⁺CD38⁻ cells (Figure 5.2D) as compared to normal donor controls. However, immunophenotypic haematopoietic stem cells were markedly deduced (Figure 5.2E).

I then performed TARGET-seq single-cell genotyping in four MPN-BP patient samples. Lineage-CD34⁺CD38⁺ cells (data not shown) were mostly clonal for all driver mutations analysed in each patient, whereas Lineage-CD34⁺CD38⁻ cells were genetically heterogeneous (Figure 5.3).

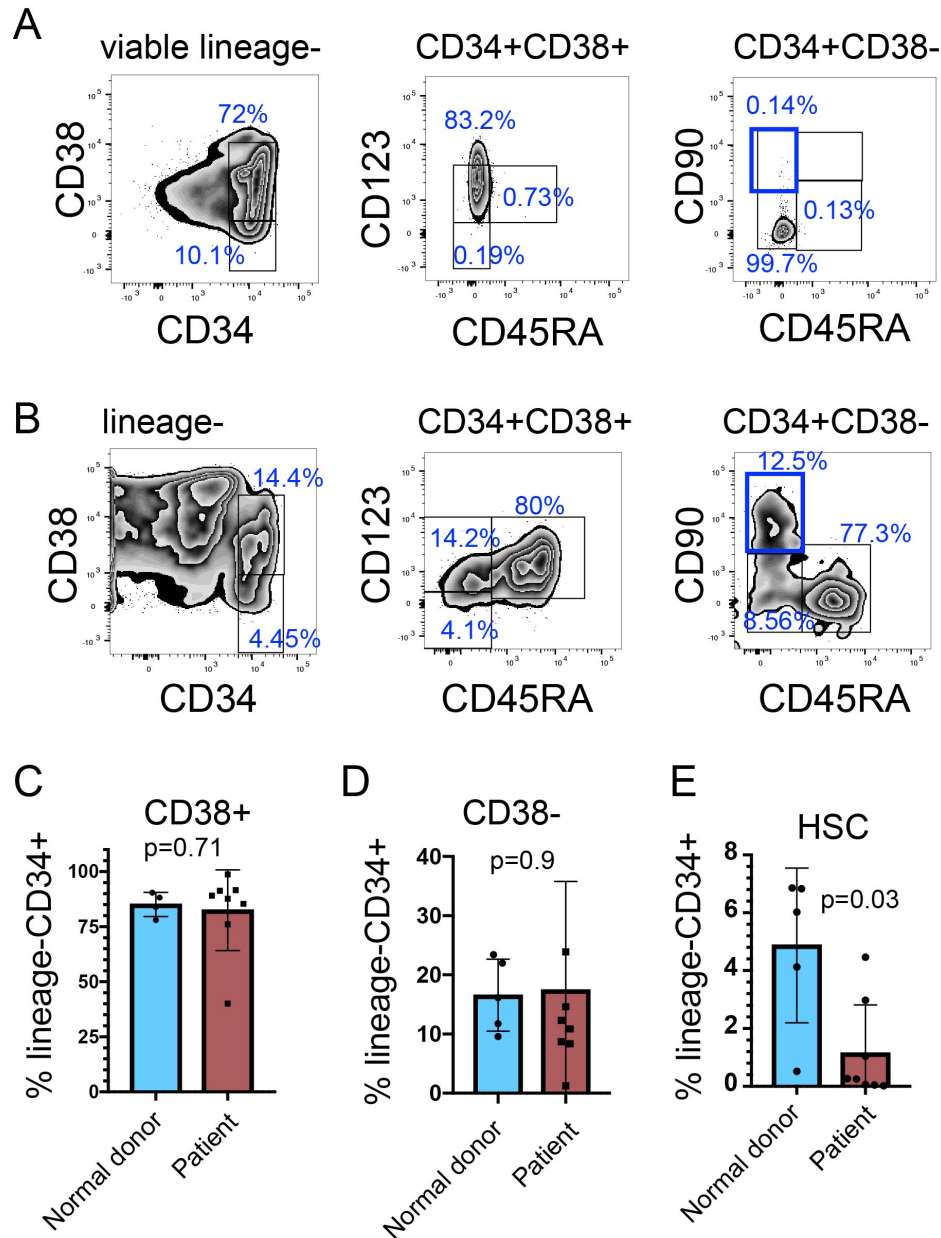


Figure 5.2. Immunophenotypic characterization of MPN-BP patient samples. (A-B) Immunophenotypic distribution of lineage- cells (left panel), lineage-CD34+CD38+ (middle panel) and lineage-CD34+CD38- (right panel) for patient GST010 (A) and SB5702 (B). Numbers indicate the percentage of cells within each parent gate. (C-E) Percentage of CD38+ (C), CD38- (D) and HSCs (CD38-CD90+CD45RA-) within lineage-CD34+ cells for patient samples (n=8) and normal donors (n=5). “p” indicates p-value from Student’s t-test with Welch’s correction for each comparison. Cells were stained with HSPC antibody panel B described in Chapter 2.

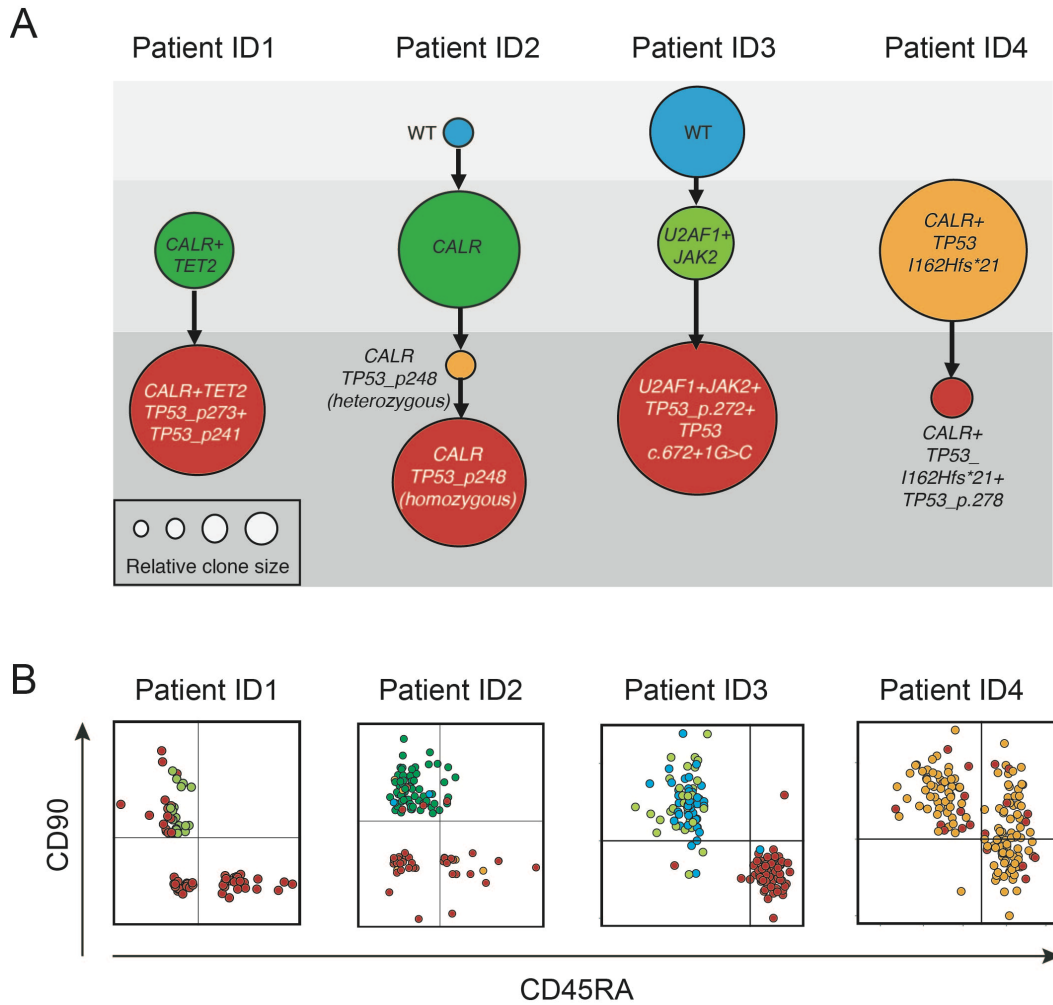


Figure 5.3. Clonal hierarchies of HSPCs from MPN-BP patients. (A) Schematic representation of the phylogenetic tree of each tumour in 4 MPN-BP patients. Each circle represents the proportion of single cells carrying a particular combination of mutations, and each circle is coloured according to the combination of mutations found in each subclone. (B) Integration of mutational information with surface proteomics for the same MPN-BP patients presented in (A) for lineage-CD34+CD38- cells. Each cell is coloured according to colour code for each genetic subclone presented in (A).

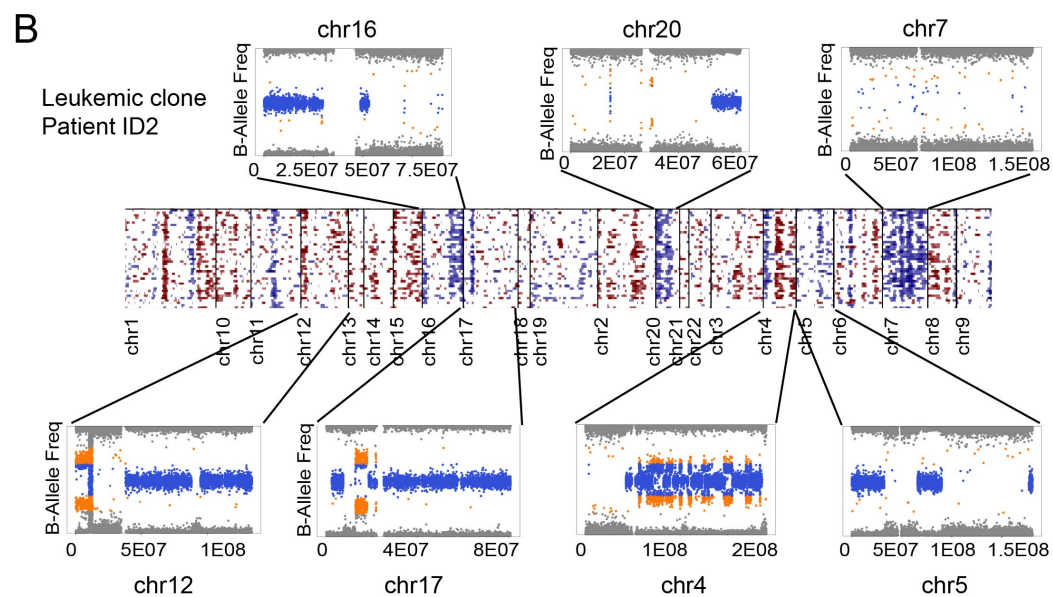
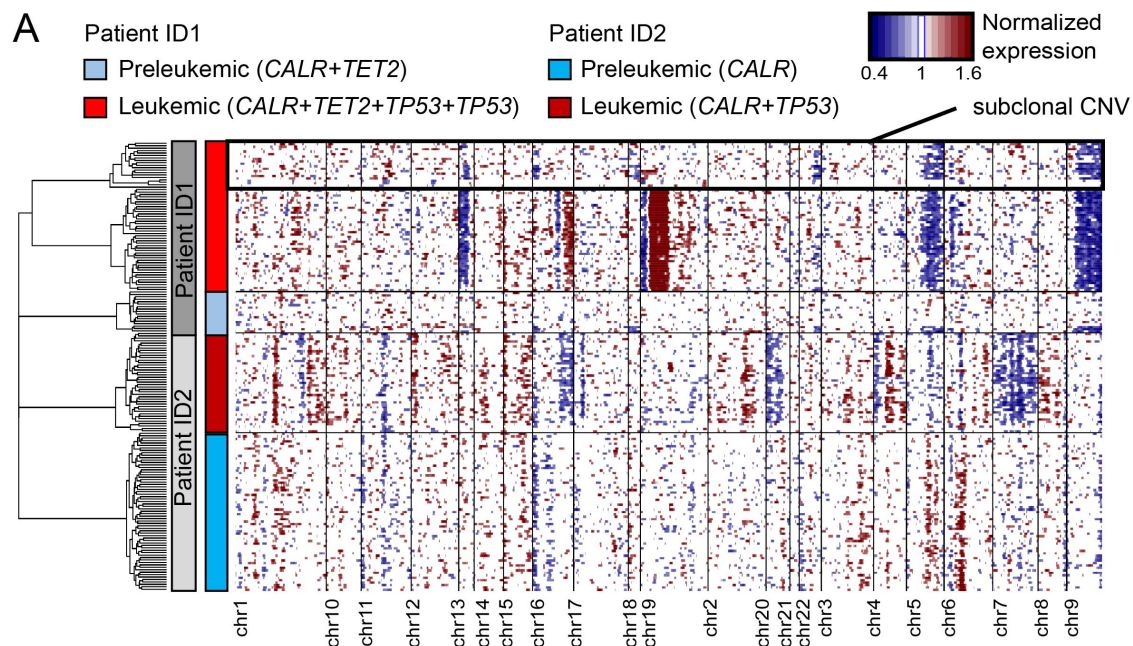
TP53 mutations were acquired in all cases in an antecedent genetic subclone carrying a MPN-associated mutation in *JAK2V617F* or *CALR*, indicating a linear pattern of clonal evolution. *TP53* missense mutations were subclonal in three out of the four patients analysed (Figure 5.3A) and both heterozygous and homozygous/hemizygous subclones were identified. Interestingly, clonal evolution was associated with the acquisition of loss of heterozygosity or two concomitant missense mutations (Figure

5.3A). Whilst it is unclear whether these different missense mutations were acquired in the same or different alleles in patients ID1, ID3 and ID4, this data suggests that complete loss of WT *TP53* activity might be required for disease progression, an observation previously described in the context of acute myeloid leukaemia (191).

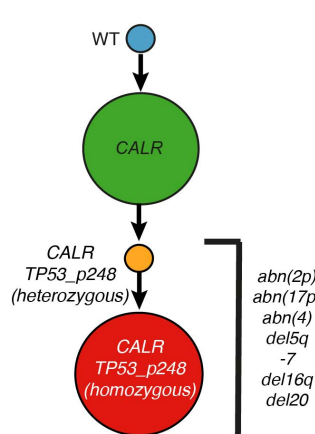
Integration of cell surface proteomics allowed mapping each mutation to different immunophenotypically defined stem and progenitor cell types (Figure 5.3B). In three out of four cases, *TP53* mutations were acquired in lineage-CD34+CD38-CD90+CD45RA- stem cells, residing at the top of the haematopoietic hierarchy. This is in contrast to the patterns of clonal evolution identified in *de novo* AML, in which leukemic transformation associated mutations are acquired in more differentiated progenitors (142). This could potentially indicate presence of different leukemic stem cell reservoirs in MPN-BP versus *de novo* AML. Although speculative, this might partially explain why commonly used AML cytotoxic therapy is mostly ineffective in MPN-BP patients, as transformation-associated mutations are present within the stem cell compartment. Alternatively, the use of less sensitive sequencing technologies in previous studies might have prevented the identification of stem cells carrying leukemic mutations in *de novo* AML.

TARGET-seq allowed the analysis of single-cell transcriptomic readouts from the same single cells, which can be used to identify chromosomal aberrations (239). This analysis identified further clonal evolution driven by the acquisition of cytogenetic abnormalities in MPN-BP patients (Figure 5.4A), which were also detected in bulk SNP Array profiling of MNCs from the same patients (Figure 5.4B), validating the CNV results obtained from single cells. Integration of single-cell mutation and CNV data indicated that acquisition of *TP53* mutations was associated with the presence of cytogenetic abnormalities, whereas *TP53* WT preleukaemic subclones were cytogenetically normal (Figure 5.4A-D). Importantly, subclonal CNVs were identified in patient ID1 (Figure 5.4A, 5.4D), suggesting that the acquisition of *TP53* mutations

leads to genetic instability and consequently, highly complex clonal evolution. *CALR*, *TP53*, *TET2* mutant and del5q+del9q genetic subclone was restricted to immunophenotypically defined haematopoietic stem cells, whereas further acquisition of del13, amp16 and amp19 was restricted to committed progenitors (MPPs and LMPPs, Figure 5.4D), further supporting the hypothesis that *TP53* leads to genetic instability and subsequent complex clonal evolution. The identification of *TP53* mutant cells lacking cytogenetic lesions will determine the impact of *TP53* mutations independently of subsequent karyotypic abnormalities.



C Patient ID2



D

Patient ID1

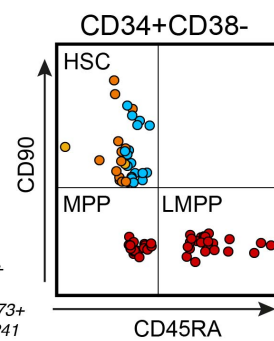
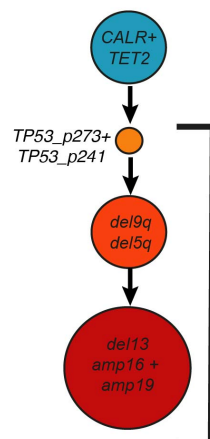


Figure 5.4. Detection of cytogenetic abnormalities in single cells using single-cell transcriptomics. (A) Heatmap of normalized expression scores calculated using inferCNV package across the entire transcriptome in single cells from patients ID1 (n=78) and ID2 (n=106). Blue colour indicates chromosomal losses and red colour, chromosomal gains. (B) Representative example of cytogenetic abnormalities detected in *TP53* mutant cells (n=41) from patient ID2 and by bulk MNC analysis of the same patient, using a SNP (Single Nucleotide Polymorphism) Array. The heatmap represented the normalized expression score for each single cell and the panels above and below the heatmap, the B-Allele Frequency of each chromosome in which cytogenetic abnormalities were detected. B-Allele Frequency values <0.20 or >0.80 are coloured in grey, in between 0.20 and 0.35 or 0.65 and 0.80, in orange, and in between 0.35 and 0.65, in blue. The numbers below each panel indicate chromosomal coordinates. (C-D) Reconstruction of clonal hierarches in patient ID2 (C) and patient ID1 (D) based on mutations and chromosomal aberrations (left panel) and immunophenotypic distribution of each genetic subclone in the lineage-CD34+CD38- compartment (D, right panel).

2.3. TRANSCRIPTOMIC AND EPIGENOMIC CHARACTERIZATION OF PRE-LEUKEMIC AND LEUKEMIC SUBCLONES

To characterize the molecular pathways dysregulated in each genetic subclone, I performed TARGET-seq analysis of 3 patient samples (Patients ID1, ID2 and ID3) and low input (500 cells) ATAC-seq analysis of different cell subpopulations (Figure 5.5). Differential expression analysis between different genetic subclones in patient ID1 identified molecular signatures driven by mutations in *TP53* (Figure 5.5.A) but also by the acquisition of chromosomal abnormalities in *TP53* mutant clones (Figure 5.5A). Subclones carrying MPN-associated mutations showed a strong enrichment in inflammatory signatures (Figure 5.5.B), in keeping with the previously described transcriptional effect of MPN-associated mutations in MPN patients (chapter 4), whereas leukemic subclones carrying *TP53* mutations were highly enriched in cell-cycle related genes, PI3K and MTORC1 signalling (Figure 5.5B).

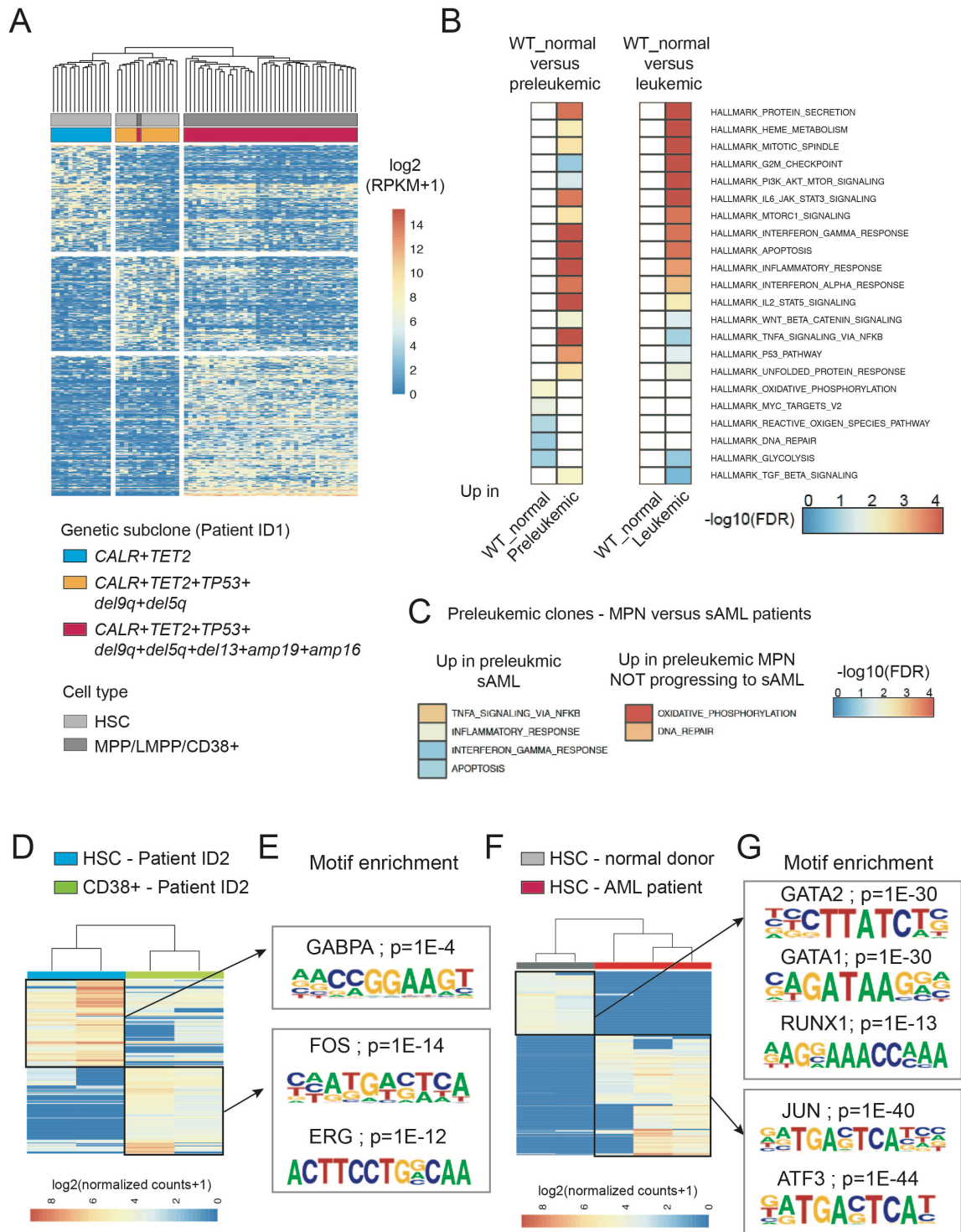


Figure 5.5. Molecular pathways dysregulated in HSPCs from MPN-BP patients. (A) Heatmap of gene expression values from 794 differentially expressed genes in between the three main genetic subclones identified in patient ID1 (Figure 5.4) (B) Pathway enrichment analysis of preleukemic (carrying MPN-associated mutation), leukemic (carrying *TP53* mutations) subclones in sAML patients (n=3) as compared to cells from normal donors (n=2), using hallmark genesets. White colour indicates non-significance. (C) Pathway enrichment analysis of preleukemic subclones (carrying MPN-associated mutations) from sAML patients (left panel)

and MPN patients who did not progress to the acute phase of the disease (right panel). (D-E) Differential analysis of chromatin accessibility (D) between HSCs and lineage-CD34+CD38+ cells in patient ID2 (n=2 replicates), and motif enrichment (E) in peaks enriched in HSCs (upper panel) or lineage-CD34+CD38+ cells (lower panel). (F-G) Differential analysis of chromatin accessibility (F) between HSCs from normal donors (n=2) and HSCs from sAML patients (n=3), and motif enrichment (G) in peaks enriched in HSCs from normal donors (upper panel) or HSCs from sAML patients (lower panel).

Interestingly, analysis of preleukemic subclones (carrying MPN-associated mutations) at the acute and chronic phase of the disease revealed a strong inflammatory signature in preleukemic cells from patients with sAML (Figure 5.5C). Inflammation has a crucial role in promoting tumorigenesis, potentially suggesting that preleukemic clones of patients who progress to the acute phase of the disease could be primed for disease transformation by upregulating inflammation associated genes signatures. Moreover, preleukemic clones from sAML patients were depleted of genes involved in oxidative phosphorylation and DNA repair, potentially indicating another pathway by which these cells might be more susceptible to undergo disease transformation. These data highlight how TARGET-seq can be implemented to characterize the cellular and molecular hierarchies of disease progression in MPN, generating testable hypothesis relating to the mechanisms of transformation that can be explored in future studies.

Epigenomic dysregulation is also an important feature driving the development of myeloid malignancies, and the high frequency of mutations in epigenetic regulators in MPN-BP suggests a crucial role in disease progression. Bulk epigenomic analysis of HSCs (mostly carrying MPN-associated mutations; Figure 5.3) and lineage-CD34+CD38+ cells (carrying *TP53* mutations) in patient ID2 identified a high enrichment of FOS and ERG transcription factor binding sites in CD38+ cells (Figure 5.5D, 5.5E). FOS is an oncogenic transcription factor required for CML progression and TKI resistance (326), and high ERG activity has been shown to confer chemoresistance in acute myeloid leukaemia (327), suggesting that the activity of these transcription factors might be driving chemoresistance in *TP53*-mutant sAML.

HSCs from patient ID2 were enriched in GABPA binding sites, which regulates stem cell maintenance and differentiation (328).

Differential analysis of chromatic accessibility between HSCs from normal donors and patient samples identified depletion of GATA1, GATA2 and RUNX1 (Figure 5.5F, 5.5G), implicated in hematopoietic differentiation, in HSCs from AML patients, and enrichment in JUN and ATF3 motifs, which cooperate to support aberrant transcriptional programs in AML (329).

This analysis provides insights into the regulatory landscape driving aberrant transcriptional programs in *TP53*-mutant secondary AML. However, bulk ATAC-sequencing requires the prospective isolation of cells by flow cytometry and therefore, the results from this analysis might be masked by the heterogeneity (genetic and molecular) found within this immunophenotypic compartments in sAML patients. Therefore, analysis of chromatin accessibility at the single-cell level will ultimately be required to precisely determine the epigenomic mechanisms of disease progression upon acquisition of a *TP53* mutation in MPN-BP.

3. DISCUSSION

Single-cell analysis of patients with blast phase MPN driven by mutations in *TP53* revealed a linear pattern of clonal evolution in the four patients analysed, with subsequent subclonal evolution through the acquisition of karyotypic abnormalities. This indicates that *TP53* mediated transformation leads to subsequent genetic instability and complex clonal evolution (Figure 5.6A). *TP53* wild-type MPN-BP patients usually have normal karyotypes or low risk cytogenetic abnormalities, and similar variant allele frequencies of driver mutations between the chronic and acute phase of the disease, as opposed to the marked increase in *TP53* mutant variant allele frequency in *TP53*-driven MPN-BP. These lines of evidence suggest that *TP53*-

mediated transformation might represent a distinct route towards leukemic transformation.

The integration of several layers of genomic information (genetic, transcriptomic) and cell surface proteomics using TARGET-seq provided a high-resolution snapshot of the clonal architecture of sAML from a preceding MPN. Future single-cell analysis of more patients, including paired chronic and acute phase samples (Figure 5.6A), will potentially allow the identification of subclones present in the chronic phase of the disease which might be “primed” to drive disease transformation. Indeed, preliminary analysis of patient samples at the acute phase of the disease indicated that a strong inflammatory signature and defects the DNA repair pathway in preleukemic clones might mediate transformation. Analysis of patient samples at the chronic phase will also determine whether *TP53* mutations can be detected at low levels years before diagnosis, and whether subpopulations of cells that are transcriptionally “primed” for transformation can be identified before disease progression. The identification and molecular characterization of these cell populations will ultimately enable to detect and target transformation events early during disease progression.

Some MPN patients carrying low level *TP53* mutations do not transform to sAML (96) during their disease course (Figure 5.6B). Therefore, TARGET-seq analysis of samples from patients carrying low level *TP53* mutations who did not transform to sAML would provide crucial insights into the molecular mechanisms leading to transformation. This analysis would allow comparison of the transcriptional signatures of *TP53* subclones from patients who do not progress to the acute phase to those who developed sAML, and will potentially help to identify patients at risk of progression. Similarly, comparison of the molecular signatures identified in MPN patients who did not progress to the acute phase of the disease (Figure 5.6C) could potentially identify predictors of transformation. Ultimately, this analysis will build a comprehensive model of the molecular steps triggering leukemic transformation, guiding the development of new

therapeutic targets to develop early therapeutic interventions. Moreover, biomarkers of disease progression identified through this analysis could be implemented in the future for routine clinical monitoring of MPN patients.

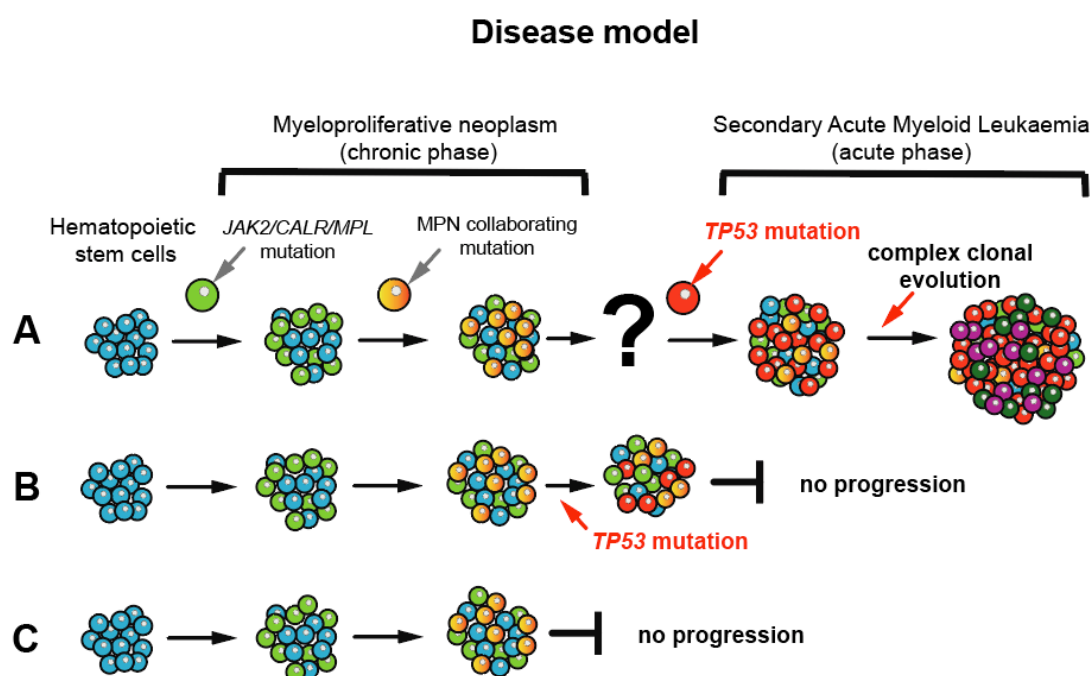


Figure 5.6. Schematic representation of MPN development and transformation to secondary Acute Myeloid Leukaemia upon acquisition of *TP53* mutations. Acquisition of JAK/STAT signalling mutations, as well as other collaborating mutations in haematopoietic stem cells, leads to MPN development. Further acquisition of *TP53* mutations and subsequent complex clonal evolution can lead to transformation to secondary acute myeloid leukaemia (A). However, some patients with low level *TP53* mutations do not progress to the acute phase of the disease (B). Other MPN patients do not progress to the acute phase of the disease (C).

Furthermore, the identification of the genetic and molecular determinants of *TP53* mediated transformation has not only crucial implications for myeloid neoplasms but also for many other types of cancers. *TP53* is the most commonly mutated gene in cancer, and it is invariably associated with poor prognosis and resistance to cytotoxic therapy. Therefore, the characterization of the role of *TP53* mediating transformation will also be of high relevance for many other types of tumours.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

The genetic and molecular heterogeneity acquired during the development of myeloid neoplasms poses many challenges for the development of effective treatments. These treatments should be able to target and eliminate disease-propagating cells. However, current therapeutic strategies are largely incapable of completely eradicating mutant cells, and as a consequence, most patients achieving remission eventually relapse. Therefore, there is an urgent need to develop new therapeutic approaches specifically targeting the malignant clone.

To achieve this goal, it is essential to characterize molecular pathways dysregulated in each genetically-distinct subclone. The first aim of this thesis was to develop a single-cell multi-omic method capable of correlating molecular (transcriptional) and genetic readouts from the same single cell. This method, TARGET-seq, was then used to study HSPCs from patients with myeloid neoplasms, identifying molecular pathways associated with clonal evolution and putative biomarkers of these genetic subclones. This molecular characterization will ultimately enable the development of novel therapeutic strategies for patients with myeloid neoplasms.

6.1. IMPLICATIONS OF THIS WORK

TARGET-seq is the first multi-omic method capable of integrating mutational, transcriptomic and cell surface proteomics from the same single cell with high sensitivity. This technology enables study of the molecular mechanisms (through transcriptional readouts) of clonal evolution, and it is a powerful tool to unravel intratumoral heterogeneity in myeloid tumours and potentially, many other types of cancers.

Concomitant analysis of mutations and whole transcriptomes from the same single cell identified distinct molecular signatures of different genetic subclones in myeloproliferative neoplasm stem and progenitor cells at the chronic and acute phase of the disease. The acquisition of poor prognosis collaborating mutations in spliceosome components and epigenetic modifiers, in addition to *JAK2V617F* mutations, increased transcriptional diversity. This analysis also identified enrichment of inflammatory signatures in cells from patient samples independently of mutational status, suggesting a marked effect of the tumour microenvironment in both mutant and WT cells from patient samples as compared to cells from aged-matched normal donors. This effect was further increased upon disease transformation, indicating that inflammatory signatures could be drivers of disease progression. TARGET-seq analysis also uniquely allowed the identification of mutation-specific biomarkers, including cell surface proteins, which would allow prospective isolation of mutant cells, paving the way for the development of immunotherapies specifically targeting the mutant clone.

6.2. LIMITATIONS OF THIS WORK

The current throughput of TARGET-seq is limited to processing approximately 1000 cells per week, meaning that only a few thousands of cells can be analysed per experiment. This throughput is relatively low compared to other recently developed droplet-based single-cell technologies, which allow processing of tens of thousands of single cells. This limitation could be solved by some of the adaptations described in this work (early barcoding of single-cell genotyping libraries), which would allow processing of approximately 3000 cells per week. However, the inherently plate-based nature of the method does not easily allow scalability of the technique to tens or hundreds of thousands of single cells. Implementing TARGET-seq in a droplet-based microfluidics platform could make the technique of comparable throughput to other droplet-based

techniques, and ultimately allow the analysis of tens to hundreds of thousand single cells.

Moreover, TARGET-seq is limited to the study of pre-selected candidate *driver* mutations identified through bulk sequencing methods, and therefore does not allow unbiased assessment of the mutational status of each single cell. Additionally, the number of mutations that can be profiled is currently limited to twelve mutations per single cell. Whilst this is sufficient to resolve genetic heterogeneity in relatively genetically simple tumours such as myeloid leukaemia, some solid tumours might require analysis of hundreds of hotspots. Therefore, its application to study other more heterogeneous tumour subtypes could be limited by the number of mutations that can be profiled, as well as the number of cells that can be processed per patient.

The analysis presented in this study was sufficiently powered to detect molecular differences between genetic subclones in HSPCs; however, to confidently characterize the effect of each mutation in each different haematopoietic cell type, higher numbers of cells would be required. Additional analysis of more differentiated progenitors as opposed to just HSPCs would also provide crucial insights into differentiation pathways dysregulated by the acquisition of somatic mutations in myeloid neoplasms.

6.3. FUTURE WORK

Single-cell multi-omic methods are ideally placed to resolve multi-layered heterogeneity of cancerous tissues, and to identify molecular mechanisms of disease initiation, progression and relapse. The integration of different multi-omic readouts from the same single cell gives a much more comprehensive picture of the coordinated regulation of each of these layers to better understand cellular function.

TARGET-seq currently allows integration of genetic, transcriptomic and cell surface proteomic readouts from the same single cell. However, the number of mutations that

can be profiled using TARGET-seq is relatively low; increasing the multiplexing capacity of the method would allow the assessment of much more complex clonal hierarchies than the ones presented in this work. Moreover, the cell surface proteomic readouts obtained from TARGET-seq are currently limited by the number of different fluorochromes that can be detected by flow cytometry. This could be expanded by adapting newly developed technologies allowing detection of epitopes by sequencing (CITE-seq) (330), which would theoretically enable the detection of over a hundred epitopes per single cell. Such approach would also validate potential cell surface markers identified through transcriptional readouts from TARGET-seq analysis of the same single cells.

The integration of epigenomic readouts with transcriptomic and mutational readouts obtained through TARGET-seq will also be important to understand the transcriptional consequences of epigenomic dysregulation upon clonal evolution. Whilst the most commonly used single cell epigenomic technique (single cell ATAC-seq) (269) would in principle be incompatible with TARGET-seq (due to the required protease digestion, which would make most of the gDNA accessible for amplification), the implementation of single-cell bisulphite conversion (252) within the TARGET-seq workflow could allow the integration of epigenomic readouts.

Additionally, the development of new computational tools to integrate several layers of genomic information from assays performed independently from the same pool of single cells also shows great potential to identify regulatory relationships between these different layers (257). The development of new computational methods to integrate epigenomic and high throughput proteomic data with TARGET-seq readouts would allow to understand the coordinated regulation of each of these genomic layers upon acquisition of somatic mutations. Ultimately, this would provide crucial insights into the regulatory networks influencing disease development and progression.

Moreover, TARGET-seq analysis of patients with MPN and sAML revealed novel molecular signatures and pathways potentially influencing disease biology which could be explored in future studies. For example, understanding the effects of the MPN microenvironment in non-clonally involved cells would be essential to develop effective therapeutic approaches, as these effects have prognostic value in myeloid neoplasms (247) and might mediate disease progression. Using mouse models to perform competitive transplantation of mutant and WT cells could provide insights into the functional mechanisms mediating inflammatory phenotypes of WT cells in MPN and sAML, and could determine whether these WT cells are then more likely to acquire leukemogenic potential.

The preliminary analysis of patients with MPN-BP revealed previously uncharacterized patterns of clonal evolution given by the acquisition of *TP53* mutations in HSCs. The data presented in this study supports a mechanism by which the initial acquisition of a *TP53* mutation, followed by loss of the wild-type *TP53* allele, gives rise to genetic instability and highly complex clonal evolution driven by chromosomal abnormalities. However, a comprehensive single-cell analysis of more patients, and analysis of paired samples from the chronic and acute phase of the disease, including patients who did not progress to MPN-BP despite long follow-up, will be essential to generate a model of *TP53* mediated transformation in myeloid neoplasms. These analyses could potentially identify cellular subpopulations “primed” to drive disease progression. Identifying the molecular pathways driving this transformation in such “primed” populations early during disease progression could help stratify patients at risk, and contribute to the development early therapeutic interventions. Importantly, *TP53* is the gene most frequently mutated in cancer and is universally associated with poor prognosis; characterizing the mechanisms of *TP53*-mediated transformation will likely provide crucial insights into *TP53* driven carcinogenesis in many other types of tumours.

Single-cell analysis of *JAK2V617F* mutant MPN patients also identified putative biomarkers of malignant cells, such as *G6B*, *IL8*, *RXFP1*, *STAT1* or *GAS2*. However, transcriptional readouts obtained from single cells do not take into account any posttranscriptional mechanisms potentially affecting those mRNA species. Further validation at the protein level will be essential to establish their utility as biomarkers of mutant cells, to ultimately enable their implementation for disease monitoring or as novel immunotherapy targets. Indeed, preliminary studies are promising in this regard. *G6B*, a cell surface marker and potential immunotherapy target identified in this work, has successfully been used to selectively isolate *JAK2V617F* mutant cells in myelofibrosis patients in follow-up work from this thesis (263), and novel bispecific antibodies targeting CD34 and *G6B* are being developed as an immunotherapy approach in myelofibrosis (263).

Moreover, functional validation using genetic tools such as CRISPR/Cas9 will be essential to establish the causal role of these targets in disease maintenance or progression. The development of new tool compounds inhibiting these molecules could also lead to the new therapeutic strategies to specifically target and eliminate mutant cells, whilst maintaining normal haematopoiesis.

The work presented in this thesis demonstrates the potential of single cell technologies to study intratumoral heterogeneity and identify novel cellular hierarchies and molecular pathways amenable of therapeutic targeting. Single cell technologies are currently well established in research laboratories and in commercially available platforms, and are constantly experiencing new technological developments to make them more cost-effective and high-throughput. In the future, these techniques could be integrated into diagnostic pathways in the clinical setting to characterize the cellular composition of the peripheral blood or bone marrow of patients with myeloid neoplasms. This could assist disease classification based on genetic, transcriptomic and/or epigenomic diversity, identify subpopulations of cells that might predict disease

progression or monitor patient's response to treatment, and will pave the way for the rational development of new therapeutic strategies.

BIBLIOGRAPHY

1. Rodriguez-Meira A, Buck G, Clark SA, Povinelli BJ, Alcolea V, Louka E, et al. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Mol Cell*. 2019;73(6):1292-305 e8.
2. Spangrude GJ, Heimfeld S, Weissman IL. Purification and characterization of mouse hematopoietic stem cells. *Science*. 1988;241(4861):58-62.
3. Baum CM, Weissman IL, Tsukamoto AS, Buckle AM, Peault B. Isolation of a candidate human hematopoietic stem-cell population. *Proc Natl Acad Sci U S A*. 1992;89(7):2804-8.
4. Morrison SJ, Weissman IL. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity*. 1994;1(8):661-73.
5. Morrison SJ, Uchida N, Weissman IL. The biology of hematopoietic stem cells. *Annu Rev Cell Dev Biol*. 1995;11:35-71.
6. Goodell MA, Rosenzweig M, Kim H, Marks DF, DeMaria M, Paradis G, et al. Dye efflux studies suggest that hematopoietic stem cells expressing low or undetectable levels of CD34 antigen exist in multiple species. *Nat Med*. 1997;3(12):1337-45.
7. Anjos-Afonso F, Currie E, Palmer Hector G, Foster Katie E, Taussig David C, Bonnet D. CD34- Cells at the Apex of the Human Hematopoietic Stem Cell Hierarchy Have Distinctive Cellular and Molecular Signatures. *Cell Stem Cell*. 2013;13(2):161-74.
8. Jacobsen SEW, Nerlov C. Haematopoiesis in the era of advanced single-cell technologies. *Nat Cell Biol*. 2019;21(1):2-8.
9. Sanjuan-Pla A, Macaulay IC, Jensen CT, Woll PS, Luis TC, Mead A, et al. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*. 2013;502:232.
10. Carrelha J, Meng Y, Kettyle LM, Luis TC, Norfo R, Alcolea V, et al. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature*. 2018;554:106.
11. Haas S, Hansson J, Klimmeck D, Loeffler D, Velten L, Uckelmann H, et al. Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors. *Cell Stem Cell*. 2015;17(4):422-34.
12. Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C, et al. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev*. 1997;11(6):774-85.
13. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*. 2015;163(7):1663-77.

14. Drissen R, Buza-Vidas N, Woll P, Thongjuea S, Gambardella A, Giustacchini A, et al. Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat Immunol.* 2016;17(6):666-76.
15. Pellin D, Loperfido M, Baricordi C, Wolock SL, Montepeloso A, Weinberg OK, et al. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nature Communications.* 2019;10(1):2395.
16. Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood.* 2016;128(8):e20-31.
17. Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology.* 2017;19:271.
18. Skoda RC, Duek A, Grisouard J. Pathogenesis of myeloproliferative neoplasms. *Exp Hematol.* 2015;43(8):599-608.
19. Campbell PJ, Green AR. The Myeloproliferative Disorders. *New England Journal of Medicine.* 2006;355(23):2452-66.
20. Titmarsh GJ, Duncombe AS, McMullin MF, O'Rourke M, Mesa R, De Vocht F, et al. How common are myeloproliferative neoplasms? A systematic review and meta-analysis. *American Journal of Hematology.* 2014;89(6):581-7.
21. Hoglund M, Sandin F, Simonsson B. Epidemiology of chronic myeloid leukaemia: an update. *Ann Hematol.* 2015;94 Suppl 2:S241-7.
22. Srour SA, Devesa SS, Morton LM, Check DP, Curtis RE, Linet MS, et al. Incidence and patient survival of myeloproliferative neoplasms and myelodysplastic/myeloproliferative neoplasms in the United States, 2001-12. *Br J Haematol.* 2016;174(3):382-96.
23. Barbui T, Thiele J, Gisslinger H, Kvasnicka HM, Vannucchi AM, Guglielmelli P, et al. The 2016 WHO classification and diagnostic criteria for myeloproliferative neoplasms: document summary and in-depth discussion. *Blood Cancer Journal.* 2018;8(2):15.
24. Spivak JL. Polycythemia vera: myths, mechanisms, and management. *Blood.* 2002;100(13):4272-90.
25. Beer PA, Green AR. Pathogenesis and management of essential thrombocythemia. *Hematology Am Soc Hematol Educ Program.* 2009:621-8.
26. Mesa RA, Schwager S, Radia D, Cheville A, Hussein K, Niblack J, et al. The Myelofibrosis Symptom Assessment Form (MFSAF): an evidence-based brief inventory to measure quality of life and symptomatic response to treatment in myelofibrosis. *Leuk Res.* 2009;33(9):1199-203.
27. Palandri F, Catani L, Testoni N, Ottaviani E, Polverelli N, Fiacchini M, et al. Long-term follow-up of 386 consecutive patients with essential thrombocythemia: safety of cytoreductive therapy. *Am J Hematol.* 2009;84(4):215-20.

28. O'Sullivan J, Mead AJ. Heterogeneity in myeloproliferative neoplasms: Causes and consequences. *Adv Biol Regul.* 2019;71:55-68.
29. Deininger MW, Goldman JM, Melo JV. The molecular biology of chronic myeloid leukemia. *Blood.* 2000;96(10):3343-56.
30. Sharma N, Magistroni V, Piazza R, Citterio S, Mezzatesta C, Khandelwal P, et al. BCR/ABL1 and BCR are under the transcriptional control of the MYC oncogene. *Mol Cancer.* 2015;14:132.
31. Kralovics R, Passamonti F, Buser AS, Teo SS, Tiedt R, Passweg JR, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N Engl J Med.* 2005;352(17):1779-90.
32. Levine RL, Wadleigh M, Cools J, Ebert BL, Wernig G, Huntly BJ, et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell.* 2005;7(4):387-97.
33. Baxter EJ, Scott LM, Campbell PJ, East C, Fourouclas N, Swanton S, et al. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *The Lancet.* 2005;365(9464):1054-61.
34. James C, Ugo V, Le Couédic J-P, Staerk J, Delhommeau F, Lacout C, et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature.* 2005;434(7037):1144-8.
35. Pikman Y, Lee BH, Mercher T, McDowell E, Ebert BL, Gozo M, et al. MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. *PLoS Med.* 2006;3(7):e270.
36. Klampfl T, Gisslinger H, Harutyunyan AS, Nivarthi H, Rumi E, Milosevic JD, et al. Somatic Mutations of Calreticulin in Myeloproliferative Neoplasms. *New England Journal of Medicine.* 2013;369(25):2379-90.
37. Nangalia J, Massie CE, Baxter EJ, Nice FL, Gundem G, Wedge DC, et al. Somatic CALR Mutations in Myeloproliferative Neoplasms with Nonmutated JAK2. *New England Journal of Medicine.* 2013;369(25):2391-405.
38. Rawlings JS, Rosler KM, Harrison DA. The JAK/STAT signaling pathway. *J Cell Sci.* 2004;117(Pt 8):1281-3.
39. Shaheen M, Broxmeyer HE. Chapter 16 - Cytokine/Receptor Families and Signal Transduction. In: Hoffman R, Benz EJ, Silberstein LE, Heslop HE, Weitz JI, Anastasi J, et al., editors. *Hematology (Seventh Edition)*: Elsevier; 2018. p. 163-75.
40. Vainchenker W, Kralovics R. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. *Blood.* 2017;129(6):667-79.
41. Liu P, Cheng H, Roberts TM, Zhao JJ. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat Rev Drug Discov.* 2009;8(8):627-44.
42. Milosevic Feenstra JD, Nivarthi H, Gisslinger H, Leroy E, Rumi E, Chachoua I, et al. Whole-exome sequencing identifies novel MPL and JAK2 mutations in triple-negative myeloproliferative neoplasms. *Blood.* 2016;127(3):325-32.

43. Rampal R, Al-Shahrour F, Abdel-Wahab O, Patel JP, Brunel JP, Mermel CH, et al. Integrated genomic analysis illustrates the central role of JAK-STAT pathway activation in myeloproliferative neoplasm pathogenesis. *Blood*. 2014;123(22):e123-33.
44. Tefferi A, Lasho TL, Tischer A, Wassie EA, Finke CM, Belachew AA, et al. The prognostic advantage of calreticulin mutations in myelofibrosis might be confined to type 1 or type 1-like CALR variants. *Blood*. 2014;124(15):2465.
45. Lim K-H, Lin H-C, Chiang Y-H, Wang W-T, Lin C-S, Chang Y-C, et al. Calr Mutations in Essential Thrombocythemia: Frequency, Clinical Correlation and Screening By High-Resolution Melting Analysis. *Blood*. 2014;124(21):1838.
46. Pardanani AD, Levine RI Fau - Lasho T, Lasho T Fau - Pikman Y, Pikman Y Fau - Mesa RA, Mesa Ra Fau - Wadleigh M, Wadleigh M Fau - Steensma DP, et al. MPL515 mutations in myeloproliferative and other myeloid disorders: a study of 1182 patients. (0006-4971 (Print)).
47. Silvennoinen O, Hubbard SR. Molecular insights into regulation of JAK2 in myeloproliferative neoplasms. *Blood*. 2015;125(22):3388-92.
48. Min X, Ungureanu D, Maxwell S, Hammaren H, Thibault S, Hillert EK, et al. Structural and Functional Characterization of the JH2 Pseudokinase Domain of JAK Family Tyrosine Kinase 2 (TYK2). *J Biol Chem*. 2015;290(45):27261-70.
49. Babon JJ, Lucet IS, Murphy JM, Nicola NA, Varghese LN. The molecular regulation of Janus kinase (JAK) activation. *Biochem J*. 2014;462(1):1-13.
50. Najfeld V, Montella L, Scalise A, Fruchtmann S. Exploring polycythaemia vera with fluorescence in situ hybridization: additional cryptic 9p is the most frequent abnormality detected*. *British Journal of Haematology*. 2002;119(2):558-66.
51. Vannucchi AM, Antonioli E, Guglielmelli P, Rambaldi A, Barosi G, Marchioli R, et al. Clinical profile of homozygous JAK2 617V>F mutation in patients with polycythemia vera or essential thrombocythemia. *Blood*. 2007;110(3):840-6.
52. Godfrey AL, Chen E, Pagano F, Ortmann CA, Silber Y, Bellosillo B, et al. JAK2V617F homozygosity arises commonly and recurrently in PV and ET, but PV is characterized by expansion of a dominant homozygous subclone. *Blood*. 2012;120(13):2704-7.
53. Li J, Kent DG, Godfrey AL, Manning H, Nangalia J, Aziz A, et al. JAK2V617F homozygosity drives a phenotypic switch in myeloproliferative neoplasms, but is insufficient to sustain disease. *Blood*. 2014;123(20):3139-51.
54. Passamonti F, Elena C Fau - Schnittger S, Schnittger S Fau - Skoda RC, Skoda Rc Fau - Green AR, Green Ar Fau - Girodon F, Girodon F Fau - Kiladjian J-J, et al. Molecular and clinical features of the myeloproliferative neoplasm associated with JAK2 exon 12 mutations. (1528-0020 (Electronic)).
55. Yao H, Ma Y, Hong Z, Zhao L, Monaghan SA, Hu MC, et al. Activating JAK2 mutants reveal cytokine receptor coupling differences that impact outcomes in myeloproliferative neoplasm. *Leukemia*. 2017;31(10):2122-31.
56. He X, Chen Z, Jiang Y, Qiu X, Zhao X. Different mutations of the human c-mpl gene indicate distinct haematopoietic diseases. *J Hematol Oncol*. 2013;6:11-.

57. Michalak M, Corbett EF, Mesaeli N, Nakamura K, Opas M. Calreticulin: one protein, one gene, many functions. *Biochem J*. 1999;344 Pt 2:281-92.
58. Corbett EF, Oikawa K, Francois P, Tessier DC, Kay C, Bergeron JJM, et al. Ca²⁺ regulation of interactions between endoplasmic reticulum chaperones. *Journal of Biological Chemistry*. 1999;274(10):6203-11.
59. Elf S, Abdelfattah NS, Chen E, Perales-Paton J, Rosen EA, Ko A, et al. Mutant Calreticulin Requires Both Its Mutant C-terminus and the Thrombopoietin Receptor for Oncogenic Transformation. *Cancer Discov*. 2016;6(4):368-81.
60. Chachoua I, Pecquet C, El-Khoury M, Nivarthi H, Albu RI, Marty C, et al. Thrombopoietin receptor activation by myeloproliferative neoplasm associated calreticulin mutants. *Blood*. 2016;127(10):1325-35.
61. Imai M, Araki M, Komatsu N. Somatic mutations of calreticulin in myeloproliferative neoplasms. *Int J Hematol*. 2017;105(6):743-7.
62. Grinfeld J, Nangalia J, Baxter EJ, Wedge DC, Angelopoulos N, Cantrill R, et al. Classification and Personalized Prognosis in Myeloproliferative Neoplasms. *New England Journal of Medicine*. 2018;379(15):1416-30.
63. Tefferi A, Lasho TL, Guglielmelli P, Finke CM, Rotunno G, Elala Y, et al. Targeted deep sequencing in polycythemia vera and essential thrombocythemia. *Blood Adv*. 2016;1(1):21-30.
64. Tefferi A, Lasho TL, Finke CM, Elala Y, Hanson CA, Ketterling RP, et al. Targeted deep sequencing in primary myelofibrosis. *Blood Advances*. 2016;1(2):105.
65. Patel KP, Newberry KJ, Luthra R, Jabbour E, Pierce S, Cortes J, et al. Correlation of mutation profile and response in patients with myelofibrosis treated with ruxolitinib. *Blood*. 2015;126(6):790.
66. Delhommeau F, Dupont S, Della Valle V, James C, Trannoy S, Masse A, et al. Mutation in TET2 in myeloid cancers. *N Engl J Med*. 2009;360(22):2289-301.
67. Moran-Crusio K, Reavie L, Shih A, Abdel-Wahab O, Ndiaye-Lobry D, Lobry C, et al. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell*. 2011;20(1):11-24.
68. Ko M, Bandukwala HS, An J, Lamperti ED, Thompson EC, Hastie R, et al. Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proceedings of the National Academy of Sciences*. 2011;108(35):14566.
69. Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*. 2010;468:839.
70. Ortmann CA, Kent DG, Nangalia J, Silber Y, Wedge DC, Grinfeld J, et al. Effect of Mutation Order on Myeloproliferative Neoplasms. *New England Journal of Medicine*. 2015;372(7):601-12.
71. Yang L, Rau R, Goodell MA. DNMT3A in haematological malignancies. *Nat Rev Cancer*. 2015;15(3):152-65.

72. Russler-Germain DA, Spencer DH, Young MA, Lamprecht TL, Miller CA, Fulton R, et al. The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers. *Cancer Cell*. 2014;25(4):442-54.
73. Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet*. 2011;44(1):23-31.
74. Abdel-Wahab O, Pardanani A, Patel J, Wadleigh M, Lasho T, Heguy A, et al. Concomitant analysis of EZH2 and ASXL1 mutations in myelofibrosis, chronic myelomonocytic leukemia and blast-phase myeloproliferative neoplasms. *Leukemia*. 2011;25(7):1200-2.
75. Abdel-Wahab O, Adli M, LaFave LM, Gao J, Hricik T, Shih AH, et al. ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer Cell*. 2012;22(2):180-93.
76. Yang Y, Akada H, Nath D, Hutchison RE, Mohi G. Loss of Ezh2 cooperates with Jak2V617F in the development of myelofibrosis in a mouse model of myeloproliferative neoplasm. *Blood*. 2016;127(26):3410.
77. Vannucchi AM, Lasho TL, Guglielmelli P, Biamonte F, Pardanani A, Pereira A, et al. Mutations and prognosis in primary myelofibrosis. *Leukemia*. 2013;27:1861.
78. Papaemmanuil E, Cazzola M, Boultonwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384-95.
79. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64-9.
80. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*. 2011;44(1):53-7.
81. Malcovati L, Papaemmanuil E, Bowen DT, Boultonwood J, Della Porta MG, Pascutto C, et al. Clinical significance of SF3B1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms. *Blood*. 2011;118(24):6239.
82. Matera AG, Wang Z. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*. 2014;15:108.
83. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nature Genetics*. 2011;44:53.
84. Meggendorfer M, Roller A, Haferlach T, Eder C, Dicker F, Grossmann V, et al. SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). *Blood*. 2012;120(15):3080.
85. Tefferi A, Finke CM, Lasho TL, Wassie EA, Knudson R, Ketterling RP, et al. U2AF1 mutations in primary myelofibrosis are strongly associated with anemia and

thrombocytopenia despite clustering with JAK2V617F and normal karyotype. *Leukemia*. 2013;28:431.

86. Wahl MC, Will CL, Luhrmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell*. 2009;136(4):701-18.

87. Wu S, Romfo CM, Nilsen TW, Green MR. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature*. 1999;402(6763):832-5.

88. Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, et al. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Research*. 2015;25(1):14-26.

89. Okeyo-Owuor T, White BS, Chatrikhi R, Mohan DR, Kim S, Griffith M, et al. U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. *Leukemia*. 2015;29(4):909-17.

90. Shirai CL, Ley JN, White BS, Kim S, Tibbitts J, Shao J, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell*. 2015;27(5):631-43.

91. Fei DL, Zhen T, Durham B, Ferrarone J, Zhang T, Garrett L, et al. Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. *Proc Natl Acad Sci U S A*. 2018;115(44):E10437-E46.

92. Yip BH, Steeples V, Repapi E, Armstrong RN, Llorian M, Roy S, et al. The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes. *The Journal of Clinical Investigation*. 2017;127(6):2206-21.

93. Jutzi JS, Bogeska R, Nikoloski G, Schmid CA, Seeger TS, Stegelmann F, et al. MPN patients harbor recurrent truncating mutations in transcription factor NF-E2. *J Exp Med*. 2013;210(5):1003-19.

94. Grand FH, Hidalgo-Curtis CE, Ernst T, Zoi K, Zoi C, McGuire C, et al. Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. *Blood*. 2009;113(24):6182.

95. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013;502(7471):333-9.

96. Kubesova B, Pavlova S, Malcikova J, Kabathova J, Radova L, Tom N, et al. Low-burden TP53 mutations in chronic phase of myeloproliferative neoplasms: association with age, hydroxyurea administration, disease type and JAK2 mutational status. *Leukemia*. 2018;32(2):450-61.

97. Hsu JI, Dayaram T, Tovy A, De Braekeleer E, Jeong M, Wang F, et al. PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. *Cell Stem Cell*. 2018;23(5):700-13.e6.

98. Mead AJ, Mullally A. Myeloproliferative neoplasm stem cells. *Blood*. 2017;129(12):1607-16.

99. Lundberg P, Takizawa H, Kubovcakova L, Guo G, Hao-Shen H, Dirnhofer S, et al. Myeloproliferative neoplasms can be initiated from a single hematopoietic stem cell expressing JAK2-V617F. *J Exp Med*. 2014;211(11):2213-30.
100. Jamieson CHM, Gotlib J, Durocher JA, Chao MP, Mariappan MR, Lay M, et al. The JAK2 V617F mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proceedings of the National Academy of Sciences*. 2006;103(16):6224.
101. Mullally A, Poveromo L, Schneider RK, Al-Shahrour F, Lane SW, Ebert BL. Distinct roles for long-term hematopoietic stem cells and erythroid precursor cells in a murine model of Jak2V617F-mediated polycythemia vera. *Blood*. 2012;120(1):166-72.
102. Chen E, Schneider RK, Breyfogle LJ, Rosen EA, Poveromo L, Elf S, et al. Distinct effects of concomitant Jak2V617F expression and Tet2 loss in mice promote disease progression in myeloproliferative neoplasms. *Blood*. 2015;125(2):327.
103. Sashida G, Wang C, Tomioka T, Oshima M, Aoyama K, Kanai A, et al. The loss of Ezh2 drives the pathogenesis of myelofibrosis and sensitizes tumor-initiating cells to bromodomain inhibition. *The Journal of Experimental Medicine*. 2016;213(8):1459.
104. Shimizu T, Kubovcakova L, Nienhold R, Zmajkovic J, Meyer SC, Hao-Shen H, et al. Loss of Ezh2 synergizes with JAK2V617F in initiating myeloproliferative neoplasms and promoting myelofibrosis. *The Journal of Experimental Medicine*. 2016;213(8):1479.
105. Yang Y, Akada H, Nath D, Hutchison RE, Mohi G. Loss of Ezh2 cooperates with Jak2V617F in the development of myelofibrosis in a mouse model of myeloproliferative neoplasm. *Blood*. 2016;127(26):3410-23.
106. Jamieson CHM, Gotlib J, Durocher JA, Chao MP, Mariappan MR, Lay M, et al. The JAK2V617F mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proceedings of the National Academy of Sciences*. 2006;103(16):6224.
107. Ishii T, Bruno E, Hoffman R, Xu M. Involvement of various hematopoietic-cell lineages by the JAK2V617F mutation in polycythemia vera. *Blood*. 2006;108(9):3128.
108. Delhommeau F, Dupont S, Tonetti C, Masse A, Godin I, Le Couedic JP, et al. Evidence that the JAK2 G1849T (V617F) mutation occurs in a lymphomyeloid progenitor in polycythemia vera and idiopathic myelofibrosis. *Blood*. 2007;109(1):71-7.
109. Reinisch A, Thomas D, Corces MR, Zhang X, Gratzinger D, Hong WJ, et al. A humanized bone marrow ossicle xenotransplantation model enables improved engraftment of healthy and leukemic human hematopoietic cells. *Nat Med*. 2016;22(7):812-21.
110. Nielsen I, Hasselbalch HC. Acute leukemia and myelodysplasia in patients with a Philadelphia chromosome negative chronic myeloproliferative disorder treated with hydroxyurea alone or with hydroxyurea after busulphan. *Am J Hematol*. 2003;74(1):26-31.
111. Hernandez-Boluda JC, Alvarez-Larran A, Gomez M, Angona A, Amat P, Bellosillo B, et al. Clinical evaluation of the European LeukaemiaNet criteria for clinicohaematological response and resistance/intolerance to hydroxycarbamide in essential thrombocythaemia. *Br J Haematol*. 2011;152(1):81-8.

112. Alvarez-Larran A, Pereira A, Cervantes F, Arellano-Rodrigo E, Hernandez-Boluda JC, Ferrer-Marin F, et al. Assessment and prognostic value of the European LeukemiaNet criteria for clinicohematologic response, resistance, and intolerance to hydroxyurea in polycythemia vera. *Blood*. 2012;119(6):1363-9.
113. Sever M, Newberry KJ, Verstovsek S. Therapeutic options for patients with polycythemia vera and essential thrombocythemia refractory/resistant to hydroxyurea. *Leuk Lymphoma*. 2014;55(12):2685-90.
114. Kiladjian JJ, Giraudier S, Cassinat B. Interferon-alpha for the therapy of myeloproliferative neoplasms: targeting the malignant clone. *Leukemia*. 2015;30:776.
115. Hasselbalch HC, Holmström MO. Perspectives on interferon-alpha in the treatment of polycythemia vera and related myeloproliferative neoplasms: minimal residual disease and cure? *Semin Immunopathol*. 2019;41(1):5-19.
116. Biron CA. Interferons alpha and beta as immune regulators - A new look. *Immunity*. 2001;14(6):661-4.
117. Chawla-Sarkar M, Lindner DJ, Liu YF, Williams BR, Sen GC, Silverman RH, et al. Apoptosis and interferons: role of interferon-stimulated genes as mediators of apoptosis. *Apoptosis*. 2003;8(3):237-49.
118. Mullally A, Bruedigam C, Poveromo L, Heidel FH, Purdon A, Vu T, et al. Depletion of Jak2V617F myeloproliferative neoplasm-propagating stem cells by interferon-alpha in a murine model of polycythemia vera. *Blood*. 2013;121(18):3692-702.
119. Kiladjian J-J, Cassinat B, Chevret S, Turlure P, Cambier N, Roussel M, et al. Pegylated interferon-alfa-2a induces complete hematologic and molecular responses with low toxicity in polycythemia vera. *Blood*. 2008;112(8):3065.
120. Verger E, Cassinat B, Chauveau A, Dosquet C, Giraudier S, Schlageter MH, et al. Clinical and molecular response to interferon-alpha therapy in essential thrombocythemia patients with CALR mutations. *Blood*. 2015;126(24):2585-91.
121. Ishii T, Xu M, Zhao Y, Hu WY, Ciurea S, Bruno E, et al. Recurrence of clonal hematopoiesis after discontinuing pegylated recombinant interferon- α 2a in a patient with polycythemia vera. *Leukemia*. 2006;21:373.
122. Mazur EM, Rosmarin AG, Sohl PA, Newton JL, Narendran A. Analysis of the mechanism of anagrelide-induced thrombocytopenia in humans. *Blood*. 1992;79(8):1931-7.
123. Harrison CN, Campbell PJ, Buck G, Wheatley K, East CL, Bareford D, et al. Hydroxyurea Compared with Anagrelide in High-Risk Essential Thrombocythemia. *New England Journal of Medicine*. 2005;353(1):33-45.
124. Quintas-Cardama A, Vaddi K, Liu P, Manshour T, Li J, Scherle PA, et al. Preclinical characterization of the selective JAK1/2 inhibitor INCB018424: therapeutic implications for the treatment of myeloproliferative neoplasms. *Blood*. 2010;115(15):3109-17.

125. Harrison C, Kiladjian J-J, Al-Ali HK, Gisslinger H, Waltzman R, Stalbovskaya V, et al. JAK Inhibition with Ruxolitinib versus Best Available Therapy for Myelofibrosis. *New England Journal of Medicine*. 2012;366(9):787-98.
126. Harrison CN, Vannucchi AM, Kiladjian JJ, Al-Ali HK, Gisslinger H, Knoop L, et al. Long-term findings from COMFORT-II, a phase 3 study of ruxolitinib vs best available therapy for myelofibrosis. *Leukemia*. 2016;30:1701.
127. Quintás-Cardama A, Kantarjian H, Cortes J, Verstovsek S. Janus kinase inhibitors for the treatment of myeloproliferative neoplasias and beyond. *Nature Reviews Drug Discovery*. 2011;10:127.
128. Deininger M, Radich J, Burn TC, Huber R, Paranagama D, Verstovsek S. The effect of long-term ruxolitinib treatment on JAK2p.V617F allele burden in patients with myelofibrosis. *Blood*. 2015;126(13):1551-4.
129. Guglielmelli P, Rotunno G, Bogani C, Mannarelli C, Giunti L, Provenzano A, et al. Ruxolitinib is an effective treatment for CALR-positive patients with myelofibrosis. *Br J Haematol*. 2016;173(6):938-40.
130. Salit RB, Deeg HJ. Role of hematopoietic stem cell transplantation in patients with myeloproliferative disease. *Hematol Oncol Clin North Am*. 2014;28(6):1023-35.
131. Kröger NM, Deeg JH, Olavarria E, Niederwieser D, Bacigalupo A, Barbui T, et al. Indication and management of allogeneic stem cell transplantation in primary myelofibrosis: a consensus process by an EBMT/ELN international working group. *Leukemia*. 2015;29:2126.
132. Deschler B, Lubbert M. Acute myeloid leukemia: epidemiology and etiology. *Cancer*. 2006;107(9):2099-107.
133. Estey E, Döhner H. Acute myeloid leukaemia. *The Lancet*. 2006;368(9550):1894-907.
134. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2017;129(4):424-47.
135. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391.
136. Boddü P, Kantarjian HM, Garcia-Manero G, Ravandi F, Verstovsek S, Jabbour E, et al. Treated secondary acute myeloid leukemia: a distinct high-risk subset of AML with adverse prognosis. *Blood Adv*. 2017;1(17):1312-23.
137. Cancer Genome Atlas Research N, Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-74.
138. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine*. 2016;374(23):2209-21.

139. Mrózek K, Heerema NA, Bloomfield CD. Cytogenetics in acute leukemia. *Blood Reviews*. 2004;18(2):115-36.
140. Corces-Zimmerman MR, Hong WJ, Weissman IL, Medeiros BC, Majeti R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc Natl Acad Sci U S A*. 2014;111(7):2548-53.
141. Shlush LI, Zandi S, Mitchell A, Chen WC, Brandwein JM, Gupta V, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*. 2014;506(7488):328-33.
142. Jan M, Snyder TM, Corces-Zimmerman MR, Vyas P, Weissman IL, Quake SR, et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med*. 2012;4(149):149ra18.
143. Miyamoto T, Weissman IL, Akashi K. AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation. *Proc Natl Acad Sci U S A*. 2000;97(13):7521-6.
144. Genovese G, Kähler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoun SF, et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*. 2014;371(26):2477-87.
145. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *The New England journal of medicine*. 2014;371(26):2488-98.
146. Steensma DP. Clinical consequences of clonal hematopoiesis of indeterminate potential. *Blood advances*. 2018;2(22):3404-10.
147. Corces MR, Chang HY, Majeti R. Preleukemic Hematopoietic Stem Cells in Human Acute Myeloid Leukemia. *Front Oncol*. 2017;7:263.
148. Thomas D, Majeti R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood*. 2017;129(12):1577-85.
149. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*. 2012;481:506.
150. Shlush LI, Mitchell A, Heisler L, Abelson S, Ng SWK, Trotman-Grant A, et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*. 2017;547(7661):104-8.
151. Dohner H, Weisdorf DJ, Bloomfield CD. Acute Myeloid Leukemia. *N Engl J Med*. 2015;373(12):1136-52.
152. Grimwade D, Freeman SD. Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for "prime time"? *Blood*. 2014;124(23):3345-55.
153. Perl AE. The role of targeted therapy in the management of patients with AML. *Blood Advances*. 2017;1(24):2281.

154. Yen K, Travins J, Wang F, David MD, Artin E, Straley K, et al. AG-221, a First-in-Class Therapy Targeting Acute Myeloid Leukemia Harboring Oncogenic IDH2 Mutations. *Cancer Discovery*. 2017;7(5):478.
155. Greenberg P, Cox C, LeBeau MM, Fenaux P, Morel P, Sanz G, et al. International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood*. 1997;89(6):2079-88.
156. Kennedy JA, Atenafu EG, Messner HA, Craddock KJ, Brandwein JM, Lipton JH, et al. Treatment outcomes following leukemic transformation in Philadelphia-negative myeloproliferative neoplasms. *Blood*. 2013;121(14):2725-33.
157. McNamara CJ, Panzarella T, Kennedy JA, Arruda A, Claudio JO, Daher-Reyes G, et al. The mutational landscape of accelerated- and blast-phase myeloproliferative neoplasms impacts patient outcomes. *Blood advances*. 2018;2(20):2658-71.
158. Finazzi G, Caruso V, Marchioli R, Capnist G, Chisesi T, Finelli C, et al. Acute leukemia in polycythemia vera: an analysis of 1638 patients enrolled in a prospective observational study. *Blood*. 2005;105(7):2664-70.
159. Tefferi A, Mudireddy M, Mannelli F, Begna KH, Patnaik MM, Hanson CA, et al. Blast phase myeloproliferative neoplasm: Mayo-AGIMM study of 410 patients from two separate cohorts. *Leukemia*. 2018;32(5):1200-10.
160. Abdel-Wahab O, Manshouri T, Patel J, Harris K, Yao J, Hedvat C, et al. Genetic analysis of transforming events that convert chronic myeloproliferative neoplasms to leukemias. *Cancer Res*. 2010;70(2):447-52.
161. Rampal R, Ahn J, Abdel-Wahab O, Nahas M, Wang K, Lipson D, et al. Genomic and functional analysis of leukemic transformation of myeloproliferative neoplasms. *Proc Natl Acad Sci U S A*. 2014;111(50):E5401-10.
162. Zhang S-J, Rampal R, Manshouri T, Patel J, Mensah N, Kayserian A, et al. Genetic analysis of patients with leukemic transformation of myeloproliferative neoplasms shows recurrent SRSF2 mutations that are associated with adverse outcome. *Blood*. 2012;119(19):4480.
163. Theocharides A, Boissinot M, Girodon F, Garand R, Teo S-S, Lippert E, et al. Leukemic blasts in transformed JAK2-V617F-positive myeloproliferative disorders are frequently negative for the JAK2-V617F mutation. *Blood*. 2007;110(1):375.
164. Chen J, Kao YR, Sun D, Todorova TI, Reynolds D, Narayanagari SR, et al. Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level. *Nat Med*. 2019;25(1):103-10.
165. Yogarajah M, Tefferi A. Leukemic Transformation in Myeloproliferative Neoplasms: A Literature Review on Risk, Characteristics, and Outcome. *Mayo Clin Proc*. 2017;92(7):1118-28.
166. Hultcrantz M, Kristinsson SY, Andersson TM, Landgren O, Eloranta S, Derolf AR, et al. Patterns of survival among patients with myeloproliferative neoplasms diagnosed in Sweden from 1973 to 2008: a population-based study. *J Clin Oncol*. 2012;30(24):2995-3001.

167. Tam CS, Nussenzweig RM, Popat U, Bueso-Ramos CE, Thomas DA, Cortes JA, et al. The natural history and treatment outcome of blast phase BCR-ABL-myeloproliferative neoplasms. *Blood*. 2008;112(5):1628-37.
168. Mesa RA, Li CY, Ketterling RP, Schroeder GS, Knudson RA, Tefferi A. Leukemic transformation in myelofibrosis with myeloid metaplasia: a single-institution experience with 91 cases. *Blood*. 2005;105(3):973-7.
169. Perez C, Pascual M, Martin-Subero JI, Bellosillo B, Segura V, Delabesse E, et al. Aberrant DNA methylation profile of chronic and transformed classic Philadelphia-negative myeloproliferative neoplasms. *Haematologica*. 2013;98(9):1414-20.
170. Wang JC, Chen W, Nallusamy S, Chen C, Novetsky AD. Hypermethylation of the P15INK4b and P16INK4a in agnogenic myeloid metaplasia (AMM) and AMM in leukaemic transformation. *British Journal of Haematology*. 2002;116(3):582-6.
171. Griffiths EA, Gore SD. Epigenetic therapies in MDS and AML. *Adv Exp Med Biol*. 2013;754:253-83.
172. Issa J-PJ, Kantarjian HM, Kirkpatrick P. Azacitidine. *Nature Reviews Drug Discovery*. 2005;4(4):275-6.
173. Quintás-Cardama A, Tong W, Kantarjian H, Thomas D, Ravandi F, Kornblau S, et al. A phase II study of 5-azacitidine for patients with primary and post-essential thrombocythemia/polycythemia vera myelofibrosis. *Leukemia*. 2008;22:965.
174. Yang AS, Doshi KD, Choi SW, Mason JB, Mannari RK, Gharybian V, et al. DNA methylation changes after 5-aza-2'-deoxycytidine therapy in patients with leukemia. *Cancer Res*. 2006;66(10):5495-503.
175. Badar T, Kantarjian HM, Ravandi F, Jabbour E, Borthakur G, Cortes JE, et al. Therapeutic benefit of decitabine, a hypomethylating agent, in patients with high-risk primary myelofibrosis and myeloproliferative neoplasm in accelerated or blastic/acute myeloid leukemia phase. *Leuk Res*. 2015;39(9):950-6.
176. Rampal RK, Mascarenhas JO, Kosiorek HE, Price L, Berenzon D, Hexner E, et al. Safety and efficacy of combined ruxolitinib and decitabine in accelerated and blast-phase myeloproliferative neoplasms. *Blood Advances*. 2018;2(24):3572.
177. Chihara D, Kantarjian HM, Newberry KJ, Ravandi F, Daver NG, Bose P, et al. Survival Outcome of Patients with Acute Myeloid Leukemia Transformed from Myeloproliferative Neoplasms. *Blood*. 2016;128(22):1940.
178. Harutyunyan A, Klampfl T, Cazzola M, Kralovics R. p53 Lesions in Leukemic Transformation. *New England Journal of Medicine*. 2011;364(5):488-90.
179. Bejar R, Stevenson K, Abdel-Wahab O, Galili N, Nilsson B, Garcia-Manero G, et al. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med*. 2011;364(26):2496-506.
180. Haase D, Stevenson KE, Neuberg D, Maciejewski JP, Nazha A, Sekeres MA, et al. TP53 mutation status divides myelodysplastic syndromes with complex karyotypes into distinct prognostic subgroups. *Leukemia*. 2019;33(7):1747-58.

181. Haferlach C, Dicker F, Herholz H, Schnittger S, Kern W, Haferlach T. Mutations of the TP53 gene in acute myeloid leukemia are strongly associated with a complex aberrant karyotype. *Leukemia*. 2008;22(8):1539-41.
182. Lundberg P, Karow A, Nienhold R, Looser R, Hao-Shen H, Nissen I, et al. Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood*. 2014;123(14):2220-8.
183. Sterkers Y, Preudhomme C, Lai JL, Demory JL, Caulier MT, Wattel E, et al. Acute myeloid leukemia and myelodysplastic syndromes following essential thrombocythemia treated with hydroxyurea: high proportion of cases with 17p deletion. *Blood*. 1998;91(2):616-22.
184. Bjorkholm M, Derolf AR, Hultcrantz M, Kristinsson SY, Ekstrand C, Goldin LR, et al. Treatment-related risk factors for transformation to acute myeloid leukemia and myelodysplastic syndromes in myeloproliferative neoplasms. *J Clin Oncol*. 2011;29(17):2410-5.
185. Aubrey BJ, Strasser A, Kelly GL. Tumor-Suppressor Functions of the TP53 Pathway. *Cold Spring Harb Perspect Med*. 2016;6(5).
186. Zhou X, Cao B, Lu H. Negative auto-regulators trap p53 in their web. *Journal of Molecular Cell Biology*. 2017;9(1):62-8.
187. Riley T, Sontag E, Chen P, Levine A. Transcriptional control of human p53-regulated genes. *Nature Reviews Molecular Cell Biology*. 2008;9:402.
188. Rucker FG, Schlenk RF, Bullinger L, Kayser S, Teleanu V, Kett H, et al. TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. *Blood*. 2012;119(9):2114-21.
189. Hou HA, Chou WC, Kuo YY, Liu CY, Lin LI, Tseng MH, et al. TP53 mutations in de novo acute myeloid leukemia patients: longitudinal follow-ups show the mutation is stable during disease evolution. *Blood Cancer Journal*. 2015;5:e331.
190. Zhao Z, Zuber J, Diaz-Flores E, Lintault L, Kogan SC, Shannon K, et al. p53 loss promotes acute myeloid leukemia by enabling aberrant self-renewal. *Genes Dev*. 2010;24(13):1389-402.
191. Loizou E, Banito A, Livshits G, Ho Y-J, Koche RP, Sanchez-Rivera FJ, et al. A gain-of-function p53 mutant oncogene promotes cell fate plasticity and myeloid leukemia through the pluripotency factor Foxh1. *Cancer Discovery*. 2019;CD-18-1391.
192. Boettcher S, Miller PG, Sharma R, McConkey M, Leventhal M, Krivtsov AV, et al. A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science*. 2019;365(6453):599-604.
193. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: New online mutation analysis and recommendations to users. *Human Mutation*. 2002;19(6):607-14.
194. Harutyunyan A, Klampfl T, Cazzola M, Kralovics R. p53 lesions in leukemic transformation. *N Engl J Med*. 2011;364(5):488-90.

195. Welch JS, Petti AA, Miller CA, Fronick CC, O’Laughlin M, Fulton RS, et al. TP53 and Decitabine in Acute Myeloid Leukemia and Myelodysplastic Syndromes. *New England Journal of Medicine*. 2016;375(21):2023-36.
196. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168(4):613-28.
197. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*. 2015;27(1):15-26.
198. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. *Nat Med*. 2015;21(8):846-53.
199. Polyak K, Haviv I, Campbell IG. Co-evolution of tumor cells and their microenvironment. *Trends Genet*. 2009;25(1):30-8.
200. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer*. 2011;11:726.
201. Povinelli BJ, Rodriguez-Meira A, Mead AJ. Single cell analysis of normal and leukemic hematopoiesis. . *Molecular Aspects of Medicine*. 2017(1872-9452 (Electronic)).
202. Adolfsson J, Månsson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, et al. Identification of Flt3+ Lympho-Myeloid Stem Cells Lacking Erythro-Megakaryocytic Potential: A Revised Road Map for Adult Blood Lineage Commitment. *Cell*. 2005;121(2):295-306.
203. Bengtsson M, Stahlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res*. 2005;15(10):1388-92.
204. Ståhlberg A, Kubista M. Technical aspects and recommendations for single-cell qPCR. *Molecular Aspects of Medicine*. 2018;59:28-35.
205. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*. 2013;31:748.
206. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009;6(5):377-82.
207. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*. 2018;13:599.
208. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell*. 2017;65(4):631-43 e4.
209. Natarajan KN, Miao Z, Jiang M, Huang X, Zhou H, Xie J, et al. Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol*. 2019;20(1):70.

210. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods*. 2017;14(4):381-7.
211. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*. 2013;10(11):1096-8.
212. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
213. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-14.
214. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*. 2001;30(4):892-7.
215. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*. 2012;30(8):777-82.
216. Harbers M, Kato S, de Hoon M, Hayashizaki Y, Carninci P, Plessy C. Comparison of RNA- or LNA-hybrid oligonucleotides in template-switching reactions for high-speed sequencing library preparation. *BMC Genomics*. 2013;14:665-.
217. Kapteyn J, He R, McDowell ET, Gang DR. Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics*. 2010;11:413.
218. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research*. 2014;24(12):2033-40.
219. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*. 2018;174(5):1293-308.e36.
220. Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol*. 2018.
221. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature Communications*. 2019;10(1):3120.
222. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, et al. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci U S A*. 2018;115(39):9726-31.
223. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, et al. Whole-genome multiple displacement amplification from single cells. *Nature Protocols*. 2006;1(4):1965-70.

224. Hosokawa M, Nishikawa Y, Kogawa M, Takeyama H. Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Scientific Reports*. 2017;7(1):5199.
225. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*. 2014;512(7513):155-60.
226. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. *Genome Biol*. 2016;17(1474-760X (Electronic)):86.
227. Ross EM, Markowitz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*. 2016;17(1):69.
228. Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods*. 2017;14(5):491-3.
229. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*. 2012;148(5):873-85.
230. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*. 2012;148(5):886-95.
231. Potter NE, Ermini L, Papaemmanuil E, Cazzaniga G, Vijayaraghavan G, Tittley I, et al. Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Research*. 2013;23(12):2115-25.
232. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature Genetics*. 2014;46:116.
233. Paguirigan AL, Smith J, Meshinchi S, Carroll M, Maley C, Radich JP. Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. *Sci Transl Med*. 2015;7(281):281re2.
234. Pellegrino M, Sciambi A, Treusch S, Durruthy-Durruthy R, Gokhale K, Jacob J, et al. High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Research*. 2018;28(9):1345-52.
235. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20(5):257-72.
236. Shema E, Bernstein BE, Buenrostro JD. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat Genet*. 2019;51(1):19-25.
237. Hu Y, An Q, Sheu K, Trejo B, Fan S, Guo Y. Single Cell Multi-Omics Technology: Methodology and Application. *Frontiers in Cell and Developmental Biology*. 2018;6:28.
238. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*. 2016;539(7628):309-13.

239. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014;344(6190):1396-401.
240. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*. 2017;355(6332):eaai8478.
241. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods*. 2015;12(6):519-22.
242. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*. 2016;26:304.
243. Han KY, Kim KT, Joung JG, Son DS, Kim YJ, Jo A, et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res*. 2018;28(1):75-87.
244. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol*. 2015;33(3):285-9.
245. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat Methods*. 2016;13(10):833-6.
246. Warfvinge R, Geirons Ulfsson L, Sommarin MNE, Lang S, Karlsson C, Roschupkina T, et al. Single-cell molecular analysis defines therapy response and immunophenotype of stem cell subpopulations in CML. *Blood*. 2017:blood-2016-07-728873.
247. Giustacchini A, Thongjuea S, Barkas N, Woll PS, Povinelli BJ, Booth CAG, et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med*. 2017;23(6):692-702.
248. Nam AS, Kim KT, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*. 2019.
249. Petti AA, Williams SR, Miller CA, Fiddes IT, Srivatsan SN, Chen DY, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun*. 2019;10(1):3660.
250. van Galen P, Hovestadt V, Wadsworth li MH, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell*. 2019;176(6):1265-81 e24.
251. Kong SL, Li H, Tai JA, Courtois ET, Poh HM, Lau DP, et al. Concurrent Single-Cell RNA and Targeted DNA Sequencing on an Automated Platform for Comeasurement of Genomic and Transcriptomic Signatures. *Clin Chem*. 2019;65(2):272-81.

252. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*. 2018;9(1):781.
253. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. 2018;361(6409):1380.
254. Rooijers K, Markodimitraki CM, Rang FJ, de Vries SS, Chialastri A, de Luca KL, et al. Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nature Biotechnology*. 2019;37(7):766-72.
255. van Steensel B, Delrow J, Henikoff S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nature Genetics*. 2001;27(3):304-8.
256. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*. 2019;177(7):1873-87.e17.
257. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, III, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-902.e21.
258. Watcham S, Kucinski I, Gottgens B. New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. *Blood*. 2019;133(13):1415-26.
259. Dahlin JS, Hamey FK, Pijuan-Sala B, Shepherd M, Lau WWY, Nestorowa S, et al. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood*. 2018;131(21):e1-e11.
260. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, Panero R, Patel SH, Jankovic M, et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature*. 2018;553(7687):212-6.
261. Notta F, Zandi S, Takayama N, Dobson S, Gan OI, Wilson G, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*. 2016;351(6269):aab2116.
262. Psaila B, Barkas N, Iskander D, Roy A, Anderson S, Ashley N, et al. Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol*. 2016;17:83.
263. Psaila B, Wang G, Rodriguez Meira A, Heuston EF, Li R, Sullivan J, et al. Single-cell analyses reveal aberrant pathways for megakaryocyte-biased hematopoiesis in myelofibrosis and identify mutant clone-specific targets. *bioRxiv*. 2019:642819.
264. Wilson Nicola K, Kent David G, Buettner F, Shehata M, Macaulay Iain C, Calero-Nieto Fernando J, et al. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell*. 2015;16(6):712-24.
265. Smith CC, Paguirigan A, Jeschke GR, Lin KC, Massi E, Tarver T, et al. Heterogeneous resistance to quizartinib in acute myeloid leukemia revealed by single-cell analysis. *Blood*. 2017;130(1):48-58.

266. McMahon CM, Ferng T, Canaani J, Wang ES, Morrisette JJ, Eastburn DJ, et al. Clonal selection with Ras pathway activation mediates secondary clinical resistance to selective FLT3 inhibition in acute myeloid leukemia. *Cancer Discov.* 2019.
267. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet.* 2016;48(10):1193-203.
268. Farlik M, Halbritter F, Müller F, Choudry FA, Ebert P, Klughammer J, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell.* 2016;19(6):808-22.
269. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature.* 2015;523:486.
270. Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Parks B, et al. A single cell framework for multi-omic analysis of disease identifies malignant regulatory signatures in mixed phenotype acute leukemia. *bioRxiv.* 2019.
271. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology.* 2019;37(8):925-36.
272. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology.* 2019;37(8):916-24.
273. Hamblin A, Burns A, Tham C, Clifford R, Robbe P, Timbs A, et al. Development and Evaluation of the Clinical Utility of a Next Generation Sequencing (NGS) Tool for Myeloid Disorders. *Blood.* 2014;124(21):2373.
274. Woll Petter S, Kjällquist U, Chowdhury O, Doolittle H, Wedge David C, Thongjuea S, et al. Myelodysplastic Syndromes Are Propagated by Rare and Distinct Human Cancer Stem Cells In Vivo. *Cancer Cell.* 2014;25(6):794-808.
275. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods.* 2017;14(10):959-62.
276. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2012;29(1):15-21.
277. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30(7):923-30.
278. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods.* 2015;12:357.
279. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology.* 2015;33:290.

280. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. 2017;14:417.
281. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res*. 2012;40(22):11189-201.
282. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*. 2017;18(1):123.
283. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189-96.
284. Xie H, Xu J, Hsu Jessie H, Nguyen M, Fujiwara Y, Peng C, et al. Polycomb Repressive Complex 2 Regulates Normal Hematopoietic Stem Cell Function in a Developmental-Stage-Specific Manner. *Cell Stem Cell*. 2014;14(1):68-80.
285. Zhang X, Su J, Jeong M, Ko M, Huang Y, Park HJ, et al. DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nature genetics*. 2016;48(9):1014-23.
286. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47-e.
287. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*. 2015;33:155.
288. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*. 2015;16(1):241.
289. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*. 2017.
290. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*. 2014;9(6):e98679.
291. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*. 2019;9(1):9354.
292. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. *Science*. 2015;349(6255):1483-9.
293. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*. 2018;15(2):81-94.
294. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*. 2014;158(4):929-44.

295. Potter NE, Ermini L, Papaemmanuil E, Cazzaniga G, Vijayaraghavan G, Tittley I, et al. Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Res.* 2013;23(12):2115-25.
296. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. *Science.* 2013;339(6127):1546.
297. Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science.* 2018;362(6418):1060.
298. Wang L, Fan J, Francis JM, Georgiou G, Hergert S, Li S, et al. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. *Genome Res.* 2017.
299. Crowe JS, Cooper HJ, Smith MA, Sims MJ, Parker D, Gewert D. Improved cloning efficiency of polymerase chain reaction (PCR) products after proteinase K digestion. *Nucleic Acids Res.* 1991;19(1):184.
300. Gierahn TM, Wadsworth li MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods.* 2017;14:395.
301. Rajan P, Elliott DJ, Robson CN, Leung HY. Alternative splicing and biological heterogeneity in prostate cancer. *Nature Reviews Urology.* 2009;6(8):454-60.
302. Louka E, Povinelli B, Meira AR, Buck G, Ashley N, Hamblin A, et al. Heterogeneous disease-propagating stem cells in juvenile myelomonocytic leukemia. *bioRxiv.* 2019:628479.
303. Dimitriou M, Woll PS, Mortera-Blanco T, Karimi M, Wedge DC, Doolittle H, et al. Perturbed hematopoietic stem and progenitor cell hierarchy in myelodysplastic syndromes patients with monosomy 7 as the sole cytogenetic abnormality. *Oncotarget.* 2016;7(45):72685-98.
304. Jiang L, Li Z, Rui L. Leptin Stimulates Both JAK2-dependent and JAK2-independent Signaling Pathways. *The Journal of Biological Chemistry.* 2008;283(42):28066-73.
305. Ruiz-Pérez MV, Henley AB, Arsenian-Henriksson M. The MYCN Protein in Health and Disease. *Genes (Basel).* 2017;8(4):113.
306. Mao Z, Liu C, Lin X, Sun B, Su C. PPP2R5A: A multirole protein phosphatase subunit in regulating cancer development. *Cancer Lett.* 2018;414:222-9.
307. Sanjuan-Pla A, Macaulay IC, Jensen CT, Woll PS, Luis TC, Mead A, et al. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature.* 2013;502(7470):232-6.
308. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* 2006;20(9):1123-36.
309. Chowdhury O, Sullivan J, Barkas N, Buck G, Hamblin A, Tefferi A, et al. Spliceosome Mutations Are Common in MPN-Associated Myelofibrosis with RBC-

Transfusion-Dependence and Correlate with Response to Pomalidomide. *Blood*. 2018;132(Suppl 1):3037.

310. Tefferi A, Finke CM, Lasho TL, Hanson CA, Ketterling RP, Gangat N, et al. U2AF1 mutation types in primary myelofibrosis: phenotypic and prognostic distinctions. *Leukemia*. 2018;32(10):2274-8.

311. Hoermann G, Greiner G, Valent P. Cytokine Regulation of Microenvironmental Cells in Myeloproliferative Neoplasms. *Mediators Inflamm*. 2015;2015:869242-.

312. Heidel FH, Bullinger L, Feng Z, Wang Z, Neff TA, Stein L, et al. Genetic and pharmacologic inhibition of β -catenin targets imatinib-resistant leukemia stem cells in CML. *Cell stem cell*. 2012;10(4):412-24.

313. Keller MD, Rampal RK, Shank K, de Stanchina E, Rose K, Amakye D, et al. Improved Efficacy Of Combination Of JAK2 and Hedgehog Inhibitors In Myelofibrosis. *Blood*. 2013;122(21):666.

314. Chen E, Beer PA, Godfrey AL, Ortmann CA, Li J, Costa-Pereira AP, et al. Distinct Clinical Phenotypes Associated with *JAK2V617F* Reflect Differential STAT1 Signaling. *Cancer Cell*. 2010;18(5):524-35.

315. Hantschel O, Gstoettenbauer A, Colinge J, Kaupe I, Bilban M, Burkard TR, et al. The chemokine interleukin-8 and the surface activation protein CD69 are markers for Bcr-Abl activity in chronic myeloid leukemia. *Mol Oncol*. 2008;2(3):272-81.

316. Mazharian A, Wang YJ, Mori J, Bem D, Finney B, Heising S, et al. Mice lacking the ITIM-containing receptor G6b-B exhibit macrothrombocytopenia and aberrant platelet function. *Sci Signal*. 2012;5(248):ra78.

317. Newland SA, Macaulay IC, Floto AR, de Vet EC, Ouwehand WH, Watkins NA, et al. The novel inhibitory receptor G6B is expressed on the surface of platelets and attenuates platelet function in vitro. *Blood*. 2007;109(11):4806.

318. Tefferi A, Vaidya R, Caramazza D, Finke C, Lasho T, Pardanani A. Circulating interleukin (IL)-8, IL-2R, IL-12, and IL-15 levels are independently prognostic in primary myelofibrosis: a comprehensive cytokine profiling study. *J Clin Oncol*. 2011;29(10):1356-63.

319. Luc S, Huang J, McElldoon JL, Somuncular E, Li D, Rhodes C, et al. Bcl11a Deficiency Leads to Hematopoietic Stem Cell Defects with an Aging-like Phenotype. *Cell Reports*. 2016;16(12):3181-94.

320. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*. 2016;17(1):239.

321. Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*. 2018;36:1197.

322. Riley T, Sontag E, Chen P, Levine A. Transcriptional control of human p53-regulated genes. *Nat Rev Mol Cell Biol*. 2008;9(5):402-12.

323. Brady CA, Jiang D, Mello SS, Johnson TM, Jarvis LA, Kozak MM, et al. Distinct p53 transcriptional programs dictate acute DNA-damage responses and tumor suppression. *Cell*. 2011;145(4):571-83.
324. Jiang D, Brady CA, Johnson TM, Lee EY, Park EJ, Scott MP, et al. Full p53 transcriptional activation potential is dispensable for tumor suppression in diverse lineages. *Proc Natl Acad Sci U S A*. 2011;108(41):17123-8.
325. Goardon N, Marchi E, Atzberger A, Quek L, Schuh A, Soneji S, et al. Coexistence of LMPP-like and GMP-like Leukemia Stem Cells in Acute Myeloid Leukemia. *Cancer Cell*. 2011;19(1):138-52.
326. Kesarwani M, Kincaid Z, Gomaa A, Huber E, Rohrabough S, Siddiqui Z, et al. Targeting c-FOS and DUSP1 abrogates intrinsic resistance to tyrosine-kinase inhibitor therapy in BCR-ABL-induced leukemia. *Nature medicine*. 2017;23(4):472-82.
327. Aqaq N, Yassin M, Yassin AA, Ershaid N, Katz-Even C, Zipin-Roitman A, et al. An ERG enhancer-based reporter identifies leukemia cells with elevated leukemogenic potential driven by ERG-USP9X feed-forward regulation. *Cancer Research*. 2019;canres.3215.2018.
328. Yu S, Cui K, Jothi R, Zhao D-M, Jing X, Zhao K, et al. GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells. *Blood*. 2011;117(7):2166-78.
329. Sykes SM, Di Marcantonio D, Martinez E, Huhn J, Gupta A, Mistry R. JUN and ATF3 Regulate the Transcriptional Output of the Unfolded Protein Response to Support Acute Myeloid Leukemia. *Blood*. 2018;132(Suppl 1):1327.
330. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*. 2017;14:865.

APPENDIX 1

BARCODED OLIGO-DT PRIMERS USED FOR 3'-TARGET-SEQ

Set of 96 oligodT-ISPCR barcoded primers used for 3'-biased high-throughput TARGET-seq.

Oligo_Name	Sequence
HT96_A1	aagcagtgggatcaacgcagagttgctGTGTCGTGCATnnnnnnnvttttttttttttttttttttttttttt
HT96_B1	aagcagtgggatcaacgcagagttgctGGTAGTGAGCTAnnnnnnnvtttttttttttttttttttttttttt
HT96_C1	aagcagtgggatcaacgcagagttgctGAGGAACAGAGTnnnnnnnvtttttttttttttttttttttttttt
HT96_D1	aagcagtgggatcaacgcagagttgctAGACGTCGATACnnnnnnnvtttttttttttttttttttttttttt
HT96_E1	aagcagtgggatcaacgcagagttgctGAGCTACGTCTAnnnnnnnvtttttttttttttttttttttttttt
HT96_F1	aagcagtgggatcaacgcagagttgctCTGATCGTAGCAnnnnnnnvtttttttttttttttttttttttttt
HT96_G1	aagcagtgggatcaacgcagagttgctCTTCAGCAATGCnnnnnnnvtttttttttttttttttttttttttt
HT96_H1	aagcagtgggatcaacgcagagttgctACAGTGCTCATGnnnnnnnvtttttttttttttttttttttttttt
HT96_A2	aagcagtgggatcaacgcagagttgctTACGGTACGTAGnnnnnnnvtttttttttttttttttttttttttt
HT96_B2	aagcagtgggatcaacgcagagttgctACGTCTCTGCATnnnnnnnvtttttttttttttttttttttttttt
HT96_C2	aagcagtgggatcaacgcagagttgctAAGTGCCTCAGTnnnnnnnvtttttttttttttttttttttttttt
HT96_D2	aagcagtgggatcaacgcagagttgctGAATCCTTACGCnnnnnnnvtttttttttttttttttttttttttt
HT96_E2	aagcagtgggatcaacgcagagttgctGACTCTATGGTGnnnnnnnvtttttttttttttttttttttttttt
HT96_F2	aagcagtgggatcaacgcagagttgctATGCTCTGTTGCnnnnnnnvtttttttttttttttttttttttttt
HT96_G2	aagcagtgggatcaacgcagagttgctTGGCATAGCTGAnnnnnnnvtttttttttttttttttttttttttt
HT96_H2	aagcagtgggatcaacgcagagttgctGAACACCATAGCnnnnnnnvtttttttttttttttttttttttttt
HT96_A3	aagcagtgggatcaacgcagagttgctAACGTGCGTATGnnnnnnnvtttttttttttttttttttttttttt
HT96_B3	aagcagtgggatcaacgcagagttgctGCATTGGATACCnnnnnnnvtttttttttttttttttttttttttt
HT96_C3	aagcagtgggatcaacgcagagttgctTCAGAGCGCATTnnnnnnnvtttttttttttttttttttttttttt
HT96_D3	aagcagtgggatcaacgcagagttgctAGTCTCCTACGTnnnnnnnvtttttttttttttttttttttttttt
HT96_E3	aagcagtgggatcaacgcagagttgctACCTTGA ACTGGnnnnnnnvtttttttttttttttttttttttttt
HT96_F3	aagcagtgggatcaacgcagagttgctGTCAGAACTACCnnnnnnnvtttttttttttttttttttttttttt
HT96_G3	aagcagtgggatcaacgcagagttgctATCTGCTATCCGnnnnnnnvtttttttttttttttttttttttttt
HT96_H3	aagcagtgggatcaacgcagagttgctGTGTAGTGTGGAAnnnnnnnvtttttttttttttttttttttttttt
HT96_A4	aagcagtgggatcaacgcagagttgctCTTGGTTCGTTGnnnnnnnvtttttttttttttttttttttttttt
HT96_B4	aagcagtgggatcaacgcagagttgctACTGTCTGTGTCnnnnnnnvtttttttttttttttttttttttttt
HT96_C4	aagcagtgggatcaacgcagagttgctTTCGACCAACGAnnnnnnnvtttttttttttttttttttttttttt
HT96_D4	aagcagtgggatcaacgcagagttgctAGAGGACGAGATnnnnnnnvtttttttttttttttttttttttttt
HT96_E4	aagcagtgggatcaacgcagagttgctAAGGTCCAAGCAnnnnnnnvtttttttttttttttttttttttttt
HT96_F4	aagcagtgggatcaacgcagagttgctAGTGTGACGAACnnnnnnnvtttttttttttttttttttttttttt
HT96_G4	aagcagtgggatcaacgcagagttgctCAGACATCGCAAnnnnnnnvtttttttttttttttttttttttttt
HT96_H4	aagcagtgggatcaacgcagagttgctGGATAGTTAGCCnnnnnnnvtttttttttttttttttttttttttt
HT96_A5	aagcagtgggatcaacgcagagttgctAACGGAATGGCTnnnnnnnvtttttttttttttttttttttttttt
HT96_B5	aagcagtgggatcaacgcagagttgctACTCGTCTCGATnnnnnnnvtttttttttttttttttttttttttt

HT017	aagcagtggatcaacgcagagttgctAGACGGAGTTGTGGvtttttttttttttttttttttttttttttt
HT018	aagcagtggatcaacgcagagttgctACATGAGTGTTCGvtttttttttttttttttttttttttttttt
HT019	aagcagtggatcaacgcagagttgctAATCCTTCGTGTGGvtttttttttttttttttttttttttttttt
HT020	aagcagtggatcaacgcagagttgctCTCTCCAAGACTGGvtttttttttttttttttttttttttttttt
HT021	aagcagtggatcaacgcagagttgctCCAAGCGAACCTTGvtttttttttttttttttttttttttttttt
HT022	aagcagtggatcaacgcagagttgctTACTACGGTTACGGvtttttttttttttttttttttttttttttt
HT023	aagcagtggatcaacgcagagttgctCACGTGCATACTGGvtttttttttttttttttttttttttttttt
HT024	aagcagtggatcaacgcagagttgctTATAGGCAACGCGGvtttttttttttttttttttttttttttttt
HT025	aagcagtggatcaacgcagagttgctATATGAAGGCTTGGvtttttttttttttttttttttttttttttt
HT026	aagcagtggatcaacgcagagttgctCCTTAGATCCTCCGvtttttttttttttttttttttttttttttt
HT027	aagcagtggatcaacgcagagttgctTACATAAGCCTTCGvtttttttttttttttttttttttttttttt
HT028	aagcagtggatcaacgcagagttgctGCCAGGAAGTCTTGvtttttttttttttttttttttttttttttt
HT029	aagcagtggatcaacgcagagttgctAGCAAGACTAGCGGvtttttttttttttttttttttttttttttt
HT030	aagcagtggatcaacgcagagttgctAGCCTTGGATCTGGvtttttttttttttttttttttttttttttt
HT031	aagcagtggatcaacgcagagttgctTGAGATCCACCTCGvtttttttttttttttttttttttttttttt
HT032	aagcagtggatcaacgcagagttgctGTCCAGTAGTATGGvtttttttttttttttttttttttttttttt
HT033	aagcagtggatcaacgcagagttgctTTATACTCCTTGCGvtttttttttttttttttttttttttttttt
HT034	aagcagtggatcaacgcagagttgctTCCTTGGTACGTGGvtttttttttttttttttttttttttttttt
HT035	aagcagtggatcaacgcagagttgctAGGCAATCAGGTGGvtttttttttttttttttttttttttttttt
HT036	aagcagtggatcaacgcagagttgctCTAAGCGGTCTTGvtttttttttttttttttttttttttttttt
HT037	aagcagtggatcaacgcagagttgctGAGGCTTAACCTTGGvtttttttttttttttttttttttttttttt
HT038	aagcagtggatcaacgcagagttgctTGACAGGACTTCGGvtttttttttttttttttttttttttttttt
HT039	aagcagtggatcaacgcagagttgctGGAGCACTAGCTTGvtttttttttttttttttttttttttttttt
HT040	aagcagtggatcaacgcagagttgctATGTTATCTCGTGGvtttttttttttttttttttttttttttttt
HT041	aagcagtggatcaacgcagagttgctAGCTAATGTTGTGvtttttttttttttttttttttttttttttt
HT042	aagcagtggatcaacgcagagttgctGGAACGATATTCAGvtttttttttttttttttttttttttttttt
HT043	aagcagtggatcaacgcagagttgctTATGGTGGACCTGGvtttttttttttttttttttttttttttttt
HT044	aagcagtggatcaacgcagagttgctGCATGTATTACCGvtttttttttttttttttttttttttttttt
HT045	aagcagtggatcaacgcagagttgctACACGCTACTTCGGvtttttttttttttttttttttttttttttt
HT046	aagcagtggatcaacgcagagttgctTTGTCAAGTGGCGGvtttttttttttttttttttttttttttttt
HT047	aagcagtggatcaacgcagagttgctAGGTATGTATACGGvtttttttttttttttttttttttttttttt
HT048	aagcagtggatcaacgcagagttgctGCAAGTACACATAGvtttttttttttttttttttttttttttttt
HT049	aagcagtggatcaacgcagagttgctTTGCGAGTCTCCTGvtttttttttttttttttttttttttttttt
HT050	aagcagtggatcaacgcagagttgctCAATCCACACTCCGvtttttttttttttttttttttttttttttt
HT051	aagcagtggatcaacgcagagttgctACACTTCGCCTTGGvtttttttttttttttttttttttttttttt
HT052	aagcagtggatcaacgcagagttgctGTATATAACTCCGGvtttttttttttttttttttttttttttttt
HT053	aagcagtggatcaacgcagagttgctCGTCATTCTTGCGGvtttttttttttttttttttttttttttttt
HT054	aagcagtggatcaacgcagagttgctACAGAATCGTCCGGvtttttttttttttttttttttttttttttt
HT055	aagcagtggatcaacgcagagttgctTACTGGCTAGCGGvtttttttttttttttttttttttttttttt
HT056	aagcagtggatcaacgcagagttgctGTAATGCCGTGTAGvtttttttttttttttttttttttttttttt
HT057	aagcagtggatcaacgcagagttgctCGCGATAACGATCGvtttttttttttttttttttttttttttttt
HT058	aagcagtggatcaacgcagagttgctTCTCTGGTAGCTAGvtttttttttttttttttttttttttttttt

HT101	aagcagtggatcaacgcagagttgctACCGTTAAGTCGCCvtttttttttttttttttttttttttttttt
HT102	aagcagtggatcaacgcagagttgctATCTGTGGTTCTACvtttttttttttttttttttttttttttttt
HT103	aagcagtggatcaacgcagagttgctACCAAGGAACATACvtttttttttttttttttttttttttttttt
HT104	aagcagtggatcaacgcagagttgctTTCGTTAACTGTCCvtttttttttttttttttttttttttttttt
HT105	aagcagtggatcaacgcagagttgctATGCTCTAACCTCCvtttttttttttttttttttttttttttttt
HT106	aagcagtggatcaacgcagagttgctTGA CTCAAGCGTGCvtttttttttttttttttttttttttttttt
HT107	aagcagtggatcaacgcagagttgctTTAACCGGAGTTCCvtttttttttttttttttttttttttttttt
HT108	aagcagtggatcaacgcagagttgctGACACGTGTTCTCCvtttttttttttttttttttttttttttttt
HT109	aagcagtggatcaacgcagagttgctTGTGCGGTTTCGTTvtttttttttttttttttttttttttttttt
HT110	aagcagtggatcaacgcagagttgctAAGCGCCATGGTTvtttttttttttttttttttttttttttttt
HT111	aagcagtggatcaacgcagagttgctTCAGAGGTAGTTACvtttttttttttttttttttttttttttttt
HT112	aagcagtggatcaacgcagagttgctGTTGGTTAGTCTTCvtttttttttttttttttttttttttttttt
HT113	aagcagtggatcaacgcagagttgctCATTACTATACCGCvtttttttttttttttttttttttttttttt
HT114	aagcagtggatcaacgcagagttgctCGACA ACTACTGCCvtttttttttttttttttttttttttttttt
HT115	aagcagtggatcaacgcagagttgctATACCTCAAGGTCCvtttttttttttttttttttttttttttttt
HT116	aagcagtggatcaacgcagagttgctCAAGGTCTTCCTCCvtttttttttttttttttttttttttttttt
HT117	aagcagtggatcaacgcagagttgctCCAATTGAGCGCTCvtttttttttttttttttttttttttttttt
HT118	aagcagtggatcaacgcagagttgctACACTACCGTGTCCvtttttttttttttttttttttttttttttt
HT119	aagcagtggatcaacgcagagttgctAATAGTTGCGTTGCvtttttttttttttttttttttttttttttt
HT120	aagcagtggatcaacgcagagttgctCAATTCAGTAGCACvtttttttttttttttttttttttttttttt
HT121	aagcagtggatcaacgcagagttgctAACCACGATAGGCCvtttttttttttttttttttttttttttttt
HT122	aagcagtggatcaacgcagagttgctGCTAGAGACTCTTCvtttttttttttttttttttttttttttttt
HT123	aagcagtggatcaacgcagagttgctGTTAGGAAGTTCACvtttttttttttttttttttttttttttttt
HT124	aagcagtggatcaacgcagagttgctTTGATCCGTTCTTCvtttttttttttttttttttttttttttttt
HT125	aagcagtggatcaacgcagagttgctGGTTAGTAGATTCCvtttttttttttttttttttttttttttttt
HT126	aagcagtggatcaacgcagagttgctCCGGA CTAACTTCvtttttttttttttttttttttttttttttt
HT127	aagcagtggatcaacgcagagttgctTAGTTGGTACCTTCvtttttttttttttttttttttttttttttt
HT128	aagcagtggatcaacgcagagttgctAGAGCGATCGTTGCvtttttttttttttttttttttttttttttt
HT129	aagcagtggatcaacgcagagttgctAAGAGTTCTACCACvtttttttttttttttttttttttttttttt
HT130	aagcagtggatcaacgcagagttgctATAATAAGCACGCCvtttttttttttttttttttttttttttttt
HT131	aagcagtggatcaacgcagagttgctGACCTAACTAGCGCvtttttttttttttttttttttttttttttt
HT132	aagcagtggatcaacgcagagttgctCTTCTAGAGGTTCCvtttttttttttttttttttttttttttttt
HT133	aagcagtggatcaacgcagagttgctAGCATGGACTCTCCvtttttttttttttttttttttttttttttt
HT134	aagcagtggatcaacgcagagttgctTTCCAGATACCTTCvtttttttttttttttttttttttttttttt
HT135	aagcagtggatcaacgcagagttgctATAAGACACTCTGCvtttttttttttttttttttttttttttttt
HT136	aagcagtggatcaacgcagagttgctGAATTCTCGATTCCvtttttttttttttttttttttttttttttt
HT137	aagcagtggatcaacgcagagttgctGTCAATTCACCTACvtttttttttttttttttttttttttttttt
HT138	aagcagtggatcaacgcagagttgctGGTATTAAGACGTCvtttttttttttttttttttttttttttttt
HT139	aagcagtggatcaacgcagagttgctTCTCACTTCACTCCvtttttttttttttttttttttttttttttt
HT140	aagcagtggatcaacgcagagttgctTGTTACCTGCGTCvtttttttttttttttttttttttttttttt
HT141	aagcagtggatcaacgcagagttgctTGTA ACTAGTTGCCvtttttttttttttttttttttttttttttt
HT142	aagcagtggatcaacgcagagttgctGGTCTACTGTTCCGvtttttttttttttttttttttttttttttt

HT353	aagcagtggatcaacgcagagttgctCCATGGCCACTTCTvtttttttttttttttttttttttttttttt
HT354	aagcagtggatcaacgcagagttgctCTTCGGTCTCTTCTvtttttttttttttttttttttttttttttt
HT355	aagcagtggatcaacgcagagttgctAGTTGTGCCACATTvtttttttttttttttttttttttttttttt
HT356	aagcagtggatcaacgcagagttgctCCGGCGCCAATATTvtttttttttttttttttttttttttttttt
HT357	aagcagtggatcaacgcagagttgctGGCGAGAGGTTATTvtttttttttttttttttttttttttttttt
HT358	aagcagtggatcaacgcagagttgctCCACCGAAGGAGTTvtttttttttttttttttttttttttttttt
HT359	aagcagtggatcaacgcagagttgctGCCTTCCAATAGTTvtttttttttttttttttttttttttttttt
HT360	aagcagtggatcaacgcagagttgctTTGCCGTACCGGTTvtttttttttttttttttttttttttttttt
HT361	aagcagtggatcaacgcagagttgctGGTAGCCTGTGGTTvtttttttttttttttttttttttttttttt
HT362	aagcagtggatcaacgcagagttgctGCTCGTCATGCGTTvtttttttttttttttttttttttttttttt
HT363	aagcagtggatcaacgcagagttgctGTATTGTAGCCGTTvtttttttttttttttttttttttttttttt
HT364	aagcagtggatcaacgcagagttgctTGGTTCATGTCGTTvtttttttttttttttttttttttttttttt
HT365	aagcagtggatcaacgcagagttgctAGGAGGTGCATGTTvtttttttttttttttttttttttttttttt
HT366	aagcagtggatcaacgcagagttgctAAGGAGCAGGTGTTvtttttttttttttttttttttttttttttt
HT367	aagcagtggatcaacgcagagttgctGAGAACCGTGTGTTvtttttttttttttttttttttttttttttt
HT368	aagcagtggatcaacgcagagttgctCCGGATGCACTGTTvtttttttttttttttttttttttttttttt
HT369	aagcagtggatcaacgcagagttgctGGACGGAGACACTTvtttttttttttttttttttttttttttttt
HT370	aagcagtggatcaacgcagagttgctCCTCACCTACGCTTvtttttttttttttttttttttttttttttt
HT371	aagcagtggatcaacgcagagttgctATTGTGGCGCGCTTvtttttttttttttttttttttttttttttt
HT372	aagcagtggatcaacgcagagttgctAAGTCTGTCTGCTTvtttttttttttttttttttttttttttttt
HT373	aagcagtggatcaacgcagagttgctGGAGTTAACACCTTvtttttttttttttttttttttttttttttt
HT374	aagcagtggatcaacgcagagttgctGCGAGATGTGCCTTvtttttttttttttttttttttttttttttt
HT375	aagcagtggatcaacgcagagttgctTCGAGCACATCCTTvtttttttttttttttttttttttttttttt
HT376	aagcagtggatcaacgcagagttgctGCATCCTTATCCTTvtttttttttttttttttttttttttttttt
HT377	aagcagtggatcaacgcagagttgctCATCCGGAGATCTTvtttttttttttttttttttttttttttttt
HT378	aagcagtggatcaacgcagagttgctGTTACACCGATCTTvtttttttttttttttttttttttttttttt
HT379	aagcagtggatcaacgcagagttgctAGTTGACGCATCTTvtttttttttttttttttttttttttttttt
HT380	aagcagtggatcaacgcagagttgctCTTGTCACAGTCTTvtttttttttttttttttttttttttttttt
HT381	aagcagtggatcaacgcagagttgctGCTGTGAGGTCCTTvtttttttttttttttttttttttttttttt
HT382	aagcagtggatcaacgcagagttgctCTCCACTGACTCTTvtttttttttttttttttttttttttttttt
HT383	aagcagtggatcaacgcagagttgctGAACATTGGCTCTTvtttttttttttttttttttttttttttttt
HT384	aagcagtggatcaacgcagagttgctGGCGAGTACCTCTTvtttttttttttttttttttttttttttttt

APPENDIX 2

PRIMERS USED IN THE PRE-AMPLIFICATION OF GDNA, MRNA AND CDNA AMPLICONS DURING RT AND PCR STEPS

Primers used in the pre-amplification of genomic DNA amplicons during PCR.

Region (hg19 coordinates)	Gene	Forward Primer	Reverse Primer
chr9:5073560-5073881	<i>JAK2</i>	CCAAAGCACATTGTATCCTCATCT	CACTGACACCTAGCTGTGATCCT
chr9:5070785-5071086	intronic	ACATAAGATCTGATAATCTGGCAGT	TGATGACAAATTCCTTAGGAGAGACT
chr9:4929565-4929848	intronic	CAAGTCAGGCCACTATCGCAG	TCAAGCTGAACTTCCAGAGAGAC
chr4:106164704-106165026	<i>TET2</i>	GGATGGAATGGTGATCACG	AAGTGCACGCTGAACTCTCTTC
chr4:106158477-106158816	<i>TET2</i>	TCAATATGATTTCCCATCTTGACG	ACAAAGAGCTAAATAGGC GAATTG
chr4:106157895-106158185	<i>TET2</i>	CCAGGGAGGAAGTCACACTCAG	TCCATGGTCTCAATGATGCTCTT
chr7:148523482-148523807	<i>EZH2</i>	GATAGCACTCTCCAAGCTGC	ACACATGGAAACCTTTTGA GAACTG
chr21:36252744-36253049	<i>RUNX1</i>	TAAGACAGACCGAGTTTCTAGGGA	ATCCCTGATGTCTGCATTTGTCC
chr9:139408930-139409229	<i>NOTCH1</i>	GATGTGTCCCCATGATCGGC	ACAGGGAAATCGAGGTTGACG
chr17:7577986-7578353	<i>TP53</i>	GAGGCCCTTAGCCTCTGTAAG	GAGAGACGACAGGGCTGTGT
chr10:89717567-89717920	<i>PTEN</i>	GCATTTCTGTGAAATAATACTGGT	ATGTCTCACCAATGCCAGAGTA
chr21:44514921-44514470	<i>U2AF1</i>	TTAAAGCGTGGATGGCAAGC	TGTACTATGTTCTAAGCAC TGCC
chr1:115258608-115258908	<i>NRAS</i>	GGGTAAAGATGATCCGACAAGTG	AGAACCAAATGGAAGGTCACAC
chr11:119148790-119149097	<i>CBL</i>	GGAAACAAGTCTTCACTTTTCTGT	AGGCCACCCCTTGTATCA GTAA
chr17:74732650-74733069	<i>SRSF2</i>	CATTTCCCCAGTCGCGAGG	CTTCGTTGCTTTACGACAA
chr2:198267075-198267591	<i>SF3B1</i>	ATCTCCCCAAATCAGTAGCCCAA	ATTACCAACTCATGACTGTCCTTTC
chr21:44514679-44514940	<i>U2AF1</i>	GGTGGTTGGAAGGAGACAT	TTAAAGCGTGGATGGCAAGC
chr4:106158202-106158534	<i>TET2</i>	TTTCACGCCAAGTCGTTATTGA	TGTCTCAGTACATTTCTG GCACT
chr20:31023100-31023500	<i>ASXL1</i>	TTGATGACGAATTAGGCTTTGGT	GTGAGGACTCAGTGCTTCAGAG
chr4:106196173-106196596	<i>TET2</i>	CCTGTTTCTGTTCTCTCTTACCCTG	GGAGCAGTTGTCCACTGATAGG
chr20:31022184-31022750	<i>ASXL1</i>	AGGTCAGATCACCCAGTCAG	ATCTCCTAGCCCATCTGTGAGT
chr19:13054450-13054735	<i>CALR</i>	CCTGGTCCTGGTCCTGATGT	AGGCCTCTCTACAGCTCGTC

chr17:7576957-7577205	<i>TP53</i>	GTGAATCTGAGGCATA ACTGCAC	GGACCTGATTCCTTACT GCCT
chr17:7577404-7577637	<i>TP53</i>	TGGAAGAAATCGGTAA GAGGTGG	CCTCATCTTGGGCCTGTG TTAT
chr4:106196888-106197332	<i>TET2</i>	CCACCCAATCTGAGCA ATCCA	GATTCTTTAAAGGGGTTG TGCC
chr21:44524258-44524657	<i>U2AF1</i>	AGTCGATCACCTGCCT CACTA	GGAGGTGCTTAATACCAC GGAA
chr20:31022872-31023195	<i>ASXL1</i>	GAAGTGAATGTGAGTC TGGCAC	TGTTAGAAACGAGAGCCT TGGT
chr17:7578037-7578319	<i>TP53</i>	ACCCATTACTTTGCAC ATCTCA	GGCCTCTGATTCCTCACT GATTG
chr17:7578295-7578657	<i>TP53</i>	AGCAATCAGTGAGGAA TCAGAGGC	GGAGGTGCTTACGCATGT TTG

Primers used in the pre-amplification of mRNA amplicons during RT.

Region (hg19 coordinates)	Gene	Forward Primer	Reverse Primer
chr9:5069183-5077565	<i>JAK2</i>	TAAATGCTGTCCCCCAA AGC	CCATGCCAACTGTTTAG CAAC
chr4:106182946-106182966	<i>TET2</i>	CTAAGTGGGTGGTTCGC AGA	TGGTGCCATAAGAGTGG ACAG
chr21:36171691-36259428	<i>RUNX1</i>	ACTCAATCGGCTTGTG TGAT	CAATGGATCCCAGGTAT TGG
chr9:139409753-139407553	<i>NOTCH1</i>	TACGAGTGTGCCTGTGA GC	GTTGAGACATGGGTTGG ACG
chr17:7578509-7577087	<i>TP53</i>	TTGCCAACTGGCCAAGA CCT	TTGCGGAGATTCTCTTC CTCTG
chr10:89692891-89720702	<i>PTEN</i>	AGCTGGAAAGGGACGA ACTG	CCTCTGGTCCTGGTATG AAGA
chr21:44527600-44513229	<i>U2AF1</i>	TCGGCACCGAGAAAGA CAAA	GCTCAGAATCGCCCAGA TCT
chr1:115256481-115259413	<i>NRAS</i>	CCTTCGCCTGTCCTCAT GTA	CCGACTGATTACGTAGC GGG
chr7:148516691-148524358	<i>EZH2</i>	ATGCTGGTAACACTGTG GTC	ATAAAGAAAGCCGCCCA CCT
chr11:119077274-119077294	<i>CBL</i>	TCAGAATCCTGATCTGA CTGGC	GGATCTACCACGATGG GTTGAG
chr17:74732406-74733069	<i>SRSF2</i>	CTTCGTGCGGATCTGGA CTT	CTTCGTTGCTTTACAG ACAA
chr2:198266772-198267754	<i>SF3B1</i>	ACCATAAGGAGTTGCTG CTTCA	CGTGGTCATTGAACCGC TATTG
chr4:106158202-106164865	<i>TET2</i>	TTTCACGCCAAGTCGTT ATTTGA	AGAGTTTGTGAGCCAGA GACAG
chr4:106194049-106194068	<i>TET2</i>	ACGCAAGCCAGGCTAAA CAG	GGAGCAGTTGTCCACTG ATAGG
chr20:31021661-31022750	<i>ASXL1</i>	AGTGTTACACCGAAAA GCC	ATCTCCTAGCCCCTCTG TGAGT
chr19:13054355-13054735	<i>CALR</i>	AGTCTGGCACCCTTT GACA	AGGCCTCTCTACAGCTC GTC
chr17:7576881-7578451	<i>TP53</i>	GGTTTCTTCTTTGGCTG GGGAG	TGGCCATCTACAAGCAG TCACA
chr17:7578226-	<i>TP53</i>	TCATCCAAATACTCCAC	CAGTCAGATCCTAGCGT

7579900		ACGCA	CGAG
---------	--	-------	------

Primers used in the pre-amplification of cDNA amplicons during PCR.

Region (hg19 coordinates)	Gene	Forward Primer	Reverse Primer
chr9:5069183-5077565	<i>JAK2</i>	AAGCAGTGGTATCAACG CAGAGTTAAATGCTGTC CCCCAAAGC	AAGCAGTGGTATCAACGC AGAGTCCATGCCAACTGT TTAGCAAC
chr4:106182946-106182966	<i>TET2</i>	AAGCAGTGGTATCAACG CAGAGTCTAAGTGGGTG GTTTCGAGA	AAGCAGTGGTATCAACGC AGAGTTGGTGCCATAAGA GTGGACAG
chr21:36252944-36171710	<i>RUNX1</i>	AAGCAGTGGTATCAACG CAGAGTGGCTGGCAATG ATGAAACT	AAGCAGTGGTATCAACGC AGAGTCAATGGATCCCAG GTATTGG
chr9:139409753-139407553	<i>NOTCH1</i>	AAGCAGTGGTATCAACG CAGAGTTACGAGTGTGC CTGTGAGC	AAGCAGTGGTATCAACGC AGAGTGTTTCAGACATGGG TTGGACG
chr17:7578509-7577087	<i>TP53</i>	AAGCAGTGGTATCAACG CAGAGTTTGCCAACTGG CCAAGACCT	AAGCAGTGGTATCAACGC AGAGTTTGCGGAGATTCT CTTCCTCTG
chr10:89692891-89720702	<i>PTEN</i>	AAGCAGTGGTATCAACG CAGAGTAGCTGGAAAGG GACGAACTG	AAGCAGTGGTATCAACGC AGAGTCCTCTGGTCCTGG TATGAAGA
chr21:44527600-44513229	<i>U2AF1</i>	AAGCAGTGGTATCAACG CAGAGTTCGGCACCAGAG AAAGACAAA	AAGCAGTGGTATCAACGC AGAGTGCTCAGAATCGCC CAGATCT
chr1:115256481-115259413	<i>NRAS</i>	AAGCAGTGGTATCAACG CAGAGTCCTTCGCCTGT CCTCATGTA	AAGCAGTGGTATCAACGC AGAGTCCGACTGATTACG TAGCGGG
chr7:148516691-148524358	<i>EZH2</i>	AAGCAGTGGTATCAACG CAGAGTATGCTGGTAAC ACTGTGGTC	AAGCAGTGGTATCAACGC AGAGTATAAAGAAAGCCG CCACCT
chr11:119077274-119077294	<i>CBL</i>	AAGCAGTGGTATCAACG CAGAGTTCAGAATCCTG ATCTGACTGGC	AAGCAGTGGTATCAACGC AGAGTGGATCTACCACGA TGGGTTTCA
chr17:74732406-74733069	<i>SRSF2</i>	CTTCGTGCGGATCTGGA CTT	CTTCGTTCGCTTTCACGAC AA
chr2:198266772-198267754	<i>SF3B1</i>	AAGCAGTGGTATCAACG CAGAGTACCATAAGGAG TTGCTGCTTCA	AAGCAGTGGTATCAACGC AGAGTCGTGGTCATTGAA CCGCTATTG
chr4:106158202-106164865	<i>TET2</i>	TTTCACGCCAAGTCGTTA TTTGA	AGAGTTTGTGAGCCAGAG ACAG
chr4:106194049-106194068	<i>TET2</i>	ACGCAAGCCAGGCTAAA CAG	GGAGCAGTTGTCCACTGA TAGG
chr20:31021661-31022750	<i>ASXL1</i>	AGTGTTTACACCGAAAA GCC	ATCTCCTAGCCCATCTGTG AGT
chr19:13054355-13054735	<i>CALR</i>	AGTCTGGCACCATCTTGACA	AGGCCTCTCTACAGCTCG TC

chr17:7576881-7578451	<i>TP53</i>	GGTTTCTTCTTTGGCTGG GGAG	TGGCCATCTACAAGCAGT CACA
chr17:7578226-7579900	<i>TP53</i>	TCATCCAAATACTCCACA CGCA	CAGTCAGATCCTAGCGTC GAG

APPENDIX 3

TARGET-SPECIFIC BARCODING PRIMERS USED FOR FIRST BARCODING PCR, TO GENERATE SINGLE-CELL TARGETED GENOTYPING LIBRARIES

Primers used for the barcoding of gDNA amplicons for targeted NGS during first barcoding PCR.

Region (hg19 coordinates)	Gene	Forward Primer	Reverse Primer
chr9:5073648-5073817	<i>JAK2</i>	ACACTGACGACATGG TTCTACATTATGGAC AACAGTCAAACAACA A	TACGGTAGCAGAGACTTG GTCTAAAGGCATTAGAAA GCCTGTAGT
chr9:5070785-5071086	<i>intronic</i>	ACACTGACGACATGG TTCTACAACATAAGA TCTGATAATCTGGCA GT	TACGGTAGCAGAGACTTG GTCTTGATGACAAATTCTT AGGAGAGACT
chr9:4929565-4929728	<i>intronic</i>	ACACTGACGACATGG TTCTACACAAGTCAG GCCACTATCGCAG	TACGGTAGCAGAGACTTG GTCTTCTTCCAAGGAACA GTGGCTC
chr4:106164707-106164982	<i>TET2</i>	ACACTGACGACATGG TTCTACATGGAATGG TGATCCACGCAG	TACGGTAGCAGAGACTTG GTCTTCCTATCAGTGGCC GCAAAG
chr4:106158477-106158630	<i>TET2</i>	ACACTGACGACATGG TTCTACATCAATATGA TTTCCCATCTTGCGAG	TACGGTAGCAGAGACTTG GTCTTGAGCCTTTACAAAT TGCTGCC
chr4:106157896-106158056	<i>TET2</i>	ACACTGACGACATGG TTCTACACAGGGAGG AAGTCACACTCAG	TACGGTAGCAGAGACTTG GTCTCCAGGTTCCACCTT AATTGGC
chr7:148523485-148523807	<i>EZH2</i>	ACACTGACGACATGG TTCTACAAGCACTCT CCAAGCTGCTTTA	TACGGTAGCAGAGACTTG GTCTACACATGGAAACCTT TTAGAAACTG
chr21:36252803-36253027	<i>RUNX1</i>	ACACTGACGACATGG TTCTACAGGTTTGTT GCCATGAAACGTG	TACGGTAGCAGAGACTTG GTCTATCCCTGATGTCTG CATTTGTCC
chr9:139408930-139409229	<i>NOTCH1</i>	ACACTGACGACATGG TTCTACAGATGTGTC CCCATGATCGGC	TACGGTAGCAGAGACTTG GTCTACAGGGAAATCGAG GTTGACG
chr17:7578138-7578319	<i>TP53</i>	ACACTGACGACATGG TTCTACATTAACCCC TCCTCCCAGAGAC	TACGGTAGCAGAGACTTG GTCTGGCCTCTGATTCTT CACTGATT

chr10:89717566-89717800	<i>PTEN</i>	ACACTGACGACATGG TTCTACAGGCATTTC CTGTGAAATAATACT GG	TACGGTAGCAGAGACTTG GTCTTCCCAATGAAAGTAA AGTACAAACC
chr21:44514679-44514567	<i>U2AF1</i>	ACACTGACGACATGG TTCTACAATGTCTCC TTCCAACCCACC	TACGGTAGCAGAGACTTG GTCTACTGAGCTTAACAAC GGGCA
chr1:115258620-115258906	<i>NRAS</i>	ACACTGACGACATGG TTCTACATCCGACAA GTGAGAGACAGGA	TACGGTAGCAGAGACTTG GTCTAACCAAATGGAAGG TCACACT
chr11:119148790-119149026	<i>CBL</i>	ACACTGACGACATGG TTCTACAGGAAACAA GTCTTCACTTTTTCTG T	TACGGTAGCAGAGACTTG GTCTTCGCTGTTTAGATCC GTACCTG
chr17:74732805-74733040	<i>SRSF2</i>	ACACTGACGACATGG TTCTACACGCGGACC TTTGTGAGGTCTG	TACGGTAGCAGAGACTTG GTCTCTGAGGACGCTATG GATGCC
chr2:198267118-198267465	<i>SF3B1</i>	ACACTGACGACATGG TTCTACATTAAACATG GACAGGCTGTGTG	TACGGTAGCAGAGACTTG GTCTCTTTTGCTGTTGTAG CCTCTGC
chr20:31023120-31023312	<i>ASXL1</i>	ACACTGACGACATGG TTCTACAGGTGGCTC ATGCCCTCCTAT	TACGGTAGCAGAGACTTG GTCTCTTTCTCCCACTCCT CTCCAAC
chr21:44514908-44514714	<i>U2AF1</i>	ACACTGACGACATGG TTCTACAATGGCAAG CACTTCTGTTTTTG	TACGGTAGCAGAGACTTG GTCTAGACATTTACTACCT CGTGTGC
chr4:106158368-106158530	<i>TET2</i>	ACACTGACGACATGG TTCTACAAGACACCA ACCAAAAGAACAGC	TACGGTAGCAGAGACTTG GTCTTCAGTACATTTCTGG CACTTACCT
chr4:106196181-106196594	<i>TET2</i>	ACACTGACGACATGG TTCTACATGTTCTCTC TTACCCTGTCCAC	TACGGTAGCAGAGACTTG GTCTAGCAGTTGTCCACT GATAGGTT
chr20:31022184-31022500	<i>ASXL1</i>	ACACTGACGACATGG TTCTACAAGGTCAGA TCACCCAGTCAG	TACGGTAGCAGAGACTTG GTCTCCTCTGCCACCTCC CTCAT
chr17:7578137-7578304	<i>TP53</i>	ACACTGACGACATGG TTCTACACTTAACCC CTCCTCCCAGAGAC	TACGGTAGCAGAGACTTG GTCTACTGATTGCTCTTAG GTCTGGC
chr17:7578333-7578626	<i>TP53</i>	ACACTGACGACATGG TTCTACACAACCAGC CCTGTCGTCTCTC	TACGGTAGCAGAGACTTG GTCTGCCGTCTTCCAGTT GCTTTATC
chr17:7576969-7577194	<i>TP53</i>	ACACTGACGACATGG TTCTACACATAACTG CACCTTGGTCTCC	TACGGTAGCAGAGACTTG GTCTCCTTACTGCCTCTTG CTTCTCT

chr17:7577407-7577631	<i>TP53</i>	ACACTGACGACATGG TTCTACAAAGAAATC GGTAAGAGGTGGGC	TACGGTAGCAGAGACTTG GTCTCTTGGGCCTGTGTT ATCTCCTA
chr21:44524366-44524585	<i>U2AF1</i>	ACACTGACGACATGG TTCTACAACGTCTTT GAAAAGAACATGAAG T	TACGGTAGCAGAGACTTG GTCTGTCATGCTGCTGAC ATATTCCAT
chr4:106196929-106197304	<i>TET2</i>	ACACTGACGACATGG TTCTACATGAACATC ATTCACCTTCTCACA	TACGGTAGCAGAGACTTG GTCTCTCACGCTTTGCAC ACTCAATG
chr19:13054451-13054727	<i>CALR</i>	ACACTGACGACATGG TTCTACACTGGTCCT GGTCCTGATGTCG	TACGGTAGCAGAGACTTG GTCTCTACAGCTCGTCCTT GGCCTG
chr20:31022983-31023122	<i>ASXL1</i>	ACACTGACGACATGG TTCTACATAGAGAAA GGAAGTGGCCAAGC	TACGGTAGCAGAGACTTG GTCTACCAAGCCCTAATTC GTCATCA

Primers used for the barcoding of cDNA amplicons for targeted NGS during first barcoding PCR.

Region (hg19 coordinates)	Gene	Forward Primer	Reverse Primer
chr9:5072596-5077484	<i>JAK2</i>	ACACTGACGACATGGT TCTACATCTGGATAAA GCACACAGAAACT	TACGGTAGCAGAGACTT GGTCTTCCAAATTTTACA AACTCCTGAACC
chr4:106164077-106180795	<i>TET2</i>	ACACTGACGACATGGT TCTACACTAAGTGGGT GGTTCGCAGA	TACGGTAGCAGAGACTT GGTCTGACAGGCGCAAG TTCTCTCT
chr21:36252903-36206883	<i>RUNX1</i>	ACACTGACGACATGGT TCTACAGCAATGATGA AACTACTCGGCT	TACGGTAGCAGAGACTT GGTCTGTTTCTGCCGAT GTCTTCGAG
chr9:139409119-139407966	<i>NOTCH1</i>	ACACTGACGACATGGT TCTACACAACATCGAT GAGTGTGCGG	TACGGTAGCAGAGACTT GGTCTACAGTTGGTCCC ACTCCACC
chr17:7578376-7577598	<i>TP53</i>	ACACTGACGACATGGT TCTACACATGAGCGCT GCTCAGATAG	TACGGTAGCAGAGACTT GGTCTTTGTAGTGGATG GTGGTACAGT
chr10:89711991-89720673	<i>PTEN</i>	ACACTGACGACATGGT TCTACATCCAATGTTC AGTGCGGAA	TACGGTAGCAGAGACTT GGTCTACCCAAAAGTGA AACATTTTGTCC
chr21:44514792-44513229	<i>U2AF1</i>	ACACTGACGACATGGT TCTACACCCGTGACG GACTTCAGAGA	TACGGTAGCAGAGACTT GGTCTGCTCAGAATCGC CCAGATCT

chr1:115256530-115259325	<i>NRAS</i>	ACACTGACGACATGGT TCTACAGTCCAGCTGT ATCCAGTATGTCC	TACGGTAGCAGAGACTT GGTCTCGGCTGTGGTCC TAAATCTGT
chr7:148523550-148524292	<i>EZH2</i>	ACACTGACGACATGGT TCTACAATTGCACTTA CGATGTAGGAAGC	TACGGTAGCAGAGACTT GGTCTTTCCAGATAAGG GCACAGCAG
chr11:119077285-119149285	<i>CBL</i>	ACACTGACGACATGGT TCTACAAGAAGATGGT GGAGAAGTGCTG	TACGGTAGCAGAGACTT GGTCTTACCACGATGGG TTCAGTACCT
chr17:74732494-74733069	<i>SRSF2</i>	ACACTGACGACATGGT TCTACAGAGACCTGGA ACGACTCCGA	TACGGTAGCAGAGACTT GGTCTCTTCGTTTCGCTTT CACGACAA
chr2:198266818-198267713	<i>SF3B1</i>	ACACTGACGACATGGT TCTACACTGATGGTCC GAACTTTCTGCT	TACGGTAGCAGAGACTT GGTCTCTAGAGTGGAAAG GCCGAGAGAT
chr21:44514640-44515568	<i>U2AF1</i>	ACACTGACGACATGGT TCTACAGTGGGGAAC GTGTACGTCAA	TACGGTAGCAGAGACTT GGTCTTCTGGAAATGGG CTTCAAATGC
chr4:106158317-106164005	<i>TET2</i>	ACACTGACGACATGGT TCTACAAACTTGATAG CCACACCCCAG	TACGGTAGCAGAGACTT GGTCTCCCTTCTGTCCA AACCTACACA
chr4:106194052-106196541	<i>TET2</i>	ACACTGACGACATGGT TCTACACAAGCCAGGC TAAACAGTTGG	TACGGTAGCAGAGACTT GGTCTTTCTGATTCAAAA GCCCAGGGT
chr20:31022184-31022525	<i>ASXL1</i>	ACACTGACGACATGGT TCTACACGAAAAGCCA CAGCCCCTA	TACGGTAGCAGAGACTT GGTCTCTCACCACCATC ACCACTGC
chr21:44515821-44527579	<i>U2AF1</i>	ACACTGACGACATGGT TCTACAAGTGTTCTG CATCTCCACATC	TACGGTAGCAGAGACTT GGTCTTCGGCACCGAGA AAGACAA
chr19:13054391-13054735	<i>CALR</i>	ACACTGACGACATGGT TCTACAACGATGAGGC ATACGCTGAG	TACGGTAGCAGAGACTT GGTCTAGGCCTCTCTAC AGCTCGTC
chr17:7577580-7578396	<i>TP53</i>	ACACTGACGACATGGT TCTACATAGTGGATGG TGGTACAGTCAG	TACGGTAGCAGAGACTT GGTCTCCATGAGCGCTG CTCAGATAG
chr17:7576900-7577604	<i>TP53</i>	ACACTGACGACATGGT TCTACAGAGAGGAGCT GGTGTTGTTGG	TACGGTAGCAGAGACTT GGTCTGCTCTGACTGTA CCACCATCC