Understanding *Tet2*-deficient haematopoietic stem cells using multimodal computational approaches

Monika Gońka

PhD

University of York
Biology

September 2025

Abstract

Haematopoietic stem cells (HSCs) are ultimately responsible for the production of all blood cells and can self-renew to create equally potent daughter HSCs to sustain the population for the lifetime of the organism. HSC fate is dependent on intrinsic and extrinsic signals; however, various mutations in genes regulating HSC behaviour can lead to haematological malignancy development. *TET2* is one of the most commonly mutated genes in haematological malignancies, with *TET2* loss of function mutations leading to increased HSC self-renewal and expansion of potentially malignant cell clones. The molecular mechanisms behind this process remain unclear.

To study regulatory networks governing HSC expansion in *TET2*-mutated haematological malignancies I took several approaches. I first examined chromatin accessibility using an assay for transposase-accessible chromatin (ATAC) sequencing dataset of HSCs isolated from a *Tet2* loss of function mouse model and performed integrative analyses with an RNA sequencing dataset[1] of HSCs isolated from the same mouse model[2]. Two integrative computational approaches were applied – a closest gene analysis and a transcription factor regulatory network, which highlighted several candidates with potentially important roles in the regulation of HSC self-renewal. Next, I investigated changes in gene expression in human haematopoietic stem and progenitor cells isolated from patients diagnosed with myelodysplastic syndrome (MDS) resulting from *TET2* mutations. Comparative data analyses of human and mouse datasets identified several candidates working across species in the context of haematological malignancies. In the final chapter, I attempted to understand the cellular ecosystem in which HSCs reside by profiling the interactome of mouse HSCs with the bone marrow extracellular fluid, generating a new catalogue of pathways altered in the *Tet2*-deficient niche. Altogether, these analyses span a spectrum of bioinformatic approaches aiming to study *Tet2*-mediated changes in HSCs and provide a valuable resource for further *in vitro* and *in vivo* studies.

List of Contents

Abstract **2**

List of Contents **3**

List of Tables **7**

List of Figures **8**

Preface **12**

Acknowledgements **15**

Declaration **17**

1. Introduction **18**

1.1. Haematopoietic stem cells *18*

1.1.1. Blood cell production 18

1.1.2. Haematopoietic stem cells and haematopoiesis 19

1.1.3. Is it a tree or is it a cloud? 21

1.1.4. Adult HSCs are mainly dormant 22

1.1.5. Signalling pathways mediating HSC fate 24

1.1.6. What about the environment? 25

1.1.7. Epigenetic control of HSC 26

1.1.8. DNA methylation 28

1.1.9. TET2 - methylcytosine dioxygenase 29

1.2. Haematological malignancies of the myeloid lineage *30*

1.2.1. Myeloid malignancies 30

1.2.2. Disrupted TET2 activity in blood cancers 32

1.2.3. TET2 partner protein function is disturbed in haematological malignancies 35

1.2.4. TET2 and Myelodysplastic Syndrome 36

1.2.5. The bone marrow microenvironment can drive malignancy development 37

1.3. Computational approaches to study haematological malignancy development *38*

1.3.1. Bioinformatics enable high throughput analyses of cell state 38

1.3.2. Exploration of cell transcriptome 38

1.3.3. Assessment of the chromatin state 41

1.3.4. Examination of cell proteome 42

1.3.5. Integrative analyses of different cell modalities 42
1.3.6. Regulatory network analyses 44
1.3.7. Understanding TET2 biology with single-cell ATAC and RNA sequencing 45
1.3.8. Bioinformatic analyses assist in deciphering MDS development 46
1.3.9. Computational studies of the BM extracellular fluid will help defining niche interactome 47
1.4. Thesis aims 47
2. Materials and methods 49
2.1. Mice 50
2.2. Human samples 50
2.3. Cell isolation 50
2.3.1. Isolation of mouse HSCs for scATACseq 50
2.3.2. Isolation of mouse HSCs for scRNAseq 51
2.3.3. Isolation of mouse HSCs for quantitative RT-qPCR 51
2.3.4. Isolation of human bone marrow haematopoietic cells for bulk RNAseq 52
2.3.5. Isolation of mouse HSCs for TMT mass-spectrometry proteomics 53
2.3.6. Isolation of BM extracellular fluid for TMT mass-spectrometry proteomics 53
2.4. Libraries preparation and sequencing 54
2.4.1. scATACseq library preparation 54
2.4.2. scRNAseq library preparation (mouse) 54
2.4.3. Bulk RNAseq library preparation (human) 55
2.5. Proteomic sample preparation and analysis of mouse HSCs 55
2.6. Proteomic sample preparation and analysis of mouse BM ECF 55
2.7. Quantitative RT-PCR profiling 56
2.8. Quality control 57
2.8.1. scATACseq 58
2.8.2. scRNAseq (mouse) 60
2.8.3. Bulk RNAseq (human) 61
2.9. Normalisation and dimensionality reduction 61
2.9.1. scATACseq 61
2.9.2. scRNAseq (mouse) 62
2.9.3. Bulk RNAseq (human) 63
2.10. Motif and feature enrichment 63

2.10.1. Peak Calling, motif enrichment and closest gene analysis (scATACseq)..... 63
2.10.2. Finding marker genes (scRNAseq)..... 64
2.11. Gene Ontology term enrichment, KEGG pathway and GSEA analysis..... 64
2.12. Closest genes integrative scATACseq and scRNAseq data analysis 64
2.13. Transcription factor regulatory network analysis 65
2.14. Shortest path analyses 65
2.15. Integrative analysis of mouse and human RNA sequencing data 66
2.16. Integrative human bulk RNAseq and mouse proteome analysis 66
2.17. Mapping selected differentially expressed genes against the Nestorowa dataset 66
2.18. BM ECF – HSC ligand-receptor analysis..... 66
2.19. Protein complexes analysis 67
2.20. Protein network analyses with HuRi, Genemania and String 67
2.21. Mapping ECF proteins against the Baccin et al. dataset 67
2.22. HSC in vitro culture with paquinimod and recombinant S100A9 67
2.23. Flow cytometry analysis of S100A8/9 expression in mouse HSC 68
2.24. HSC gel culture with VWF 68
Results..... 69
3. Chromatin accessibility and gene expression analyses of *Tet2*$^{-/-}$ LT-HSC to discover potential self-renewal regulators 69
3.1. Chromatin accessibility measurements indicates increased accessibility in Tet2$^{-/-}$ HSC 70
3.2. Single-cell gene expression profiling reveals increased expression in Tet2$^{-/-}$ HSC 73
3.3. Integrative analyses of chromatin accessibility and gene expression point to new potential regulators of HSC self-renewal 76
3.4. Transcription factor network analysis explores potential regulatory interactions governing HSC self-renewal..... 83
3.5. Quantification of transcript levels in vitro (RT-qPCR) and exploration of the expression of selected targets in human blood cells 86

3.6. Discussion 89

4. Transcriptomic changes in primary human haematopoietic cells carrying *TET2* mutations 93

4.1. Transcriptomic analyses of patient haematopoietic cells from TET2-mutated myelodysplastic syndromes 94

4.2. Comparative analysis with human cord blood-derived $CD34^+$ cells with CRISPR/Cas9 induced TET2 knock-out 104

4.3. Integrative analysis with mouse single-cell RNA sequencing data shows little overlap of differentially expressed genes 110

4.4. Discussion 113

5. Exploring extracellular mechanisms governing HSC self-renewal 116

5.1. Integrative analysis of mouse proteome and human TET2-mutated MDS datasets point to several targets potentially involved in the TET2-mutated MDS development 117

5.2. Exploring the function of calprotectin in $Tet2^{-/-}$ HSCs 120

5.3. Bone marrow extracellular fluid proteome analysis 125

5.4. Ligand-receptor analyses of differentially abundant HSC and ECF proteins..... 130

5.5. Discussion 137

6. Discussion 140

6.1. A summary of presented analyses 140

6.2. Gene regulation in Tet2-deficient mouse HSCs 141

6.3. Molecular repercussions of TET2 mutations in HSPCs in patients diagnosed with MDS 142

6.4. Disturbed signalling between Tet2-deficient HSCs and their bone marrow niche 143

6.5. Current understanding of HSC biology following TET2 mutations..... 144

6.6. Future directions 146

Abbreviations 147

References 151

List of Tables

Table 2.1 Antibody panel used to isolate HSCs for the scATACseq assay..... 51
Table 2.2 Antibody panel used to isolate human haematopoietic cells for bulk RNA sequencing. 52
Table 2.3 Numbers of sorted human haematopoietic cell fractions..... 53
Table 2.4 Mouse TaqMan assays used in the RT-qPCR experiments..... 57
Table 2.5 Flow cytometry antibody panel used to check expression of S100A8/A9. 68

List of Figures

Figure 1.1 A brief overview of the history of research of blood. 18
Figure 1.2 Identification of HSC markers and fluorescence-activated cell sorting (FACS) enabled prospective identification of human and mouse haematopoietic stem cells. 20
Figure 1.3 HSCs sustain blood production by maintaining a balance between self-renewal of their own population and differentiation into mature cells..... 23
Figure 1.4 An overview of known cytokines and growth factors governing HSC quiescence. 25
Figure 1.5 DNA methylation is associated with repressed gene expression..... 28
Figure 1.6 TET2 protein structure..... 30
Figure 1.7 An overview of the main milestones in bioinformatics..... 38
Figure 2.1 An overview of used datasets and analyses presented in this thesis 49
Figure 2.2 Quality control assessment by CellRanger (10x Genomics Cell Ranger, 10xgenomics.com) 59
Figure 2.3 Data quality control by ArchR..... 60
Figure 2.4 scRNAseq quality control. 61
Figure 2.5 scATACseq data analysis 62
Figure 2.6 scRNAseq Elbow Plot 63
Figure 3.1 Peak calling analyses indicate increased chromatin accessibility in the Tet2$^{-/-}$ HSC samples in comparison to their wild-type counterparts..... 71
Figure 3.2 scATACseq data analyses suggests a set of targets deregulated in Tet2-deficient HSC..... 72
Figure 3.3 Analysis of cell transcriptome at the single-cell level reveals increased gene expression in Tet2$^{-/-}$ HSC in comparison to their WT counterparts 74
Figure 3.4 Clustering of scRNAseq data identifies a Tet2-/- HSC subpopulation with a high MoIO and Repopsig scores 75
Figure 3.5 The integrative analyses of scATACseq and scRNAseq datasets indicates a set of targets deregulated in both modalities..... 77
Figure 3.6 Overlapping scATACseq/scRNAseq targets show little relatedness to TET2 but high inner connectivity..... 79
Figure 3.7 Shortest path analysis demonstrates connectivity of studied targets to TET2..... 81
Figure 3.8 A number of genes identified with the shortest paths analyses are associated with haematopoietic or lymphoid organ development and regulation of cell population proliferation 82

Figure 3.9 Transcription factor network analysis represents potential regulatory interactions governed by targets identified with scATACseq and scRNAseq assays 83
Figure 3.10 A median log2FC of expression of sets of genes regulated by each TF identified with the transcription factor network analysis is negative for most of the targets 84
Figure 3.11 Transcription factor network analysis indicates convoluted patterns of gene expression in Tet2$^{-/-}$ HSC 85
Figure 3.12 Selected targets identified with the scATACseq and scRNAseq analyses are expressed across murine haematopoietic differentiation 87
Figure 3.13 RT-qPCR analysis indicates no significant differences in expression of selected genes between Tet2$^{-/-}$, Tet2$^{+/-}$ and WT HSC 88
Figure 3.14 Selected targets show high and medium-level tissue expression and tissue specificity in human blood cells 89
Figure 4.1 An overview of the bulk RNA sequencing data analysis of bone marrow samples of patients diagnosed with MDS and healthy bone marrow controls 94
Figure 4.2 Bulk RNA sequencing data analysis of CD34^{+} bone marrow cells indicates upregulated gene expression in patients diagnosed with MDS in comparison to healthy bone marrow control. 95
Figure 4.3 Interferon response is upregulated and DNA repair downregulated in CD34^{+} cells isolated from patients diagnosed with MDS with TET2 mutations. 96
Figure 4.4 Bulk RNA sequencing data analysis of STRC bone marrow cells indicates upregulated gene expression in patients diagnosed with MDS in comparison to healthy bone marrow control. 98
Figure 4.5 Bulk RNA sequencing data analysis of HSPCs indicates downregulated gene expression in patients diagnosed with MDS with TET2 mutations in comparison to healthy bone marrow control. 99
Figure 4.6 Interferon and inflammatory response are predicably deregulated in HSPC isolated from patients diagnosed with MDS with TET2 mutations. 100
Figure 4.7 Zinc ion transport is upregulated in HSPCs isolated from patients diagnosed with MDS with TET2 mutations 101
Figure 4.8 Targets involved in the calcium signalling are downregulated in TET2-mutated HSPC. 102
Figure 4.9 A comparison of changes in gene expression across all studied cell fractions – CD34^{+}, STRC, HSPC 103
Figure 4.10 Bulk RNA sequencing analysis of Nakauchi et al. dataset shows decrease in the gene expression upon TET2 knock-out in cord blood-derived CD34^{+} cells 105
Figure 4.11 Integrative analysis of Nakauchi et al. dataset and bulk RNAseq HSPC MDS dataset indicate an overlap of downregulated genes 106

Figure 4.12 Reactive oxygen species pathway is downregulated across all studied datasets 107
Figure 4.13 RAGE receptor binding is deregulated in TET2-mutated HSPCs isolated from patients diagnosed with MDS and TET2KO cord blood-derived $CD34^+$ cells. 108
Figure 4.14 Expression of RAGE ligands is decreased in human HSPC in patients diagnosed with MDS with TET2 mutations 109
Figure 4.15 Human-mouse integrative analysis shows little overlap of differentially expressed genes 111
Figure 4.16 Expression of multimerin-1 is upregulated in patients diagnosed with MDS with TET2 mutations 112
Figure 5.1 An overlap of targets from the HSC proteome dataset and the RNAseq analysis of HSPC isolated from patients diagnosed with TET2-mutated MDS highlights several proteins potentially involved in the TET2-mutated MDS development 117
Figure 5.2 Identified consistently up- or down-regulated targets in both human HSPC TET2-mutated MDS and mouse HSC proteome datasets might be potential candidates for further studies 119
Figure 5.3 Exploration of calprotectin complex-TET2 connectivity 121
Figure 5.4 HSCs in vitro cell cultures with a S100A9 inhibitor and a recombinant mouse S100A9 protein show no differences between the numbers of $Tet2^{-/-}$ and WT cells 123
Figure 5.5 Flow cytometry gating strategy for the analysis of S100A8 and S100A9 expression in $Tet2^{-/-}$ and WT HSCs 124
Figure 5.6 Flow cytometry analysis suggests potential difference in the S100A8 expression between $Tet2^{-/-}$ and WT SLAM cells 125
Figure 5.7 An overview of an experimental set up for the downstream proteome analyses 126
Figure 5.8 Proteins in the bone marrow extracellular fluid are upregulated in $Tet2^{-/-}$ mice in comparison to WT controls 127
Figure 5.9 Analyses of the bone marrow extracellular fluid proteome of $Tet2^{-/-}$ mice identify dysregulation in abundances of proteins important in the haematopoietic system 128
Figure 5.10 Differentially abundant proteins in the extracellular fluid bone marrow of $Tet2^{-/-}$ mice are mainly secreted by neutrofils and dentritic cells 130
Figure 5.11 Ligand-receptor analyses of differentially abundant HSC and BM ECF proteins identify a set of interactions disturbed in the Tet2-deficient niche 131
Figure 5.12 Analysis of $Tet2^{-/-}$ HSC interactome with the BM extracellular fluid shows expected enrichment in growth factor binding, integrin binding and cytokine receptor binding 132
Figure 5.13 Shortest path analysis reveals potentially interesting direct interactions between HSC and ECF DA proteins 134

Figure 5.14 Direct ligand-receptor analyses of differentially abundant HSC and ECF proteins reveal a set of interactions potentially disturbed in the BM niche upon loss of function in Tet2 ..... *135*
Figure 5.15 VWF and other proteins mediating cell adhesion might regulate Tet2$^{-/-}$ HSC self-renewal *136*

Preface

Below is a chapter by chapter summary of the contributions that comprise the results of this thesis:

Chapter 3.

All work was conducted under the supervision of Prof David Kent. Mouse bone marrow cells for scATACseq and scRNAseq were isolated and sorted by Dr Daniel Bode and Dr Alyssa Cull in collaboration with the University of York Technology Facility. 10x Genomics scATACseq (www.10xgenomics.com) and SmartSeqII[3] scRNAseq libraries were prepared by Dr Daniel Bode and Dr Sally James at the University of York Technology Facility. scRNAseq alignment, reads sorting and gene counts were computed by Dr Daniel Bode, Dr John Davey and Dr Iwo Kucinski. Initial code for the scRNAseq data analysis was received from Dr Daniel Bode. The initial scRNAseq dataset was analysed and published in Dr Daniel Bode's PhD thesis[1]. The downstream scRNAseq pipeline was optimised with assistance from Dr Sandy MacDonald (Data Science Group, University of York). scATACseq alignment was run by Dr Sophia Ahmed (Data Science Group, University of York) and the downstream pipeline was optimised with assistance from Bart Theeuwes, Dr MS Vijayabaskar, and Dr Nicola Wilson (University of Cambridge) and Dr Andrew Holding (University of York). Following scATACseq and scRNAseq dataset alignments all further analyses were performed by Monika Gońka. Single-cell multi-modal integrative analyses were conducted with advice from the Prof Bertie Göttgens' research group (Cambridge Stem Cell Institute, University of Cambridge). Network analyses were performed in collaboration with Prof Benjamin Hall (University College London). RNA isolation and reverse transcription for the quantitative RT-PCR and qRT-PCR were performed by Alexander Hogg and Joanna Milek (Figure 3.13), all qRT-PCR data was analysed by Monika Gońka. A portion of work presented in this Chapter is a part of a preprint publication on which I am a co-author (Jassinskaja et al. 2025, bioRxiv).

Chapter 4.

All work was conducted under the supervision of Prof David Kent. Cell isolation by fluorescence-activated cell sorting and RNA library preparation were performed in collaboration with Dr Nicholas Wong, Tony Chen, Emily Tsao, Pratik Joshi, Amanda Tajik and Prof Kristin Hope at the Princess Margaret Cancer Centre, University Health Network, Toronto, Canada. All permissions/certificates to work with human bone marrow samples were provided by Prof Kristin Hope's group, University Health Network, Toronto. Library preparation was performed by members of the Next-Generation Sequencing Facility at the Centre for Applied Genomics at Hospital for Sick Children in Toronto. Dr Soheil Jahangiri (University Health Network, Toronto) ran alignment, quality control, PCA[4], normalisation and differential expression analyses. Network analyses were performed in collaboration with Prof Benjamin Hall (University College London). All downstream analyses were performed by Monika Gońka. Funding for this project was provided in part (travel and accommodation costs) by an EHA Research Mobility Grant awarded by the European Hematology Association (project title: "*Molecular programs of blood stem cell fate: how to break the self-renewal loop?*").

Chapter 5.

All work was conducted under the supervision of Prof David Kent. Cell isolation, raw data processing and differential abundance analyses for *Tet2*$^{-/-}$ and WT CD45^{+}Lineage^{-}cKit^{+}CD150^{+}CD48^{-} tandem mass tag (TMT) mass spectrometry samples were performed by Dr Daniel Bode and the University of York Technology Facility and that dataset was analysed and published in Dr Daniel Bode's PhD thesis[1] and later reanalysed by Dr Maria Jassinskaja. Cell isolation, raw data processing and differential abundance analysis for *Tet2*$^{-/-}$ and WT bone marrow extracellular fluid (ECF) TMT mass spectrometry samples were performed by Dr Sudip Ghosh and Dr Maria Jassinskaja (Figure 5.7) at the University of York and Lund University. All downstream analyses were performed by Monika Gońka. S100A8/9 *in vitro* validation experiment and flow cytometry data analysis from this experiment were performed by Alexander Hogg and Samuel Elberfeld (Figure 5.5, Figure 5.6). Cell culture with paquinimod and the recombinant S100A9 experiment

was performed by Ewan Taylor-Smith (Figure 5.4). Cell culture on hydrogels functionalized with VWF experiment was performed by Joanna Milek and Ellie Bennett, flow cytometry data from this experiment were analysed by Dr Maria Jassinskaja (Figure 5.15C,D). A portion of work presented in this Chapter is a part of a preprint publication on which I am a co-author (Jassinskaja et al. 2025, bioRxiv).

I have responsibly used Generative Artificial Intelligence (AI) during this project to aid coding and as a search engine (Chat GPT, OpenAI, https://chatgpt.com and Perplexity, Perplexity AI, Inc., https://www.perplexity.ai, accessed: 2023-2025).

Acknowledgements

I would like to thank my supervisors, Prof David Kent and Prof Ian Hitchcock for giving me the opportunity to do this PhD and for their continuous support over the last four years. I am particularly grateful to Prof David Kent for his invaluable mentorship throughout all these years, thank you for believing in me.

I am really grateful to Dr James Hewitson for his feedback on my project during the thesis advisory panel meetings. I would like to thank Dr Maria Jassinskaja for all her guidance and help and Dr Daniel Bode for his support. I would also like to thank Dan, Maria, Dr Sudip Ghosh and Dr Jenny Hansson for letting me use the ATAC sequencing, RNA sequencing and proteomics data. I am also immensely grateful for the support and advice I have received from Prof Benjamin Hall over the last two years. I would like to give a special thanks to members of the University of York Biology Data Science Hub - Dr Sandy MacDonald, Dr Matthew Care, Dr Alastair Droop and Dr Sophia Ahmed for all their help and feedback. I also want to thank all Kent Lab members for their support, in particular Alexander Hogg, Ewan Taylor-Smith, Sam Elberfeld, Joanna Milek and Ellie Bennett for performing wet-lab experiments to validate my computational analyses. I am also very grateful for the feedback that I have received from Dr Andrew Holding, Dr Veronique Voisin, Andy Zeng and Dr Alyssa Cull. I would like to extend my appreciation to the members of Prof Berthold Göttgens' and Prof Kristin Hope's labs, in particular Bart Theeuwes, Dr MS Vijayabaskar, Dr Nicola Wilson, Dr Soheil Jahangiri, Dr Nicholas Wong, Emily Tsao, Tony Chen and Pratik Joshi. I would also like to thank the European Hematology Association for the EHA Research Mobility Grant award that facilitated the collaboration with the University Health Network. I would like to thank members of the Next-Generation Sequencing Facility at the Centre for Applied Genomics at the Hospital for Sick Children in Toronto and Lund University. I also greatly appreciate all help from all members of the University of York Technology Facility and Biological Services Facility, in particular Dr Sally James, Dr Lesley Gilbert, Dr Karen Hogg, Dr Graeme Park and Dr Karen Hodgkinson. I am also grateful for computational support from the University of York IT Services. Last but not least, I would like to thank my family and friends, this would not be possible without you.

It is with great appreciation that I thank the UKRI Biotechnology and Biological Sciences Research Council and the White Rose BBSRC Doctoral Training program for funding this project.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author except as declared in the Preface and specified in the text. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references. Some analyses presented in Chapter 3 and Chapter 5 are part of a publication on which I am a co-author and which is currently under review (Jassinskaja et al., "*Selection of pre-leukemic hematopoietic stem cells driven by distinct extracellular matrix molecules*", bioRxiv).

1. Introduction

1.1. Haematopoietic stem cells

1.1.1. Blood cell production

For centuries blood has been revered as a critical bodily fluid, often considered a source of magical powers[5,6]. The ruby liquid vital for life was used to treat various diseases and to ward off evil spirits. In the 17th century, William Harvey first described the circulatory system and was the first to hypothesise that blood has biological, not magical functions[7]. Approximately 200 years later, scientists began to recognise and characterise the diverse roles of blood, such as oxygen transport and immune response. However, it was not until the 20th century when researchers identified the source of all those magical blood cell – the soft material in the cavities of bones called bone marrow (BM).

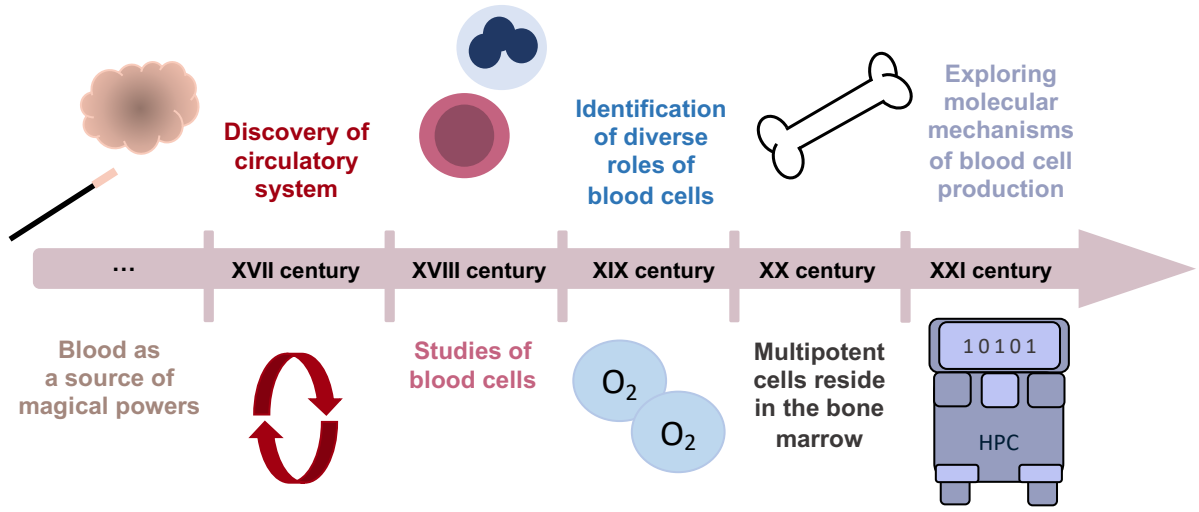


Figure 1.1 A brief overview of the history of research of blood. *In the past blood was often considered as a source of magical powers until the 17th century when William Harvey described the circulatory system. In the 18th century scientists began investigating blood cells and in the 19th century different roles of blood cells were discovered. In the 20th researchers found that multipotent blood cells reside in the bone marrow cavities. Nowadays, scientists explore molecular mechanisms of blood cell production.*

In 1951 Jacobson et al. performed a sinister but ground-breaking experiment – they lethally irradiated mice and then rescued them by introducing BM cells into

their bloodstream[8]. This experiment pointed to a hypothesis that bones contain multipotent cells capable of reconstituting all blood. Moreover, it also showed that BM contained haematopoietic progenitor cells of different blood lineages, such as lymphoid that include cells involved in a long-term immunity and myeloid that include cells responsible for oxygen transport around the body, platelets participating in blood clotting and cells that form a front-line defence against pathogens. A decade later Ernest McCulloch and James Till devised functional experimental strategies which confirmed the existence of the multipotent blood cells in the BM[9]. They performed BM transplantation experiments that showed a formation of single-cell derived colonies of clonal origin in mice's spleen with cells from different blood lineages[9]. They called these multipotent cells – haematopoietic stem cells (HSCs) (Figure 1.1) as the process of blood cell production was referred to as haematopoiesis (Greek *haima* – blood, *poiēsis* – to make)[10].

1.1.2. Haematopoietic stem cells and haematopoiesis

Following the functional description of HSCs, scientists were determined to isolate these cells and study their properties. If HSCs could reconstitute all blood cells, their isolation and *in vitro* amplification could potentially cure blood diseases, such as anaemias and blood cancers. Moreover, HSC isolation would provide an opportunity to study molecular mechanisms mediating blood production and help understanding blood cancer development and progression. However, that proved to be challenging as researchers discovered that HSCs are very rare and around trillion of new cells are produced daily in the BM[11].

Enriched populations of HSCs were first prospectively isolated from the mouse bone marrow in 1980s using fluorescence activated cell sorting (FACS), a technique that sorts cells using fluorescent markers targeting cell subpopulations of interest (Figure 1.2). The first landmark study to detail the purification and characterisation of mouse HSCs was performed by Spangrude et al. in 1988[12]. It involved flushing the tibiae and femora from mice, incubating the cells with fluorescently-labelled antibodies, sorting cells with FACS, followed by a functional test – the spleen colony assay. Cells were isolated using antibodies targeting CD90,

Sca-1 and markers known to be expressed on mature blood cells. HSCs were sorted based on low expression of CD90, positive expression of Sca1 and the absence of mature blood cell associated markers ($Lin^{-}Sca\text{-}1^{+}CD90^{low}$ phenotype). Until this point, HSCs were never purified to such a high degree of enrichment, but the majority of cells in the fraction were still not functional HSCs. The race commenced to find more specific markers and isolate more enriched populations of HSCs. Soon enough the search for human HSC markers has also begun with subsequent studies identifying a series of cell surface and intracellular proteins that could distinguish HSC from all other BM cells due to their unique expression in the HSC compartment[13,14].

Although there are many markers and competing strategies now in existence, for the purposes of my thesis, I will be using the mouse HSC phenotype $CD45^{+}CD150^{+}CD48^{-}EPCR^{+}$[15], and human HSC as $Lin^{-}CD34^{+}CD38^{-}CD45RA^{-}CD49f^{+}$[16].

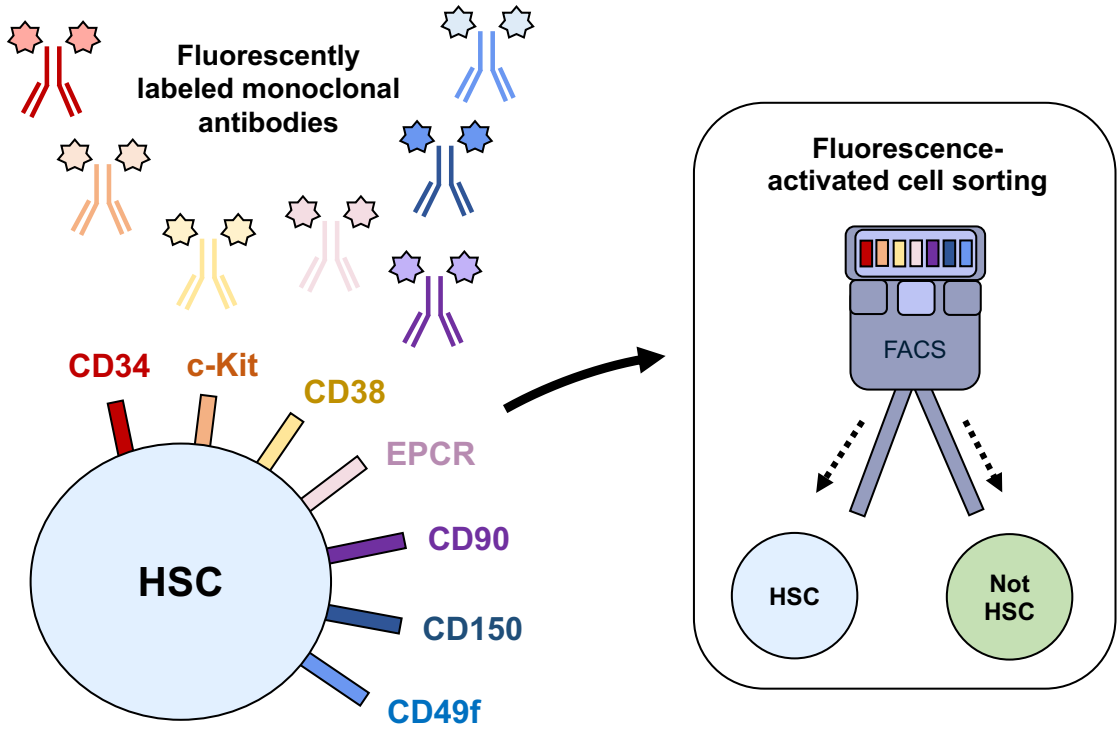


Figure 1.2 Identification of HSC markers and fluorescence-activated cell sorting (FACS) enabled prospective identification of human and mouse haematopoietic stem cells. *A selection of markers commonly used in mouse and in human HSC isolation: CD34, c-Kit, CD38, EPCR, CD90, CD150, CD49f. Many more markers are used across the world. Fluorescently labeled monoclonal antibodies enable isolation of cells with a phenotype of interest, following staining with selected antibodies cells can be sorted using FACS.*

The discoveries of various markers detected on HSCs and blood cells revealed a presence of a plethora of blood progenitor cells - immature cells arising from HSCs and developing into functional blood cells[13]. This naturally led to attempts on conceptual modelling the process of haematopoiesis. Gradually the haematopoietic tree model emerged, a hierarchical diagram depicting the HSC at the top (the root of the tree) branching into more and more differentiated blood cells. Early models consisted of two main tree branches – myeloid and lymphoid, with the common myeloid progenitor (CMP) and common lymphoid progenitor (CLP) cells at the apex of each of these branches respectively.

1.1.3. Is it a tree or is it a cloud?

By the end of millennium, a tree-like diagram representing haematopoiesis was widely accepted. However, it soon turned out that the model needed to be adjusted to account for major novel findings, including the discovery that not all HSCs are equivalent. Descriptions of intermediate-term HSCs (IT-HSC), short-term HSCs (ST-HSC) and multipotent progenitor cells (MPP2, MMP3, MPP4) which differed in their self-renewal potential were identified and subsequently added to the tree. Shortly thereafter, another tree upgrade seemed necessary, as new single-cell analyses of cellular transcriptomes offered novel approaches in studying haematopoiesis. Advanced bioinformatic algorithms were developed, where single-cell transcriptomes were treated as snapshots of cellular state that could be mapped along haematopoietic differentiation trajectories. This resulted in a continuum “cloud” model representing transcriptional landscapes of haematopoietic cell subpopulations. In 2019 a novel model of haematopoiesis that portrays complex differentiation landscapes was proposed[17]. This model not only accounted for HSC heterogeneity and different possible cell transitions, but also shows the exponential increase in blood cell numbers – a principal characteristic of haematopoiesis that was not depicted with the first tree models[17].

A tree-to-cloud model evolution represents a fundamental change in how we understand haematopoiesis nowadays. The initial tree model was constructed based on the identification of cell surface markers detected on haematopoietic cells, which

led to categorising cells into distinct subpopulations and subsequently creating artificially homogenous cell groups, but the reality is that all of these populations are highly heterogenous from a functional and molecular standpoint. HSCs represented as a single root cell responsible for reproducing all blood cells not only assumed homogeneity of cell populations but was also inferred from transplant experiments which naturally do not represent homeostatic blood production[18]. Overall, a tree model proved oversimplified and unlikely. Nevertheless, it should be noted that the tree model provided invaluable progress in grasping the concept of cell lineage specification and differentiation into mature blood cells.

1.1.4. Adult HSCs are mainly dormant

What makes HSCs extraordinary is their capacity to sustain blood production throughout the whole life of an organism. That is possible due to their unique ability to maintain a balance between self-renewal of their own population and differentiation into mature blood cells[19,20]. This means that at any given cell division, HSC can produce either two equally potent daughter HSCs (symmetric self-renewal), one daughter HSC and one non-HSC (asymmetric self-renewal), or no daughter HSCs (symmetric differentiation) (Figure 1.3). Moreover, HSCs could also die, resulting in a loss to the stem cell pool or stay in the G_0 cell cycle phase and remain quiescent[21] (Figure 1.3). The mechanisms of this dynamic, but precisely regulated process have been thoroughly studied since the 20th century.

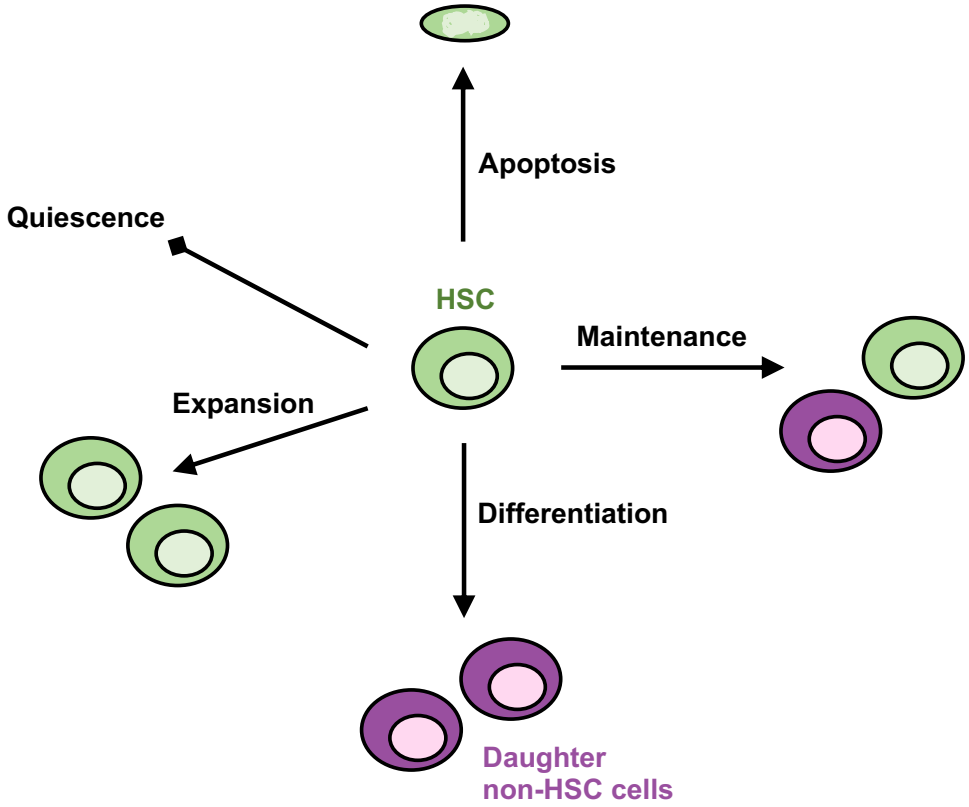


Figure 1.3 HSCs sustain blood production by maintaining a balance between self-renewal of their own population and differentiation into mature cells. *HSC can divide in a symmetrical division and produce two HSCs or two daughter non-HSC cells. HSC can also divide asymmetrically to produce one HSC and one daughter non-HSC cell. HSC can also die in apoptosis or remain quiescent in the G_0 cell cycle phase. HSC = haematopoietic stem cell.*

The initial understanding of the blood cell production assumed that HSCs alone drive haematopoiesis on a daily basis[17]. Sitting at the top of the haematopoietic tree, HSC could enter the cell cycle at any time to generate blood cells if needed, otherwise remaining quiescent. However, interesting studies using histone 2B-green fluorescent protein to label cell cycle estimated that HSCs only divide four times before losing their long-term self-renewal potential[22–24]. Moreover, recent experiments using barcoding showed intriguing results. Surprisingly, in homeostasis HSCs are mainly in a dormant state, “sleeping” in the bone marrow and haematopoiesis is driven primarily by multipotent progenitor cells (MPPs). Moreover, it has been suggested that the longer an HSC remains dormant in the bone marrow, the better its ability to reconstitute the blood system[22,24].

1.1.5. Signalling pathways mediating HSC fate

Why are HSCs predominantly in a dormant state? A reasonable rationale is that because they have capacity to regenerate all blood cells, HSCs are very precious and need to be protected at all costs. Cell divisions come with increased DNA repair processes and a high risk of introducing mutations, so to avoid unnecessary danger HSCs mainly exist in the G_0 cell cycle phase and "leave all the work" to their daughter cells – MPPs. A lot of effort was invested in understanding mechanisms mediating the quiescent HSC behaviour. Firstly Wnt[25,26], Notch[27–29] and Hedgehog[30,31] developmental pathways were discovered to govern HSC fate. Soon many cytokines and growth factors were shown to regulate HSC behaviour, among which the most important ones seem to be CXCL12[32], SCF[33], TGFβ[34], Ang1[35] and TPO[36]. Together they orchestrate HSC self-renewal and entry into the cell cycle. An overview of the main known cytokines and growth factors regulating HSC quiescence is presented in Figure 1.4.

Another threat for HSCs comes from reactive oxygen species (ROS). ROS induce oxidative stress, loss of the HSC quiescence and might lead to HSC exhaustion[37]. Therefore, ROS are kept under control to avoid any DNA damage in HSC[38]. One of the main mechanisms helping maintain low ROS level is a hypoxic environment in the BM[37,39,40]. For this reason, dormant HSCs express Hypoxia-Inducible Factor 1α (HIF1α) and MEIS1 Homeobox protein which regulate oxygen level and promote anaerobic glycolytic metabolism in quiescent HSCs[12]. It seems that the glycolytic metabolic state of dormant HSCs is critical for their proper function. That is aligned with the observed high level of non-active mitochondria in the dormant HSCs[41,42], as following HSC activation that would allow for tricarboxylic acid cycle and oxidative phosphorylation to efficiently provide necessary metabolites and energy needed for HSC proliferation[10].

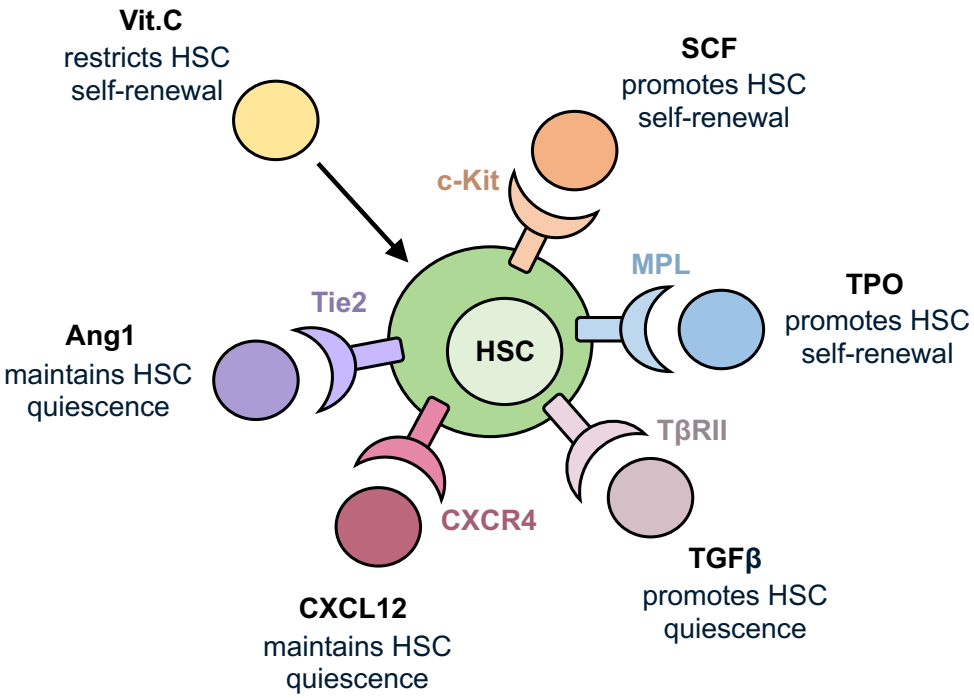


***Figure 1.4 An overview of known cytokines and growth factors governing HSC quiescence**; simplified model, not to scale. Vit.C = Vitamin C, SCF = Stem Cell Factor, TPO = Thrombopoietin, MPL = Myeloproliferative Leukemia protein, TGFβ = Transforming Growth Factor Beta, TβRII = Transforming growth factor beta receptor II, CXCL12 = C-X-C Motif Chemokine Ligand 12, CXCR4 = C-X-C chemokine receptor type 4, Ang1 = Angiopoietin-1, Tie = tyrosine kinase with immunoglobulin like and EGF like domains 1.*

1.1.6. What about the environment?

In 1978, Raymond Schofield proposed that HSCs were localised in a specified microenvironment in the BM which plays an important role in HSC function[43]. Until then most studies focused on closely examining the HSCs, not investigating the function of BM in the HSC behaviour. However, researchers soon realised the BM microenvironment might play as important role as the HSC intrinsic regulation itself. The exploration of the BM environment and search for the “HSC niche” had begun.

Identification of HSC markers enabled their separation from blood progenitor cells and immunofluorescence staining of the haematopoietic tissue, with SLAM family proteins discovery by Kiel et al. in 2005 allowing one of the first imaging analyses of the HSC environment with high purity[44]. A discovery of HoxB5 expression in long-term self-renewing HSCs by Chen et al. allowed to distinguish

HSCs from other BM cells and facilitate *in situ* imaging analysis of mouse bones that highlighted perivascular location of HSCs in the BM[45]. Subsequent *in vivo* label-retention assays performed by Saçma et al. indicated perisinusoidal location of the most quiescent HSCs in mouse[46]. Nowadays researchers investigate the niche at even greater detail, a recent landmark study by Christodoulou et al. used live-animal *in vivo* time-lapse imaging to examine motility of long-term HSCs at steady state and their responses following activation[47]. Numerous other studies have aimed to determine HSC niche in the bone marrow[48]. The exact HSC location in the BM still remains debatable, nevertheless many cell subtypes regulating HSC activity were identified using fluorescently labelled antibodies and microscopy, such as endothelial cells, regulatory T cells, sympathetic nervous system nerves, Schwann cells, osteoblasts, macrophages, megakaryocytes, adipolineage cells and perivascular mesenchymal stem cells[49]. Today it seems undoubtable that HSC cellular decisions are dependent on both intrinsic cellular mechanisms and extrinsic regulation by the BM microenvironment[50,51].

1.1.7. Epigenetic control of HSC

Cytokines, reactive oxygen species, metabolic pathways and the BM cells are not the only players directing HSC function. Another key factor for HSC self-renewal and differentiation is epigenetic control – regulations of gene expression that do not change the DNA code. In order to exit quiescent and become active, HSCs must initialise transcription of genes responsible for proliferation and lineage commitment. Expression of these genes is regulated by various epigenetic mechanisms. Among the most important epigenetic modulators controlling HSC differentiation are DNA methylation, non-coding RNA and histone modifications[52]. Together they remodel chromatin structure and modify its accessibility allowing efficient binding of transcription factors to enhancers and promoters of genes regulating HSC differentiation.

Histones are proteins that bind DNA to organise DNA structure in the cell nucleus[53]. Histones can be modified in many ways, among the modifications seem to play the most important role in haematopoiesis are histone methylation and

histone acetylation[54,55]. Addition and removal of an acetyl group to lysine in histone tails impairs DNA-histone interaction, due to neutralisation of electrostatic charge[56]. This weakened connection results in open chromatin which is one of the first steps leading to active transcription[56]. Histone acetylation is mediated by histone acetyltransferases (HATs) and histone deacetylases (HDACs) enzymes[55,56]. Histones can also be methylated, which takes place at lysine and arginine residues[57]. Depending on the position, histone methylation may lead to activation or repression of transcription, with arginine methylation usually leading to the activation of transcription, whereas a result of the methylation of lysine depending on the $-CH_3$ location[57]. Histone methylation is mediated by histone methyltransferases (HMTs) and histone demethylases (KDMs)[55,57].

RNA transcripts that do not code proteins – noncoding RNAs (ncRNAs) play various regulatory functions, among the ncRNAs that seem to play the most important role in HSC self-renewal and haematopoietic differentiation are microRNAs and long noncoding RNAs (lncRNAs)[58]. microRNAs primarily mediate transcripts degradation by binding to mRNA molecules[59]. lncRNAs interact with DNA, RNA and proteins and can silence or activate transcription both at the epigenetic level and the posttranscriptional stage[60]. Among the first noncoding RNAs identified to regulate HSC activity were e.g. miR-223 found by Johnnidis et al. in 2008 to supress haematopoietic progenitor proliferation using miR-223 mutant mice[61], miR-155 discovered by O'Connell et al. in the same year to be overexpressed in patients diagnosed with acute myeloid leukaemia (AML) - a type of blood cancer[62], miR-126, firstly shown to control HSC proliferation using lentiviral knockdown by Lechman et al. in 2012[63], and many others such as e.g. miR-125b, miR-196a, miR-196b, miR-130a, let-7d, miR-148b, and miR-351 initially found to be upregulated in haematopoietic stem and progenitor cells (HSPCs) in 2010 by Petriv et al. using qPCR experiments[64]. Since the first investigations in the early 2010s numerous other studies have set out to examine roles of noncoding RNAs in the HSC biology with new discoveries made in the field every year[65,66].

1.1.8. DNA methylation

DNA methylation is the most well-studied epigenetic modification in haematopoietic cells[67–70]. Not only does it play an important role in HSC self-renewal but also influences lineage commitment[55]. DNA methylation is regulated by several enzymes that add and remove the methyl group (CH_3) at DNA cytosines[69]. This process mainly occurs at CpG dinucleotide sites, usually found in CpG islands located near gene promoter regions[71]. CpG dinucleotides that are not part of the CpG islands are predominantly methylated, however the majority of cytosines in CpG islands are unmethylated[71,72]. Unmethylated cytosines at gene promoters, enhancers and transcription factor binding sites are associated with active transcription[71]. DNA methylation leads to silenced expression by physically blocking transcription factor binding or by recruiting proteins that repress transcription[73]. Different cell types exhibit different methylation patterns which suggest that methylation might in a way program cell functions[74].

DNA methylation is mediated by DNA methyltransferases (DNMT), enzymes that transfer CH_3 to the 5th carbon of cytosine using S-adenosyl methionine as a source of the methyl group. This process results in generating 5-methylcytosine (5mC) (Figure 1.5). There are 3 DNMT enzymes that control methylation of DNA – DNMT1 maintaining methylation during replication and two de novo methyltransferases: DNMT3A and DNMT3B.

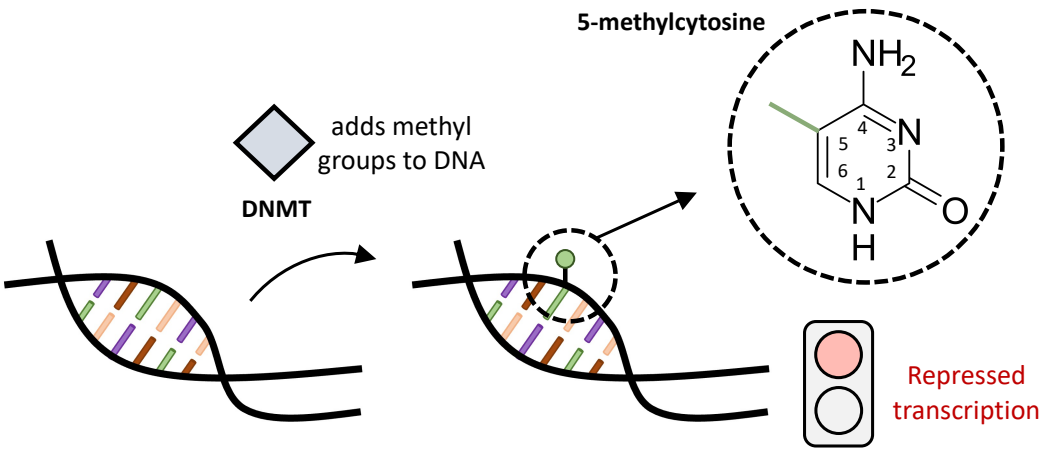


Figure 1.5 DNA methylation is associated with repressed gene expression. *DNA methyltransferases (DNMT) add methyl groups at the 5th carbon of cytosine which results in 5-methylcytosine. DNA methylation leads to repressed transcription (represented here as a red signal light).*

The removal of CH_3 group from 5mC is required to enable transcription factors to bind to DNA and initiate transcription. This process is called DNA demethylation and can be passive or active. In a passive DNA demethylation, the methyl group is lost during replication as the synthesized DNA does not have 5mC. In the active DNA demethylation CH_3 group is removed primarily via the base excision repair (BER) mechanism using the TET–TDG–BER system. In this process 5mC is firstly oxidised to 5-hydroxymethylcytosine (5hmC) by methylcytosine dioxygenase enzymes from the TET family[75,76]. 5hmC can then be converted by TET enzymes to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), and both 5fC and 5caC can be excised by thymine DNA glycosylase (TDG) and replaced by unmodified cytosine in a BER process[77]. 5hmC can also be deaminated by APOBEC leading to BER and unmodified cytosine[77].

1.1.9. TET2 - methylcytosine dioxygenase

TETs are an epigenetic enzyme family consisting of three proteins: TET1, TET2 and TET3. All these methylcytosine dioxygenases are involved in DNA demethylation, however TET2 in particular seems to play an important role in the haematopoietic system. All TETs can catalyse the conversion of 5mC into 5hmC, 5hmC to 5fC and 5fC to 5caC, although TET2 prefers 5mC as a substrate[78,79]. TET2 is a 223 kDa protein containing a catalytic region at the carboxy-terminal protein end. The catalytic region consists of cysteine-rich domain (CR), double-stranded β-helix domain (DSBH) and a low complexity insert (LCI) (Figure 1.6). DSBH enables binding of two cofactors necessary for the proper TET2 function - α-ketoglutarate (α-KG) and Fe(II). The exact mechanisms of action of other TET2 domains are still being investigated, with the latest studies revealing the role of CR domain in binding chromatin preferentially at the histone H3 tail[80].

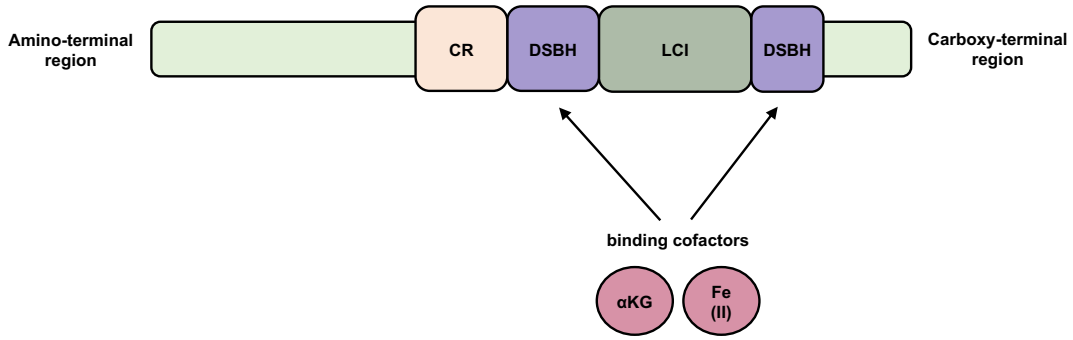


Figure 1.6 TET2 protein structure. *TET2 contains a catalytic region at the carboxy-terminal protein end which includes cysteine-rich domain (CR), double-stranded β helix domains (DSBH), low complexity insert (LCI) and binding sites for alpha ketoglutarate (α-KG) and iron ion (Fe (II)). Figure adapted (with changes) from Zhang et al., Signal Transduction and Targeted Therapy 2023[78], under http://creativecommons.org/licenses/by/4.0/.*

In the early 2010s several independent studies reported loss of function mutations in *TET2* in patients diagnosed with haematological malignancies. In 2009 Delhommeau et al., Langemeijer et al., Jankowska et al. and Tefferi et al. identified somatic mutations in *TET2* in patients diagnosed with myeloproliferative neoplasms (MPN), MDS, chronic myelomonocytic leukaemia (CMML) and AML[81–85]. Since then, TET2 has been intensively investigated by researchers, both in the context of blood malignancies and homeostasis. Bibliometric analyses of scientific articles published since 2009 up to 2021 indicated extensive exploration of TET2 role in clonal haematopoiesis, AML, and inflammation[86].

1.2. Haematological malignancies of the myeloid lineage

1.2.1. Myeloid malignancies

There are three main types of haematological malignancies: leukaemias, lymphomas and myelomas. Depending on the type of malignant blood cells we can distinguish lymphoid and myeloid malignancies, and the latter includes MDS, AML and MPNs. Myeloid malignancies are diagnosed based on the percentage and type of malignant cells, cell morphology and presence of specific mutations[87]. These malignancies can be diagnosed at any age and, whereas some are currently

curable, for many patients chemo- and radiotherapy treatment causes great suffering and gives a poor prognosis.

HSCs maintain a balance between self-renewal of their own population and differentiation. Disruption of this balance can result in the uncontrolled proliferation of a mutant, immature cell population which can lead to blood cancer development[88]. The majority of myeloid malignancies originate from haematopoietic stem and progenitor cells which accumulated gene mutations and epigenetic alterations[89]. This malignant transformation usually takes years and is different in each patient as mutations may occur in various orders[90]. This causes substantial heterogeneity between patients and major difficulties with finding appropriate treatments[91]. Understanding distinct molecular mechanisms governing malignant transformations is crucial as it will enable precise targeting of disturbed biological pathways and development of more personalized therapies.

DNA sequencing analyses of blood cancer cells not only identified commonly mutated genes, but also suggested a clonal background of these diseases[92]. Many studies identified subclones of blood cancer cells sharing mutational profiles implying patterns in disease progression[93–98]. This indicated that despite extensive heterogeneity, the evolution in haematological malignancies is not random. In 2012 Jan et al. examined mutation acquisition patterns in HSCs isolated from patients diagnosed with AML using exome sequencing and they were the first to suggest clonal evolution of preleukaemic HSCs[99]. These analyses also suggested that different molecular subtypes harbour distinct sets of so called ‘driver mutations’, which were defined as mutations in genes crucial for haematopoietic cell function whose deregulation initiate cancer progression. Further studies of mutational profiles of patients diagnosed with blood cancers help assessing molecular risk factors, as shown e.g. by Metzeler et al. who analysed recurrent mutations in a cohort of 664 patients diagnosed with AML and explored prognostic value of different driver mutations[100]. Measurements of the variant allele frequency (VAF) – a fraction of sequencing reads with the studied mutation of all reads within that locus, indicated that some mutations are present in most cells, whereas some only in few cells, which helped understanding which mutations occur early in the disease development as these mutations have higher VAF. In 2014 several research groups detected

presence of driver mutations at low VAF in older individuals without any symptoms of haematological malignancies, this discovery was called clonal haematopoiesis of intermediate potential (CHIP, often shortened to CH) [101–104].

Nowadays, genomic landscapes of haematological malignancies are studied at a high resolution, e.g. with simultaneous analyses of DNA sequencing and cell proteome which help deciphering branching patterns of mutations acquisition in different malignant cell subclones, as performed by Morita et al[105]. Moreover, analyses of evolution of distinct clones at a single cell level reveal combinations of mutations that support further malignancy progression[106]. Altogether, this provide an opportunity to not only identify patients with a higher risk of blood cancer development, but also enable better understanding the molecular processes of blood cancer progression, proving that a singular mutation will not give rise to haematological malignancy, but rather a hierarchical selection of a mutated cell clones takes place in the patient's BM leading to the blood cancer expansion[90].

Mutations in various transcription factors (*CBFβ*, *CEBPα*, *GATA2*, *NPM1*, *RUNX1*, *TP53,* among others) and signalling proteins (*FLT3*, *JAK2*, *KIT*, *MPL*, among others) are detected in patients diagnosed with myeloid malignancies[107]. These mutations include insertions and deletions, chromosomal translocations, point mutations and copy number variations. Importantly, epigenetic control of haematopoietic cells is often dysregulated during myeloid malignancy development and progression[108]. Over the last ~15 years functions of various epigenetic writers, readers and erasers, such as *ASXL1*, *DNMT3A*, *MLL*, *IDH1* and *IDH2* (citric acid cycle enzymes that generate α-ketoglutarate) and *TET2* were found to be altered in blood cancer cells[107,109–115]. Current studies focus on understanding molecular consequences of these epigenetic mutations[107].

1.2.2. Disrupted TET2 activity in blood cancers

TET2 is primarily expressed in haematopoietic tissues[116–118] and its function is frequently disturbed in patients diagnosed with myeloid malignancies. *TET2* mutations are detected in 13%-20% of patients diagnosed with MPN, 18%-33% of

patients diagnosed with MDS, 22%-56% of patients diagnosed with CMML and 7.3%-23% of patients diagnosed with AML[119]. These mutations commonly include deletions, substitutions, and frameshift mutations[120]. Several gene mutations often cooccur with mutations in *TET2*, among these are mutations in genes encoding signalling proteins (FLT3-ITD, NRAS, KIT, JAK2), histone modifications (EZH2, BCOR), RNA splicing (SF3B1), DNA methylation (DNMT3A, TET1, TET3), TP53[121] and oncogenic-fusion genes (AML1-ETO)[122]. Importantly, *TET2* is one of the most common mutations detected in healthy persons with age-related clonal haematopoiesis. Moreover, clonal analyses indicated that *TET2* is often mutated at the early stages of the malignancy, but *TET2*-mutated haematopoietic cells do not drive the cancer development, indicating that additional mutations seem to be necessary for the disease progression. However, since the type and order of subsequent mutations is different for almost every patient, to fully understand myeloid cancer pathogenesis we should study molecular mechanisms of expansion of each malignant cell clone. Therefore, current research focuses on explaining the process of acquisition of additional mutations by *TET2*-mutated HSPCs leading to the malignancy development[116,119].

The discovery of TET2 dioxygenase function[123,124] and its mutations in the myeloid diseases', initiated studies on understanding TET2 function in haematopoietic cells. In 2011, Ko et al. generated a *Tet2*-deficient mouse model with deletion of exons 8, 9 and 10 in the catalytic double-stranded β-helix domain of the *Tet2* gene[2]. These mice displayed a CMML-like phenotype providing an excellent model to study the origins of myeloid malignancies. They showed that *Tet2* knock-out HSCs have a potential for the multilineage reconstitution, show a myeloid-biased cell differentiation and have a competitive advantage over wild-type HSCs *in vivo*[2]. Importantly, the absence of *Tet2* also increased a number of HSPC populations in the BM, indicating that HSC self-renewal is regulated by TET2[2]. Soon further studies using mice models revealed a pleiotropic role of TET2 in haematopoietic cells, indicating that TET2 controls lineage commitment at various blood differentiation stages[119]. Interestingly, study led by Ito et al. showed that the loss of TET2 catalytic activity is responsible for the myeloid bias, but it is the disturbed non-catalytic function of TET2 that drives increased HSC self-renewal[125]. Importantly, due to its

enzymatic-independent activity TET2 also plays a role in histone modifications[116]. Moreover, latest study by Flores et al. showed that mice with mutated catalytic function of TET2 developed myeloid malignancies, older mice with *Tet2* knockout develop myeloid malignancies, however young mice with *Tet2* knockout developed both lymphoid and myeloid malignancies[126]. It seems that mice age is important in regard to TET2 function. A recent study by Holmes et al. highlighted a difference in numbers of Lin-Sca1+c-kit+ HSPCs in wild-type (C57BL/6) mouse model dependent not only on mice age but also sex[127]. They showed that males have more HSPCs than females, moreover, there seem to be a sex-dependent difference in HSPC numbers in *Tet2* transgenic mice between the age of 6 weeks and 4 months[127].

Crucially, recent epigenomic analyses indicated that *TET2* knockout results in genome-wide increase of methylation at enhancer sides. This led to the hypothesis that TET2 shapes methylation profiles at enhancers and promotors of the key genes responsible for haematopoietic cell fate. Indeed, several studies showed that loss of TET2 results in enhancer hypermethylation and changes of transcription factors activities at cell-type specific enhancers[70,128–133]. Altogether, demethylation process initiated by *TET2* plays a crucial role in HSC cellular decisions, which was well documented for myeloid cells, where TET2 and key differentiation TFs bind enhancers of genes regulating cell fate[116,128]. Most notably, a product of TET2 activity - 5hmC is an important epigenetic mark of activated enhancers and promoters[119,129]. Deletion of *TET2* not only leads to DNA hypermethylation, but also reduces the level of 5hmC in the BM[119]. Moreover, 5-hydroxymethylation profiles are cell-type specific[134]. Altogether, TET2 deficiency dysregulates gene expression by altering methylation state of enhancers and disturbing 5hmC landscape[134].

Interestingly, as recently showed by Tian et al., DNA methylation not only regulates chromatin state but also chromosome organisation, with Tet-triple knockout (*Tet1/2/3*$^{-/-}$) weakening chromatin compartmentalisation in mouse embryonic stem cells[135]. Another study by Hong et al. showed that TET2 modulates spatial relocalisation of heterochromatin which contributes to HSPC aging[136]. Recent studies exploring TET2 non-enzymatic function in the regulation of transcription via histone modifications broaden our understanding of TET2 activity in haematopoietic cells[116].

1.2.3. TET2 partner protein function is disturbed in haematological malignancies

Intriguingly, other members of the TET family - TET1 and TET3 have a CXXC domain that enables them to bind to DNA, however TET2 lost this domain during evolution and depends on partner proteins on DNA binding. Over the last years researchers identified many proteins interacting with TET2, such as TFs, histone modifiers and signalling molecules[116], with the majority of them binding to the C-terminal region. Importantly, various TFs recruit TET2 to target genes which drives transcription[116,137–140]. Interestingly, studies showed that TFs display different binding affinities to methylated DNA - some prefer to bind methylated DNA, some 5hmC, some unmethylated DNA and other are not affected by the methylation status[116]. Intriguingly, all these groups of TFs seem to cooperate, successively binding DNA – firstly TFs recruit TET proteins to oxidate 5mC to 5hmC, then 5hmC-reader TFs join to support further demethylation and finally methylation-sensitive TFs recruit co-activators and co-repressors[116].

Aside from the proteins mentioned above, many other factors influence TET2 function. For instance, *TET2* expression is repressed by micro-RNA[116,141], e.g. miR-22 targets TET2 leading to leukaemia development[142]. Additionally, numerous post-translational modifications regulate TET2 activity, including phosphorylation, O-GlcNAcylation, ubiquitination and acetylation[116]. Moreover, the access to the cofactors - Fe^{2+}, α-ketoglutarate and O_2 also determines TET2 function[143,144].

Interestingly, although most research focuses on TET2 function on DNA, the latest studies aim to develop approaches to examine RNA 5-methylcytidine dioxygenases and investigate epitranscriptomic regulation of RNA[145]. Importantly, Zou et al. have recently showed that TET2 oxidizes chromatin-associated retrotransposon RNA 5-methylcytosine via MBD6, leading to H2AK119ub ubiquitination, closed chromatin state and decreased transcription[146]. Another study by Li et al. showed the importance of TET2-mediated demethylation of mRNA in

leukaemic cell homing and self-renewal[147]. Taken together this underlines how much is left to discover about the TET2 interactome.

1.2.4. TET2 and Myelodysplastic Syndrome

A study of 698 patients diagnosed with MDS showed that the average number of mutations was higher in older patients and these patients displayed a distinct genetic landscape than younger patients, suggesting an age-dependent mechanisms of disease development[148]. As *TET2* is often mutated in older individuals, this highlights the need for studies of TET2 in the MDS pathogenesis. In this thesis, the transcriptional landscape of patients diagnosed with MDS carrying *TET2* mutations was analysed. *TET2* is commonly mutated in MDS and patients with MDS also display a lower *TET2* expression[149] with concomitant downregulation of TET2 in high-risk MDS correlated with a shorter survival[150]. Additionally, patients diagnosed with MDS carrying mutations in *TET2* had a higher probability to transform into secondary AML[150]. Interestingly, recent DNA and RNA sequencing analyses of 1665 patients diagnosed with MDS suggested a compensatory mechanism, where following the lack of TET2 activity, TET3 seem to be upregulated, and patients with *TET2* mutations and a low *TET3* expression had a poor prognosis[149]. Intriguingly, Sun et al. showed that even a single post-translational modification of TET2 protein can disturb gene expression, giving MDS malignant cells clonal advantage[150,151]. Altogether, TET2 deficiency seems to promote leukemogenesis in patients diagnosed with MDS[150].

Aging, clonal haematopoiesis and MDS pathogenesis are closely associated with inflammatory signalling. Inflammation is also linked to transformation of MDS to secondary AML[152–154]. Moreover, TET2 and inflammation are closely related. Following inflammatory stress, the frequency of *Tet2*-knockout myeloid cells and HSPCs increases[155]. TET2 also seem to regulate the expression of inflammatory cytokines, such as IL6, in myeloid cells[87,156]. Altogether, the inflammatory environment further supports malignant cell expansion and MDS progression. Further studies of dysregulated signalling pathways will help justifying the expansion of *Tet2*$^{-/-}$ HSPCs upon inflammation.

Currently DNA hypomethylating agents, such as 5-azacytydine and decitabine are used for the treatment of MDS[157–159]. However, due to the high patient heterogeneity, it is challenging to design a strategy that would eradicate all malignant subclones. Moreover, the same genetic disease background regularly presents different phenotypes. Interestingly, a study by Schiroli et al. indicated that pre-malignant cell clones at similar differentiation status have diverse epigenetic states which can prone some clones to expand[160]. Therefore, extensive research is needed to develop therapies for patients diagnosed with haematological malignancies with *TET2* mutations. Importantly, TET2-targeted therapies should be tailored for specific cancer types, as TET function seem to differ in different cancers[116]. Moreover, in depth analyses of different *TET2* mutations in MDS and their molecular consequences and prognostic significance are needed[161,162]. For example, a recent study by Jin et al. discovered an unexpected TET2 activity upon mutations at residues involved in the 5-methylcytosine interactions, where TET2 specificity was changed to using thymine as a substrate, instead of 5-methylcytosine, further highlighting the need for more studies of *TET2*-mediated development of MDS[163].

1.2.5. The bone marrow microenvironment can drive malignancy development

It is a subject of intense research to what degree extrinsic factors are important for disease progression[164]. MDS pathogenesis is associated with significant bone marrow environment deregulation. Malignant cells can modify the niche to create a new environment where they expand and evolve at the same time suppressing the function of healthy haematopoietic cells[165–167]. Moreover, the inflammatory environment further supports the malignancy development through NF-κB, IL6, TGFβ1, IL1β and TNF signalling[165,168]. Altogether stress associated with the malignancy development hijacks niche interactome, altering communication between HSPC and the niche. Further studies are needed to delineate the niche cellular composition and the network of molecular interactions in the BM to design niche-oriented therapies.

1.3. Computational approaches to study haematological malignancy development

1.3.1. Bioinformatics enable high throughput analyses of cell state

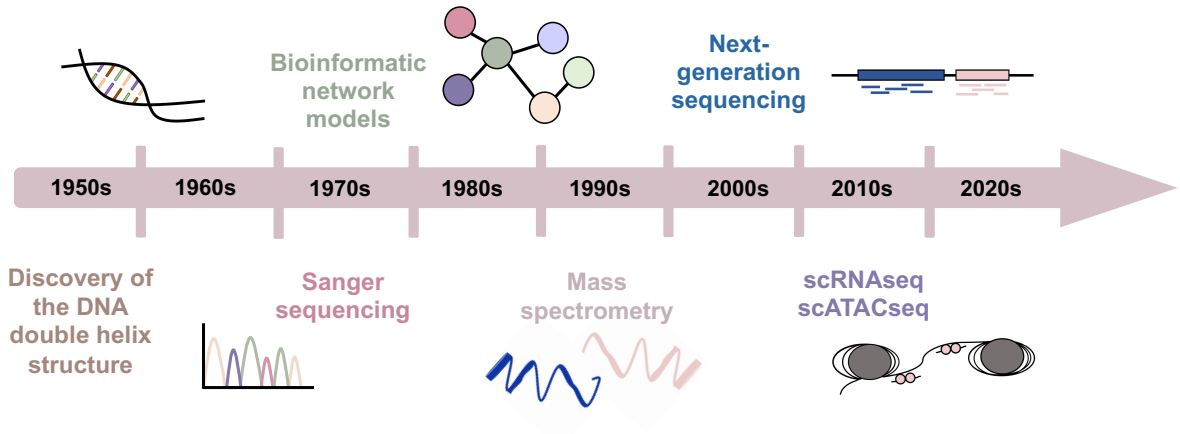


***Figure 1.7 An overview of the main milestones in bioinformatics**, simplified model; scRNAseq = single-cell RNA sequencing, scATACseq = single-cell ATAC sequencing.*

Scientists were eager to decipher the DNA code since the discovery of its structure in 1953[169]. That became possible in the 1970s when Frederick Sanger invented a DNA sequencing method[170]. Soon, the analysis of cellular proteome and chromatin accessibility turned out to be attenable[171,172]. The bioinformatic era began (Figure 1.7). Over the past few years, a plethora of new technologies enabling studies of several cell modalities in one assay were invented, such as e.g. G&T-seq[173], CUT&Tag[174], snNMTseq[175], CITE-seq[176], DOGMA-seq[177], NEAT-seq[178] among others, as summarised in Jassinskaja et al.[179]. Altogether this provides powerful opportunities to better understand HSC behaviour in homeostasis and malignancy development.

1.3.2. Exploration of cell transcriptome

At the end of 20th century scientists acknowledged a key role for RNA in a cell[180]. Quantification of gene expression with northern blotting and reverse transcription quantitative polymerase chain reaction (RT-qPCR) became widely used. In 1995 Schena et al. developed a first high-capacity system to study mRNA

expression patterns using complementary DNA microarrays[181]. However, microarrays used already known sequences as probes and therefore did not enable discoveries of new transcripts. This limitation was overcome with a milestone reached at the beginning of the 21st century - the invention of RNA sequencing (RNAseq) technology, followed soon by RNAseq analysis at the single-cell level (scRNAseq)[182].

A standard RNAseq experiment consists of RNA extraction, library preparation, sequencing and bioinformatic data analysis. Firstly, cells of interest need to be isolated, lysed and the mRNA obtained is then reverse transcribed into cDNA. At this step, adaptor and unique molecular identifiers (UMI) are also added to enable sequencing and molecular identification following sequencing. Then the cDNA is amplified and sequenced using next-generation sequencing platforms. Over the past decade, a multitude of library preparation protocols and sequencing technologies were invented, ranging from more expensive plate-based sequencing that enables detection of ~5,000 genes, to droplet-based technologies detecting ~3,000 genes in thousands of cells in one experiment. The key steps of computational pipelines for the analysis of RNA sequencing data are alignment to a reference genome, quality control (QC), normalization, dimensionality reduction, clustering, and differential expression analysis. Firstly, sequenced reads need to be mapped to a reference genome. Among the commonly used aligners are STAR[183] for bulk RNAseq and 10x Genomics CellRanger and STARsolo[184] for scRNAseq. RNA aligners should be splice-aware to account for reads spanning exon-exon junctions. Importantly, alignment usually requires a lot of computational power therefore sensitivity versus speed should be considered when choosing the aligner. There are many quality controls throughout the bulk and single-cell RNAseq data analyses. In case of bulk RNAseq among the most important QC steps are read quality check, e.g. with FastQC[185] which determines quality per-base, sequence length distribution, GC content and adapter content among others, and in the post-alignment QC assessment of the number of detected genes and detection of outliers in the principal component analysis (PCA)[4]. In the scRNAseq low-quality cells are filtered out, e.g. based on the number of genes detected per cell and the mitochondrial gene percentage. Later, PCA[4] and Uniform Manifold Approximation and Projection (UMAP[186]) visualisation help detecting batch effects. Another

important step of the RNA sequencing pipelines is data normalization to regress out technical noise, e.g. caused by the sequencing depth. In bulk RNAseq Reads Per Kilobase per Million mapped reads (RPKM) and Transcripts Per Million (TPM) are commonly used. scRNAseq data are more noisy, therefore more advanced tools are needed, such as e.g. SCTransform[187] using regularized negative binomial regression modelling.

Next-generation RNA sequencing measures expression of thousands of genes and therefore results in highly dimensional datasets. To process this amount of data various dimensionality reduction tools are used, such as UMAP[186] commonly applied for single-cell RNAseq or PCA[4] used for both bulk and single-cell RNAseq. Lower dimensionality datasets can then be visualised, which enables a first-pass exploratory analysis of the data. The next step is often sample/cell clustering with many clustering methods available to choose from, such as k-means clustering, hierarchical clustering, or graph-based community-detection including Louvain[188] and Leiden[189] algorithms. Clustering groups samples/cells into similar populations allows to explore data heterogeneity. Subsequently, a differential expression (DE) analysis is usually performed. The aim of DE analysis is to determine significant changes in gene expression between experimental conditions of interest. Among the most common DE tools for bulk RNAseq are DESeq2[190] and edgeR[191] and Seurat's FindMarkers[187] for scRNAseq. Finally, pathway analyses, such as KEGG[192], Reactome[193,194], Gene Ontology[195] (GO) and gene set enrichment (GSEA)[196,197] are usually performed to draw meaningful conclusions from the lists of differentially expressed genes.

There are many issues with RNA sequencing experiments, among the biggest technical limitations are dropout and batch effects. Dropout is characterised by not detecting expression of all genes, which results in scRNAseq data sparsity. Batch effects are biases introduced during wet-lab preparation of samples before sequencing. Careful planning is needed to minimise the risk of batch effects, including strategies such as randomised experimental design.

1.3.3. Assessment of the chromatin state

Soon after developing RNAseq technology a lot of effort was put into developing novel sequencing methods to explore other cell dimensions. In 2007 Chromatin Immunoprecipitation sequencing (ChIPseq)[198] was invented, a method based on the commonly used ChIP assay. ChIP assay detects genomic regions bound by the protein of interest using crosslinking and immunoprecipitation. With ChIPseq these interactions can be studied at a genome-wide scale. Invention of ChiPseq was soon followed by the development of DNase-seq in 2010, this technique built on commonly used in the past mapping of DNase I hypersensitive (HS) sites and facilitated examination of DNA regulatory elements across the genome[199]. Another genome-wide technique was developed in 2012 - MNase-seq provided an opportunity to examine nucleosome occupancy[200]. More recent designs of Cleavage Under Targets and Release Using Nuclease (CUT&RUN)[201] and Cleavage Under Targets and Tagmentation (CUT&Tag)[202] assays further improved our capability to profile epigenome, both, unlike ChiPseq, requiring less cells and providing better resolution. However, ChiPseq, CUT&Run and CUT&Tag technologies require initial decision on which protein binding to analyse, which enables in-depth studies of binding sites of selected proteins of interest, but cannot be used for exploratory analyses. That was made possible in 2013 with the development of Assay for Transposase-Accessible Chromatin using sequencing (ATACseq)[203].

ATACseq provides an opportunity to determine open chromatin regions and study gene regulation. This assay uses Tn5 transposase which cuts and tags accessible chromatin regions, these fragments are amplified and sequenced. ATACseq reads are then aligned to a reference genome and so-called accessible chromatin “peaks” are determined with algorithms such as e.g. MACS2[204]. Commonly used workflows for scATACseq data analysis include Signac[205] and ArchR[206]. Depending on the experimental aim, different samples/cell types can be compared to check for differences in the chromatin state. Peak annotation and TF motif enrichment analyses facilitate studies of transcription factor binding sites, altogether leading to more comprehensive studies of gene regulation[207].

1.3.4. Examination of cell proteome

Analyses of chromatin accessibility and transcriptome broaden our understanding of gene regulation. However, scATACseq and scRNAseq do not measure the abundance of actual proteins in a cell. For this aim mass-spectrometry-based (MS) proteomic technology is currently commonly used. In a standard MS-based proteomics protocol proteins are extracted, digested into peptides, separated using liquid chromatography and then analysed with a mass spectrometer. Proteome data analysis involves identifying peptides, matching them to proteins and quantifying protein abundances. Generated data are normalized and finally protein functions can be studied, e.g. using pathway analyses.

There are several limitations with the proteomics assays. Firstly, proteome coverage with current methods is not sufficient, and the low-abundance proteins are often not detected. Moreover, studies of proteome at the single-cell level has been challenging as proteins cannot be amplified like nucleic acids[208]. Altogether this made is difficult to explore proteome of HSCs which are very heterogenous and rare cell population. Nonetheless new proteomic technologies are constantly being developed and novel low-input mass-spectrometry based methods enable exploration of HSC proteomic landscape[1,209]. The field is evolving rapidly, including increased automation of sample preparation, innovations in liquid chromatography, more sensitive mass spectrometers and data-independent acquisition strategy[210]. Notably, spatial proteome analyses are now possible[211], including studies of formalin-fixed paraffin-embedded tissues[210]. Moreover, highly anticipated exploration of proteome at the single-cell level has now become reality[210,212]. Hopefully, the next years will make the cutting-edge proteomic technology widely accessible, fostering a deep understanding of biological processes.

1.3.5. Integrative analyses of different cell modalities

Studies of transcriptome, chromatin accessibility or proteome help us to better understand cell biology, however the greatest advantage comes with an integrative analysis of different cell modalities. The power that comes from analysing more than

one cell modality can be demonstrated by integrated analyses of scRNAseq and scATACseq datasets which can link open chromatin regions (identified with scATACseq) with cellular transcriptome (from the scRNAseq assay) and provide insights into DNA regulatory elements[213].

Overall, different cellular layers can be simultaneously studied using wet-lab and computational-based integration. Wet-lab approaches to analyse distinct cellular layers use protocols that capture information from different cell modalities in one experiment, e.g. snNMT-seq[175] that analyses chromatin accessibility, DNA methylation and cell transcriptome at the same time, or G&Tseq[173] that profiles genome and transcriptome from the same cell. Commonly used bioinformatic approaches rely on making shared projections of cells derived from different sequencing assays and transferring labels across[214,215]. Variation across distinct modalities can be examined using e.g. linear approaches[213], as presented in MOFA+[213,216], where latent factors are used to explain global variability facilitating reconstruction of low-dimensional projections. Seurat[187], a workflow commonly used to analyse scRNAseq data, has implemented weighted nearest neighbour (WNN) approach which constructs a graph using cell-specific modality weights to deduce neighbours[213,217]. Subsequent analyses typically include differential analyses aiming to compare cell subpopulations of interest[213]. Importantly, simply accumulating more data will not provide any novel insights without robust and thorough downstream analyses which can be supported by information curated in online databases[218]. Comprehensive analyses utilising knowledge on e.g. transcription factor binding sites, signalling pathways or ligand-receptor interactions accessed through research repositories can assist in decoding complex molecular interactions governing cellular activity[218].

It is necessary to highlight that computational integration of unpaired datasets, that are still widely generated[213,219], can be arduous as sequencing technologies differ in their sensitivities, resolution and signal-to-noise ratio[179,220]. Moreover, each cellular layer is heterogenous, noisy and have missing datapoints. Nevertheless, although integrative analyses of different cell modalities come with various challenges they also enable to get a far more comprehensive portrayal of cell behaviour than singular analysis of RNA sequencing data[214].

1.3.6. Regulatory network analyses

Over the recent years computational modelling of molecular interactions with networks have become increasingly popular[221]. One of the first logic-based biological models were genetic nets developed by Stuart Kauffmann in 1969[222,223]. Since then, the field has advanced rapidly as simulation of biological interactions in a form of a network provides possibilities to find patterns unobtainable when focusing on individual genes or proteins. Networks in biomedicine consist of nodes that represent proteins or genes connected by edges representing predicted interactions. There are various approaches currently used in the field, from models built manually out of targets identified in multi-omics data analyses[223], to curated disease maps[224] or networks reconstructed using correlations in gene expression[225]. Additionally, a wide range of types of networks exists, among which gene regulatory networks and protein signalling networks are frequently employed[226].

Gene co-expression networks are commonly constructed from single-cell RNAseq data. These networks are built using pairwise corelations in gene expression[227]. This strategy allows to examine patterns in expression in analysed cells, but it does come with pitfalls - gene co-expression networks contain many false-positive interactions, moreover, identified connections are not directional, making it more difficult to draw conclusions[225]. Adding molecular layers, e.g. information of transcription factor DNA binding from the ATACseq assay improves resolution and aids in development of more comprehensive networks. However, notably, inferring networks from unpaired omics data generated from different cells could give misleading conclusions, therefore these types of integrative analyses need to be performed carefully and strategically, bearing in mind all limitations and assumptions made along the way[228]. Commonly used frameworks for gene regulatory networks inference include GRNBoost2[229], SCENIC+[230], LEAP[231] or PECA[225,232]. Reconstructed networks can be further examined, e.g. with topological analyses such as betweenness centrality which allows to identify network master regulators. Eventually, hypotheses derived from constructed networks should be validated by wet-lab experiments[225].

To examine signalling pathways protein signalling networks are frequently built utilising next-generation sequencing data supported by prior knowledge from online databases. The source of sequencing data is usually an RNAseq dataset and prior knowledge is derived from various databases such as e.g. KEGG[192], BioGRID[233] or meta-recourses such as OmniPath[226,234]. Numerous workflows exist for this modelling approach such as e.g. NicheNet[235], HiPathia[236], CLIPPER[237] or CausalPath[238] among others, recently benchmarked by Garrido-Rodriguez et al[226]. In the assembled network, protein interactions are often condensed into discrete (-1,0,1) or Boolean (True/False) values to represent activation or inhibition allowing for further modelling of signal propagation. In addition, network representation allows to apply graph theory methods, such as clustering or shortest path analyses to better understand the complexity of analysed datasets. A limitation to these approaches is the fact that some proteins are more thoroughly researched which can introduce a bias represented by potentially artificial hubs and bottlenecks in constructed networks. Moreover, usage of databases to build signalling networks can be flawed due to a dearth of knowledge on interactions in different tissues in both humans and mice. Nevertheless, protein signalling networks provide invaluable insights into cell biology allowing robust hypotheses generation[226].

1.3.7. Understanding TET2 biology with single-cell ATAC and RNA sequencing

Following the discovery of *TET2* mutations in HSCs isolated from patients diagnosed with myeloid cancers, several *Tet2*-knock out mice models were generated[2,239,240]. The first experiments using these models aimed to examine blood counts and the potential to repopulate mouse recipients in transplantation assays. Soon, next-generation sequencing was applied to study TET2 role in haematopoietic cells. Nowadays, TET2 function is studied across many different cell and cancers types[241].

Since the discovery of TET2 researchers has rapidly broadened our understanding of its role in blood cells. Pioneering methylome profiling studies showed that TET2 deficiency results in hypermethylation of enhancers in haematopoietic cells[70,128,132]. Closer examination of TET2 binding to DNA revealed

many TET2 partners proteins[128]. Moreover, regulation of TET2 activity, e.g. by various post-translational modifications, is currently extensively investigated. Another area of active research is TET2 enzymatic independent activity[156]. In addition, mechanisms of TET-mediated oxidation of m5C mRNA are being discovered[146] building on recognised TET2 function on RNA in 2014[242]. Furthermore, emerging role of 5hmC as an epigenetic mark initiated conversion of its potential function as a diagnostic marker[123] with recent studies exploring 5hmC landscape in haematopoietic stem and progenitor cells[134]. On top of that, current studies of cell epigenetic profile help defining mechanisms governing expansion of *Tet2*-mutated clones[160].

It is a thriving time for deciphering TET2 biology and hopefully future years will enable development of TET2-targeted therapies. Despite a multitude of studies on TET2 over the last 16 years, molecular mechanisms behind the increased self-renewal of TET2-deficient HSC still remain elusive. Further analyses of chromatin state and transcriptome changes upon TET2 deficiency in haematopoietic cells will help identifying potential targets for future therapies.

1.3.8. Bioinformatic analyses assist in deciphering MDS development

Sequencing of HSPC isolated from patients diagnosed with MDS help resolving disturbed signalling pathways in malignant cells. However, due to a great heterogeneity of clonal expansion mechanisms between the patients, studies of cells with different clonal background are needed in order to design targeted therapies and eradicate mutated clones. Combined effort of researchers over the last decade help mapping various trajectories of malignant cell expansion. Current studies focus on explaining how different mutations functionally collaborate to drive malignancy progression.

TET2 is commonly mutated in patients diagnosed with MDS leading to its loss of function. Importantly, diagnosis of *TET2* mutations seem to be linked to a higher probability of MDS-AML transformation[150] but the molecular mechanisms of this transformation remain still unknown. Signalling networks offer an opportunity to

explore TET2-induced changes in HSPCs isolated from patients diagnosed with MDS. Identified targets could then be further validated with wet-lab experiments and provide a more comprehensive view of the disease development. Altogether, future studies need to focus on explaining clonal expansion mechanisms upon the loss of function of TET2 in HSPC in patients diagnosed with MDS.

1.3.9. Computational studies of the BM extracellular fluid will help defining niche interactome

HSC fate is regulated not only intrinsically but also by the BM niche. Over the last decades, many studies have focused on identifying distinct niche cell subpopulations, defining their phenotype and studying potential function in the BM. However, little attention was paid to the BM extracellular fluid (ECF).

The BM niche undergoes significant changes during aging, inflammation, and haematological malignancy development[48,243]. As BM ECF contains various proteins secreted by cells in the niche, proteomic studies of BM ECF will help deciphering dysregulated interrelationship of HSCs in the altered niche. Each malignant cell clone bearing distinct mutational background might secrete different proteins and interact with the niche in a different way. Deciphering the communication of HSC with mutations in a singular gene will help identifying ligands and receptors to directly enhance or interrupt unwanted interactions and design future therapies for patients diagnosed with haematological malignancies.

1.4. Thesis aims

HSC self-renewal is regulated by various external and internal mechanisms. However, mutations in certain genes might generate incorrect molecular instructions leading to a disturb HSC balance and malignancy development. Extensive efforts over the last decade provided lists of mutations commonly detected in patients diagnosed with haematological malignancies. Recent studies aim to explain molecular mechanisms of the disease onset and progression. Analyses of molecular consequences of each of these mutations will help in the identification of new

diagnostic strategies and therapeutic targets and hopefully contribute to the development of better treatment strategies for patients diagnosed with haematological malignancies.

The overall aim of my thesis was to better understand cell regulatory changes in haematopoietic stem cells following mutation in *TET2*. In Chapter 3 I performed integrative analysis of transcriptome and chromatin accessibility in *Tet2*-deficient mouse HSC aiming to better understand regulatory networks governing increased HSC self-renewal following *Tet2* knock out in a mouse model without any other mutations. The aim of analyses presented in Chapter 4 was to determine deregulated gene expression patterns in patients diagnosed with MDS with *TET2* mutations, identify disturbed pathways, potential connectivity to TET2 and overlap with targets defined in the mouse analyses from Chapter 3. In Chapter 5, I examined extrinsic regulation of *Tet2*-deficient mouse HSCs by analysing ligand-receptor interactions with the BM ECF.

2. Materials and methods

An overview of performed analyses presented in this thesis is summarised in Figure 2.1.

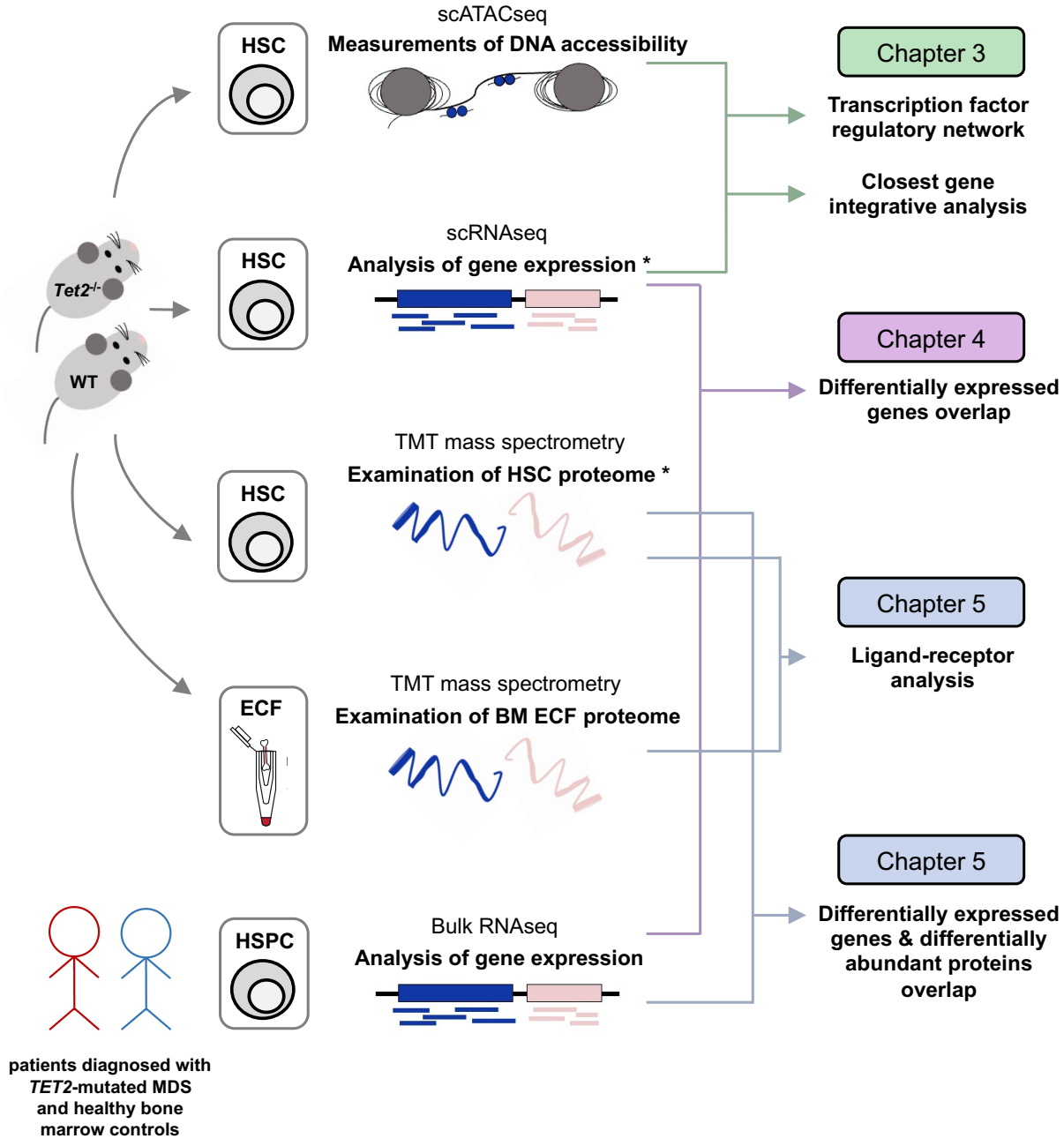


***Figure 2.1 An overview of used datasets and analyses presented in this thesis**; * scRNAseq and HSC proteome datasets were generated, analysed and published in Dr Daniel Bode's PhD thesis[1] and later reanalysed by Dr Maria Jassinskaja, scATACseq dataset was generated by Dr Daniel Bode and Dr Alyssa Cull, ECF proteome dataset was generated by Dr Sudip Ghosh and Dr Maria Jassinskaja - please see all details on contributions of Prof David Kent's lab members in the Materials and Methods chapter. HSC = haematopoietic*

stem cells, HSPC = haematopoietic stem and progenitor cells, MDS = Myelodysplastic Syndrome, scATACseq = single-cell Assay for Transposase-Accessible Chromatin with sequencing, scRNAseq = single-cell RNA sequencing, TMT = Tandem Mass Tag proteomics, BM ECF = bone marrow extracellular fluid, WT = wild type.

2.1. Mice

All presented mice experiments were performed using wild-type (C57BL/6N) and *Tet2*$^{-/-}$ and *Tet2*$^{+/-}$ Ko et al[2]. mouse model bred in-house. Mice were housed at the Biological Service Facility (BSF) at University of York. All mice were kept in specified pathogen-free conditions, and all procedures performed according to the United Kingdom Home Office regulations, in accordance with the Animal Scientific Procedure Act.

2.2. Human samples

Bulk RNA sequencing of human samples was performed at the Princess Margaret Cancer Centre, University Health Network in Toronto, Canada in collaboration with Prof Kristin Hope's lab. All permissions/certificates to work with human BM samples were provided by Prof Kristin Hope's group, University Health Network, Toronto.

2.3. Cell isolation

2.3.1. Isolation of mouse HSCs for scATACseq

Isolation of mouse HSCs for the single-cell Assay for Transposase-Accessible Chromatin with sequencing (scATACseq) was performed by Dr Daniel Bode, Dr Alyssa Cull and the University of York Technology Facility.

5 WT mice and 5 *Tet2*$^{-/-}$ mice were used, mice were males aged 12-16 weeks. Bone marrow was extracted from tibiae, femurs, hips, sternum and spines. Bones were crushed using a mortar and pestle and cell suspension was filtered through a

40 µm filter. Red blood cells were lysed using ammonium chloride (Stem Cell Technologies). Mature cells were depleted using EasySep Mouse Hematopoietic Stem and Progenitor Cell Isolation Kit (Stem Cell Technologies). Long-term haematopoietic stem cells (LT-HSCs, $CD45^+EPCR^+CD150^+CD48^-$) were isolated by fluorescence-activated cell sorting (FACS) with Beckman Coulter MoFlo Astrios cell sorter using antibodies listed in Table 2.1. 4,000 WT and 4,000 *Tet2*$^{-/-}$ LT-HSCs were sorted.

Antibody	Clone	Fluorochrome	Supplier
EPCR	RMEPCR1560	PE	STEMCELL Tech
Sca1	D7	BV421	Biolegend
CD150	TC15-12F12.2	PECy7	Biolegend
CD48	HM48-1	APC	Biolegend
CD45	30-F11	FITC	Biolegend
Viability	-	7AAD	Life Technologies

Table 2.1 Antibody panel used to isolate HSCs for the scATACseq assay.

2.3.2. Isolation of mouse HSCs for scRNAseq

Isolation of mouse HSCs for single-cell RNA sequencing was performed by Dr Daniel Bode and is described in his PhD thesis[1]. Mice were males aged 12-16 weeks.

2.3.3. Isolation of mouse HSCs for quantitative RT-qPCR

Isolation of mouse HSCs for quantitative RT-qPCR was performed by Alexander Hogg, Joanna Milek and Monika Gońka.

Cells were isolated using the same protocol as described for scATACseq (see 2.3.1 Isolation of mouse HSCs for scATACseq), with the exception that both long-term HSCs (LT-HSC, $CD45^+EPCR^+CD150^+CD48^-$) and short-term self-renewing HSCs (ST-HSC, $CD45^+EPCR^+$ $CD150^-CD48^-$) were FACS sorted and cells isolated

from *Tet2*$^{-/-}$, *Tet2*$^{+/-}$ and WT mice were analysed, a mix of male and female animals was used, aged 11-18 weeks.

2.3.4. Isolation of human bone marrow haematopoietic cells for bulk RNAseq

Cell isolation by FACS and RNA library preparation were performed in collaboration with Dr Nicholas Wong, Tony Chen, Emily Tsao, Pratik Joshi, Amanda Tajik and Prof Kristin Hope at the Princess Margaret Cancer Centre, University Health Network, Toronto, Canada.

BM samples were isolated from five patient samples, those patients were diagnosed with MDS with detectable *TET2* mutations. Samples were preserved in DMSO (following a standard cryopreservation protocol) and stored in -80^{0}C. Three healthy commercially obtained BM samples were analysed as a healthy control (Lonza human BM, cat # 2M-125C). All samples and files were deidentified.

Antibody	Clone	Fluorochrome	Supplier
CD34	581	BUV395	BD Biosciences
CD38	HB7	PE	BD Biosciences
CD45RA	HI100	FITC	BD Biosciences
CD90	5E10	PECy7	BD Biosciences
CD49f	GoH3	AF647	BD Biosciences
Viability	-	SYTOX^{TM} Blue Dead Cell Stain	Invitrogen

Table 2.2 Antibody panel used to isolate human haematopoietic cells for bulk RNA sequencing.

On the sort day, all BM samples were thawed in the water bath (37^{o}C), transferred to a buffer containing FBS, DNaseI and heparin (5 ml of FBS, 1 mg/mL in 5 mL of DNaseI and 20U/mL in 5 mL of heparin) and then topped up to 50 mL with PBS (final concentration of DNase 0.1 mg/mL, heparin 2U/mL). Cells were counted, pelleted (475 x g, 5 min) and resuspended in 100 µl of PEF (PBS; 2% FBS; 2 mM EDTA) for staining. After 30-minute incubation Lineage$^-$CD34$^+$, Lineage$^-$CD34$^+$CD38$^-$

and Lineage$^-$CD34$^+$CD38$^+$ cells were sorted using the antibody panel listed in Table 2.2 on the BD FACSAria^{TM} Fusion cell sorter. Final numbers of sorted cells are shown in Table 2.3.

Sample	CD34$^+$	CD34$^+$CD38$^-$	CD34$^+$CD38$^+$
MDS1	809	-	-
MDS2	7000	1678	2850
MDS3	7000	6927	9076
MDS4	6834	-	-
MDS5	7000	870	8287
BM1	15000	11000	26000
BM2	10000	10000	36600
BM3	9780	1900	8100

Table 2.3 Numbers of sorted human haematopoietic cell fractions.

2.3.5. Isolation of mouse HSCs for TMT mass-spectrometry proteomics

Isolation of mouse HSCs for TMT mass-spectrometry proteomics was performed by Dr Daniel Bode and is described in his PhD thesis[1]. A mix of female and male animals was used, mice were 30-weeks old.

2.3.6. Isolation of BM extracellular fluid for TMT mass-spectrometry proteomics

Isolation of bone marrow extracellular fluid (BM ECF) for TMT mass-spectrometry proteomics was performed by Dr Maria Jassinskaja and Dr Sudip Ghosh at the University of York and the Lund University following the protocol described in Jassinskaja et al., Leukemia 2024[244]. Briefly, femur and tibia bones were analysed from middle-aged (25-53 weeks) wildtype, *Tet2*$^{+/-}$ and *Tet2*$^{-/-}$ mice (four WTs, two *Tet2*$^{+/-}$, four *Tet2*$^{-/-}$). A mix of female and male animals was used. Bones were punctured at both ends and placed on in-house made '3-trap' tubes. 50 µl sterile PBS was added to the inner tube, and the samples were centrifuged at 300G for 5 min and an additional 2 min at 500G without bones before collecting the

ECF. Samples were further cleared at 2000G, concentrated with Amicon 3 K filters, snap frozen and kept at −80°C for further processing.

2.4. Libraries preparation and sequencing

2.4.1. scATACseq library preparation

10x Genomics scATACseq was prepared by Dr Daniel Bode, Dr Alyssa Cull and Dr Sally James at the University of York Technology Facility.

Cell nuclei isolation was performed using the "*Nuclei isolation for single cell ATAC sequencing*" protocol (https://cdn.10xgenomics.com/image/upload/v1660261285/support-documents/CG000169_DemonstratedProtocol_NucleiIsolation_ATAC_Sequencing_Rev_E.pdf) by 10xGenomics (www.10xgenomics.com), instructions for the "*Low Cell Input Nuclei Isolation*" were followed. 1-hour tagmentation with nuclei in 10xGenomics isolation buffer (www.10xgenomics.com) was run, and using the tagmentase enzyme provided with the 10xGenomics Chromium Next GEM Single Cell ATAC Reagent Kits v1.1 (www.10xgenomics.com), before loading all of the isolated nuclei (total loaded cell counts of between 1100 and 5500 per sample) onto the 10X Genomics Chromium Chip H chips for GEM (Gel Bead-In Emulsion) generation (www.10xgenomics.com) and subsequent library generation, following the manufacturer's recommended protocol, and using 12 PCR cycles for GEM amplification. Each ATACseq library was amplified using a unique set of indexing primers, and libraries were pooled at ratios proportional to the estimated number of cells recovered prior to sequencing. Paired-end sequencing was performed on a NextSeq 2000 at Leeds University Next Generation Sequencing facility using 50, 8, 24 and 49 base sequencing for Read 1, Index 1, Index 2 and Read 3 sequences respectively.

2.4.2. scRNAseq library preparation (mouse)

Mouse scRNAseq library preparation was performed by Dr Daniel Bode and is described in his PhD thesis[1].

2.4.3. Bulk RNAseq library preparation (human)

Freshly sorted cells were subjected to bulk RNA sequencing. RNA was extracted using the Arcturus Picopure RNA isolation kit (Applied Biosystems, cat # KIT0204). Libraries were prepared using the NEBNext Single-cell/Low Input RNA Library Prep kit, quality control was performed using Bioanalyzer system (Agilent). Library preparation was performed by members of the Next-Generation Sequencing Facility at the Centre for Applied Genomics at Hospital for Sick Children in Toronto. Constructed libraries were sequenced using the NovaseqX (full lane/10B reads), paired-end 100.

2.5. Proteomic sample preparation and analysis of mouse HSCs

Proteomic sample preparation and analysis of mouse HSCs was performed by Dr Daniel Bode and is described in his PhD thesis[1]. This dataset was reanalysed by Dr Maria Jassinskaja and generated list of differentially abundant proteins (abs($log_2FC \geq 0.5$)) was used by Monika Gońka for the downstream analyses presented in this thesis.

2.6. Proteomic sample preparation and analysis of mouse BM ECF

Sample preparation, raw data processing and differential abundance analysis for *Tet2*$^{-/-}$ and WT bone marrow extracellular fluid TMT mass spectrometry proteomics samples were performed by Dr Sudip Ghosh and Dr Maria Jassinskaja at the University of York and the Lund University following the protocol described in Jassinskaja et al., Leukemia 2024[244].

Briefly, samples were diluted in 50 mM ammonium bicarbonate (Sigma) with 0.1% RapiGest (Waters) to denature proteins and shaken on a thermomixer (Eppendorf) at 400 RPM for 15 min decreasing the temperature from 80°C to 56°C, followed by reduction of disulfide bonds with 0.1 M dithiothreitol at 56°C, cysteine

alkylation with 0.2 M iodoacetamide at room temperature in the dark, and digestion overnight at 37°C with sequencing grade modified trypsin (enzyme: protein ratio 1:50, Promega). Digested peptides were acidified with 10% trifluoroacetic acid and RapiGest was precipitated by incubation at 37°C, followed by centrifugation. Peptides were desalted with SepPak C18 cartridges (Waters), dried by vacuum centrifugation, and stored at −80°C. To generate a spectral library for the deep proteomic profile, 15-20 µg of peptides pooled from each sample were fractionated into 13 fractions by high-pH reversed-phase fractionation. Fractioned samples were dried by vacuum centrifugation and stored at −80°C.

Liquid chromatography and mass spectrometry (LC-MS) analyses were carried out on an Orbitrap Exploris 480 MS instrument with a reverse phase UltiMate 3000 UHPLC system via an EASY-Spray ion source equipped with FAIMS Pro (all Thermo Fisher Scientific). Digested peptides were loaded onto a trap cartridge (Acclaim PepMap C18, Thermo Fisher Scientific) and separated on an EASY-Spray analytical column (2 mm particle size, Thermo Fisher Scientific). Spectronaut (version 18, Biognosys AG) was used to build the spectral library. Protein quantification and statistical analysis were performed with Msstats1 (version 4.12.0). Contaminants were filtered and features were converted to MSStats format for downstream processing. Uninformative features were removed and missing values were imputed with the 'MBimpute' function within MSStats. For statistical analysis, MSStats 'group comparison' was performed between *Tet2*$^{-/-}$ or *Tet2*$^{+/-}$ and wildtype samples.

2.7. Quantitative RT-PCR profiling

RNA isolation, reverse transcription and quantitative RT-PCR was performed by Alexander Hogg, Joanna Milek and Monika Gońka. All qRT-PCR data was analysed by Monika Gońka.

RNA was extracted using the Picopure RNA Isolation Kit (ThermoFisher) according to the manufacturer's protocol. cDNA was synthesised using SuperScript IV First-Strand cDNA synthesis protocol, according to the manufacturer's protocol.

RT-qPCR was performed using the TaqMan protocol (ThermoFisher) using assays listed in Table 2.4, according to manufacturer's protocol and analysed on the StepOne Real-Time PCR (Applied Biosystems) machine.

Gene name	Assay ID (TaqMan assay, ThermoFisher)	Probe	Cat#
Dusp1	Mm00457274_g1	FAM-MGB	4453320
Egr1	Mm00656724_m1	FAM-MGB	4453320
Gapdh	Mm99999915_g1	FAM-MGB	4351368
Hes1	Mm01342805_m1	FAM-MGB	4331182
Nr4a1	Mm01300401_m1	FAM-MGB	4453320
Nr4a2	Mm00443060_m1	FAM-MGB	4453320
Nr4a3	Mm00450071_g1	FAM-MGB	4453320
Plag1	Mm00451356_m1	FAM-MGB	4453320
Rgs1	Mm00450170_m1	FAM-MGB	4453320
Smarcc1	Mm00486224_m1	FAM-MGB	4453320
Socs3	Mm01249143_g1	FAM-MGB	4453320

Table 2.4 Mouse TaqMan assays used in the RT-qPCR experiments.

Gene expression levels were normalised to the mean expression of GAPDH housekeeping gene and ΔCt values were calculated. Fold change expression across all samples and expression for each target gene was calculated and visualised using a custom Python script.

2.8. Quality control

Single-cell multi-modal bioinformatic analyses were conducted with help of Prof Bertie Göttgens' research group (Cambridge Stem Cell Institute, University of Cambridge), Prof Kristin Hope's group (University Health Network, Toronto), and help of the Data Science group led by Dr Alastair Droop (Technology Facility, University of York).

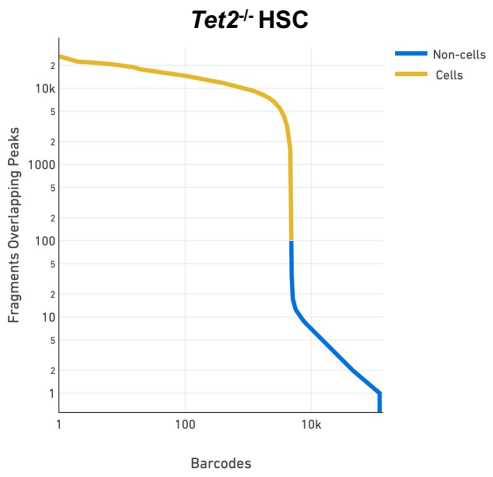
Single-cell computational analyses of mouse data were performed on the University of York Research High Performance Computing Clusters (The Viking clusters, Centos Linux 7.3 and Rocky 8.8) and MacBook Air (macOS Ventura 13.4, M1 chip). scRNAseq alignment, reads sorting and gene counts were computed by Dr Daniel Bode, Dr John Davey and Dr Iwo Kucinski. The downstream pipeline based on the Seurat[217] workflow was provided by Dr Daniel Bode[1] and later optimised by Monika Gońka with Sandy MacDonald's help (Data Science Group, University of York). scATACseq data demultiplexing and alignment was run by Dr Sophia Ahmed (Data Science Group, University of York) and the downstream pipeline was optimised by Monika Gońka with Bart Theeuwes, Dr MS Vijayabaskar, and Dr Nicola Wilson's help (University of Cambridge).

The analyses of bulk RNA sequencing data of all human samples were run using the Health Network High Performance Computing Cluster HPC4Health (https://www.hpc4health.ca/) and downstream analyses on MacBook Air (macOS Ventura 13.4, M1 chip). Dr Soheil Jahangiri (University Health Network, Toronto) run alignment, quality control, PCA[4], normalisation and differential expression analyses. All downstream analyses were performed by Monika Gońka.

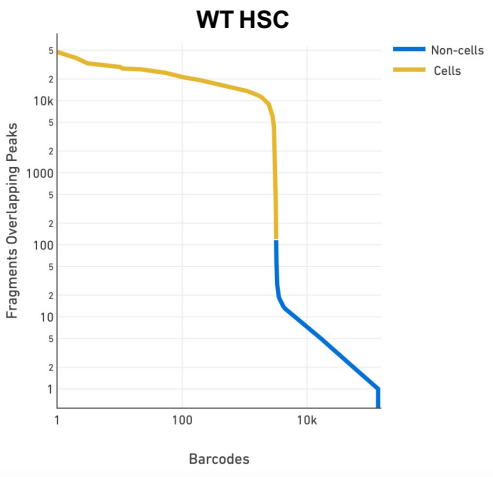
2.8.1. scATACseq

Data was demultiplexed using the BCL-to-fastq workflow with the Cell Ranger software (www.10xgenomics.com). Read alignment to a reference genome (mm10) was performed using CellRanger pipeline (CellRanger-ATAC, 10x Genomics, cellranger-atac count). Initial quality assessment indicated good data quality (Figure 2.2). For further scATACseq data analysis ArchR[206] workflow was used.

A



B



C

	Tet2$^{-/-}$ HSC	WT HSC
Estimated number of cells	4,822	3,208
Mean raw read pairs per cell	15,800.24	30,904.76
Fraction of high-quality fragments in cells	95.9%	94.0%
Fraction of transposition events in peaks in cells	63.2%	56.6%
Median high-quality fragments per cell	10,086	19,690

Figure 2.2 Quality control assessment by CellRanger (10x Genomics Cell Ranger, 10xgenomics.com)*; fragments overlapping peaks, sample order: A) Tet2*$^{-/-}$ *HSCs, B) WT HSCs; C) Summary of the quality control (CellRanger). Plots made with ArchR*[206].

Per cell quality control and doublet inference was done with ArchR[206] (Figure 2.3). TSS (transcription start site) enrichment score was analysed to assess signal-to-background ratio and exclude possible dead cells with the de-chromatinzed DNA (Figure 2.3C). Fragment size distribution plots showed periodical distribution indicating expected pattern of DNA wrapping around nucleosomes (Figure 2.3B). Because of the low intra-sample heterogeneity, the simulation of synthesized in silico doublets over the data was not used to exclude potential doublet cells (Figure 2.3A) and doublets were removed by filtering cells containing less than 40,000 fragments instead.

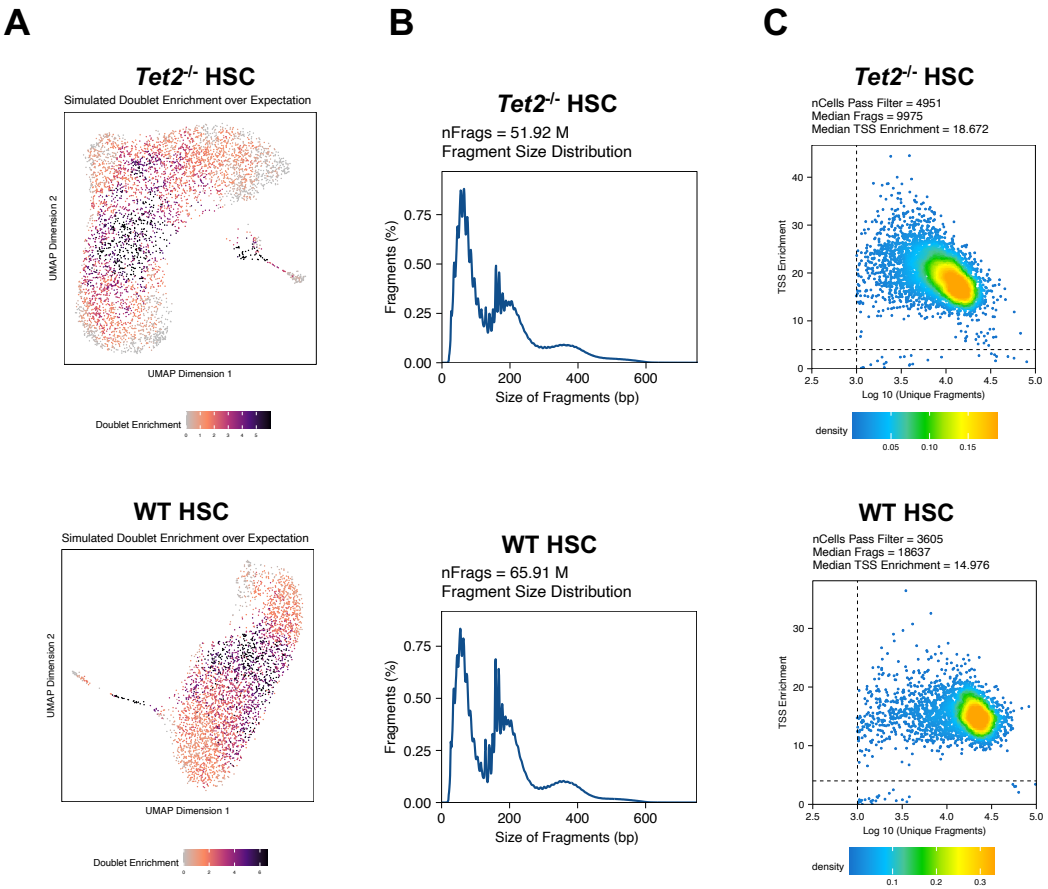


Figure 2.3 Data quality control by ArchR[206]*, sample order: Tet2*[-/-] *HSCs, WT HSCs; A) Simulated doublets enrichment, B) Fragment size distribution, C) TSS Enrichment. Plots made with ArchR*[206].

2.8.2. scRNAseq (mouse)

Sequenced reads were aligned to a reference mouse genome (mm10) using STAR aligner[183] (default settings), gene counts were computed using featureCounts[245] (default settings). To filter out low quality cells the quality control (QC) analysis was performed, results of the QC analysis are summarised in Figure 2.4. Following thresholds were applied: 2 x10^5 – 2.5 x10^6 reads per cell, <2% reads mapped to mitochondrial genome, <25% reads mapped to ERCCs spike ins, >30% reads mapped to the genome, >4000 genes detected in each cell. For downstream scRNAseq data Seurat[217] workflow was used.

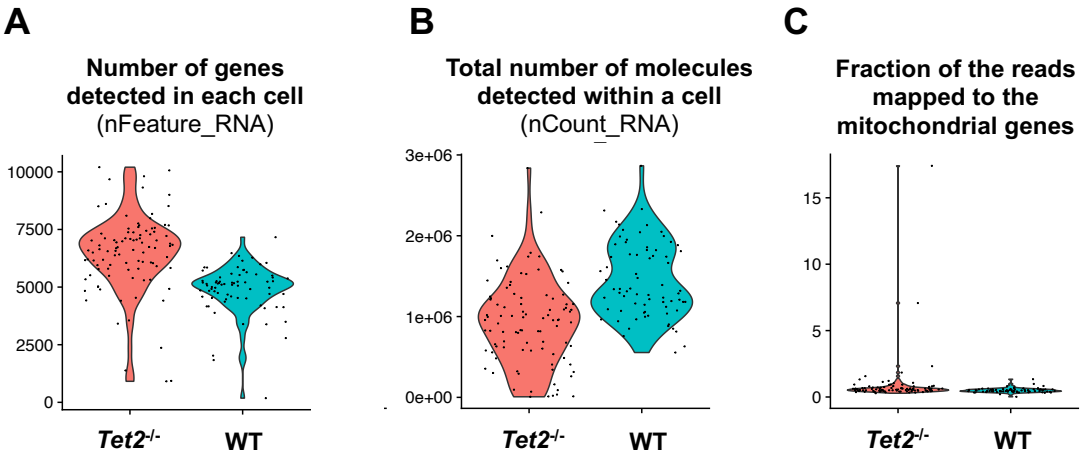


***Figure 2.4 scRNAseq quality control**, A) Number of genes detected in each cell; B) Total number of molecules detected within a cell; C) Fraction of the reads mapped to the mitochondrial genes. Plots made with Seurat[217].*

2.8.3. Bulk RNAseq (human)

Quality control was performed with MultiQC[246] general stats and FastQC[185]. Reads were mapped to GENCODE v38 using the STAR aligner[183]. The gene expression was quantified using FeatureCounts from the subread toolkit[245]. Genes with less than 10 total counts across all samples were excluded from the downstream analysis. Transcript abundances for the Nakauchi et al. dataset[134] were quantified with Kallisto[247].

2.9. Normalisation and dimensionality reduction

2.9.1. scATACseq

The term frequency-inverse document frequency (TF-IDF) normalization and the singular value decomposition (SVD) were performed on the QC-filtered data (latent semantic indexing, LSI)[248] using addIterativeLSI from the ArchR[206] package. Non-linear dimension reduction (UMAP[186]) was performed with ArchR's[206] addUMAP (Figure 2.5). Louvain clustering[188] was done using ArchR's[206] addClusters with resolution = 0.3.

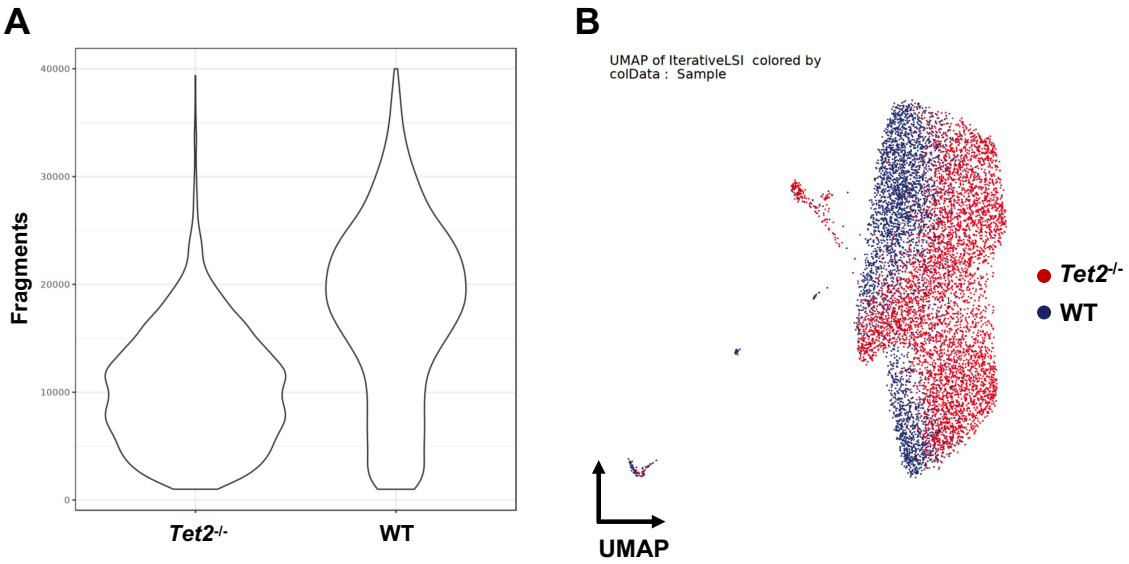


***Figure 2.5 scATACseq data analysis**; A) Fragment distribution across samples; B) UMAP[186] embedding. Plots made with ArchR[206].*

2.9.2. scRNAseq (mouse)

Data were normalised using regularised negative binomial regression (Seurat[217] SCTransform). Principal component analysis (PCA) linear dimensionality reduction[4] analysis was performed with RunPCA (default parameters), and top 5 principal components were selected as a representation of true dimensionality of analysed data. Uniform Manifold Approximation and Projection (UMAP)[186], a non-linear dimensionality reduction technique was used to determine graph-based clusters of cells using RunUMAP with default parameters. Local neighbourhoods were defined with FindNeighbors (first 5 PCs, Figure 2.6), and cells were clustered using the Louvain algorithm[188] (FindClusters).

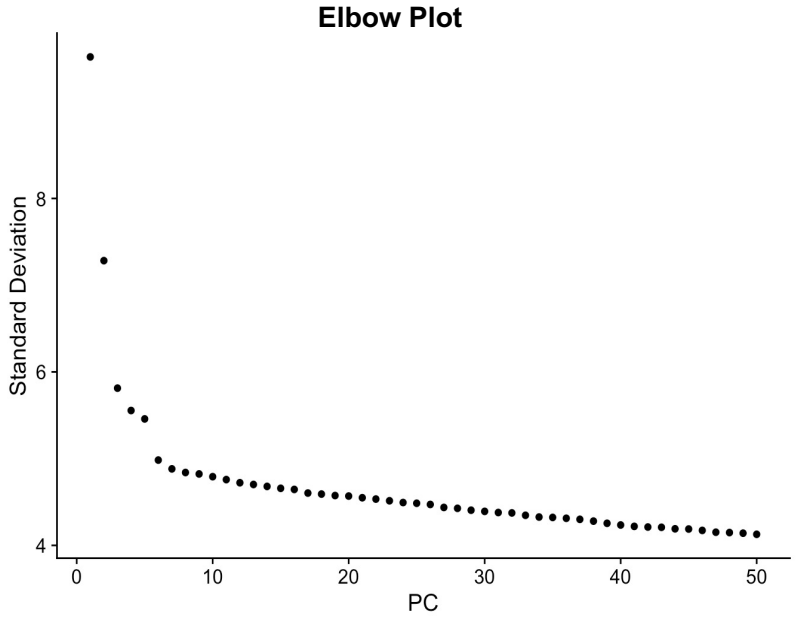


***Figure 2.6 scRNAseq Elbow Plot**, top 5 PC were selected for the downstream analysis; plot made with Seurat[217].*

2.9.3. Bulk RNAseq (human)

Normalisation and differential expression analysis based on the negative binomial distribution was run with DESeq2[190]. The counts were normalised for library size. The Limma[249] package was used to generate batch-corrected counts.

2.10. Motif and feature enrichment

2.10.1. Peak Calling, motif enrichment and closest gene analysis (scATACseq)

Pseudo-bulk replicates were created with ArchR's[206] addGroupCoverages function, replicates were made per WT/*Tet2*$^{-/-}$ sample (groupBy = "Sample"). Peaks were called on the defined groups of cells using the MACS2[204] algorithm (addReproduciblePeakSet). Marker peaks unique to individual groups were identified with ArchR's[206] getMarkerFeatures. Paired samples Wilcoxon test was used to compare WT and *Tet2*$^{-/-}$ samples (abs($\log_2$FC $\geq$ 0.5) & FDR $\leq$ 0.1, corrected for multiple testing using Benjamini-Hochberg method). Transcription factor binding motifs were annotated using ArchR's[206] addMotifAnnotations function and motif set

from the cisbp database[250] [http://cisbp.ccbr.utoronto.ca/] was used (chromVAR package[251]). Differentially accessible peaks were tested for motif enrichment with the peakAnnoEnrichment function from the ArchR[206] package. To visualise peaks BigWig files were exported with ArchR's[206] getGroupBW and then IGV[252] software was used. Closest genes to the accessible regions were identified if the distance to the transcription start site was < 100k base pairs (FDR <= 0.1 & $\log_2$FC $\geq$ 0.5).

2.10.2. Finding marker genes (scRNAseq)

Differentially expressed genes for *Tet2*$^{-/-}$ and WT cell groups were found with Seurat[217] FindMarkers. Significantly up/down regulated genes were defined if Q-value < 0.05 & abs(avg $\log_2$FC) > 1 (adjusted p-values were based on Bonferroni correction).

2.11. Gene Ontology term enrichment, KEGG pathway and GSEA analysis

Gene Ontology[195] and KEGG[192] feature enrichment were performed with Cytoscape[253]'s plugin ClueGO[254]. Information about the thresholds was found using the ClueGO online documentation available at http://www.ici.upmc.fr/cluego/ ClueGODocumentation 2019.pdf and http://www.ici.upmc.fr/cluego/ClueGO Documentation2013.pdf. Used thresholds are listed in each figure caption.

GSEA[196,197] analysis was performed using Hallmark Gene Sets (MSigDB[255,256]) and the BaderLab gene sets (Human_ GOBP_AllPathways_no_GO _iea_January_01_2024_symbol.gmt, www.baderlab.org /ChangjiangXu/GSEA).

2.12. Closest genes integrative scATACseq and scRNAseq data analysis

The lists of more/less accessible regions determined in the *Tet2*$^{-/-}$ HSC vs WT HSC pairwise testing analysis were intersected with the lists of genes showing significantly highest/lowest expression in the *Tet2*$^{-/-}$ HSC vs WT HSC scRNAseq analysis.

2.13. Transcription factor regulatory network analysis

DoRothEA[257] database was used to build the transcription factor regulatory network. More accessible transcription factor motifs detected in *Tet2*$^{-/-}$ HSC (scATACseq) that were also present in the DoRothEA[257] database, and all the differentially expressed genes (scRNAseq) that those TF regulate that were also present in the DoRothEA[257] database were filtered. Then an interaction graph containing defined regulons was constructed. Data was visualised with igraph[258], graphics[259], ggplot2[260].

2.14. Shortest path analyses

All protein network analyses were performed in collaboration with Prof Benjamin Hall (University College London).

Protein interactions were determined using the OmniPath R package[261] (following the Omnipath tutorial [https://github.com/saezlab/OmnipathR/blob/master/vignettes/omnipath_intro.Rmd]). Known signalling protein interactions were downloaded using import_omnipath_interactions function. Interaction graphs were built with interaction_graph. The shortest paths were calculated using all_shortest_paths. Graphs were visualised with the Cytoscape[253] software.

Shortest path analyses in Figure 3.7 was performed using all databases available in OmniPath[261]. Shortest path analyses in Figure 4.8 was performed using SignaLink3[262] and SIGNOR[263] databases accessed via OmniPath[261]. Shortest path analyses in Figure 5.3 was performed using all databases available in OmniPath[261]. Shortest path analyses in Figure 5.13 was performed using ligrecextra database[264] accessed via OmniPath[261]. Shortest path analyses in Figure 5.14C-F was performed using all databases available in OmniPath[261].

2.15. Integrative analysis of mouse and human RNA sequencing data

The list of targets identified with the integrative scATACseq/scRNAseq data analysis was intersected with the list of differentially expressed genes identified with the bulk RNAseq analysis of Nakauchi et al. data[134] and bulk RNAseq human MDS dataset. Human homologues to mouse genes were found using ENSEML Biomart[265].

2.16. Integrative human bulk RNAseq and mouse proteome analysis

The list of differentially expressed genes identified with the bulk RNAseq human MDS dataset was intersected with the list of differentially abundant proteins identified in the HSC proteome dataset[1]. Human homologues to mouse targets were found using ENSEML Biomart[265].

2.17. Mapping selected differentially expressed genes against the Nestorowa et al. dataset

UMAP[186] plots showing expression of selected targets in the Nestorowa et al.[266] dataset were generated using a web interface at https://blood.stemcells.cam.ac.uk/single_cell_atlas.html.

2.18. BM ECF – HSC ligand-receptor analysis

Differentially abundant ECF proteins were selected with abs(log_2FC) > 0.5 & adj pval < 0.05. Ligand-receptor interactions were determined using the OmniPath R package[261], using import_omnipath_interactions function, from the ligrecextra database[264]. Interaction graphs were built with interaction_graph and shortest paths from and to ligands and receptors were calculated using all_shortest_paths. Shortest path results were visualised with the Cytoscape software[253]. Graphs were generated with dNetInduce[267].

2.19. Protein complexes analysis

Protein complexes analysis in Figure 5.3 was run with OmniPath's[261] import_omnipath_complexes using CORUM[268], CellChatDB[269], Cellinker[270], CellPhoneDB[271], ComplexPortal[272] databases.

2.20. Protein network analyses with HuRi, Genemania and String

HuRi[273] human reference interactome database was accessed at http://www.interactome-atlas.org. Genemania and String networks were built using https://genemania.org and https://string-db.org web interfaces respectively.

2.21. Mapping ECF proteins against the Baccin et al. dataset

Mapping of proteins to the Baccin et al.[274] dataset was performed using CellHub by Nygen analytics at database.nygen.io/.

2.22. HSC *in vitro* culture with paquinimod and recombinant S100A9

This experiment was performed by Ewan Taylor-Smith. *Tet2*$^{-/-}$ and WT HSCs ($CD45^+EPCR^+CD150^+CD48^-$) were isolated like described in 2.3.1 Isolation of mouse HSCs for scATACseq for the scATACseq experiment. Cells were FACS sorted into 96-well plates, 10 cells/well, and cultured for 4 weeks using Wilkinson et al. protocol[275] with the exception that paquinimod (MedChemExpress, cat# HY-100442) and recombinant S100A9 (Biotechne, R&D Systems, cat# 2065-S9-050) were added to the cell culture medium at concentrations of 10 ng/mL ("low"), 50 ng/mL ("medium"), and 100 ng/mL ("high") in both cases. No treatment was added to the control plate wells. Cell clone sizes were estimated by a visual assessment under the microscope, based on the relative differences between wells (very small/small/medium/large clone), all estimations were performed by Ewan Taylor-Smith to ensure consistency. Cell clone sizes were determined at 2 and 3.5 weeks following the cell culture start day.

2.23. Flow cytometry analysis of S100A8/9 expression in mouse HSC

This experiment was performed by Alexander Hogg and Samuel Elberfeld. *Tet2*$^{-/-}$ and WT HSCs were isolated like described in 2.3.1 Isolation of mouse HSCs for scATACseq for the scATACseq experiment with the exception that cells were isolated from mouse spines only and instead of the FACS sorting, cells were stained with an antibody panel described in Table 2.5 and analysed with the Beckman Coulter's CytoFLEX LX flow cytometer aiming to measure expression of S100A8 and S100A9 in HSCs ($CD45^+EPCR^+CD150^+CD48^-$, ESLAM cell population). Flow cytometry data were analysed by Alexander Hogg using the FlowJo[276] software. Final plots were made by Monika Gońka using GraphPad Prism (Boston, Massachusetts USA, www.graphpad.com).

Antibody	Clone	Fluorochrome	Supplier
CD45	30-F11	BV785	BioLegend
CD48	HM48-1	BV421	BioLegend
CD150	TC15-12F12.2	PE-Cy7	BioLegend
EPCR	1560	PE	STEMCELL Technologies
Viability	-	Fixable viability dye eFluor 780	Invitrogen
S100a8	63N13G5	AlexaFluor488	Bio-techne, R&D Systems
S100a9	2B10	AlexaFluor647	BD Pharmingen

Table 2.5 Flow cytometry antibody panel used to check expression of S100A8/A9.

2.24. HSC gel culture with VWF

HSC culture gel culture with VWF was performed by Joanna Milek and Ellie Bennett, as described in the Jassinskaja et al. 2025 bioRxiv preprint publication.

Results

3. Chromatin accessibility and gene expression analyses of *Tet2*$^{-/-}$ LT-HSC to discover potential self-renewal regulators

An array of proteins take part in the addition and removal of a methyl group at the 5th position of cytosines in the DNA sequence[77]. TET dioxygenases play a crucial role in this process, converting 5-methylcytosine to 5-hydroxymethyl cytosine, and then 5-formylcytosine and 5-carboxylcytosine leading to thymine DNA glycosylase-mediated base excision repair and subsequent generation of an unmodified cytosine, eventually resulting in the DNA demethylation[77]. In 2009, several studies reported somatic mutations in *TET2* in blood cells of patients diagnosed with haematological malignancies[77,82–85,277]. Subsequent genome and transcriptome sequencing studies have generated molecular landscapes of some of the haematological malignancies with many of these being at the single-cell level[278–282]. Moreover, genomic studies showed that *TET2* mutations preceding disease development with *TET2* being one of the most frequently mutated genes in clonal haematopoiesis[283]. These data confer driver mutation status onto *TET2*, but also make it clear that a loss of function *TET2* mutation does not necessarily result in haematological disease. Therefore, studies of the cellular changes following TET2 mutations might provide invaluable insights into the origins of disease development.

Multiple studies using mouse models showed that TET2 is involved in the control of gene expression at several levels of haematopoietic differentiation hierarchy[119]. Considering the complex role of TET2 protein in the regulation of haematopoiesis, understanding molecular mechanisms following its loss of function in HSCs might help us identify targets responsible for the deviation from homeostasis in those cells and design clinical strategies to bring the equilibrium back[78,119]. In this chapter, we analysed chromatin accessibility and gene expression

of *Tet2*-mutated mouse HSCs using a mouse model with a disruption of the TET2 catalytic domain generated by Ko et al in 2011[2]. Those mice do not express a functional TET2 protein and display a CMLL-like disease phenotype, providing an excellent model to study pathways disrupted upon the absence of TET2 function. These results provide insight into the regulation of transcription in HSC upon *Tet2* mutation and will serve as a valuable resource for future research.

3.1. Chromatin accessibility measurements indicates increased accessibility in *Tet2*$^{-/-}$ HSC

To study the molecular mechanisms of HSC self-renewal we analysed chromatin accessibility of freshly isolated HSCs ($CD45^+EPCR^+CD150^+CD48^-$) from *Tet2*$^{-/-}$ (also referred in this thesis as *Tet2* knock-out, *Tet2*KO) mice with a loss of function of TET2 resulting from disruption in the *Tet2* catalytic domain[2]. scATACseq analysis indicated significant changes in the chromatin accessibility following mutation in *Tet2* ($abs(\log_2 FC \geq 0.5)$ & $FDR \leq 0.1$, libraries generated from 5 WT and 5 *Tet2*$^{-/-}$ mice, 4,000 WT HSCs and 4,000 Tet2$^{-/-}$ HSCs). Differential analysis showed increased accessibility in *Tet2*$^{-/-}$ HSC globally compared to their WT counterparts (Figure 3.1A). As expected, the analysis showed that detected peaks were skewed toward the closer distance to transcription start side (TSS) (Figure 3.1B). In order to find chromatin accessible regions unique to HSCs with no functional TET2, marker peaks analysis was performed. This analysis detected 71,589 peaks, however only ~1.2% of them were differentially accessible between *Tet2*$^{-/-}$ and WT samples (Figure 3.1C). 35% of significantly more accessible peaks in *Tet2*$^{-/-}$ cells, that were within 100kb distance to any gene, were localised in the distal regions and a significant portion (47%) in the intronic regions (Figure 3.1D).

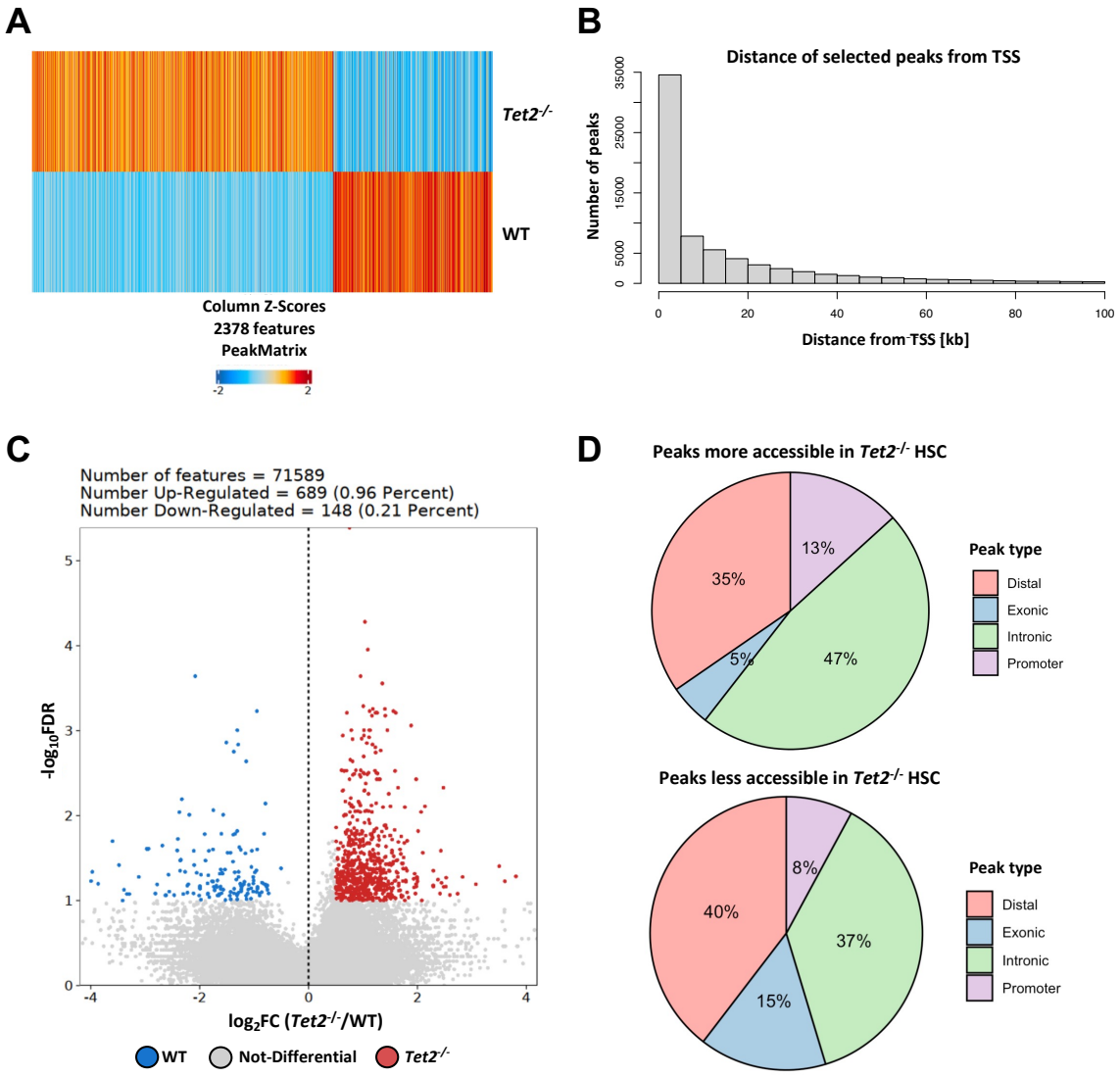


Figure 3.1 Peak calling analyses indicate increased chromatin accessibility in the Tet2$^{-/-}$ HSC samples in comparison to their wild-type counterparts*; A) Heatmap plot showing called marker peaks; B) Histogram of distance for the selected peaks from transcription start site (TSS, filtered distance to TSS < 100,000 bp), C) Volcano Plot comparing Tet2$^{-/-}$ and WT HSC samples, X axes = log$_2$FC (WT vs Tet2$^{-/-}$), Y axes = adjusted p-value; D) Pie charts showing distribution of significantly more/less accessible peak regions that are close to genes within the TSS < 100,000 bp distance. FDR = false discovery rate, FC = fold change. Plots made with ArchR[206], ggplot2[260] and formattable[284].*

A

Nearest gene	Chromosome	Start	End	log_2FC	FDR	Distance to TSS
Scube3	chr17	28140860	28141360	-3.415142	0.0998544311	1415
Capns1	chr7	30196145	30196645	3.074216	0.0641975061	1346
Mir3064	chr11	106784025	106784525	-2.327898	0.0064153238	247
Ninj1	chr13	49186622	49187122	2.296038	0.0666729931	674
Lmna	chr3	88503733	88504233	2.081160	0.0993701849	630
Erbb4	chr1	69108326	69108826	-2.009511	0.0871273425	516
Prima1	chr12	103242505	103243005	-1.994744	0.0675515673	608
Plek2	chr12	78908066	78908566	1.809460	0.0445139802	1377
Zfp133-ps	chr2	144459037	144459537	-1.652018	0.0872905490	6
Kcnk6	chr7	29233627	29234127	1.602511	0.0006162837	1354
Tekt4	chr17	25469861	25470361	-1.447094	0.0875870105	1478
Mir7000	chr2	92389119	92389619	1.423993	0.0231000015	1922
Robo3	chr9	37433788	37434288	1.402280	0.0005551848	862
Mir33	chr15	82197221	82197721	1.363698	0.0463898091	650
Cd53	chr3	106789912	106790412	1.327570	0.0482740087	12

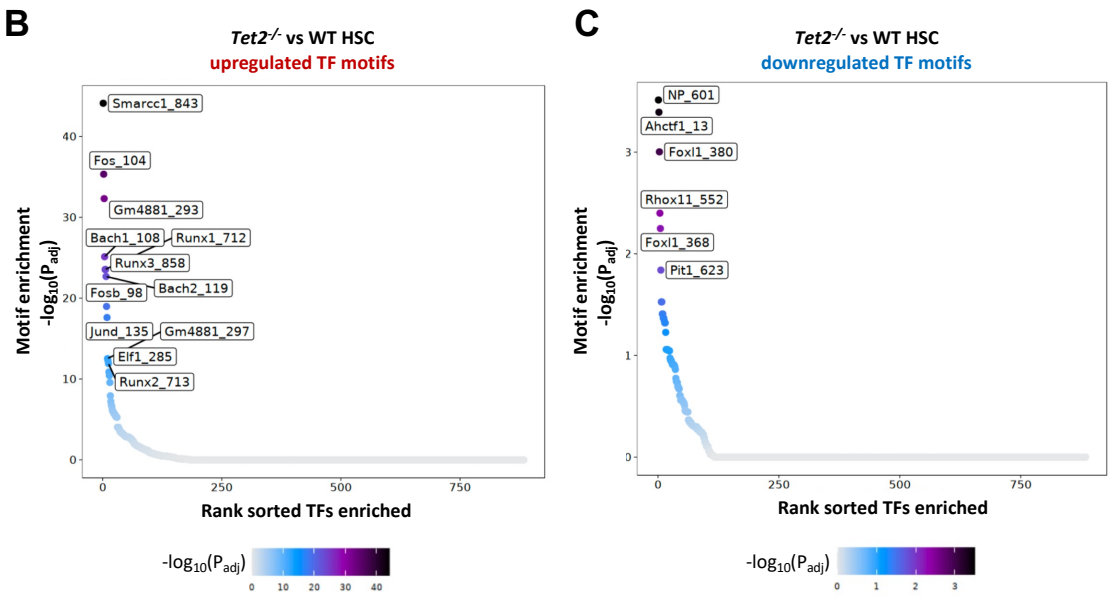


Figure 3.2 scATACseq data analyses suggests a set of targets deregulated in Tet2-deficient HSC*; A) Table presenting information on peak location, log_2FC, FDR and the distance to the transcription start site for the top 15 peaks (sorted by abs(log_2FC) value) called in the promoter regions; B-C) Transcription factor motif enrichment analysis, Tet2$^{-/-}$ vs WT pairwise testing, B) upregulated motifs in Tet2$^{-/-}$ HSCs, C) downregulated motifs Tet2$^{-/-}$ HSCs. Table made with formattable[284], plots made with ArchR[206].*

To get a better understanding of the location of detected chromatin peaks, a closest gene analysis was performed (threshold: distance to TSS < 100,000). Analysis indicated a set of interesting targets, potentially involved in the regulation of HSC self-renewal (Figure 3.2A). Next, transcription factor motif enrichment analysis

was carried out aiming to detect TFs binding in the determined marker peaks and predict what transcript factor's activities result in open chromatin states. The most enriched motifs in the peaks for each of the analysed sample were assessed. Pairwise testing between *Tet2*$^{-/-}$ and WT HSCs showed enrichment in a set of TFs, among which SMARCC1, RUNX1 and BACH1 binding motifs draw our attention, as they were shown to be involved in the regulation of self-renewal[285], myeloid differentiation[286] and cell cycle regulation[287], respectively (Figure 3.2B-C).

3.2. Single-cell gene expression profiling reveals increased expression in Tet2$^{-/-}$ HSC

To add another layer of information about the HSC molecular state we analysed cell transcriptome at the single-cell level (dataset published in Dr Daniel Bode's PhD thesis[1]), using the same *Tet2*$^{-/-}$ mouse model[2] as for the chromatin accessibility assay. Plate-based scRNAseq[3] was performed on the freshly isolated *Tet2*$^{-/-}$ and WT HSCs (CD45^{+}EPCR^{+}CD150^{+}CD48^{-})[1]. Differential gene expression analysis showed significant differences in gene expression upon *Tet2* loss of function (Figure 3.3A), indicating 131 differentially expressed genes (DEG, abs($\log_2$FC)>1 & p-value<0.05)[1]. scRNAseq data analysis indicated more upregulated genes in *Tet2*$^{-/-}$ HSCs (Figure 3.3B), suggesting more active transcription in these cells[1].

To determine function of identified differentially expressed genes, Gene Ontology[195] (GO) analysis was performed. GO analysis of biological processes showed enrichment in pathways regulating myeloid differentiation, regulation of transcription in response to stress and cellular response to calcium ion, among others (Figure 3.3C). These terms accorded with previous datasets implicating TET2 as a regulator of HSC differentiation into myeloid cells[2] and disturbed cellular response to calcium ions[288]. GO analysis of molecular function indicated e.g. enrichment in the promoter binding, indicating expected dysregulation of transcription machinery binding to promoters upon *Tet2* deficiency and enrichment in nuclear glucocorticoid receptor binding, suggesting possible disturbance of

glucocorticoid signalling leading to inflammatory response in *Tet2*-deficient HSCs (Figure 3.3D).

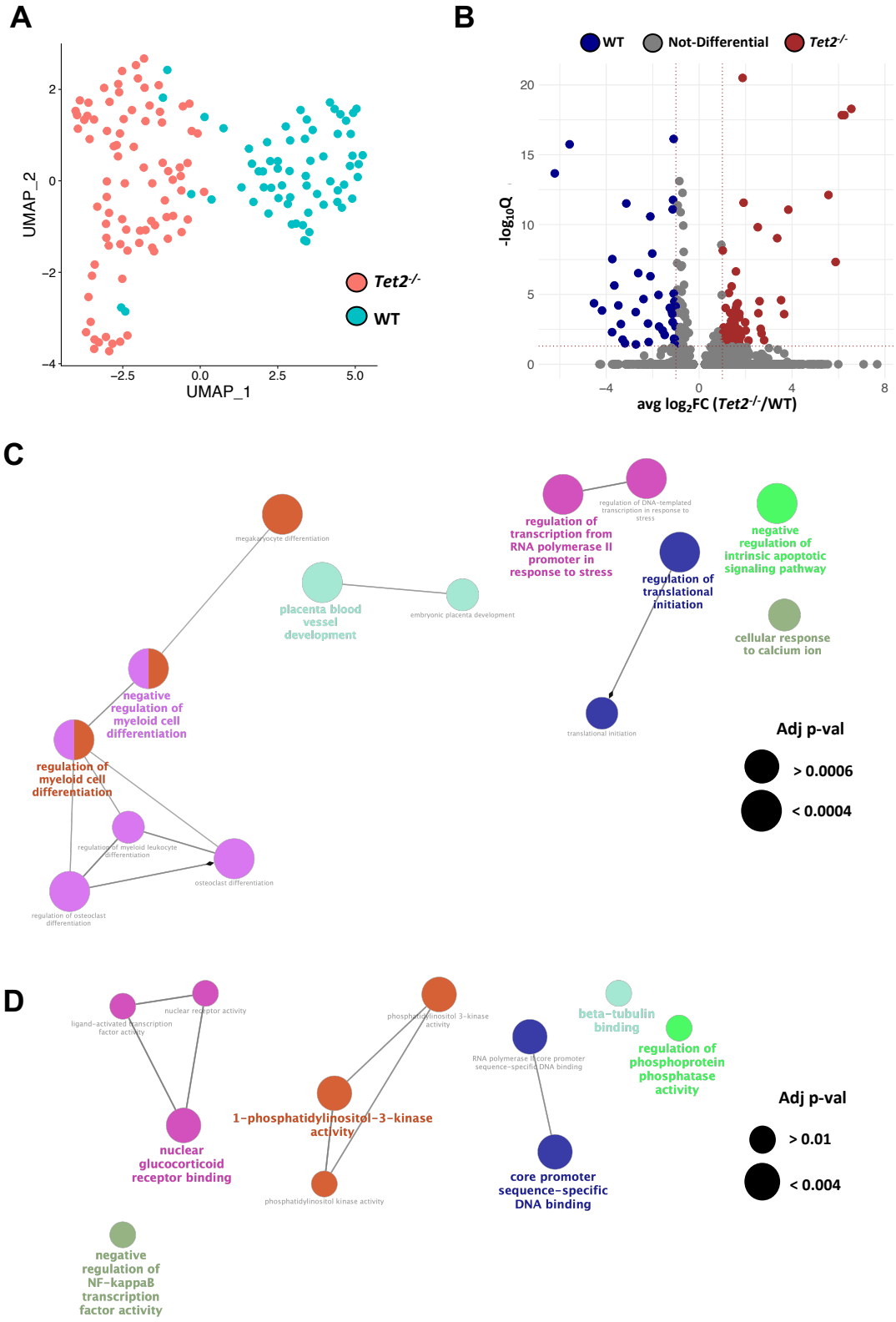


Figure 3.3 Analysis of cell transcriptome at the single-cell level reveals increased gene expression in Tet2$^{-/-}$ HSC in comparison to their WT counterparts*; A) UMAP[186] embedding with highlighted Tet2$^{-/-}$ and WT*

HSCs; B) Volcano Plot, X axes = log_2FC (Tet2$^{-/-}$ vs WT), Y axes = adjusted p-value calculated on the negative binomial distribution; C-D) Gene Ontology[195] analysis performed with ClueGO[254] node sizes reflect the statistical significance of the terms, force directed layout presented by the Kappa score, layout adjusted to minimise label overlap (see Methods for more details), C) Biological Process, thresholds = min 5 analysed genes (default = 3) / each term, all remaining thresholds were set to the default settings, D) Molecular Function, all thresholds were set to the default settings; Plots made with Seurat[217], ggplot2[260], Cytoscape[253] and ClueGO[254].

To further resolve expression profile of *Tet2*$^{-/-}$ HSCs, clustering analysis was performed using Seurat's FindClusters[217] (Louvain[188] algorithm, default resolution settings). The analysis identified three distinct subclusters of the *Tet2*$^{-/-}$ HSCs (Figure 3.4A).

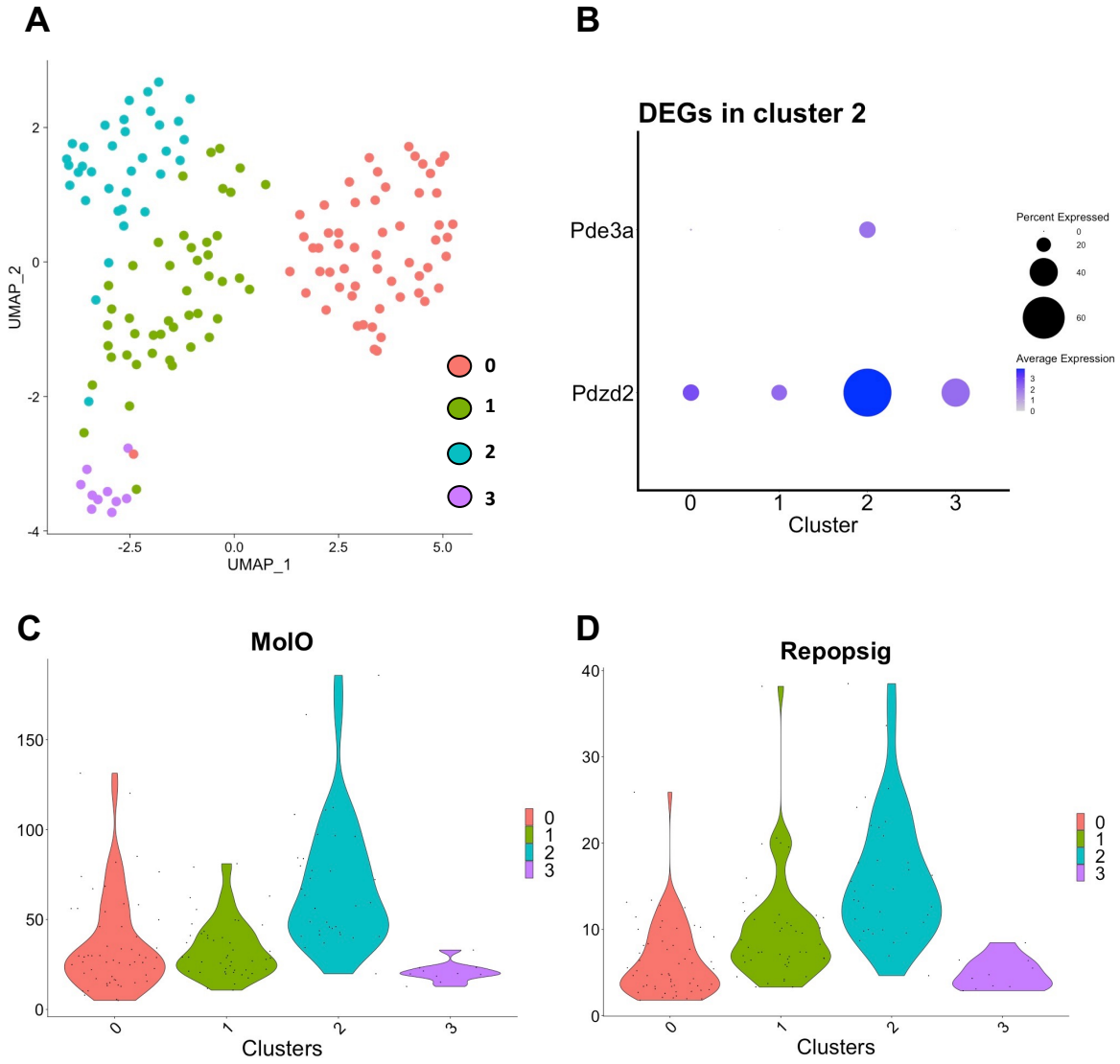


Figure 3.4 Clustering of scRNAseq data identifies a Tet2$^{-/-}$ HSC subpopulation with a high MolO and Repopsig scores; *A) Cell clustering identifies three distinct Tet2$^{-/-}$ subclusters; B) Expression of genes significantly upregulated in the second cluster (in comparison to all other clusters), Q < 0.05 & avg log_2FC > 1; C)*

MolO score calculated per each cluster; D) Repopsig score calculated per each cluster. Plots made with Seurat[217].

Several groups have previously curated gene lists to try to identify HSC related gene signatures. Two of the more prominent of these, MolO[289] and RepopSig[290], are linked to HSC functional output. Interestingly, both MolO[289] (Figure 3.4C) and Repopsig[290] (Figure 3.4D) scores seemed to be higher in the Cluster 2 – one of the *Tet2*$^{-/-}$ subclusters. Only two genes were significantly differentially expressed in the Cluster 2 ($Q < 0.05$ & avg $\log_2FC > 1$) in comparison to all other clusters (analysis performed using Seurat's FindAllMarkers[217]) (Figure 3.4B). One of them was PDZ domain-containing protein 2 (*Pdzd2*) (Figure 3.4B), gene important in steady-state haematopoiesis, whose loss in HSPC was recently shown to play a role in erythroid differentiation[291].

3.3. Integrative analyses of chromatin accessibility and gene expression point to new potential regulators of HSC self-renewal

As the scATACseq analysis suggested a more accessible chromatin state and scRNAseq analysis a higher rate of transcription in HSCs upon TET2 deficiency, we investigated whether both modalities reveal a cooperative set of targets that induce an active transcriptional cell state. For that reason, information from both cellular modalities was integrated aiming to reveal the molecular mechanisms mediating HSC fate following *Tet2* mutation. The closest gene to each peak was identified in the scATACseq peak coordinates, and the expression of these genes was assessed in scRNAseq data (threshold: distance to TSS < 100,000). When considered globally, there was no correlation between gene expression (scRNAseq) and chromatin accessibility (scATACseq) of regions close to those genes (Figure 3.5A). However, 2.6% of identified closest genes in the scATACseq analysis were significantly up or down regulated in the scRNAseq data analysis (18 out of 698 genes, Figure 3.5B). To assess whether this overlap was significantly different than expected by chance, a random bootstrapping was applied. 698 peaks were randomly pulled from all 68,532 peaks (filtered by distToTSS < 100,000) to see how many of the associated genes would overlap with the scRNA-seq DEGs, that process was

repeated 2000 times to generate a distribution. This analysis determined that observed 18 overlaps were significantly away from the random mean with p-value 0.0000061 (random overlap mean = 7.348, SD = 2.44, Z-score = 4.37) (Figure 3.5C), confirming that the observed overlap is significantly different than expected by chance.

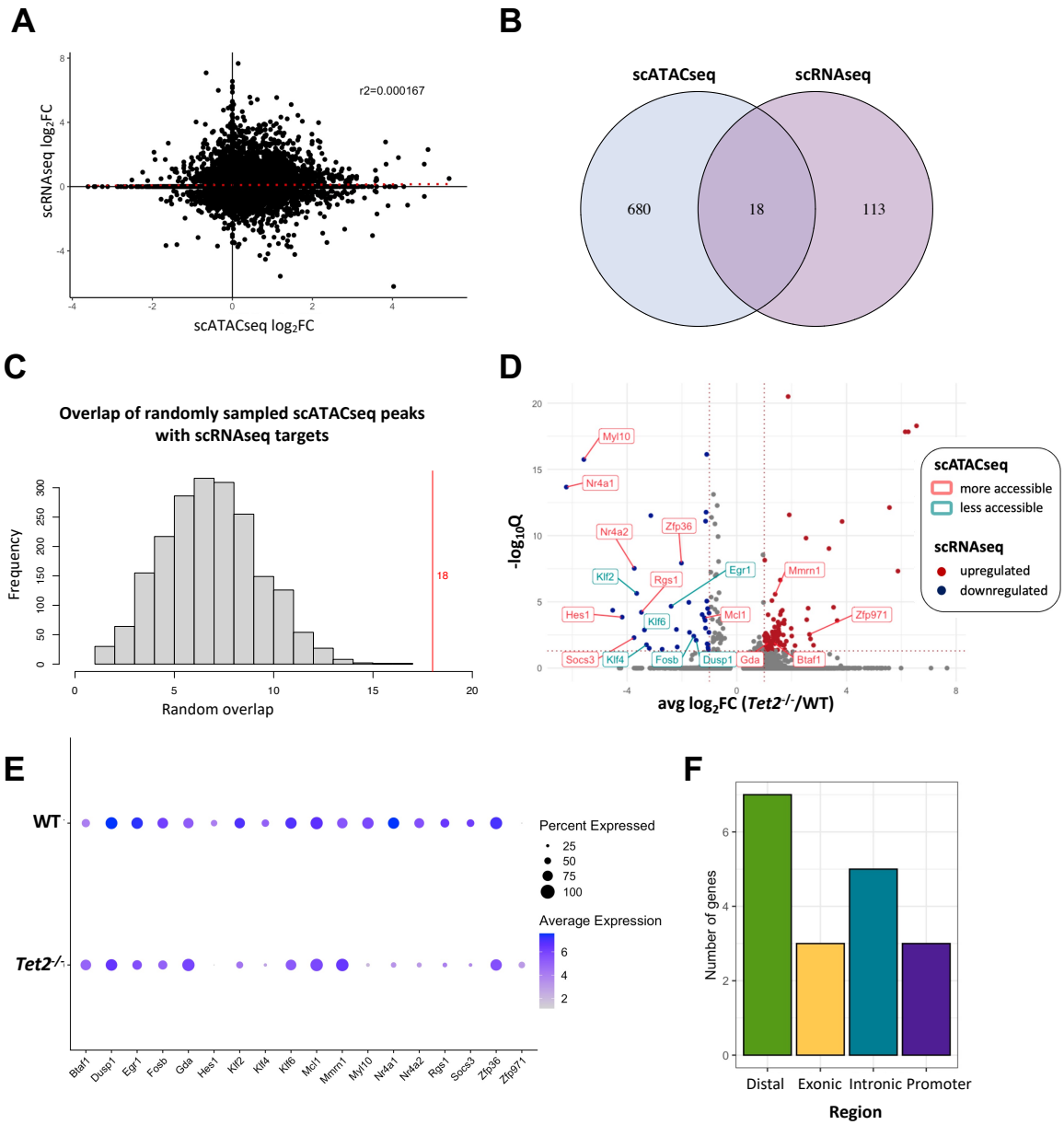


Figure 3.5 The integrative analyses of scATACseq and scRNAseq datasets indicates a set of targets deregulated in both modalities*; A) A correlation plot of all genes detected by both scATACseq and scRNAseq assays (includes not significantly differentially expressed genes); B) An overlap of targets identified by both modalities (Tet2$^{-/-}$ HSC vs WT HSC pairwise testing); C) Histogram showing an overlap of randomly sampled scATACseq peaks with scRNAseq targets; D) A Volcano plot of scRNAseq dataset with overlayed labels for the significantly accessible targets identified in the scATACseq closest gene analysis (Tet2$^{-/-}$ HSC vs WT HSC pairwise testing), blue/red dots indicate genes down/upregulated in Tet2$^{-/-}$ HSCs in scRNAseq analysis,*

green/orange labels indicate less/more accessible markers in Tet2$^{-/-}$ HSCs in the scATACseq analysis, X axes = log$_2$FC (Tet2$^{-/-}$ vs WT), Y axes = adjusted p-value; E) A Dot plot showing gene expression of identified targets of interest (the 18 targets that were significantly different between Tet2$^{-/-}$ and WT samples for both scATACseq and scRNAseq analyses); F) Distribution of peak regions for those targets. Plots made with VennDiagram[292], ggplot2[260], Seurat[217].

Interestingly, targets determined by both modalities were not consistently "up" or "down" in *Tet2*$^{-/-}$ HSC in comparison to WT HSC - integrative analysis detected genes whose expression was downregulated in *Tet2*$^{-/-}$ HSC (in comparison to WT HSC) but regions closest to those genes were more accessible in *Tet2*$^{-/-}$ HSC (Figure 3.5D), which might potentially be explained by binding of repressor TFs. Closer analysis of gene expression (as determined by the scRNAseq assay) of the overlapping targets (Figure 3.5E) suggested several potentially interesting genes for further studies. Among these we found NR4A1 receptor expressed in quiescent LT-HSC[293], HES1, a transcription factor playing a role in HSC exhaustion[294], SOCS3 suppressor of cytokine signalling and guanine deaminase GDA, previously shown to characterise HSC clones with differentiation-inactive fates[295,296]. Notably, peaks closest to identified overlapping targets were mainly detected in the distal regions for most of the targets (Figure 3.5F), suggesting potential regulation of transcription by distal enhancers or repressors. 5 out of 7 of these targets (*Rgs1*, *Socs3*, *Nr4a1*, *Mcl1* and *Myl10*) showed more accessible chromatin but downregulated gene expression indicating possible suppression of expression of these genes upon *Tet2*-deficiency.

Next, to determine whether identified targets play any collaborative role in HSC, the String (https://string-db.org) and Genemania (https://genemania.org) databases were surveyed and the connectivity of those targets was characterised. Both analyses indicated little relatedness to TET2 (Figure 3.6A,B), with a direct connection to only two targets - KLF4 (identified with the String analysis) and NR4A1 (identified with the Genemania analysis). However, both analyses showed that the targets themselves are closely connected, suggesting at least partial overlap in terms of biological roles of those genes. Importantly, it should be noted that connectivity of those targets to TET2 might be even higher but was simply not detected in those analyses due to sparse experimental data on TET2 and therefore no information of potential interactions in those databases. Additionally, most

connections identified with Genemania were classified as co-expression, which further advocate for the collaborative role of those targets in similar pathways. As expected, NR4A1 and NR4A2 nuclear receptors and KLF2, KLF4 and KLF6 transcription factors were closely connected, as they belong to the same families.

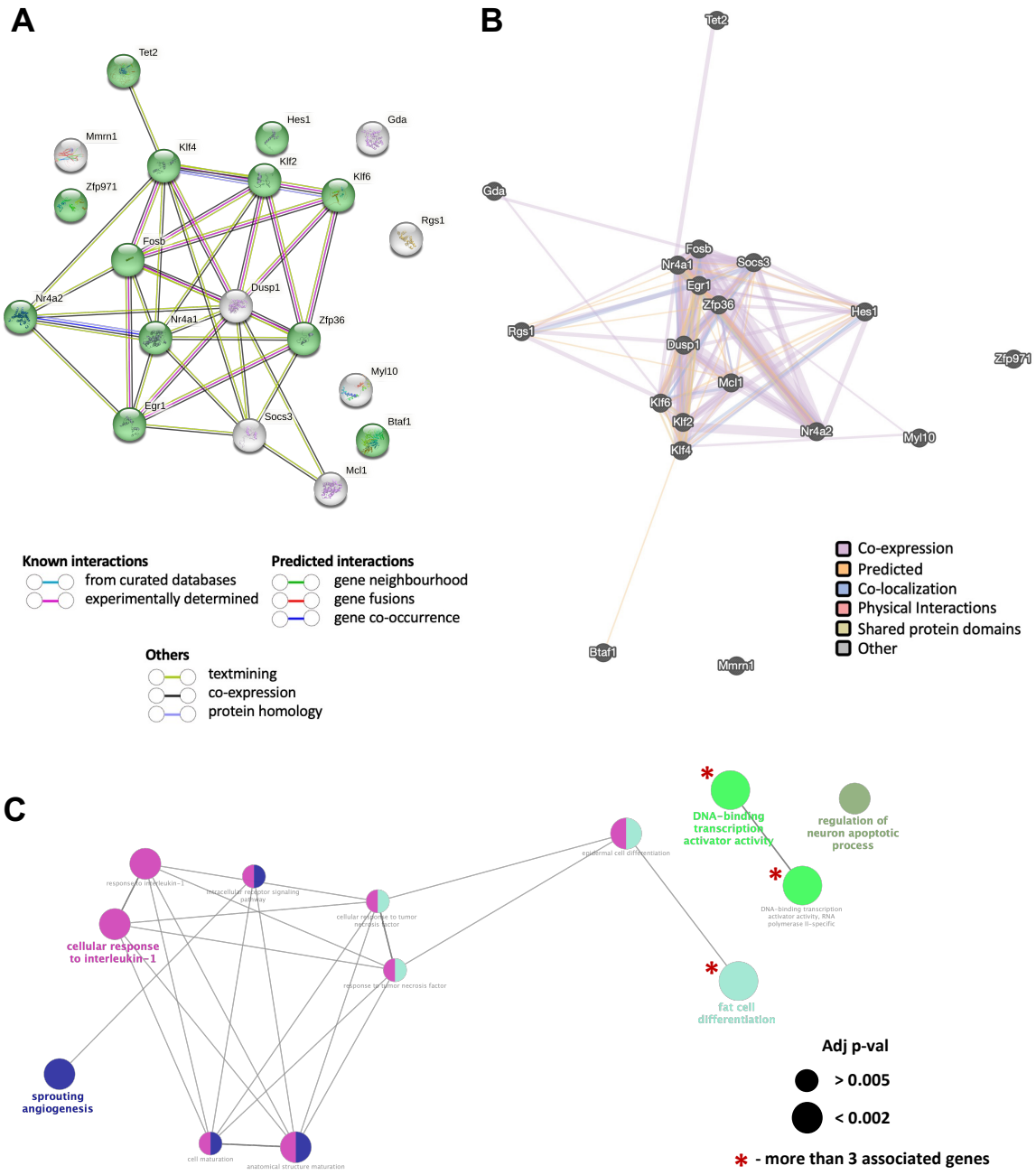


Figure 3.6 Overlapping scATACseq/scRNAseq targets show little relatedness to TET2 but high inner connectivity; *A) Static picture of interactions found for the studied targets (and Tet2) in the String database (https://string-db.org), 19 nodes; B) Interaction graph for the same targets constructed using the Genemania database (https://genemania.org), 19 nodes (target genes), 237 links in total, 85% of links were classified as co-expression); C) Gene Ontology[195] analysis performed with ClueGO[254], node sizes reflect the statistical*

significance of the terms, force directed layout presented by the Kappa score, layout adjusted to minimise label overlap (see Methods for more details), Biological Process, thresholds - the minimum percentage that the mapped identifiers are representing from the total associated genes with the term = 1% (default = 4%), all remaining thresholds were set to the default settings;. Plots made with String (https://string-db.org), Genemania (https://genemania.org), ClueGO[254], Cytoscape[253].

In order to identify biological processes in which targets identified with both scATACseq and scRNAseq analysis are involved, a GO[195] analysis was run. Notably, the GO[195] analysis indicated enrichment in the DNA-binding transcription activator activity (associated genes with this term: *Egr1*, *Fosb*, *Klf4*, *Klf6*, *Nr4a1* and *Nr4a2*), which is in line with the increase in open chromatin observed in *Tet2*-deficient HSCs (Figure 3.6C).

Since the String and Genemania analyses focused only on the direct interactions of targets of interest with TET2, the Omnipath database[261] - a meta-resource on intercellular signalling, was used to further study signalling pathways regulated by the targets identified in the integrative scATACseq / scRNAseq data analysis. Firstly, the interaction graph was built using information on protein interactions from all databases available via Omnipath[261]. Then shortest paths to and from TET2 to and from each of studied targets were determined. The analysis detected 76 target → TET2 shortest paths (Figure 3.7A) and 147 TET2 → target shortest paths (Figure 3.7B). Most intermediate genes were not differentially expressed (as showed by our scRNAseq analysis) (Figure 3.7A-B). Moreover, the shortest path analyses indicated a bottleneck at the connection to TET2. This is probably a consequence of little information about the TET2 connectivity in the Omnipath[261] databases, as *Tet2* is still not a well characterised gene. Protein kinase AMP-activated catalytic subunit alpha 1 (PRKAA1) and O-linked N-acetylglucosamine (GLCNAC) transferase (OGT) were present in the TET2 bottleneck. These connections were not unexpected to find in databases, as AMPK (human Prkaa1 ortholog) was recently shown to enhance TET2 protein stability[297] and OGT is directly interacting and associating with TET2 at transcription start site[298]. Despite the fact that TET2 was not well connected to the targets of interest, the shortest path analyses allowed us to look into potential molecular context in *Tet2*-deficient HSCs in mice.

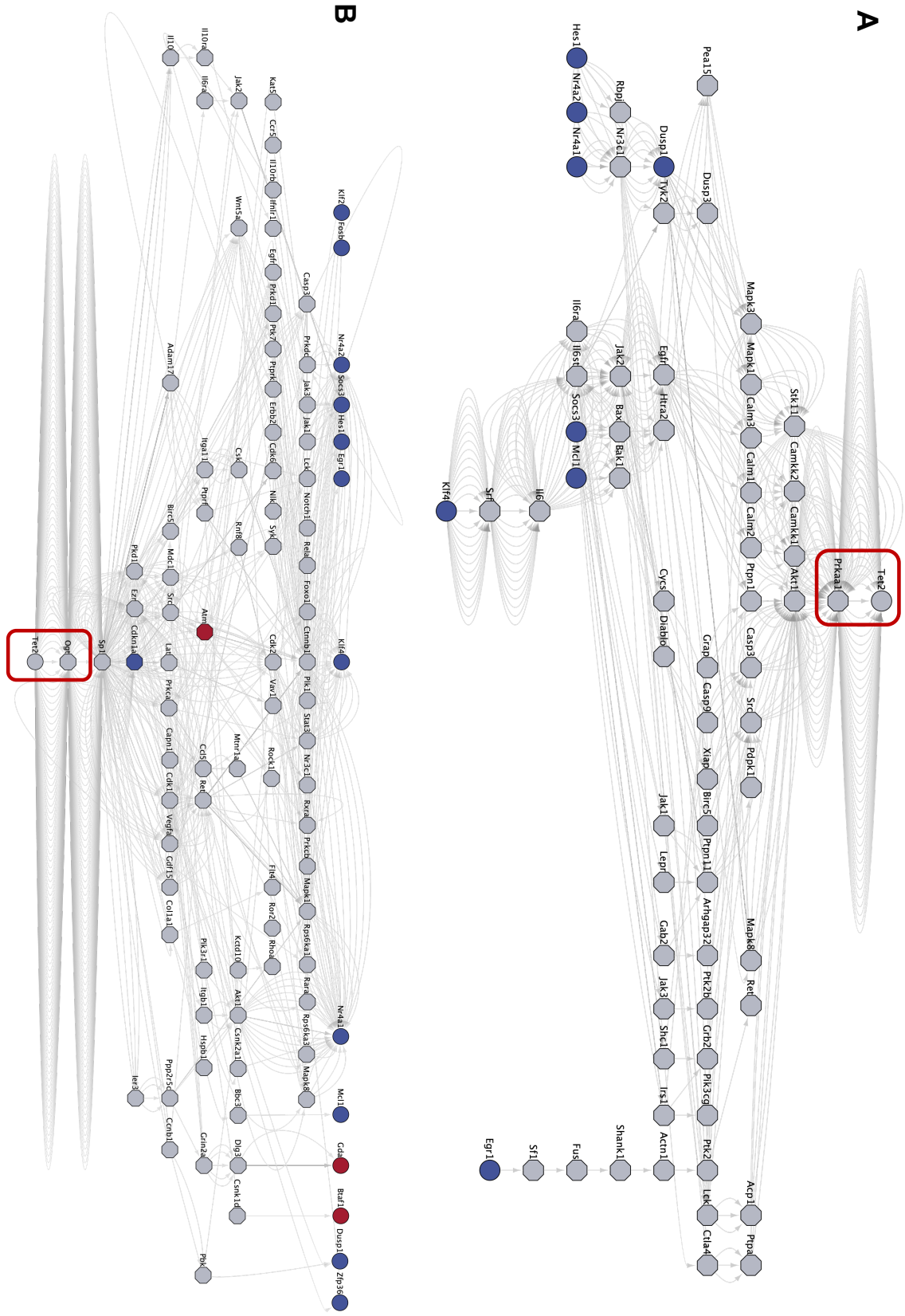


Figure 3.7 Shortest path analysis demonstrates connectivity of studied targets to TET2*; A-B) Shortest path analysis performed usign the Omnipath[261] database, ellipse node shape = Tet2 & targets identified in the scATACseq and scRNAseq data analyses, octagon node shape = intermediate targets identified with Omnipath,*

blue/red/grey = genes showing reduced/increased/no change (respectively) in expression in Tet2$^{-/-}$ HSC in comparison to WT HSC (as identified with the scRNAseq analysis). A) Directed interaction graph showing path analysis from targets identified in the scATACseq and scRNAseq data analyses to TET2; B) Directed interaction graph showing path analysis from TET2 to targets identified in the scATACseq and scRNAseq data analyses. It is important to note that causality should not be inferred from these networks, as these graphs were built on all types of interactions. Graphs were made with Cytoscape[253].

GO[195] analysis was run for all targets present in the interaction graphs. Notably, among the enriched terms we have found hematopoietic or lymphoid organ development and regulation of cell population proliferation, which imply that constructed graphs might provide reasonable insights into HSC signalling.

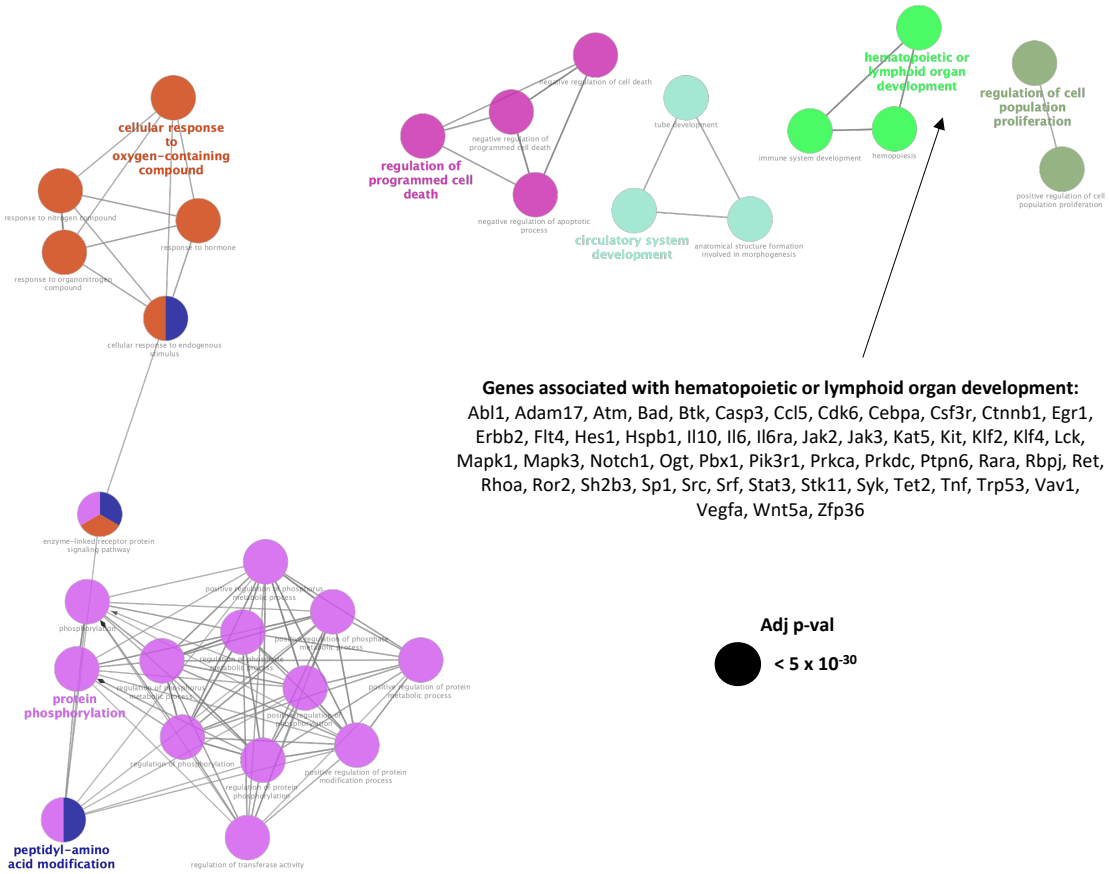


***Figure 3.8 A number of genes identified with the shortest paths analyses are associated with haematopoietic or lymphoid organ development and regulation of cell population proliferation**; Gene Ontology[195] analysis performed with ClueGO[254] for all targets present in Figure 3.7 interactions graphs, node sizes reflect the statistical significance of the terms, force directed layout presented by the Kappa score, layout adjusted to minimise label overlap (see Methods for more details), Biological Function, thresholds - min 50 analysed genes (default = 3) / each term, all remaining thresholds were set to the default settings. Graph made with ClueGO[254] and Cytoscape[253].*

3.4. Transcription factor network analysis explores potential regulatory interactions governing HSC self-renewal

To further investigate potential molecular pathways governing increased HSC self-renewal upon the loss of function of TET2, a transcription factor regulatory network was built. For this aim the DoRothEA[257] database was used, a curated collection of gene regulatory interactions. Firstly, all TFs whose binding motifs were significantly more accessible in *Tet2*$^{-/-}$ HSC (as shown with the scATACseq analysis) that were also present in the in the DoRothEA[257] database were selected. Then, a list of all differentially expressed genes (as determined by the scRNAseq analysis) that those TFs regulate was curated, using information from the DoRothEA[257] database. Finally, an interaction graph containing all defined TF-regulatory gene pairs (regulons) was constructed (Figure 3.9). DoRothEA[257] confidence level, a score based on the number of supporting evidence (A – highly reliable, E – computational predictions), of all identified TF-target interactions was A or B.

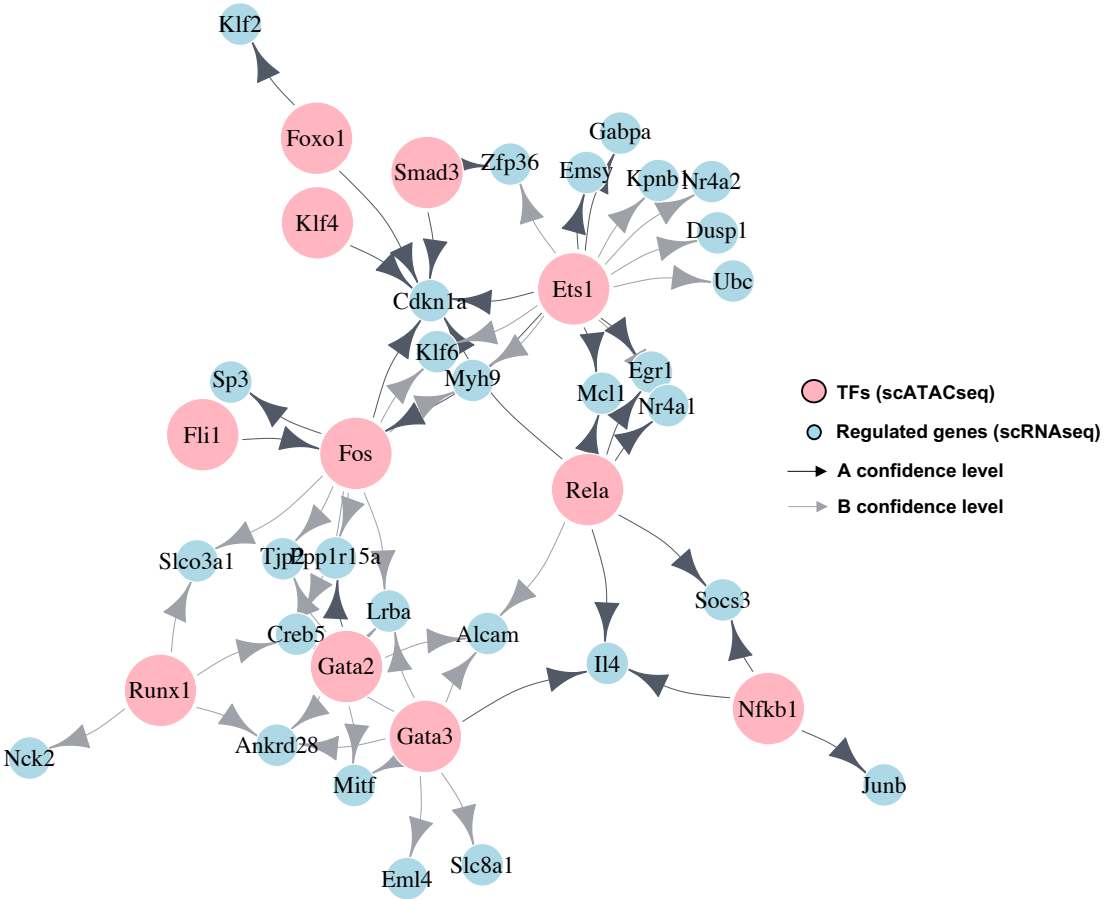


Figure 3.9 Transcription factor network analysis represents potential regulatory interactions governed by targets identified with scATACseq and scRNAseq assays*; Interaction graph built using information from the DoRothEA[257]database. Graph made with igraph[258], graphics[259], ggplot2[260].*

No correlation was found in terms of a change in gene expression of identified TF and the genes they regulate (Figure 3.10). The binding motifs sites of all TF present in this graph were more accessible in the *Tet2*$^{-/-}$ samples, however the gene expression of most of them was not significantly increased (Figure 3.10, Figure 3.11A,B). Intriguingly, gene expression of *Fos* and *Klf4* was actually decreased in *Tet2*$^{-/-}$ HSC (Figure 3.10, Figure 3.11A,B). Altogether, gene expression of most of transcription factors was not significantly different between *Tet2*$^{-/-}$ and WT samples (Figure 3.11B), however this does not necessarily indicate a negative result as expression of those TFs might be constant, but the activity increased. Interestingly, the median log_2FC of expression of sets of genes regulated by each TF was negative for the majority of the targets (Figure 3.10), indicating downregulation of expression of those genes upon the loss of function of TET2 (Figure 3.11A, C). Gene expression of *Gata2*, *Gata3* and *Runx1* was not changed upon *Tet2* deficiency, however a median expression of genes these transcription factors regulate was positive (Figure 3.10, Figure 3.11A).

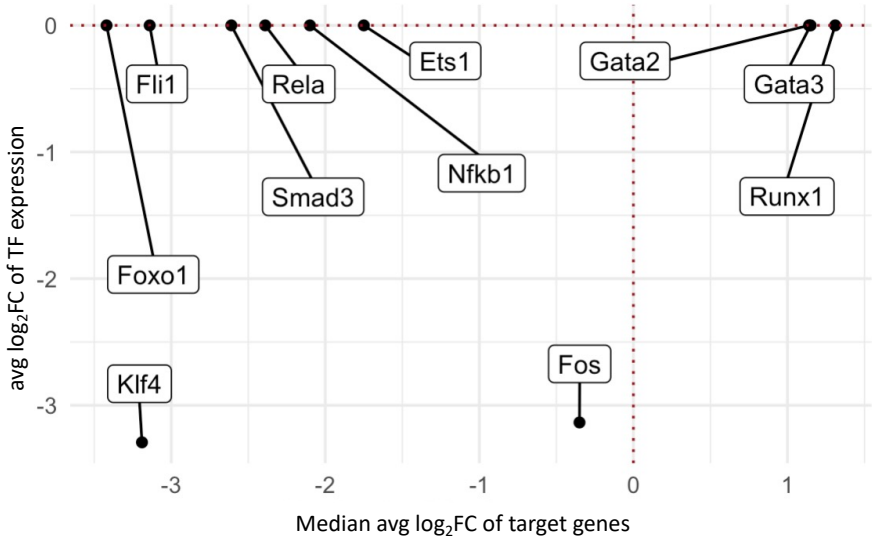


Figure 3.10 A median log2FC of expression of sets of genes regulated by each TF identified with the transcription factor network analysis is negative for most of the targets*; Median expression of set of genes that the identified transcription factors regulate. Plot made with ggplot2*[260]*.*

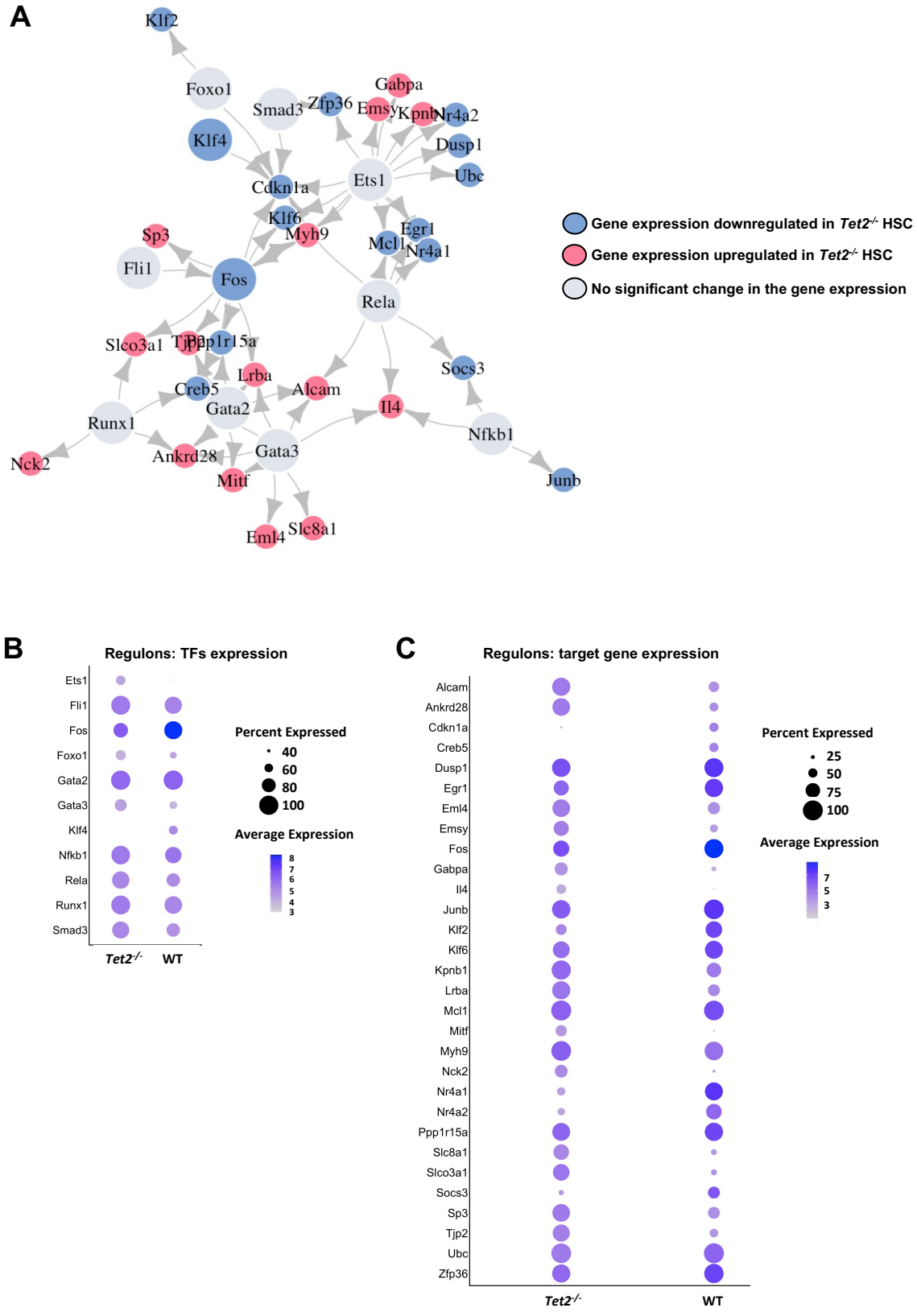


***Figure 3.11 Transcription factor network analysis indicates convoluted patterns of gene expression in Tet2$^{-/-}$ HSC**; A) Transcription factor network presented in Figure 3.9 with mapped information on gene expression; B) Expression of source nodes (TFs) as determined by scRNAseq data analysis; C) Expression of*

target nodes (regulated genes) as determined by scRNAseq data analysis. Plots made with igraph[258], graphics[259], ggplot2[260], Seurat[217].

In general, it seemed that three groups of transcriptions factors could be distinguished depending on their expression (Figure 3.11B) and expression of genes they regulate (Figure 3.11C) – TFs whose expression was downregulated (FOS and KLF4), TFs mediating increased gene expression (GATA2, GATA3, RUNX1) and TFs regulating expression of genes that were downregulated in *Tet2*$^{-/-}$ HSC (ETS1, NFKB, RELA, SMAD3, FLI1, FOXO1) (Figure 3.10). All transcription factors that were part of the first group - GATA2[299], GATA3[300], RUNX1[301] are well known self-renewal regulators. Transcription factors forming the second group – FOS[302] and KLF4[303] also play important roles in haematopoiesis. Among the third group we found e.g. NFKB, extensively researched mediator of inflammation[304] and FLI1 – a well-known regulator of HSC differentiation[305]. Interestingly, identified three groups of TFs seemed to be located at distinct regions in the network (Figure 3.11A).

3.5. Quantification of transcript levels *in vitro* (RT-qPCR) and exploration of the expression of selected targets in human blood cells

Following the analyses of scATACseq dataset, scRNAseq dataset, the integrative analyses of both modalities and a literature screen, a set of 10 genes was selected for further investigation for their potential role in TET2-deficient HSCs. Firstly, expression of selected genes across the haematopoietic differentiation was determined using the Nestorowa et al.[266] dataset (Figure 3.12). *Dusp1*, *Egr1*, and *Smarcc1* drew our attention as they seem to be expressed across all studied haematopoietic subpopulations (progenitor cells, HSPC and LT-HSC). The expression of *Nr4a1* and *Rgs1* seemed to be higher in the HSPC and LT-HSC clusters. The expression of *Hes1*, *Nr4a2*, *Nr4a3*, *Smarcc1* and *Socs3* was below the threshold for significantly variable genes between cell subpopulations.

Expression of each selected target was examined with RT-qPCR, aiming to verify the computational results. Gene expression was studied in long-term HSCs ($CD45^+EPCR^+CD48^-CD150^+$) and short-term HSCs ($CD45^+EPCR^+CD48^-CD150^-$)

isolated from *Tet2*$^{-/-}$, *Tet2*$^{+/-}$ and wild-type mice. RT-qPCR suggested a mild upregulation of *Dusp1* and *Rgs1* expression relative to the *Gapdh* housekeeping gene, however no significant changes in gene expression between *Tet2*$^{-/-}$, and *Tet2*$^{+/-}$ and WT mice were detected (Figure 3.13).

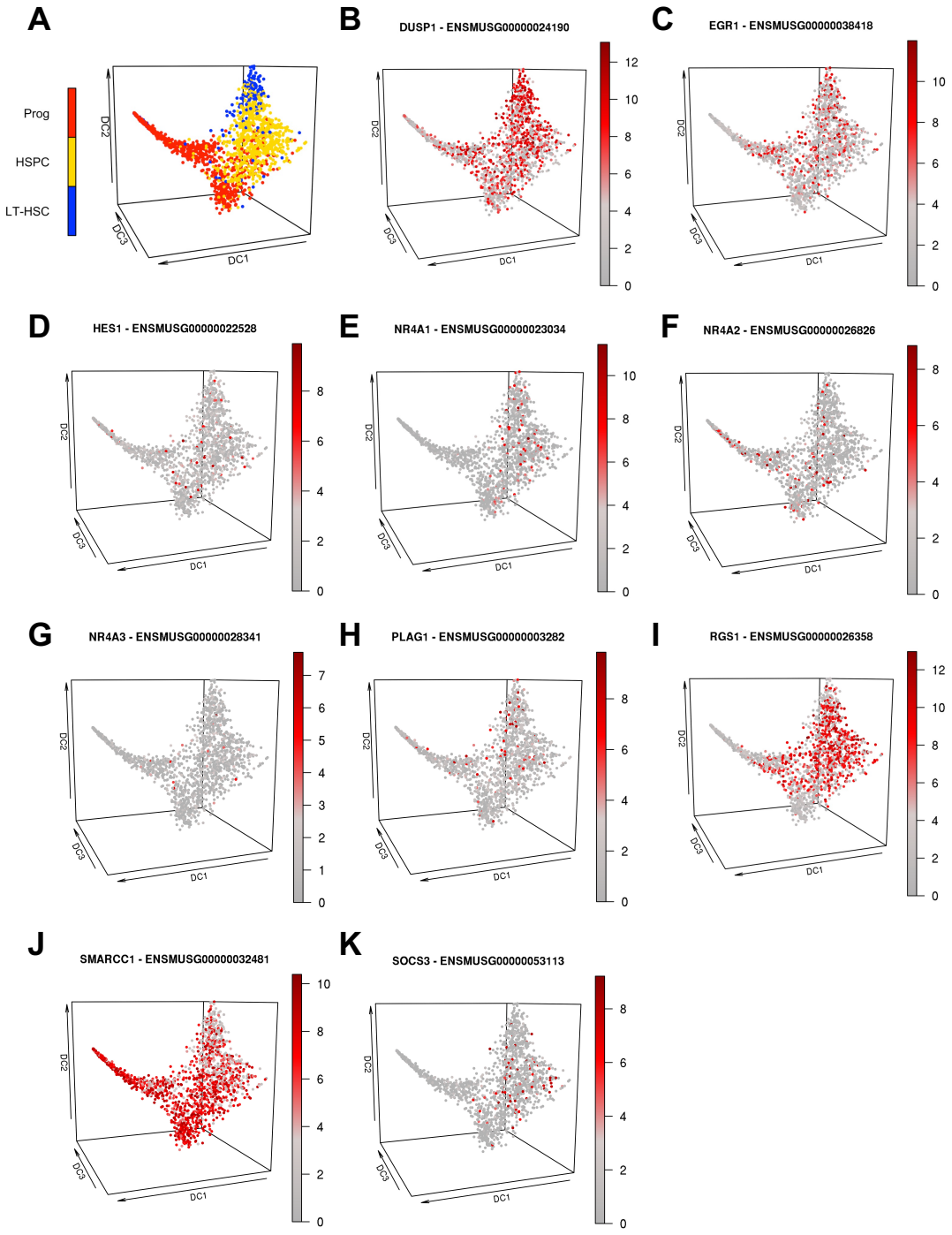


Figure 3.12 Selected targets identified with the scATACseq and scRNAseq analyses are expressed across murine haematopoietic differentiation*; A) UMAP*[186] *plot by Nestorowa et al.*[266] *with highted haematopoietic subpopulation clusters, Prog = haematopoietic progenitor cells (Lineage*$^{-}$*Sca1*$^{-}$*cKit*$^{+}$*), HSPC = haematopoietic stem and progenitor cells (here defined as Lineage*$^{-}$*Sca1*$^{+}$*cKit*$^{+}$*), LT-HSC = long-term self-*

renewing haematopoietic stem cells (here defined as Lineage⁻c-Kit⁺Sca1⁺CD34⁻Flk2⁻), B-K) Expression of selected genes of interest across haematopoietic cell subpopulations. Plots made using https://blood.stemcells.cam.ac.uk/single_cell_atlas.html.

Figure 3.13 RT-qPCR analysis indicates no significant differences in expression of selected genes between Tet2$^{-/-}$, Tet2$^{+/-}$ and WT HSC; *A) RT-qPCR results for long-term HSCs (CD45⁺EPCR⁺CD48⁻CD150⁺); B) RT-qPCR results for short-term HSCs (CD45⁺EPCR⁺CD48⁻CD150⁻); ΔCt = $Ct_{[target\ gene]}$ – $Ct_{[GAPDH]}$, Kruskal-Wallis test with Dunn post-hoc test with Bonferroni correction were used to check for differences in expression, no significant differences were found. Plots made with ggplot2[260].*

All analyses presented so far were performed using a mouse model. However, there are differences in gene expression mechanisms across species. Therefore, to further study selected targets of interest a human database was used aiming to check gene expression in the human blood cells. A human network of protein interactions was built using HuRI human reference database (HuRi, www.interactome-atlas.org) for human homologues to selected targets of interest. Analysis indicated relatively high connectivity with high and medium level tissue expression and tissue specificity for most targets (Figure 3.14), suggesting rationale for further studies of these genes.

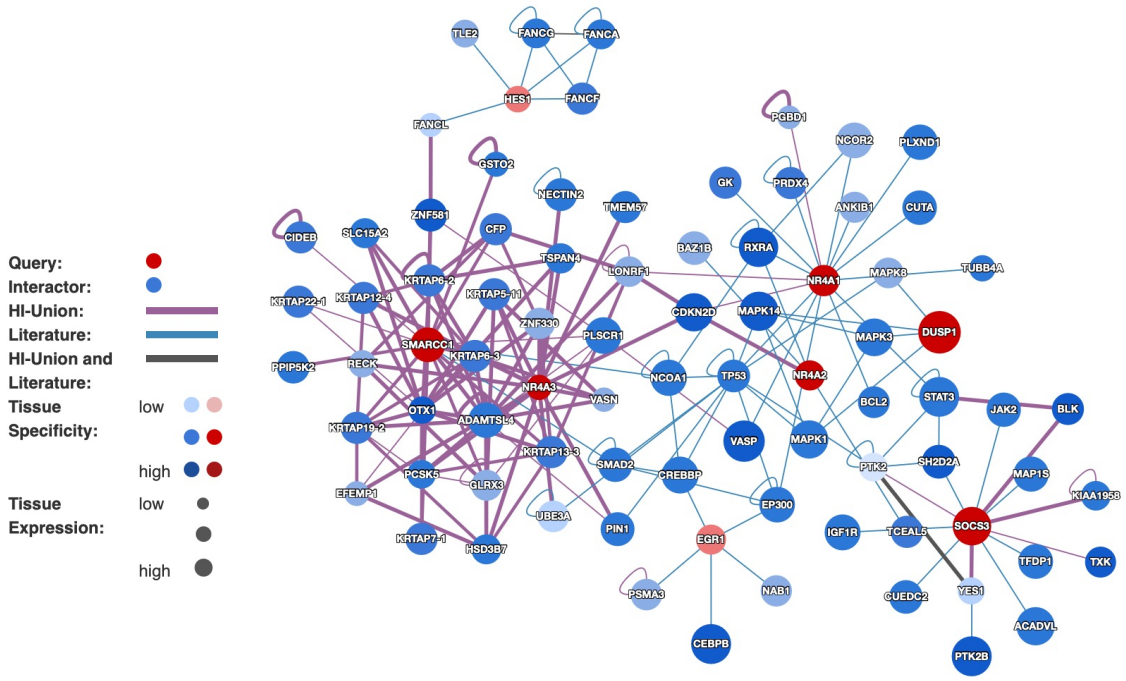


***Figure 3.14 Selected targets show high and medium-level tissue expression and tissue specificity in human blood cells**; Analysis of interactions for selected 9 targets of interest (RGS1 was not present in the HuRi database) in human blood cells, undirected graph of human interactions built using the HuRi database (www.interactome-atlas.org), tissue expression defined as "whole blood".*

3.6. Discussion

Studies using mouse models provide invaluable insights into the mechanisms driving disease development. The bioinformatic analyses presented in this chapter evaluate cell behaviour upon the loss of function of TET2, a known self-renewal regulator. We analysed DNA accessibility HSC by single-cell ATAC sequencing assay and gene expression by single-cell RNA sequencing assay aiming to determine the molecular mechanisms leading to the increased self-renewal of *Tet2*-mutated HSC.

The presented analyses include two different approaches aiming to integrate information from both assays. The first method comprised intersecting the list of genes closest to the differentially accessible regions (scATACseq assay) and the list of differentially expressed genes (scRNAseq assay). In the second approach, the TF whose binding motifs were significantly more accessible in *Tet2*$^{-/-}$ HSC (as shown

with scATACseq analysis) were selected, then the list of all differentially expressed genes (as determined by scRNAseq analysis) that those TF regulate - using the information from the DoRothEA[257] database - was curated, and finally, an interaction graph containing all defined TF-regulatory gene pairs (regulons) was constructed. There are disadvantages for each of those integration methods. In the “closest gene” analysis we naturally assume that genes closest to the accessible chromatin sites are expressed, which is not always the case, and might lead to incorrect conclusions[306]. In the “regulon” integrative approach we are obviously limited by the TF-gene regulatory pairs already published in the DoRothEA[257] database. Together this approach has identified a set of potentially interesting targets which can serve as a platform for future researchers to undertake functional validation assays to fully understand the role of TET2 in HSC regulation.

Interestingly, the less accessible peaks called in *Tet2*$^{-/-}$ cells were mainly (40%) mapped to distal regions (Figure 3.1C), which is in accordance with recent studies suggesting that TET2 regulates methylation state at the enhancer sites, facilitating binding of transcription factors[70,128]. However, overall, performed analyses indicated a more active state of *Tet2*$^{-/-}$ HSCs compared to WT HSC, reflected by an increased chromatin accessibility and increased gene expression. This is in contrast to the hypothesis that DNA methylation is positively correlated with inhibited gene expression, as DNA methylation was shown to cause inhibitory effects[128,307], therefore loss of TET2 should lead to hypermethylation, less accessible chromatin state and downregulated transcription. Indeed, recent chromatin accessibility (ATACseq) analyses performed by Rasmussen et al. of mouse MPP and AML cells reported that the majority of differential chromatin regions were less accessible in *Tet2*-deficient cells[128]. Interestingly, the exception was the GMP fraction, where the accessibility was balanced. In the same study, they compared binding sites of TET2 in haematopoietic and embryonic stem cells and showed that only ~7% regions overlapped for those two settings, indicating a highly cell-type specific binding pattern of TET2 protein. Altogether this indicates that HSC should display cell-type specific lower chromatin accessibility and repressed gene expression upon the loss of *Tet2*. However, the analyses presented in this chapter showed the opposite. Interestingly, a recent study by Zou et al. proposed that *Tet2* loss of function mutations indeed result in the DNA hypermethylation at enhancer regions, however

the increase in chromatin accessibility and transcription is not a direct result of TET2-mediated oxidation of DNA 5-methylcytosine[146]. Instead, Zou et al. proposed that chromatin-associated RNA 5-methylcysine is oxidised by TET2, which prevents methyl-CpG-binding-domain protein (MBD6)-dependent deubiquitination of Lys119 of histone H2A[146]. Therefore, loss of TET2 function leads to decreased ubiquitination and an increase in chromatin accessibility and gene expression[146], which aligns with the results presented in this chapter. Observed increased chromatin accessibility and increased gene expression can therefore be explained by the mechanism introduced by Zou et al., where following TET2 deficiency cells learn to depend on this activation pathway[146]. Drawing upon this line of thought, this could lead to increased expression of genes involved in cell proliferation, substantially altering HSC behaviour. This mechanistic interpretation of the results is reflected by e.g. observed increased expression of multimerin-1 (*Mmrn1*) in *Tet2*$^{-/-}$ HSCs (Figure 3.5D), a gene further studied in Chapter 4. Further comprehensive analyses of individual targets identified in this Chapter, genomic interactions and downstream signalling in the TET2-deficient environment, followed by experimental validation, would help to better understand biological consequences of TET2 loss of function in HSCs and to design approaches to inhibit pathways deregulated at the early stages of *Tet2*-mutated haematological malignancies, underlining relevance of presented analyses. Altogether, this highlights the importance of further studies of complex roles of TET2 in homeostasis and repercussions of TET2 loss of function in haematological malignancies.

Little overlap was found between gene expression (scRNAseq) and chromatin accessibility (scATACseq) of regions close to those genes (Figure 3.5A). This might be explained by the fact that these datasets were derived from different mice, which might have introduced a lot of variance as disease might have progressed differently. Additionally, these datasets were generated by different technologies – scATACseq was a 10xGenomics assay[308] and scRNAseq was SmartSeq2-based[3] (for details see Methods). Nevertheless, the integrative analyses pointed to several targets potentially responsible for disturbed molecular mechanisms in HSC upon *Tet2* loss, which might help in identifying targets for further *in vivo* studies.

The RT-qPCR for the selected targets did not confirm differences in the expression of transcripts of interest between *Tet2*$^{-/-}$, *Tet2*$^{+/-}$ and WT HSCs. A possible explanation for this might be that RT-qPCR was not run on the exact same samples as scRNAseq and scATACseq – each of those experiments was performed using different mice, cells were sorted on different days, samples were prepared by different scientists. Moreover, only 3-4 replicates were analysed which is likely why significance was not met for some of the studied genes, e.g. *Dusp1*, *Egr1*, *Nr4a1*. Nonetheless, these analyses constitute a good screening approach to identify new potential targets.

Pioneering work of many research groups over the last years generated methylation landscapes of haematopoietic cells in *Tet2*-deficient conditions, reporting changes in cell methylation patterns upon *Tet2* mutations[123,128,129,131–133,309]. Future analyses should prioritise understanding molecular mechanisms controlling cellular response to mutations of *Tet2*[128], taking into account that different *Tet2* mutations might cause distinct outcomes and that cell behaviour might differ in case when *Tet2* is not the first mutation[310]. Moreover, some mechanisms might be entirely different in mouse and human cells, a concept explored further in Chapter 4 where next-generation sequencing analyses are used to examine *TET2*-mediated development of myelodysplastic syndrome.

4. Transcriptomic changes in primary human haematopoietic cells carrying *TET2* mutations

MDS is associated with a poor prognosis, with a median survival of less than 3 years for high-risk patients[311]. The collaborative work of scientists and clinicians over the last years has contributed to the identification of a set of mutations present in the blood cells of patients diagnosed with MDS[312]. 2-3 driver mutations are usually detected and the elements of the DNA methylation (*DNMT3A*, *TET2*, *IDH1*, *IDH2*, *WT1*), chromatin modification (*EZH2*, *SUZ12*, *EED*, *JARID2*, *ASXL1*, *KMT2*, *KDM6A*, *ARID2*, *PHF6*, and *ATRX*) and RNA splicing (*SF3B1*, *SRSF2*, *U2AF1*, *U2AF2*, *ZRSR2*, *SF1*, *PRPF8*, *LUC7L2*) pathways seem to be commonly disturbed[312]. Importantly, mutations in certain genes, such as *ASXL1*, *IDH1* or *TP53* are associated with worse prognosis[313–316]. Notably, some of the mutations seem to cooccur while others are exclusive implying common molecular mechanisms[312]. Indeed, and in line with other myeloid malignancies, the malignant transformation in MDS follows a model of clonal evolution[90,312]. Understanding transcriptional changes in the malignant cells is critical as it will help developing drugs that precisely target these cell clones and consequently lead to better diagnosis and adequate choice of treatment protocols[90,312].

TET2 is mutated in ~15% of patients diagnosed with myeloid malignancies, and is commonly mutated early in malignancy development[87]. However, loss of *Tet2* on its own does not typically result in the blood cancer progression, as suggested by studies using mouse models[2,239]. Moreover, TET2 is commonly mutated in clonal haematopoiesis[317]. This makes *TET2* an important target for further analyses, as it seems that mutations in this gene must give mutated cells a clonal advantage in the BM niche. Therefore, in this chapter we generated and analysed bulk RNA sequencing data of HSPCs from patients diagnosed with MDS who carried *TET2* loss of function mutations, providing new insights into disrupted pathways potentially leading to malignancy progression.

4.1. Transcriptomic analyses of patient haematopoietic cells from *TET2*-mutated myelodysplastic syndromes

To study the molecular signatures associated with *TET2* mutations in human HSPCs we analysed BM samples of patients diagnosed with MDS. Cells were divided into an HSPC-enriched fraction ($CD34^+$), a short-term reconstituting cell (STRC) fraction ($CD34^+CD38^+$) and an HSPC-fraction ($CD34^+CD38^-$) and were isolated by FACS (Figure 4.1A). Bulk RNAseq was performed on freshly isolated cells and control healthy BM (see Methods).

A

Fraction	Cell phenotype	Diagnosis	Number of samples
CD34	CD34	MDS	5
CD34	CD34	Healthy control	3
STRC	$CD34^+CD38^+$	MDS	3
STRC	$CD34^+CD38^+$	Healthy control	3
HSPC	$CD34^+CD38^-$	MDS	3
HSPC	$CD34^+CD38^-$	Healthy control	3

B

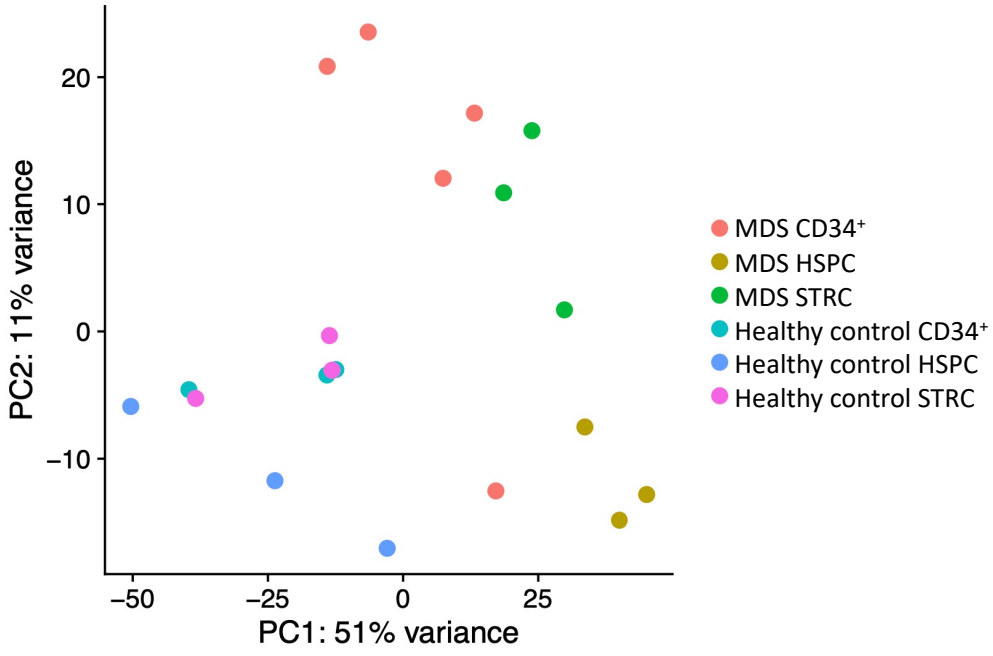


Figure 4.1 An overview of the bulk RNA sequencing data analysis of bone marrow samples of patients diagnosed with MDS and healthy bone marrow controls*; A) Table summarising all analysed samples; B) PCA[4] analysis. $CD34^+$ = HSPC-enriched fraction, STRC = a short-term reconstituting cell fraction with the $CD34^+CD38^+$ phenotype, HSPC-fraction with the $CD34^+CD38^-$ phenotype. Plot made with ggplot2[260].*

Principal component analysis clearly separated MDS samples from healthy controls, however it should be noted that all samples belonged to different sexes based on grouping (healthy/MDS samples), therefore the differences between groups could be potentially convoluted by sex differences. Also of note, the MDS samples clustered according to the cell fraction (HSPC-enriched, STRC, HSPC), suggesting that additional information might be extractable by analysing different cell fractions individually (Figure 4.1B).

Firstly, gene expression in the HSPC-enriched fraction ($CD34^+$) was compared between MDS samples and healthy controls. Differential gene expression analysis revealed significant differences between patient and healthy control samples, manifesting in 586 genes being differentially expressed (abs(log_2FC) > 1 & adjusted p-value <0.05), with the majority of genes (~78%) being upregulated in comparison to the healthy control $CD34^+$ cells (460 genes upregulated in MDS samples, 126 genes downregulated in MDS samples).

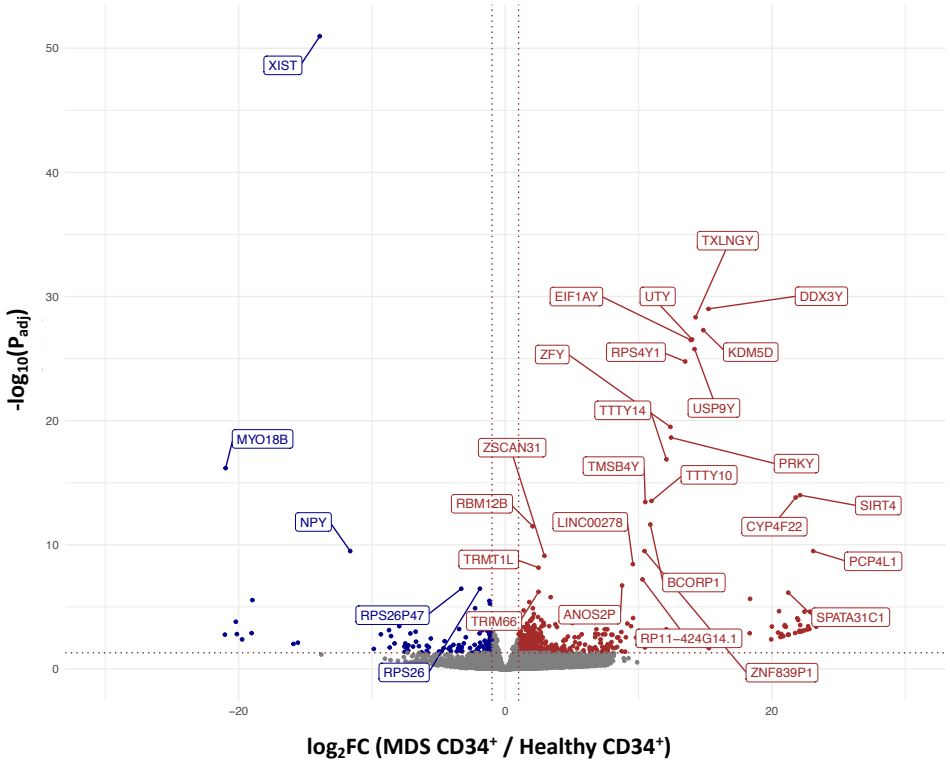


Figure 4.2 Bulk RNA sequencing data analysis of $CD34^+$ bone marrow cells indicates upregulated gene expression in patients diagnosed with MDS in comparison to healthy bone marrow control; *Volcano Plot with highlighted top 30 differentially expressed genes. Plot made with ggplot2*[260].

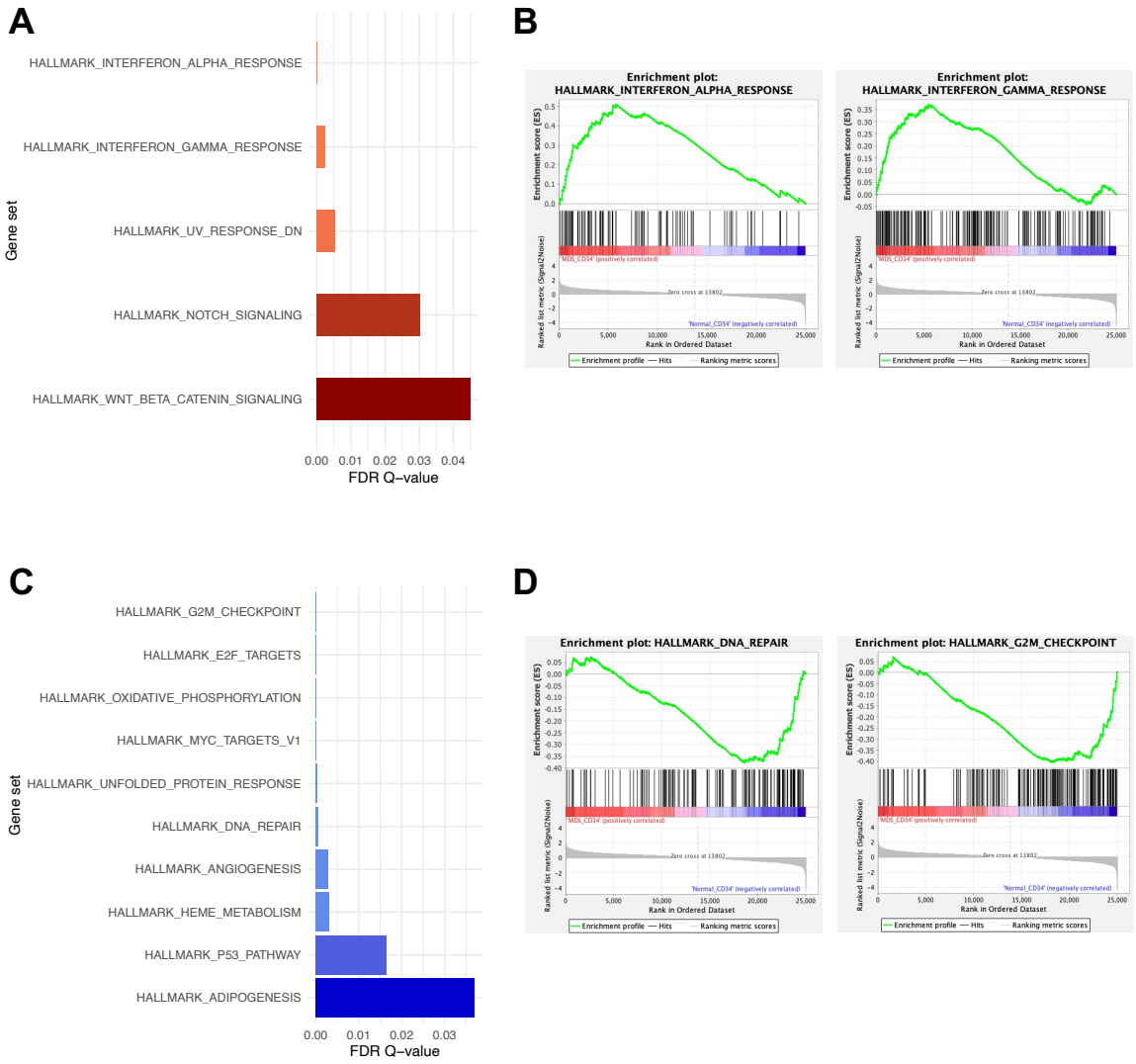


Figure 4.3 Interferon response is upregulated and DNA repair downregulated in CD34⁺ cells isolated from patients diagnosed with MDS with TET2 mutations*; A-D) GSEA[196,197] using Hallmark Gene Sets (MSigDB[255,256]), A) A plot showing 5 gene sets enriched in the MDS samples, B) Enrichment plots for selected gene sets from plot A), C) A plot showing top 10 gene sets downregulated in the MDS samples, D) Enrichment plots for selected gene sets from plot C). Plots made with ggplot2[260], GSEA[196,197].*

To study the function of identified genes GSEA[196,197] was performed using Hallmark Gene Sets (H) from the Molecular Signatures Database (MSigDB) human collection[255,256]. GSEA[196,197] identified 20 gene sets upregulated and 30 downregulated in the MDS samples in comparison to healthy samples, out of which 5 and 20 were significant at a false discovery rate (FDR) < 25% (FDR threshold commonly used[318] in GSEA[196,197]) respectively (Figure 4.3A,C). Among the terms enriched in the MDS samples we found interferon alpha and interferon gamma

response (Figure 4.3B) - pathways intensively studied over the last years in the context of loss of function TET2 mutations[319,320]. Not unexpectedly, among the terms downregulated in MDS samples we found DNA repair, which aligned well with recently discovered 5hmC accumulation at damaged DNA sites and role of TET2 in genome integrity[321]. To study gene sets curated from more sources than the well-defined Hallmark Gene Set[255,256], the BaderLab Gene Sets[322] was used. GSEA[196,197] analysis performed using BaderLab Gene Sets[322] indicated 3070 gene sets upregulated in the MDS samples and 3409 in the healthy samples, however none of these was significantly enriched at FDR < 25%.

Next, we examined gene expression in the STRC fraction ($CD34^+CD38^+$) aiming to find differences between MDS samples and healthy controls with a focus on the rapidly expanding progenitor cell fraction. Differential gene expression analysis also showed significant differences between patient and healthy control samples, with 1234 differentially expressed genes (abs(log_2FC) > 1 & adjusted p-value < 0.05). More genes (~59%) were upregulated in the MDS samples in comparison to the healthy control STRC cells (727 genes upregulated in MDS samples, 507 genes downregulated in MDS samples) (Figure 4.4). GSEA[196,197] performed using Hallmark Gene Sets (H) from the Molecular Signatures Database (MSigDB) human collection[255,256] showed 13 gene sets upregulated and 37 downregulated in the MDS samples in comparison to healthy samples, however none of them was significant at FDR < 25%. GSEA[196,197] analysis performed using BaderLab Gene Sets[322] indicated 2214 gene sets upregulated in the MDS samples and 4250 in the healthy samples, however none of these were significantly enriched at FDR < 25%.

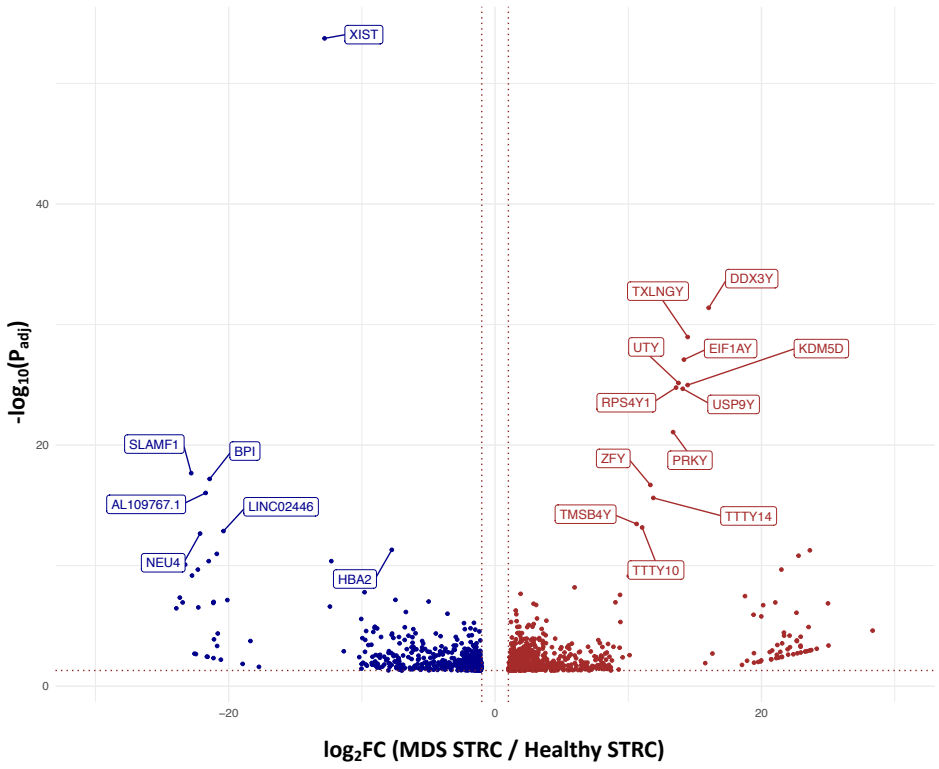


Figure 4.4 Bulk RNA sequencing data analysis of STRC bone marrow cells indicates upregulated gene expression in patients diagnosed with MDS in comparison to healthy bone marrow control; *Volcano Plot with highlighted top 20 differentially expressed genes. Plot made with ggplot2*[260].

Finally, we studied gene expression in HSPCs ($CD34^{+}CD38^{-}$) isolated from patients diagnosed with MDS carrying *TET2* mutations. The differential gene expression analysis indicated noticeable differences between MDS and healthy conditions with 2557 significantly differentially expressed genes (abs(log_2FC) > 1 & adjusted p-value < 0.05), where 56% of genes were downregulated in the MDS samples (1439 genes downregulated in MDS samples, 1118 genes upregulated in MDS samples).

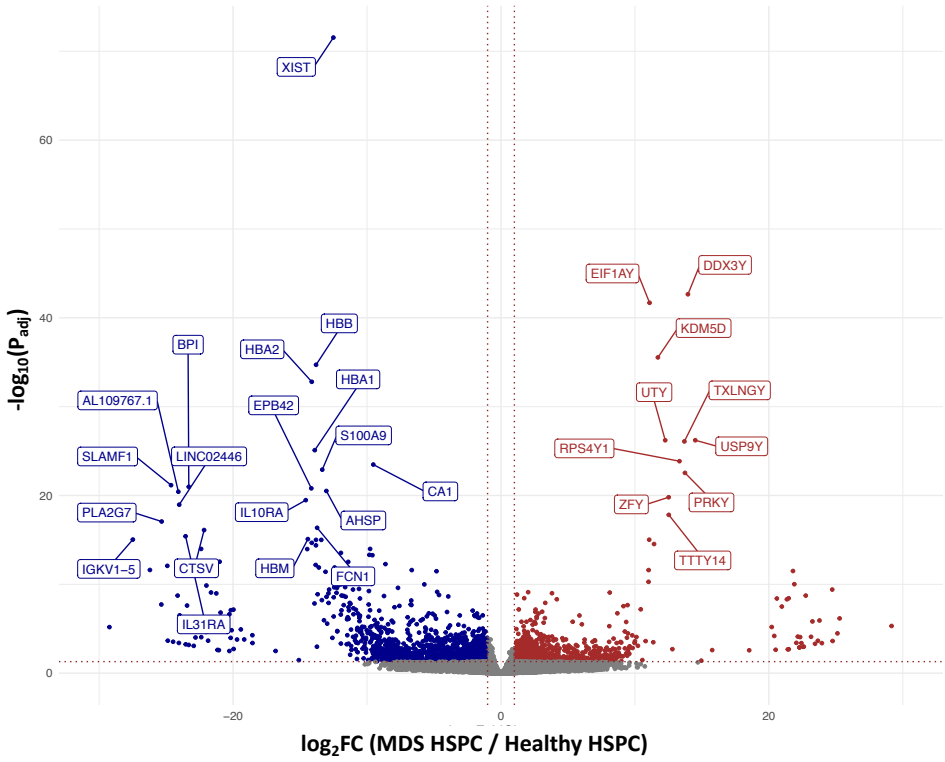


Figure 4.5 Bulk RNA sequencing data analysis of HSPCs indicates downregulated gene expression in patients diagnosed with MDS with TET2 mutations in comparison to healthy bone marrow control; *Volcano Plot with highlighted top 30 differentially expressed genes. Plot made with ggplot2*[260].

GSEA[196,197] analysis using the Hallmark Gene Set[255,256] showed 15 gene sets upregulated and 35 downregulated in the MDS samples in comparison to healthy controls, out of which 2 and 24 were significant at FDR < 25% respectively (Figure 4.6A,B). Among the enriched terms we found an interferon alpha response, like in case of bulk CD34^{+} cells. Another expected deregulation was in the inflammatory response, as inflammation is a hallmark of haematological malignancies and is believed to be regulating clonal selection and aging[152,323]. Moreover, TET2 protein regulates inflammatory response in myeloid cells by mediating transcription, controlling expression of cytokines and modulating BM microenvironment[87,156,324,325].

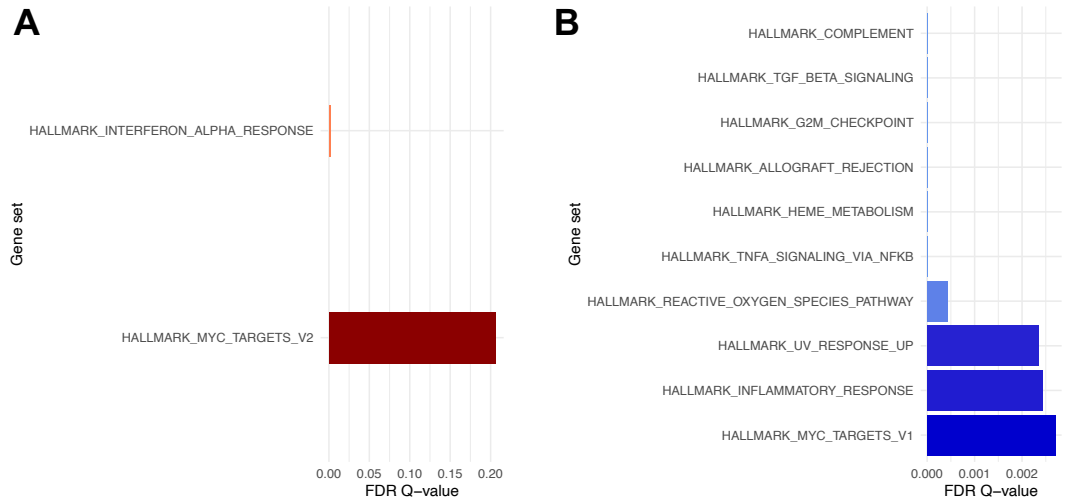


Figure 4.6 Interferon and inflammatory response are predicably deregulated in HSPC isolated from patients diagnosed with MDS with TET2 mutations; *A-B) GSEA[196,197] using Hallmark Gene Sets (MSigDB[255,256]), A) gene sets enriched in the MDS samples, B) gene sets downregulated in the MDS samples, graph showing top 10 enriched gene sets. Plot made with ggplot2[260].*

To further explore deregulated gene sets beyond the hallmark pathways, GSEA[196,197] analysis performed using BaderLab Gene Sets[322]. The analysis indicated 1915 gene sets upregulated and 4549 downregulated in the MDS samples, with respectively 13 gene sets and no gene sets significant at FDR < 25% (Figure 4.7A).

A potentially interesting gene set enriched in the MDS samples was zinc ion transport (Figure 4.7B). Zinc plays an important role in cell proliferation and differentiation[326–330]. Moreover, zinc level seems to be decreased in serum in patients diagnosed with MDS, as recently shown by Moazed et al[331]. Importantly, zinc ions also stabilize TET2 protein structure[332]. Altogether, studies of zinc cations transport might give novel insights into *TET2*-mutated HSPC behaviour in patients diagnosed with MDS.

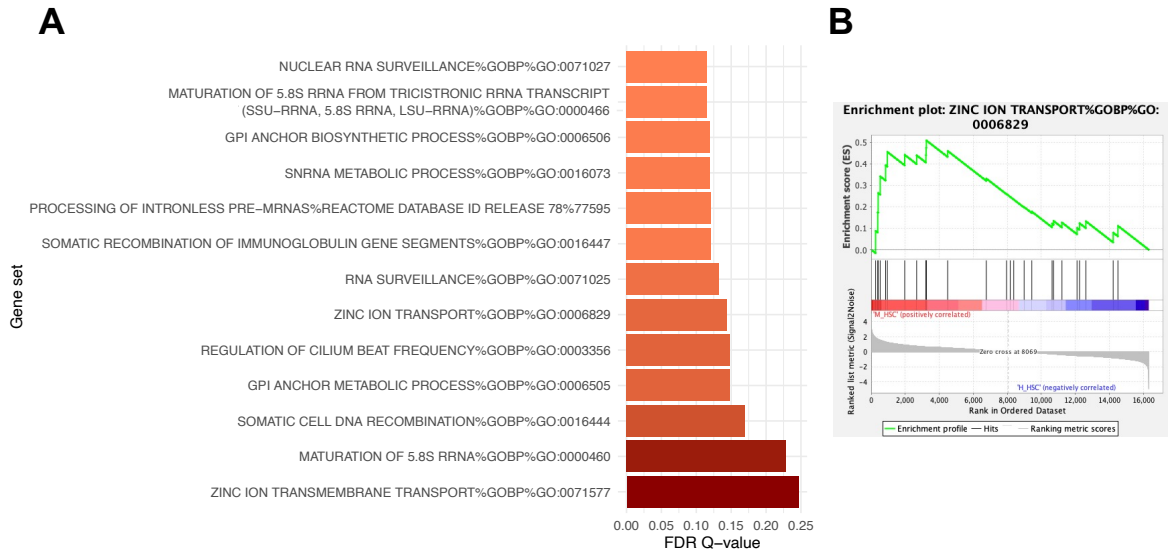


Figure 4.7 Zinc ion transport is upregulated in HSPCs isolated from patients diagnosed with MDS with TET2 mutations; *A-B) GSEA[196,197] using BaderLab Gene Sets[322], A) A plot showing 13 gene sets enriched in the MDS samples, B) Enrichment plots for the selected zinc ion transport gene set from plot A). Plots made with ggplot2[260], GSEA[196,197].*

Another ion type recently discovered to play a crucial role in TET2 biology is Ca^{2+}. Luchsinger et al. showed that low intracellular calcium concentration in HSC stabilizes TET enzymes via inhibition of calpain proteases[288]. Indeed, we have found calcium ion binding among the GO terms in the analysis of *TET2*-mutated HSPC isolated from patients diagnosed with MDS. Interestingly, most genes contributing to this GO enrichment were significantly downregulated in *TET2*-mutated HSPC (Figure 4.8A). To further study these genes, a signalling network was built using information from the SignaLink3[262] and SIGNOR[263] databases (Figure 4.8B). One of the dominant clusters were centred around calmodulin-1 (CALM1) and cadherin-1 (CDH1), which were shown to play a regulatory role in the growth of hematopoietic progenitor cells[333] and myeloid differentiation and HSPC proliferation[334], respectively. A second dominant cluster was centred around RET, a kinase receptor recently shown to mediate HSC growth[335]. No specific patterns in terms of gene up- or downregulation were found in the constructed network, but, interestingly, the proteins encoded by most of identified genes were located in the plasma membrane and the extracellular region (Figure 4.8C).

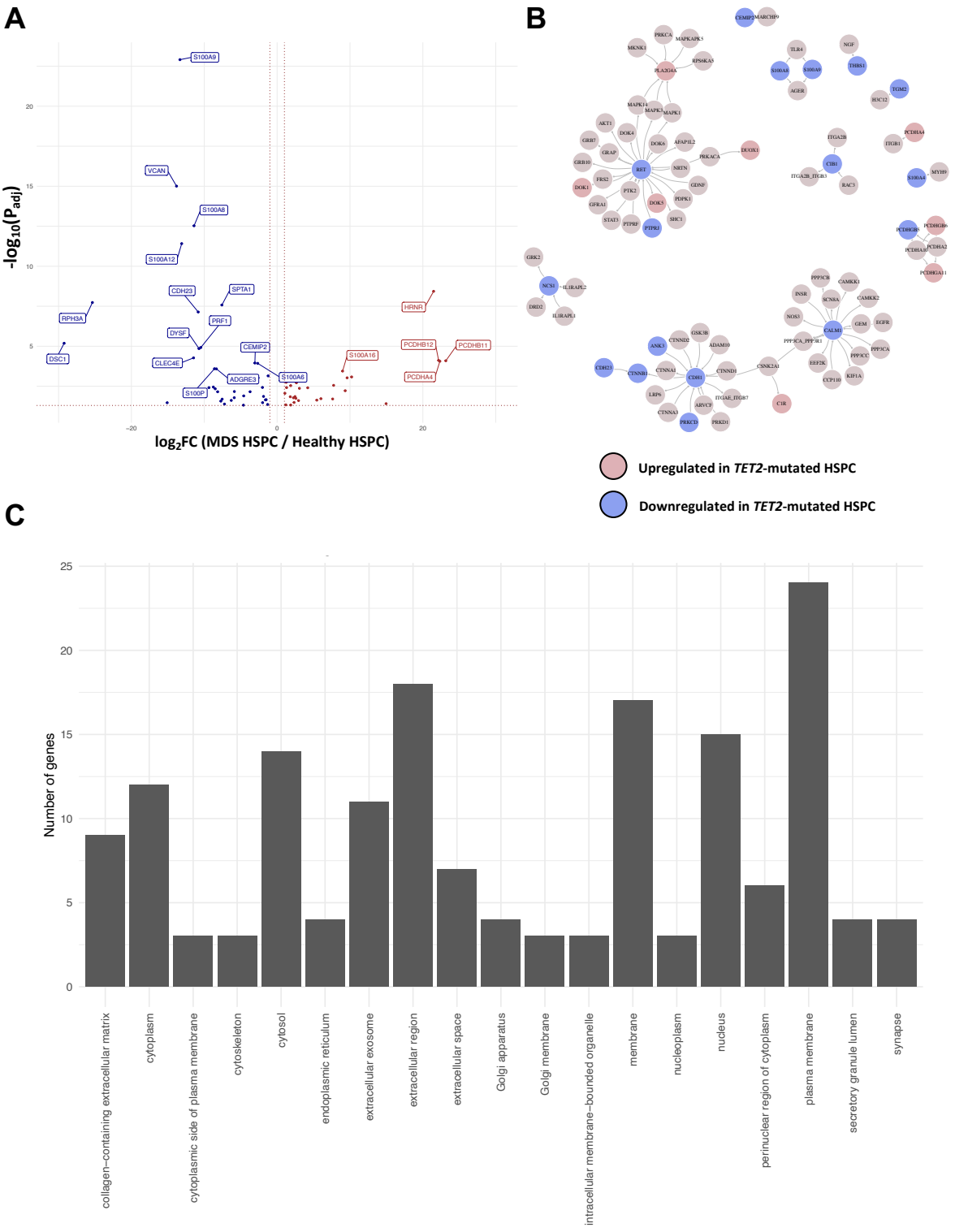


Figure 4.8 Targets involved in the calcium signalling are downregulated in TET2-mutated HSPC; *A) Volcano Plot for differentially expressed genes contributing to the "calcium ion binding" GO[195] term in the ClueGO[254] analysis of GO[195] Molecular function, thresholds - min 50 analysed genes (default = 3) / each term, all remaining thresholds were set to the default settings; B) A signalling network of these genes built using information from the SignaLink3[262] and SIGNOR[263] databases, highlighted genes were significantly up and*

downregulated in TET2-mutated HSPC in comparison to the healthy control cells; C) Localisation of identified genes, filtered locations with min 2 genes, information about localisation accessed via Uniprot[336]. Plots made with ggplot2[260], Omnipath[261], igraph[258], graphics[259], Uniprot[336].

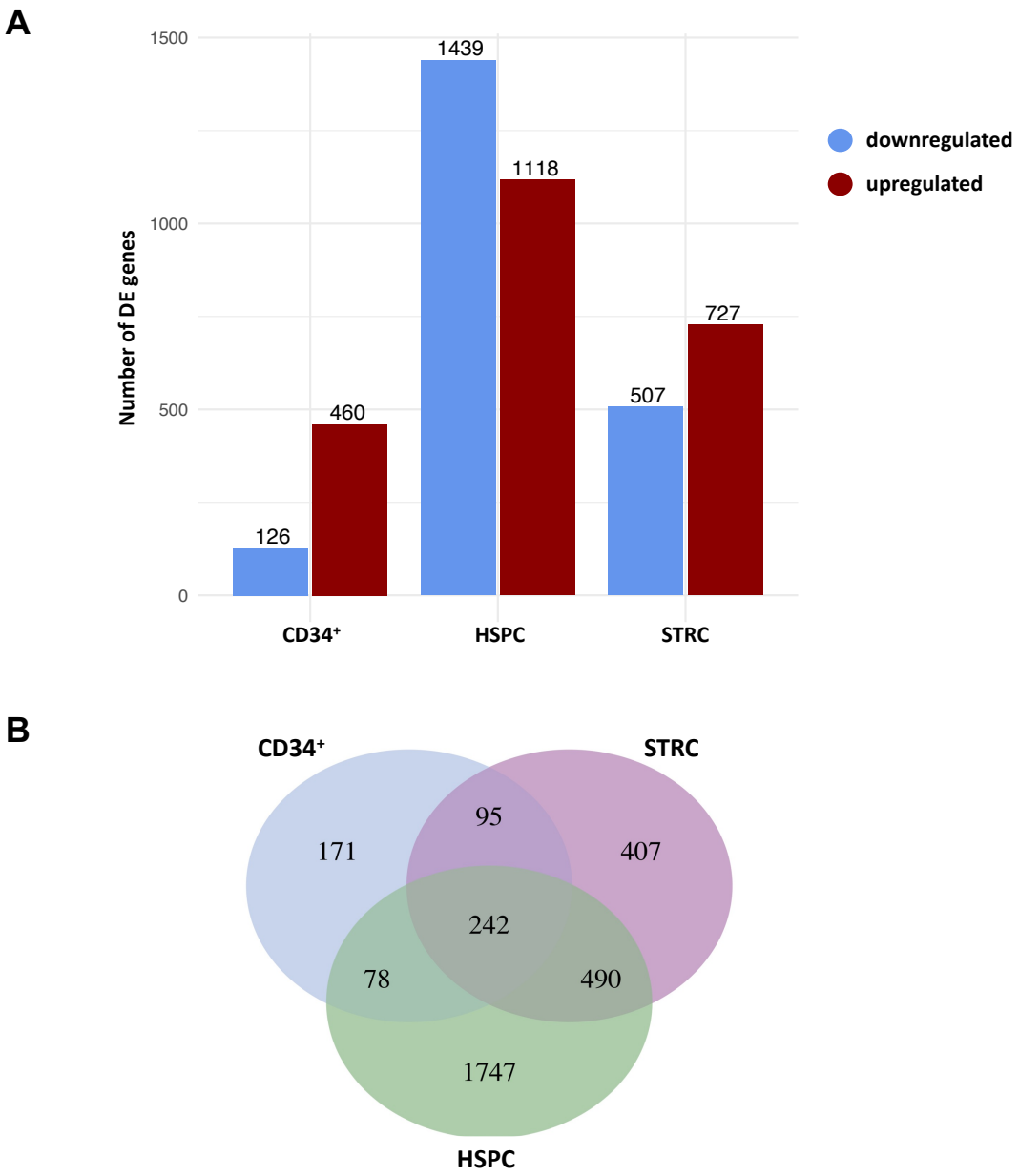


Figure 4.9 A comparison of changes in gene expression across all studied cell fractions – CD34⁺, STRC, HSPC; *A) A bar plot showing numbers of genes down- and upregulated in the MDS samples in comparison to healthy controls; B) Venn diagram showing an overlap of identified DEGs across all studied cell fractions. Plots made with ggplot2[260], VennDiagram[292].*

Finally, changes in the transcriptional landscape across all studied cell fractions – HSPC-enriched fraction ($CD34^+$), STRC fraction ($CD34^+CD38^+$) and the

HSPC-fraction ($CD34^+CD38^-$) - were examined. Interestingly, most differentially expressed genes were found in HSPCs, and least in the $CD34^+$ cells (Figure 4.9A). Many DEGs overlapped across all fractions with 242 genes being differentially expressed in all studied cell types (Figure 4.9B). Altogether, these analyses suggest that gene expression is significantly deregulated in early haematopoietic cell landscape, with both HSPCs and STRCs displaying substantial changes in cellular transcriptome. Notably, in the HSPC fraction more downregulated genes were found, whereas the opposite was a case for $CD34^+$ and STRC cells. Finally, these results highlight the need for studying HSPCs at the highest possible resolution, as the analysis of the bulk $CD34^+$ fraction did not detect as many differences as the HSPC analysis.

4.2. Comparative analysis with human cord blood-derived $CD34^+$ cells with CRISPR/Cas9 induced TET2 knock-out

To compare transcriptomic analyses of *TET2*-mutated HSPCs isolated from patients diagnosed with MDS with a more general human haematopoietic *TET2* dataset, we analysed a RNAseq dataset published by Nakauchi et al.[134] of human cord blood-derived (CB) $CD34^+$ cells with induced *TET2* knock-out (KO) by CRISPR/Cas9 and recombinant adeno-associated virus 6 (rAAV6). Differential gene expression analysis showed significant differences between *TET2*KO and AAV control conditions, with 906 significantly expressed genes ($abs(log_2FC) > 1$ & adjusted p-value < 0.05) (Figure 4.10). 79% of genes were downregulated in *TET2*KO cells in comparison to the AAV control (719 genes downregulated, 187 genes upregulated in *TET2*KO samples).

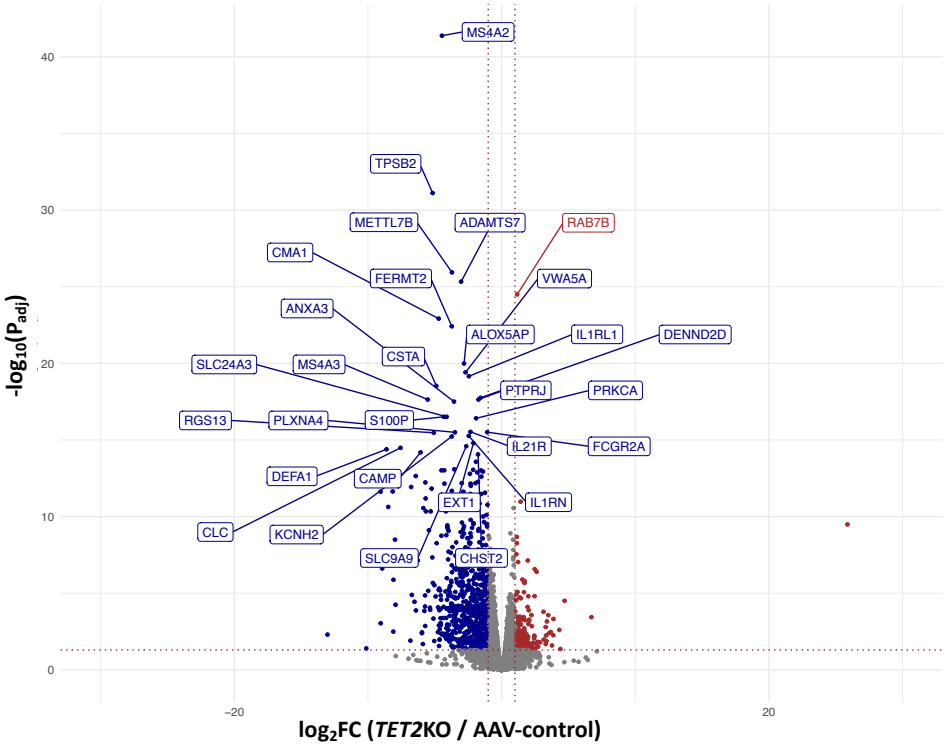


Figure 4.10 Bulk RNA sequencing analysis of Nakauchi et al.[134] dataset shows decrease in the gene expression upon TET2 knock-out in cord blood-derived CD34⁺ cells; *Volcano Plot with highlighted top 30 differentially expressed genes. Plot made with ggplot2[260].*

Next, we aimed to determine whether there was an overlap of detected changes in gene expression in the Nakauchi et al.[134] dataset and our HSPC MDS analysis. The intersection of differentially expressed genes was 171 genes (Figure 4.11A), with the majority of genes being downregulated in both the MDS and transduced CB samples (Figure 4.11B).

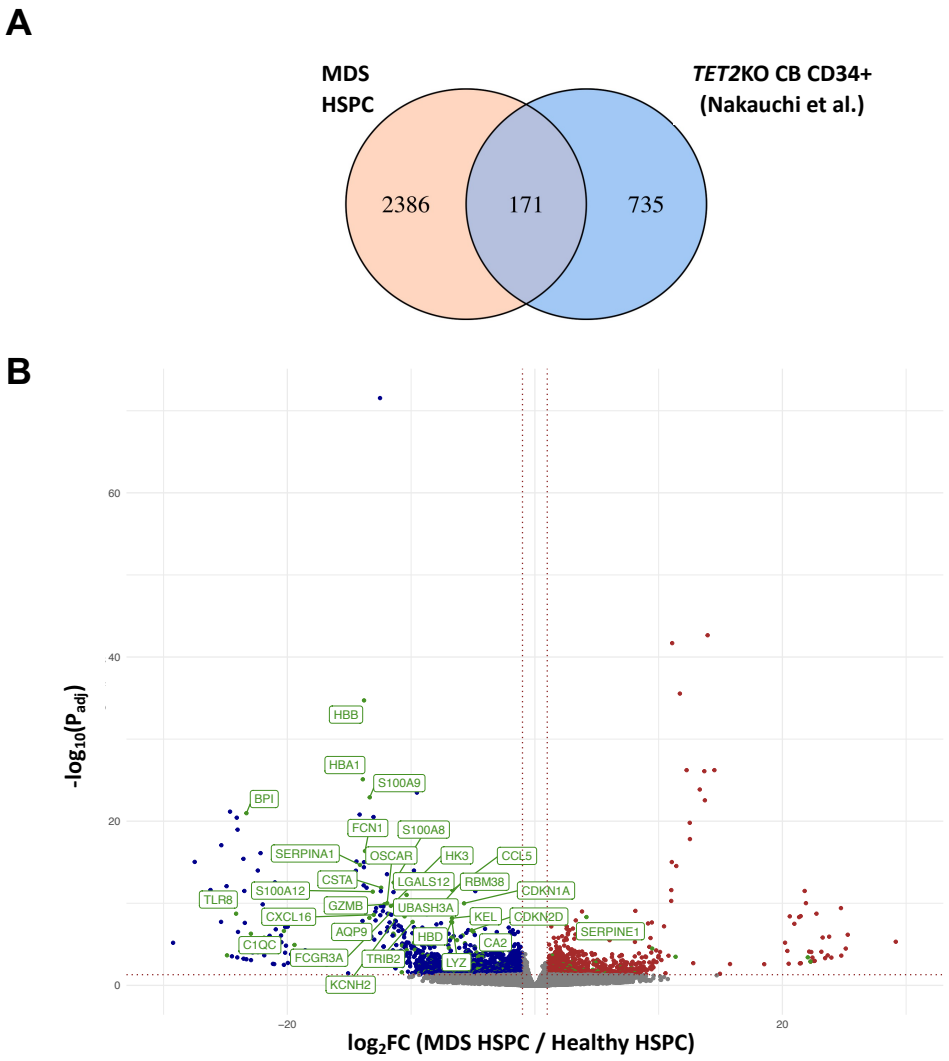
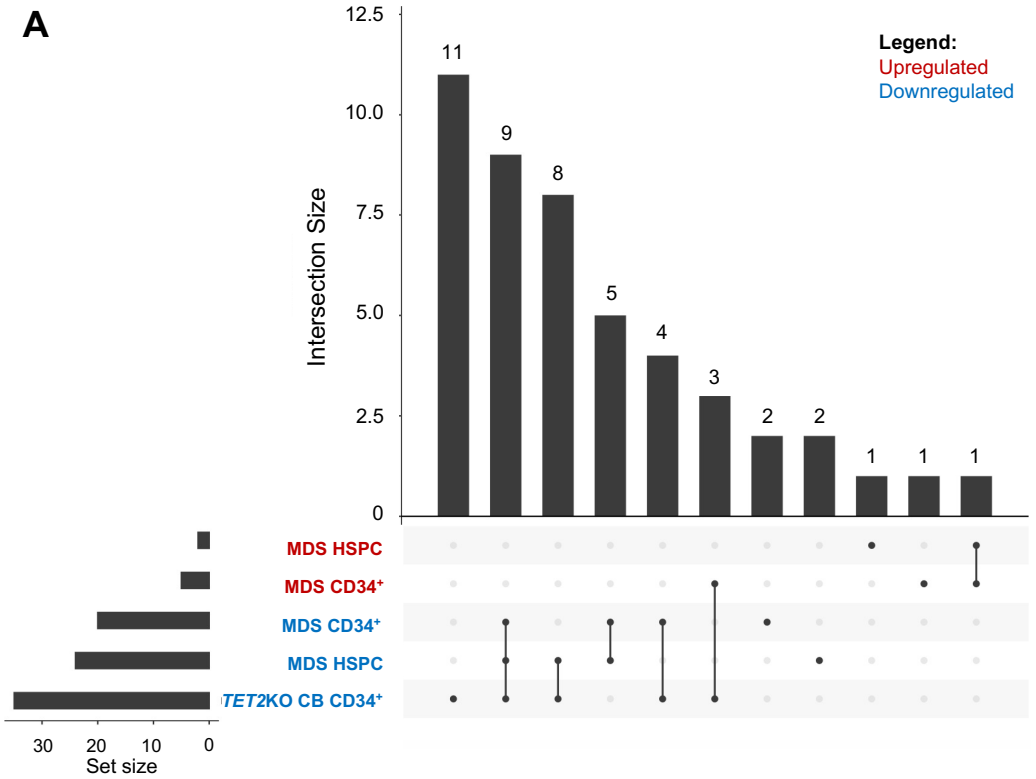


Figure 4.11 Integrative analysis of Nakauchi et al.[134] dataset and bulk RNAseq HSPC MDS dataset indicate an overlap of downregulated genes; *A) Venn Diagram showing an overlap of differentially expressed genes reported in both analyses; B) Volcano Plot of the MDS dataset with highlighted labels for the top 30 differentially expressed genes reported both in the MDS and Nakauchi et al.[134] analyses. Plots made with VennDiagram[292], ggplot2[260].*

Next, to further compare both datasets, an overlap of GSEA[196,197] results using Hallmark Gene Sets (MSigDB[255,256]) was determined. As the GSEA[196,197] analysis did not indicate any gene sets upregulated in the *TET2*KO CB CD34$^+$ samples that were significant at FDR < 25% (in comparison to the AVV control samples), no overlap was found for the gene sets upregulated in both MDS and *TET2*KO conditions (Figure 4.12A). The majority of the overlap was observed for gene sets downregulated in the *TET2*-deficient conditions (Figure 4.12A,B).



B

Gene sets downregulated in *TET2*-deficient samples
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_HEME_METABOLISM
HALLMARK_ANGIOGENESIS
HALLMARK_P53_PATHWAY
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_UV_RESPONSE_UP
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY
HALLMARK_TGF_BETA_SIGNALING
HALLMARK_XENOBIOTIC_METABOLISM

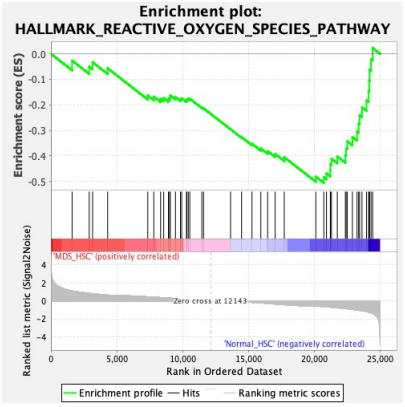


Figure 4.12 Reactive oxygen species pathway is downregulated across all studied datasets; *A) Upset plot of overlapping GSEA[196,197] gene sets in all performed analyses of CD34+ and HSPC cells; B) A list of 11 gene sets downregulated in TET2-deficient samples; C) GSEA[196,197] enrichment plot for reactive oxygen species pathway gene set (MSigDB[255,256]). Plots made with UpsetPlot[337], formattable[284], GSEA[196,197].*

Among gene sets downregulated in the *TET2*-deficient conditions we found the reactive oxygen species (ROS) pathway (Figure 4.12B,C) - a gene set containing genes upregulated by ROS. As these genes were downregulated in *TET2*-deficient MDS samples, this suggests that *TET2*-deficient cells might require less protection from ROS. Intriguingly, recent studies by Aljoufi et al. indicated comparable

engraftment of *Tet2*$^{-/-}$ mouse HSCs and WT mouse HSCs treated with 3% O_2 in the hypoxia chamber, showing reduction in *Tet2* expression in WT HSCs under physioxic conditions[338]. This would be consistent with HSCs residing in a hypoxic environment in the BM niche[339], however more functional studies would be required on the role of ROS in *TET2*-mutated MDS HSPCs.

Interestingly, a GO[195] analysis of molecular function for the 171 overlapping genes indicated enrichment in the "RAGE receptor activity" (Figure 4.13). All genes contributing to this enrichment - *FPR1*, *S100A8, S100A9* and *S100A12* were downregulated in HSPCs isolated from patients with MDS with *TET2* mutations in comparison to healthy control samples (Figure 4.14A).

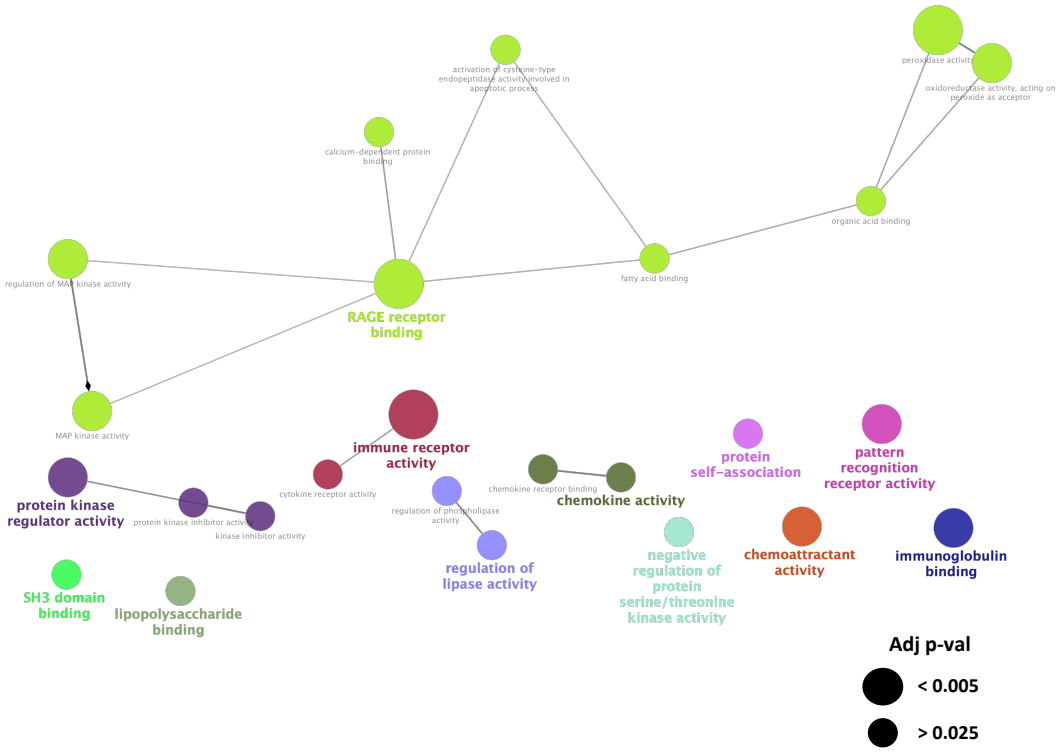


Figure 4.13 RAGE receptor binding is deregulated in TET2-mutated HSPCs isolated from patients diagnosed with MDS and TET2KO cord blood-derived CD34⁺ cells; *Gene Ontology[195] analysis performed with ClueGO[254] for the 171 overlapping genes, node sizes reflect the statistical significance of the terms, force directed layout presented by the Kappa score, layout adjusted to minimise label overlap (see Methods for more details), Molecular Function, thresholds - min 4 analysed genes (default = 3) / each term, all remaining thresholds were set to the default settings. Plots made with Cytoscape[253] and ClueGO[254].*

When we examined this family using the Genemania (https://genemania.org) database, little relatedness to *TET2* was identified, with only two connections to the

RAGE genes: *TET2-FPR1* “genetic interaction” and *TET2-S100A8* “coexpression” (Figure 4.14B). As expected, all analysed S100A family member were closely connected and showed high coexpression with FPR1. That said, as highlighted in Chapter 3, it should be noted that this analysis might be biased due to the lack of available published experimental data regarding the connectivity of *TET2* with other molecules.

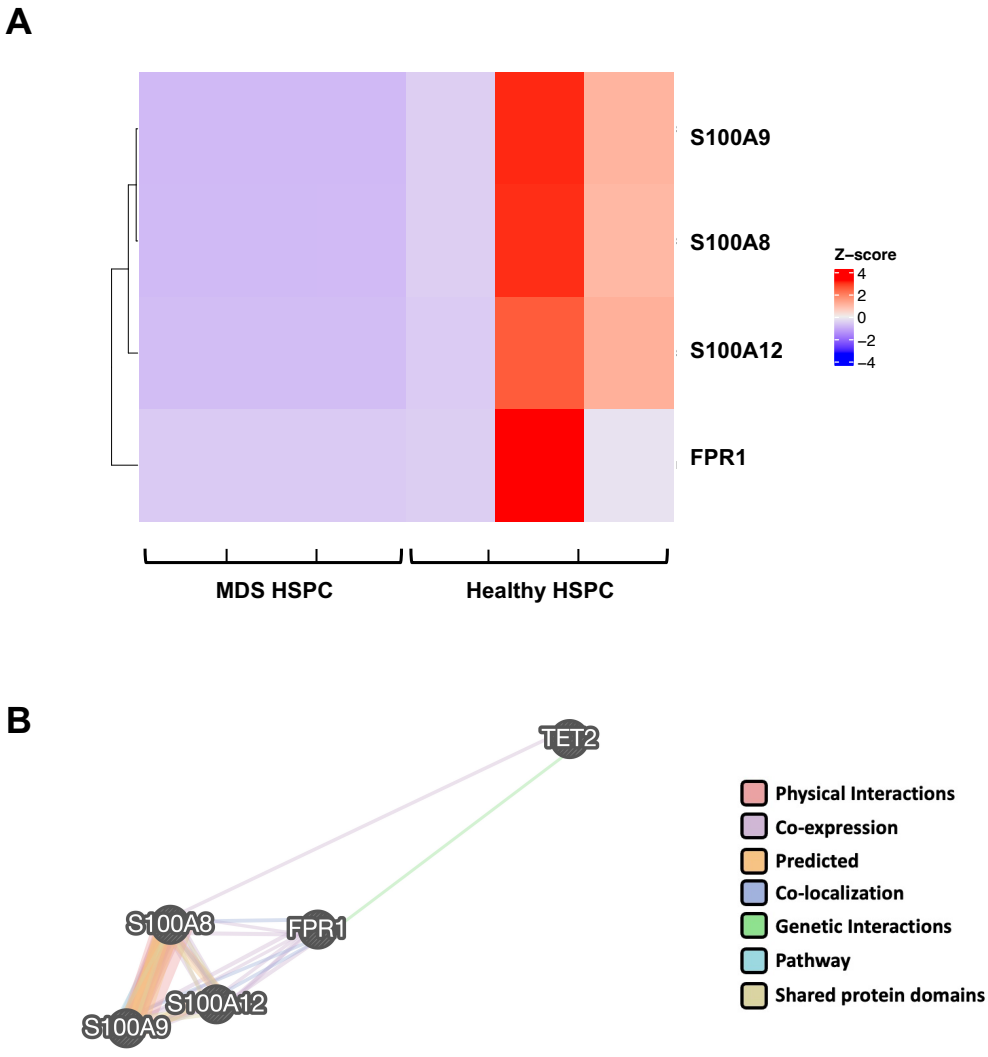


Figure 4.14 Expression of RAGE ligands is decreased in human HSPC in patients diagnosed with MDS with TET2 mutations; *A) A heatmap showing expression of RAGE ligands in HSPCs isolated from patients diagnosed with MDS with TET2 mutations; B) Interaction graph for the RAGE ligands and TET2 constructed using the Genemania database (https://genemania.org), 5 nodes, 60 links in total, ~78% of links were classified as physical interactions). Plots made with ggplot2[260], Genemania (https://genemania.org).*

Since S100A8 and S100A9 are calcium-binding proteins located in the cytoplasm and secreted by immune cells during inflammation and have been identified as biomarkers in haematological malignancies[340], it is worth considering them further. Wang et al. reported an elevated level of S100A8/A9 in BM plasma in MDS patients, which correlated with poor prognosis for those patients[341]. Krečak et al. shown calprotectin (the S100A8/S100A9 dimer and tetramer forms) to be a serum biomarker of the inflammatory state in MPN patients at diagnosis[342]. In addition to S100A8/A9 cytokine function, Song et al. recently revealed its role as a transcriptional coactivator in breast cancer[343]. Altogether, this makes calprotectin an interesting target for future studies, potentially playing a role in the development of MDS in patients diagnosed with *TET2*-mutations. S100A8 and S100A9 was further studied in Chapter 5.

4.3. Integrative analysis with mouse single-cell RNA sequencing data shows little overlap of differentially expressed genes

To study gene expression in HSCs disturbed upon *TET2* loss of function across species, an overlap of DEGs from transcriptome analyses of *Tet2*$^{-/-}$ mouse HSCs (from Chapter 3), HSPCs isolated from patients diagnosed with MDS with *TET2* mutations and cord blood-derived *TET2*KO CD34^{+} cells[134] was determined. The analysis revealed very little overlap across these datasets (Figure 4.15A), but did identify a small handful of interesting targets. For example, between mouse *Tet2*$^{-/-}$ HSCs and MDS HSPCs datasets we identified *Plag1* and *Rbm15 (*Figure 4.15B, C). *PLAG1* is a transcription factor and it drew our attention as it was recently shown to promote human HSC self-renewal[344]. *PLAG1* expression was upregulated in *TET2*-deficient cells in both mouse and human datasets (relative to wild-type / healthy controls) making it an interesting candidate for further studies on its role in TET2 biology. *RBM15* encodes an RNA-binding protein that regulates N6-methyladenosine (m6A) methylation of RNAs[336] and plays various roles in haematopoietic development[345], however its function in relation to TET2 remains unknown. Intriguingly, *RBM15* expression was downregulated in *TET2*-deficient HSPCs isolated from patients diagnosed with MDS and upregulated in *Tet2*$^{-/-}$ mouse HSC (in comparison to healthy / wild-type controls).

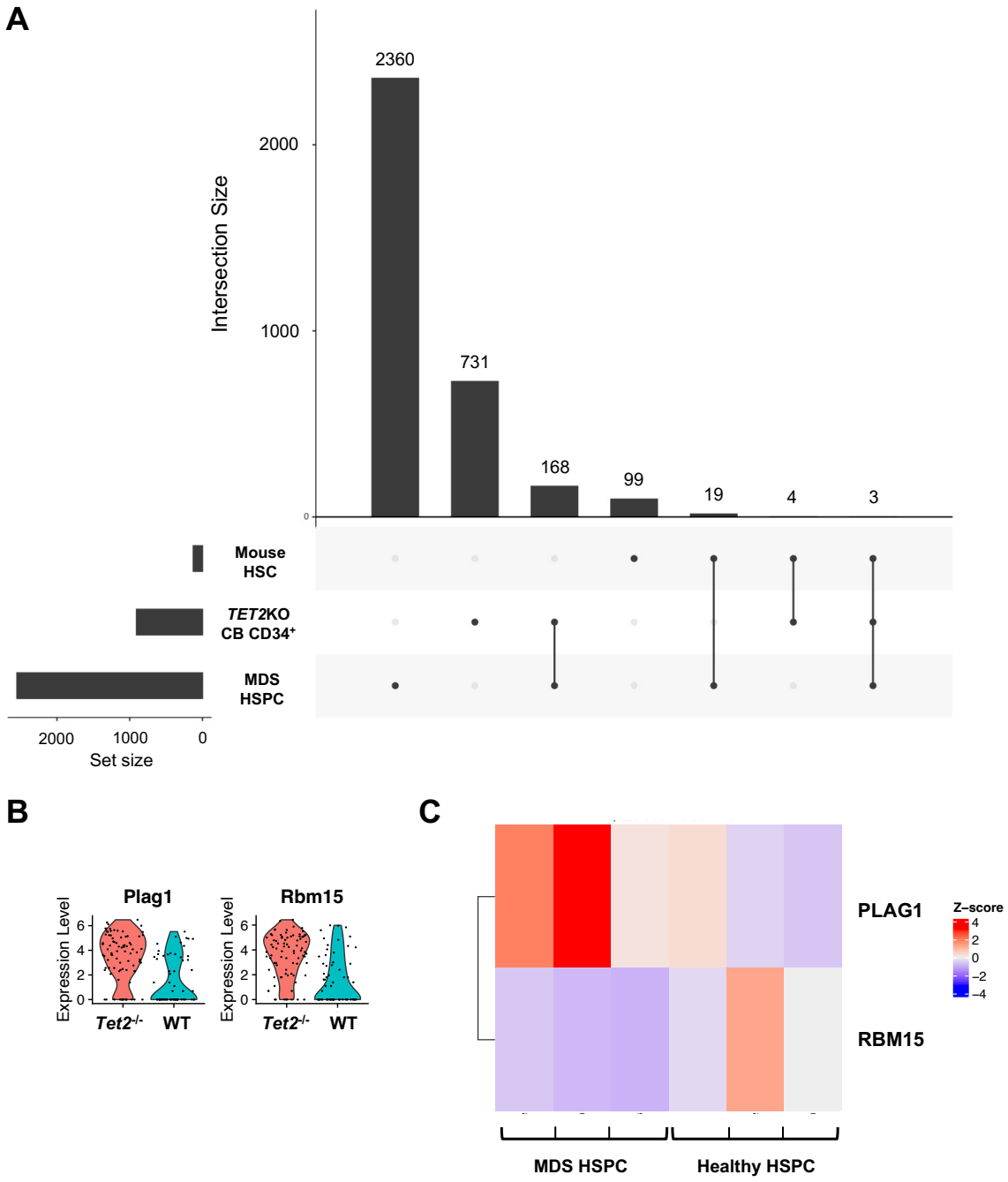


***Figure 4.15 Human-mouse integrative analysis shows little overlap of differentially expressed genes**; A) Upset plot showing an overlap of differentially expressed genes reported in human and mouse transcriptome analyses; B) Violin Plots for Plag1 and Rbm15 showing expression in mouse Tet2$^{-/-}$ HSC; C) A heatmap showing expression of the same genes in the MDS RNAseq dataset in HSPCs. Plot made with UpsetPlot[337], ComplexHeatmap[346], Seurat[217].*

Interestingly, only 3 genes (*Mmrn1*, *Cdkn1a*, *Rgs2*) were significantly differentially expressed in all datasets (mouse *Tet2*$^{-/-}$ HSCs, MDS HSPCs and *TET2*KO CB CD34^{+} cells[134]) (Figure 4.15A). This discrepancy might partially be

caused by cell types not perfectly mapping across species, e.g., mouse CD45^{+}EPCR^{+}CD150^{+}CD48^{-} HSCs are ~50% functional HSCs[15], whereas human CD34^{+}CD38^{-} cells are <2% functional HSCs[347]. Altogether, this underlines the differences between humans and mice and highlights the need for further studies of human cells. Hopefully, as sorting and labelling strategies improve in human HSCs, these questions will become more answerable.

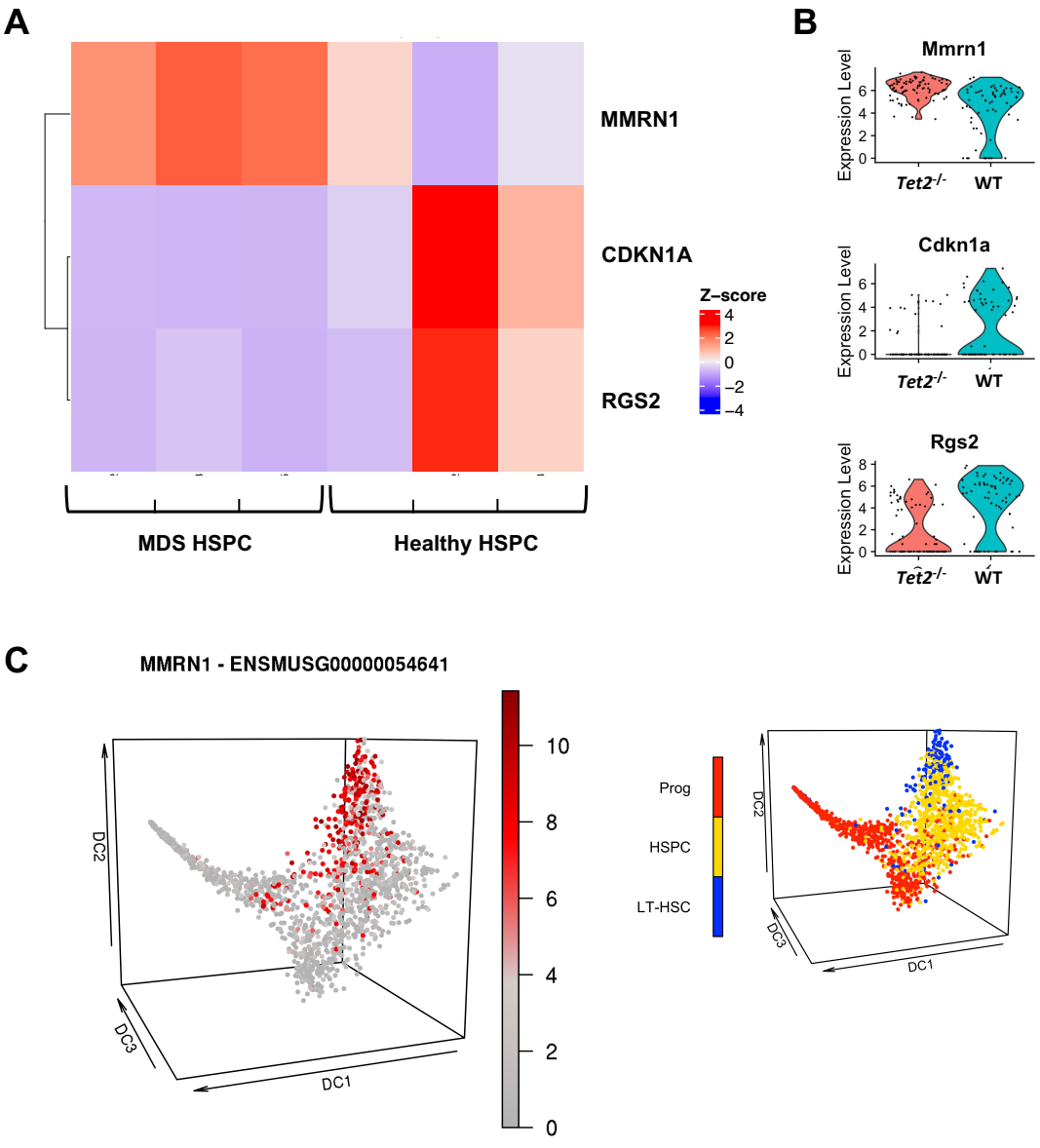


Figure 4.16 Expression of multimerin-1 is upregulated in patients diagnosed with MDS with TET2 mutations*; A) A heatmap showing expression of the overlapping genes (genes differentially expressed in mouse Tet2$^{-/-}$ HSC, TET2KO CB-derived CD34^{+} and HSPC isolated from patients diagnosed with MDS with TET2 mutations) in the MDS HSPC RNAseq dataset; B) Violin Plots for the same genes showing expression in mouse Tet2$^{-/-}$ HSC; C) Expression of Mmrn1 across haematopoietic cell subpopulations in the Nestorowa et al.[266] dataset; D) UMAP[186] plot by Nestorowa et al.[266] with highted haematopoietic subpopulation clusters, Prog = haematopoietic progenitor cells (Lineage^{-}Sca1^{-}cKit^{+}), HSPC = haematopoietic stem and progenitor cells (here*

defined as Lineage$^-$Sca1$^+$cKit$^+$), LT-HSC = long-term self-renewing haematopoietic stem cells (here defined as Lineage$^-$c-Kit$^+$Sca1$^+$CD34$^-$Flk2$^-$). Plots made with ComplexHeatmap[346], Seurat[217] and https://blood.stemcells.cam.ac.uk/ single_cell_atlas.html.

Among the overlapping differentially expressed genes we also found multimerin-1 (*Mmrn1*) (Figure 4.16A), a gene encoding secreted protein MMRN1 produced by megakaryocytes and endothelial cells[336,348,349]. scRNAseq analysis of mouse *Tet2*$^{-/-}$ HSCs showed that *Mmrn1* was upregulated in *Tet2*$^{-/-}$ HSC in comparison to WT HSC (Figure 4.16B). Moreover, analysis of the Nestorowa et al.[266] dataset indicated high *Mmrn1* expression in the LT-HSC cell cluster (Figure 4.16C,D). *Mmrn1* has also been previously shown to characterise HSC clones with differentiation-inactive fates[295,296]. These findings suggest that *Mmrn1* might be an interesting target for future studies on *Tet2*-deficient HSCs.

4.4. Discussion

Analyses of next-generation sequencing data of patients diagnosed with myeloid malignancies can help identifying pathways responsible for malignant transformation and understanding the mechanisms of clonal evolution leading to disease progression. In this chapter we generated a bulk RNA sequencing data to compare HSPCs transcriptomes of healthy controls and those generated from patients diagnosed with *TET2*-mutated MDS. We compared this to a previously published *TET2*KO cord blood CD34$^+$ cells dataset[134] and to a scRNAseq dataset[1] generated from a *Tet2*$^{-/-}$ mouse model[2].

The differential expression analyses of CD34$^+$ and HSPC (CD34$^+$CD38$^-$) revealed significant differences between patient and healthy control samples. Interestingly, in the case of CD34$^+$ cells, the analysis showed more upregulated genes (~78%), whereas in the case of HSPC most genes were downregulated (56%). Notably, overall, more genes were differentially expressed in the HSPC analysis, with 586 DEGs in the MDS/healthy comparison for CD34$^+$ cells and 2557 DEGs for the HSPC fraction (abs($\log_2$FC) > 1 & adjusted p-value < 0.05). This suggests better resolution when studying HSPC rather than CD34$^+$ cells, potentially

due to a high heterogeneity of cells with the $CD34^+$ phenotype, therefore further analyses presented in this chapter were based on the MDS/healthy comparative studies of HSPC cells.

Interestingly, many targets involved in the calcium ion binding showed significantly lower expression in *TET2*-mutated MDS HSPC samples (Figure 4.8A), such as members of the S100A family (S100A6, S100A8, S100A9, S100A12) and S100P. This attracted our attention, as Luchsinger et al. has recently shown that the low intracellular calcium concentration in mouse HSC stabilises TET enzymes by the inhibition of calpain proteases[288]. Their functional studies revealed lower Ca^{2+} concentration in mouse HSCs in comparison to progenitor cells and better engraftment of HSC cultured in lower $CaCl_2$ concentration before mouse transplants. Moreover, they showed that TET enzymes were more stable upon the inhibition of calpains, calcium-dependent proteases[288,350]. Among calcium binding proteins identified in this Chapter, members of the S100A family (S100A8, S100A9 and S100A12) are known to activate inflammatory pathways[351]. In 2013 Chen et al. proposed that S100A9 through interaction with CD33 leads to the expansion of malignant cells and MDS development[352]. Subsequent studies investigated the S100A8/9-induced NF-κB pathway[341,353,354]. S100A9 was proposed to activate the PD-1/PD-L1 signalling in the malignant BM environment[355]. In another study, Cluzeau et al. suggested that S100A9 together with tumour necrosis factor-α (TNF-α) disturb erythropoiesis in MDS[356]. However, the exact mechanism of the function of S100A proteins in HSPCs in *TET2*-mutated MDS remains unknown. At this stage, we do know that niche-derived S100A8 and S100A9 promote genotoxic stress in HSPCs and that increased S100A8 and S100A9 levels in bone marrow plasma of patients diagnosed with MDS is correlated with worse prognosis[49,341,357]. Moreover, as mentioned, there is also strong evidence supporting the calcium-TET2 link[288]. This also aligns well with increased intracellular Ca^{2+} concentration triggering HSC cell divisions as reported by Umemoto et al[350,358]. Taken together, considering these pairwise links (calcium-TET2, S100A-MDS, S100A-HSC, calcium-HSC and calcium-S100A and TET2-HSC), it could be speculated that the calcium-S100A-TET2-HSC axis plays a role in the MDS development, however further thorough studies investigating these variables in a joint experiment aiming to confirm this mechanism are needed. Altogether this suggests important role of calcium in HSC maintenance

and self-renewal. Analyses presented in this chapter confirm significant decrease in expression of calcium ion binding proteins in *TET2*-mutated human HSPC advocating for further studies on the role of calcium ions in the development of *TET2*-mutated MDS.

The analysis of an intersection of differentially expressed genes in the *Tet2*$^{-/-}$ mouse model and human *TET2*-mutated MDS samples indicated little overlap, with only 19 genes being significantly different in both disease / mutant vs healthy comparisons. This highlights the need for analyses of human cells. Currently, mice constitute an essential model and are pivotal at the early research stages, however in order to fully understand malignancy development we must study molecular mechanisms in human cells. Naturally, one of the main difficulties is huge variability caused by heterogeneity of patients' samples due to presence of different mutations, different order of mutations acquisition and therefore different mechanisms of clonal evolution of the disease. Moreover, TET2 might have different downstream targets in mice and humans and therefore the consequences of its deficiency might differ across species. Among the most promising candidates for further validation studies that overlapped between mouse and human datasets was multimerin1 (*MMRN1*), a gene significantly upregulated in HSPCs isolated from patients diagnosed with *TET2*-mutated MDS in comparison to healthy controls. Moreover, analysis of the Nestorowa et al.[266] dataset showed that *Mmrn1* is highly expressed in LT-HSCs (Figure 4.16C,D). Altogether further *in vitro* studies, such as e.g. RT-qPCR validation of differences in expression and generation of *Mmrn1*-deficient mouse models might provide useful insights into molecular mechanisms of *TET2*-dependent increases of HSPC self-renewal in MDS.

Analyses of molecular mechanisms of MDS origins and progression are fundamental for improving diagnosis and treatment for patients. Bioinformatic studies help with identification of potential clinical markers and molecular targets and development of novel therapies. Analyses presented in this chapter suggest a set of genes (*Mmrn1*, RAGE ligands) potentially deregulated in HSPCs in patients diagnosed with MDS with *TET2*-mutations, proving a valuable resource for further studies.

5. Exploring extracellular mechanisms governing HSC self-renewal

Various factors modulate the BM niche, such as inflammation, ageing or malignancy development. Over the last decades, researchers have been investigating the bone marrow microenvironment, aiming to define extrinsic factors supporting HSC quiescence, self-renewal, proliferation[165,359,360]. Since Schofield's discovery[43], many studies have focused on characterising phenotypes and roles of different cell subpopulations in the HSC maintenance, however less attention was paid to the ECM[359] and the BM ECF. Accordingly, in this chapter I aimed to study the role of the BM ECF in HSC biology.

Examining the HSC microenvironment is important as multiple studies have suggested that the niche plays a major role in the haematological malignancy development[165,359,360]. Malignant cells not only compete with the healthy blood cells in the BM, but also remodel the niche to support their own growth[49,361]. Therefore, analyses of the interactome of HSCs with mutations in genes commonly found in patients diagnosed with haematological malignancies might provide intricate details on disease evolution and lead to development of novel therapies.

However, studies of HSC communication with their niche are difficult as many new molecular partners are being identified every year. This suggests network analyses as a good tool to resolve complex HSC-niche relationships, identify patterns and key interactions. That is why in this chapter we performed proteome network analyses aiming to study intercommunication of *Tet2*-mutated HSC with the BM. The results provide an insight into molecular partners of *Tet2*-deficient HSCs.

5.1. Integrative analysis of mouse proteome and human *TET2*-mutated MDS datasets point to several targets potentially involved in the *TET2*-mutated MDS development

To investigate targets identified in the RNAseq analysis of human HSPCs isolated from patients diagnosed with *TET2*-mutated MDS (analyses presented in Chapter 4), an integrative analysis with mouse HSC proteome data was performed. A proteome dataset of *Tet2*$^{-/-}$ and WT mouse CD45^{+}CD150^{+}CD48^{-} HSC was used for this aim (dataset published and analysed in Dr Daniel Bode's PhD thesis[1] and reanalysed by Dr Maria Jassinkaja, (abs($log_2FC \geq 0.5$))).

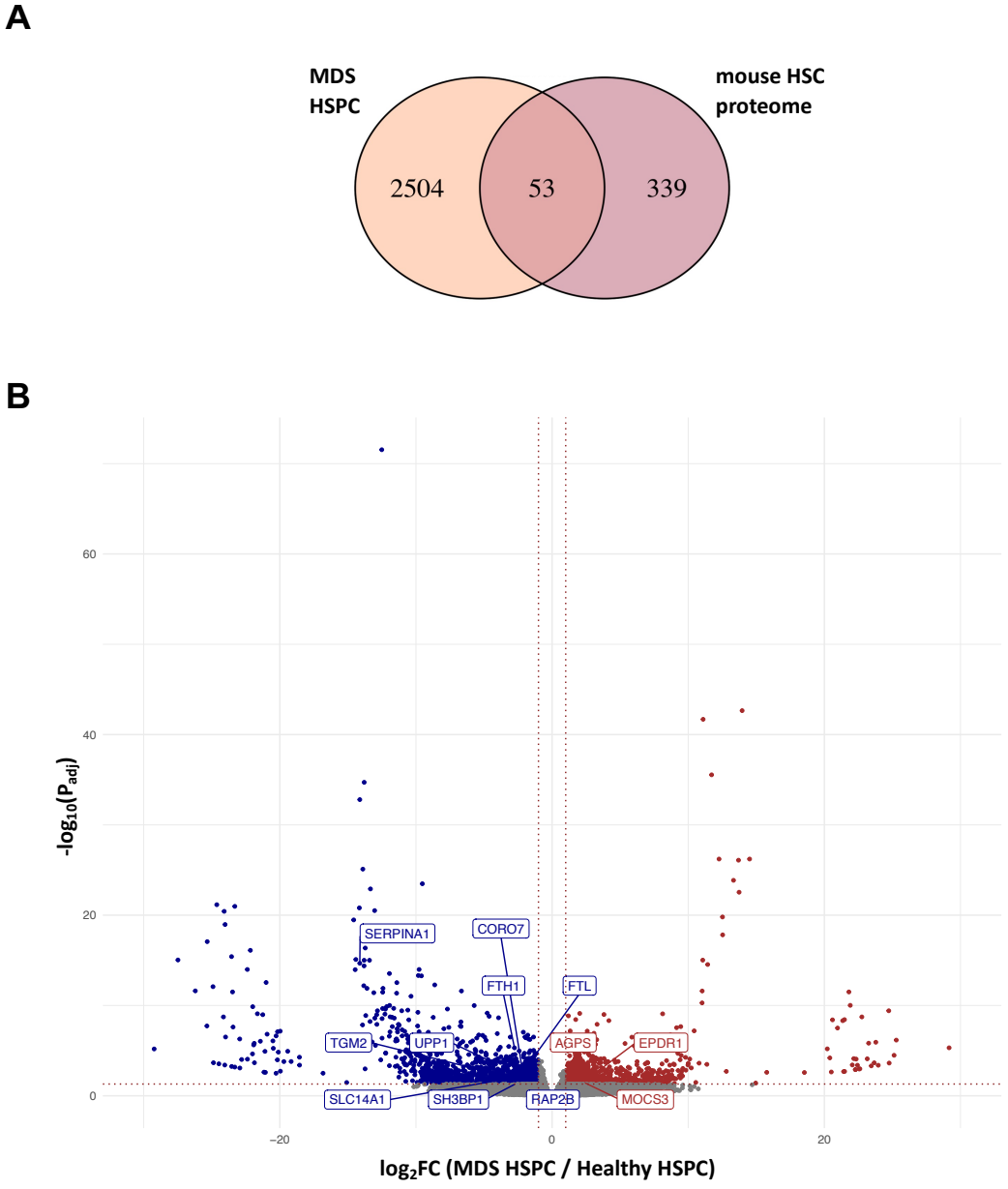


Figure 5.1 An overlap of targets from the HSC proteome dataset and the RNAseq analysis of HSPC isolated from patients diagnosed with TET2-mutated MDS highlights several proteins potentially involved

***in the TET2-mutated MDS development**. A) Venn Diagram showing an overlap of differentially expressed genes reported in mouse HSC proteome and human TET2-mutated MDS HSPC datasets; B) Volcano Plot of the human TET2-mutated MDS HSPC dataset with highlighted overlapping targets that were consistently up- or down-regulated in both mouse HSC proteome and human TET2-mutated MDS HSPC datasets. Plot made with VennDiagram[292], ggplot2[260].*

An overlap of all differentially expressed genes in the *TET2*-mutated HSPC MDS dataset and *Tet2*$^{-/-}$ mouse HSC dataset showed only 53 targets (Figure 5.1A). However, intriguingly only 12 of them were consistently up- or down-regulated in both datasets (Figure 5.1B).

We investigated further interactome of these 12 targets which revealed little connectivity (Figure 5.2A), however drew our attention to several targets that were already known to play a role in haematopoiesis, such as *SERPIN1A*[362,363] or *FTH1* and *FTL1*[364]. However, some other targets represent interesting areas for further investigation, e.g. alkylglycerone phosphate synthase (*AGPS*), coronin-7 (*CORO7*), ependymin-related protein 1 (*EPDR1*), or solute carrier family 14 member 1 (*SLC14A1*), as each of them have not only been significantly differentially expressed in both *TET2*-mutated MDS HSPC RNAseq and mouse HSC proteome datasets, but also were expressed across the HSPC landscape in the Nestorowa et al.[266] dataset (Figure 5.2B).

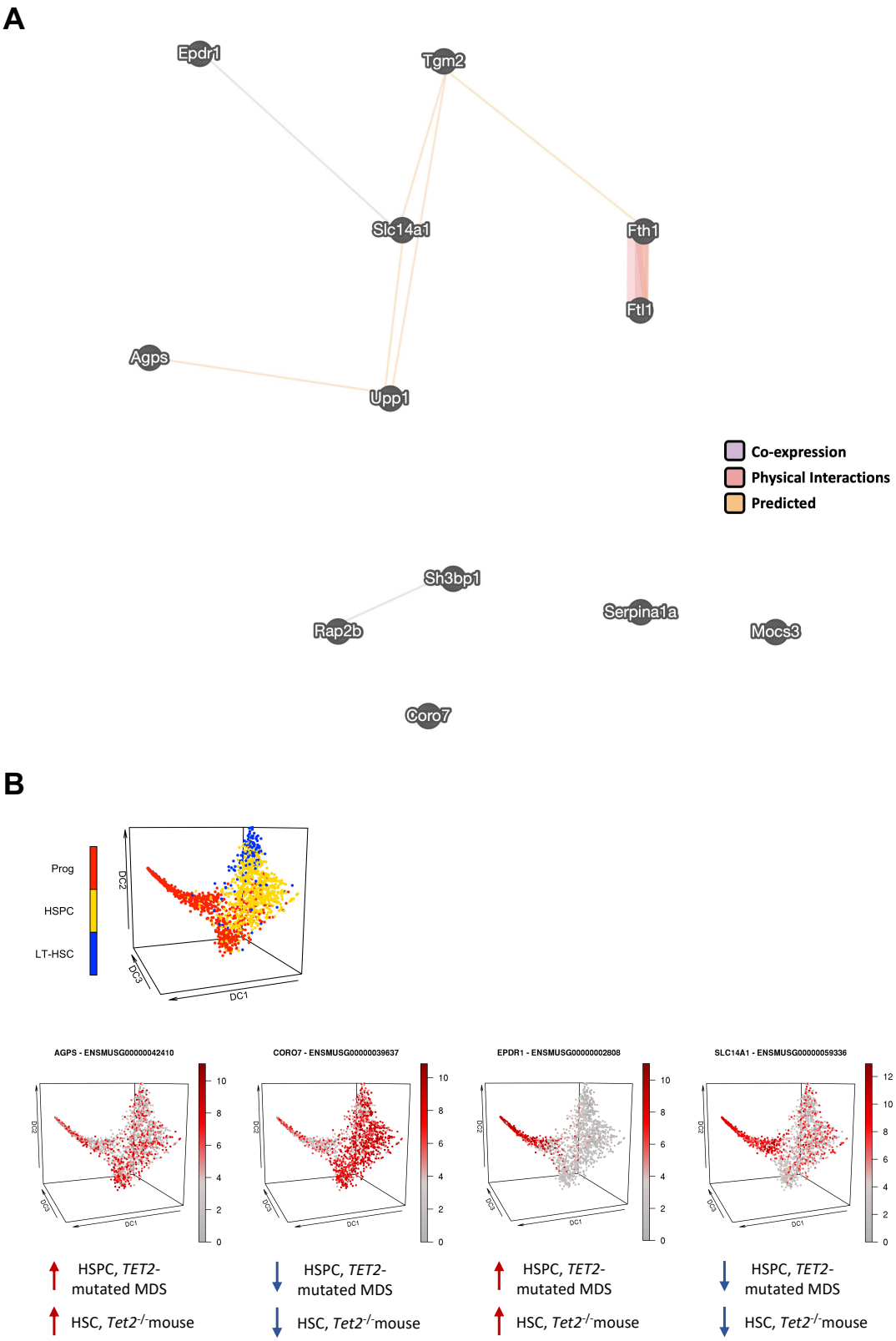


Figure 5.2 Identified consistently up- or down-regulated targets in both human HSPC TET2-mutated MDS and mouse HSC proteome datasets might be potential candidates for further studies*; A) Interaction graph for the overlapping targets constructed using the Genemania database (https://genemania.org), 12 nodes, 13 links in total, 88% of links were classified as physical interactions; B) Expression of selected targets across*

haematopoietic cell subpopulations in the Nestorowa et al.[266] dataset with information on expression in studied TET2-mutated HSPC MDS and mouse HSC proteome datasets. Plots made with Genemania (https://genemania.org) and https://blood.stemcells.cam.ac.uk/single_cell_atlas.html.

5.2. Exploring the function of calprotectin in *Tet2*$^{-/-}$ HSCs

Among more abundant proteins in *Tet2*$^{-/-}$ HSCs in comparison to WT HSCs in the Dr Daniel Bode's proteome dataset were S100A8 and S100A9[1] - proteins forming a calprotectin complex (Figure 5.3A). These proteins are RAGE ligands and were both found to be deregulated in HSPCs in patients diagnosed with MDS with *TET2* mutations in Chapter 4. Intriguingly, calprotectin expression was downregulated in HSPCs isolated from patients diagnosed with *TET2*-mutated MDS and upregulated in the proteome dataset of *Tet2*$^{-/-}$ mouse HSCs. Calprotectin is a well-known biomarker in haematological malignancies[340], however its potential link to TET2 function in HSC remains unknown.

Analysis of protein interactions did not indicate overlapping pathways for S100A8/9 and TET2 (Figure 5.3B), however as previously discussed this may be because of relatively little published information on TET2 interactions. However, the shortest path analyses revelated several downstream proteins potentially connecting S100A8/9 and TET2 (Figure 5.3C,D).

A

Protein complex name	Complex componenets	Stoichiometry	Sources
Calprotectin complex	S100A8, S100A9	6:6	CORUM, ComplexPortal, PDB, SIGNOR
iNOS-S100A8/A9 complex	NOS2, S100A8, S100A9	1:1:1	CORUM, ComplexPortal
S100A8 complex	S100A8	2	ComplexPortal, PDB
S100A9 complex	S100A9	2	ComplexPortal, PDB

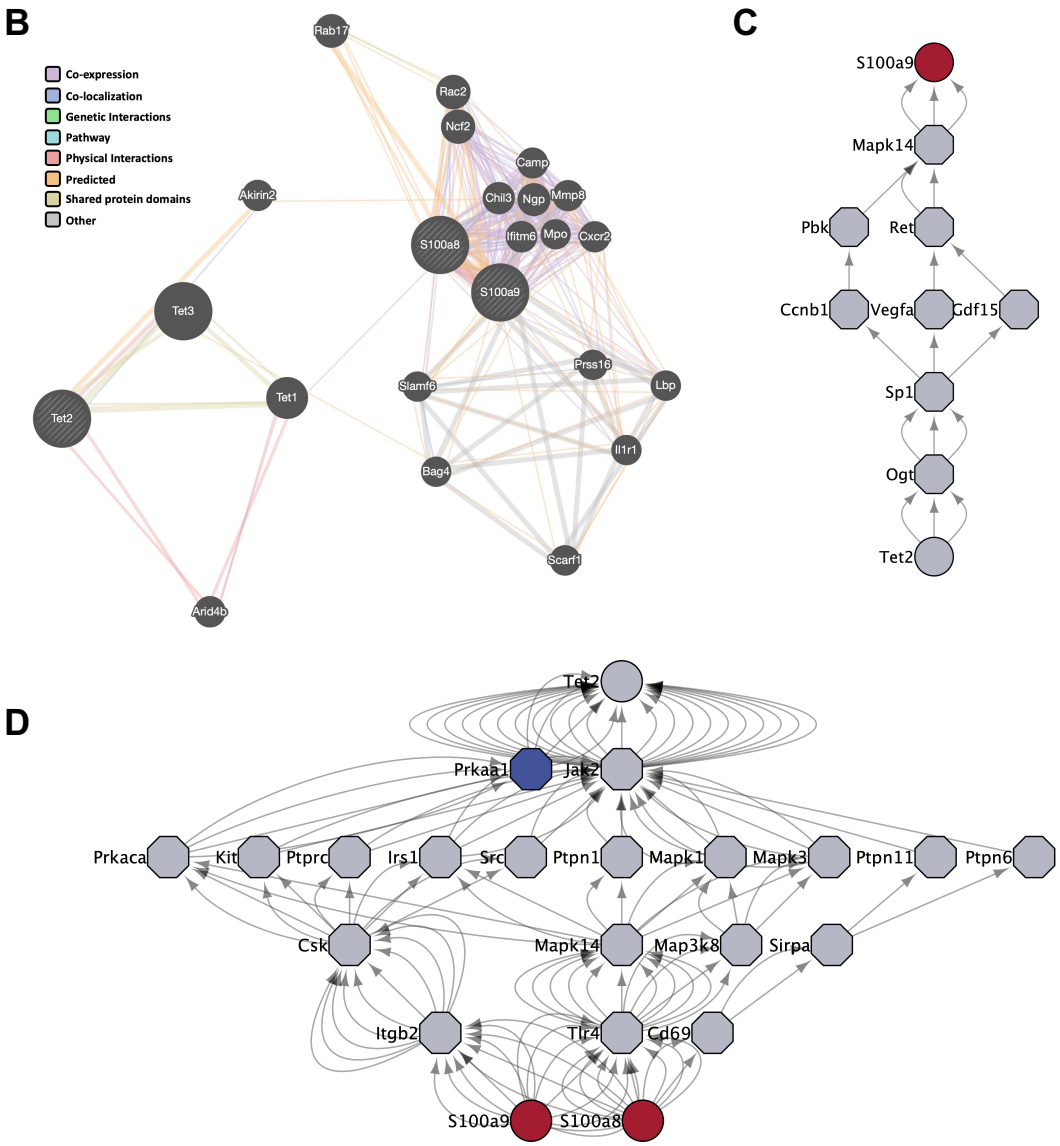


***Figure 5.3 Exploration of calprotectin complex-TET2 connectivity**; A) Table showing all protein complexes containing S100a8 and S100a9 proteins deregulated in Tet2$^{-/-}$ HSC; B) Interaction graph for S100a8, S100a9 and Tet2 and 20 related genes constructed using the Genemania database (https://genemania.org), 23 nodes, 523 links in total, 45% of links were classified as physical interactions; C-D) Shortest path analysis performed usign the Omnipath[261] database, ellipse node shape = Tet2, S100a8 and S100a9, octagon node shape =*

intermediate targets identified with Omnipath, blue/red/grey = genes showing reduced/increased/no change in (respectively) expression in Tet2$^{-/-}$ HSC in comparison to WT HSC (as identified with the proteome analysis), C) Directed interaction graph showing path analysis from Tet2 to S100a9, D) Directed interaction graph showing path analysis from S100a8 and S100a9 to Tet2. Plots made with formattable[284], Genemania (https://genemania.org), Omnipath[261], igraph[258], graphics[259], Cytoscape[253].

To further study potential function of RAGE receptor activity in *Tet2*-deficient HSCs, ESLAM HSCs were isolated from WT and *Tet2*$^{-/-}$ mice using FACS and cultured for 4 weeks *in vitro* following the HSC expansion culture PVA system[275] additionally supplemented by paquinimod (a S100A9 inhibitor) or a recombinant mouse S100A9 protein. Colony sizes were assessed in 2 and 3.5 weeks following the cell culture start and defined as very small/small/medium/large based on the relative differences between wells. In line with the competitive advantage of *Tet2*$^{-/-}$ HSCs over WT HSCs in vivo[2], *Tet2*$^{-/-}$ HSC colonies were in most cases bigger then the WT controls across all conditions (paquinimod, recombinant S100A9 and no treatment control) (Figure 5.4A,B). However, no differences were found between the numbers of *Tet2*$^{-/-}$ and WT HSCs following a treatment with paquinimod or the recombinant S100A9 in comparison to not-treated control cells (Figure 5.4A,B).

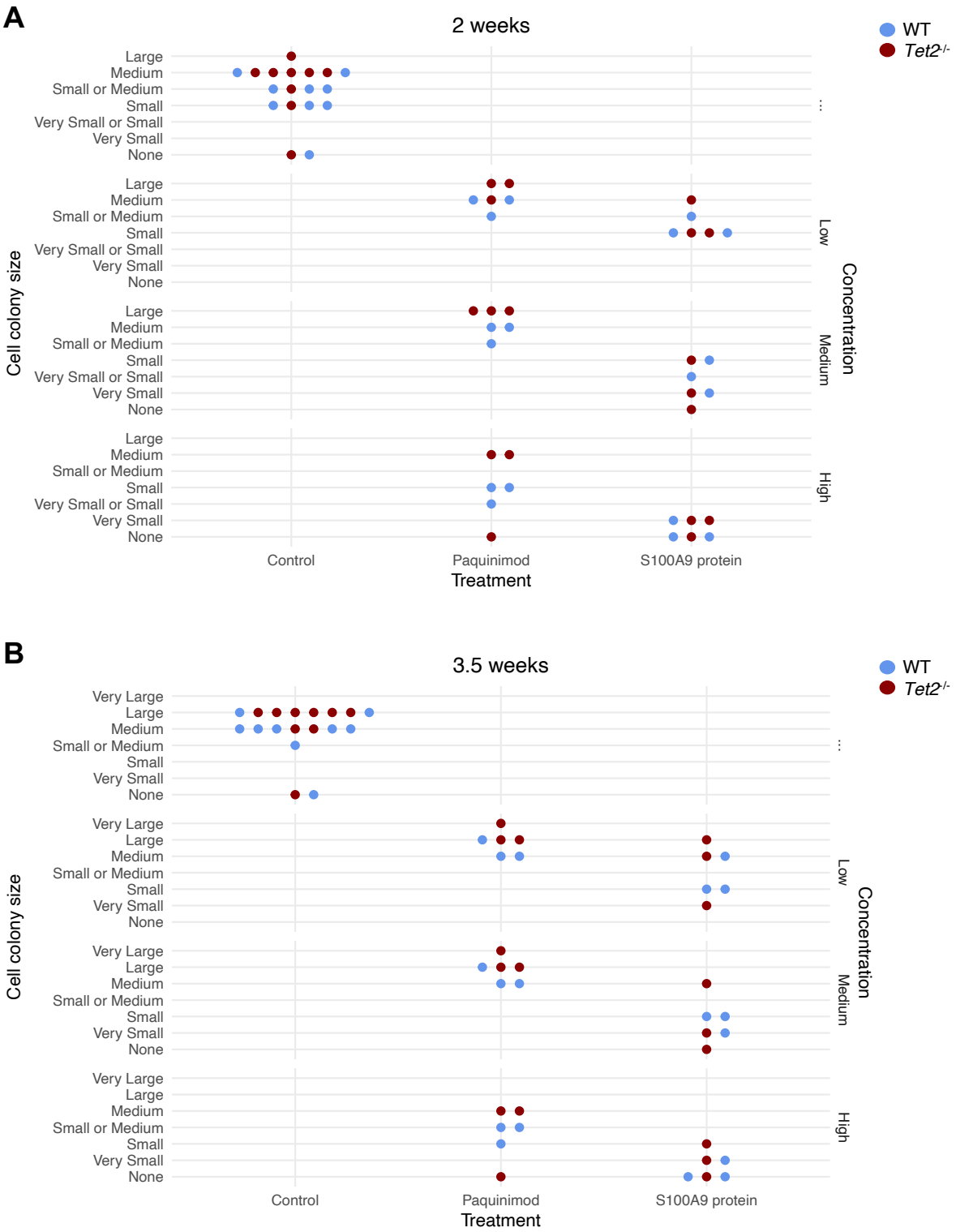


Figure 5.4 HSCs in vitro cell cultures with a S100A9 inhibitor and a recombinant mouse S100A9 protein show no differences between the numbers of Tet2$^{-/-}$ and WT cells; *A) Cell colonies sizes after 2 weeks following the cell culture start; B) Cell colonies sizes after 3.5 weeks following the cell culture start. Plots made with ggplot2*[260].

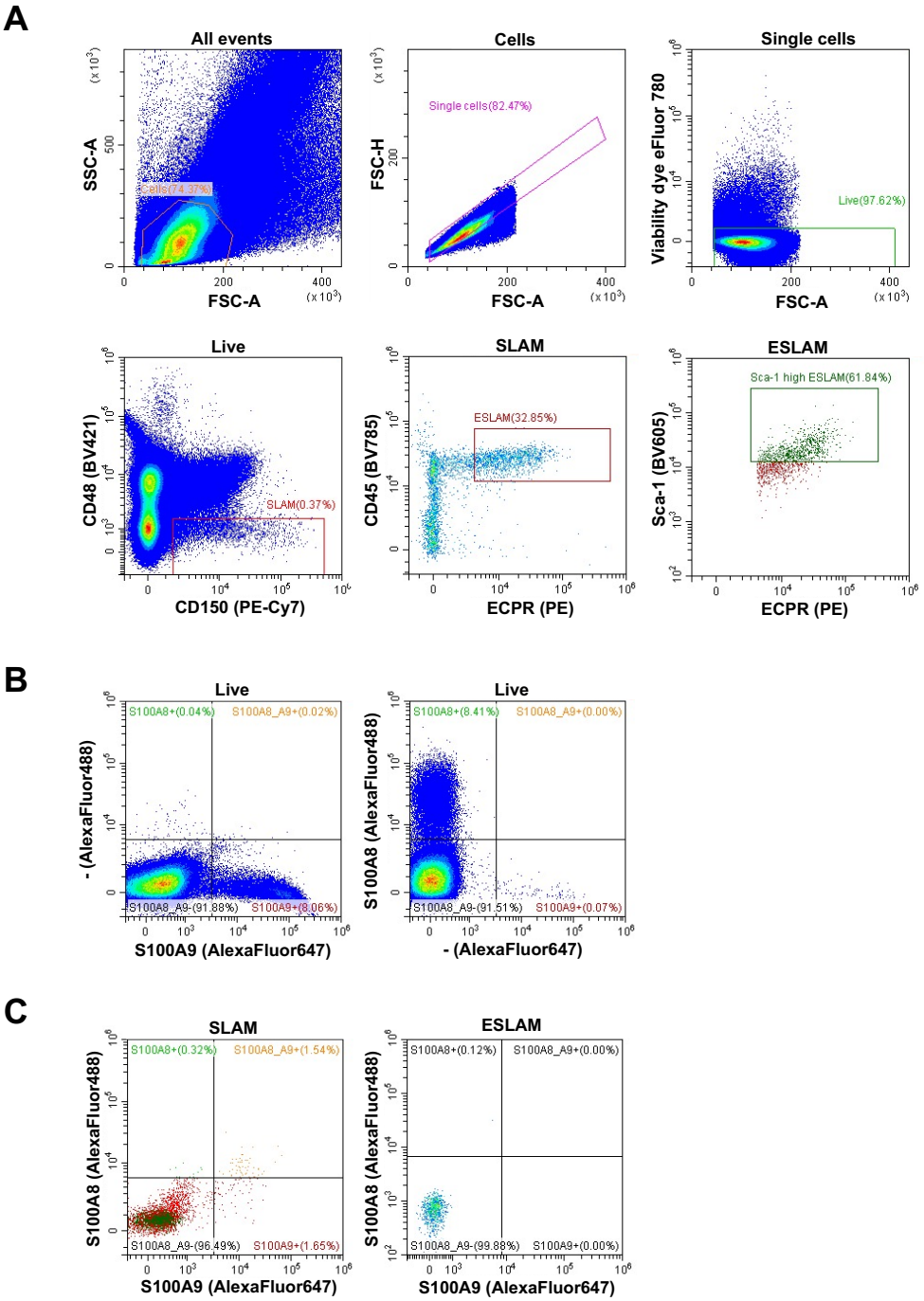


Figure 5.5 Flow cytometry gating strategy for the analysis of S100A8 and S100A9 expression in Tet2$^{-/-}$ and WT HSCs*; A) Gating of ESLAM cells (CD45^{+}EPCR^{+}CD150^{+}CD48^{-}); B) Fluorescence-minus-one (FMO) controls for S100A8 and S100A9; C) Expression of S100A8 and S100A9 on SLAM cell population (CD45^{+} CD150^{+}CD48^{-}) and ESLAM cell population (CD45^{+}EPCR^{+}CD150^{+}CD48^{-}). Plots made with FlowJo*[276].

Finally, S100A8 and S100A9 expression was determined in *Tet2*$^{-/-}$ and WT HSCs (ESLAM phenotype, CD45^{+}EPCR^{+}CD150^{+}CD48^{-}) (Figure 5.5). Flow cytometry analysis showed no S100A8 or S100A9 expression in the HSC population (Figure

5.5A). Analysis of cells with the SLAM phenotype ($CD45^+CD150^+CD48^-$) - cells without expression of EPCR, which is a marker used to enrich HSCs[365], indicated possible difference in the expression of S100A8 between *Tet2*$^{-/-}$ and WT cells (Figure 5.6B). However, it should be noted that this experiment was performed on 2 WT mice and 2 *Tet2*$^{-/-}$ mice therefore more sample replicates would provide a definite answer on a potential change in the S100A8 expression in HSCs upon loss of function of TET2.

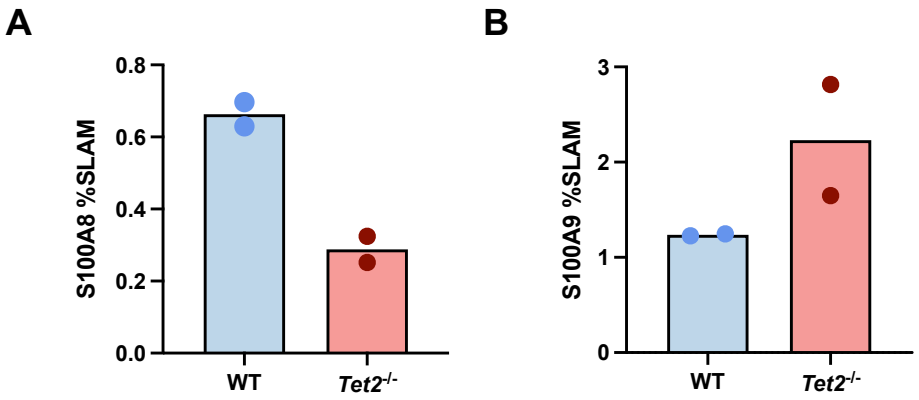


Figure 5.6 Flow cytometry analysis suggests potential difference in the S100A8 expression between Tet2$^{-/-}$ and WT SLAM cells*; A) Percentage of SLAM cells ($CD45^+CD150^+CD48^-$) that express S100A8; B) Percentage of SLAM cells ($CD45^+CD150^+CD48^-$) that express S100A9. Plots made with GraphPad Prism (Boston, Massachusetts USA, www.graphpad.com).*

5.3. Bone marrow extracellular fluid proteome analysis

Next, we aimed to explore extracellular interactions of *Tet2*-mutated HSCs. For this aim we analysed proteome of the bone marrow extracellular fluid (BM ECF), following the protocol described in Jassinskaja et al., Leukemia 2024[244], and proteome of *Tet2*$^{-/-}$ mouse HSC (dataset published in Dr Daniel Bode's PhD thesis[1] and reanalysed by Dr Maria Jassinkaja, ($abs(log_2FC \geq 0.5)$)) (Figure 5.7A). In the HSC proteome dataset three different HSC subpopulations were present: ESLAM cells, $CD45^+CD150^+CD48^-$ cells (here referred to as "$CD150^+$" cells) and $CD45^+CD48^-CD150^-$ cells (here referred to as "$CD150^-$" cells), isolated from *Tet2*$^{-/-}$ and WT mice. In the BM ECF proteome assay *Tet2*$^{-/-}$, *Tet2*$^{+/-}$ and WT samples were analysed (four WT mice, two *Tet2*$^{+/-}$ mice, four *Tet2*$^{-/-}$ mice). PCA[4] analysis showed

that samples seem to cluster according to the genotype, both for the HSC (Figure 5.7B) and BM ECF (Figure 5.7C) analysis. For the downstream analyses, CD150^{+} cells were used from the HSC dataset and further ECF analyses were focused on *Tet2*$^{-/-}$ and WT samples.

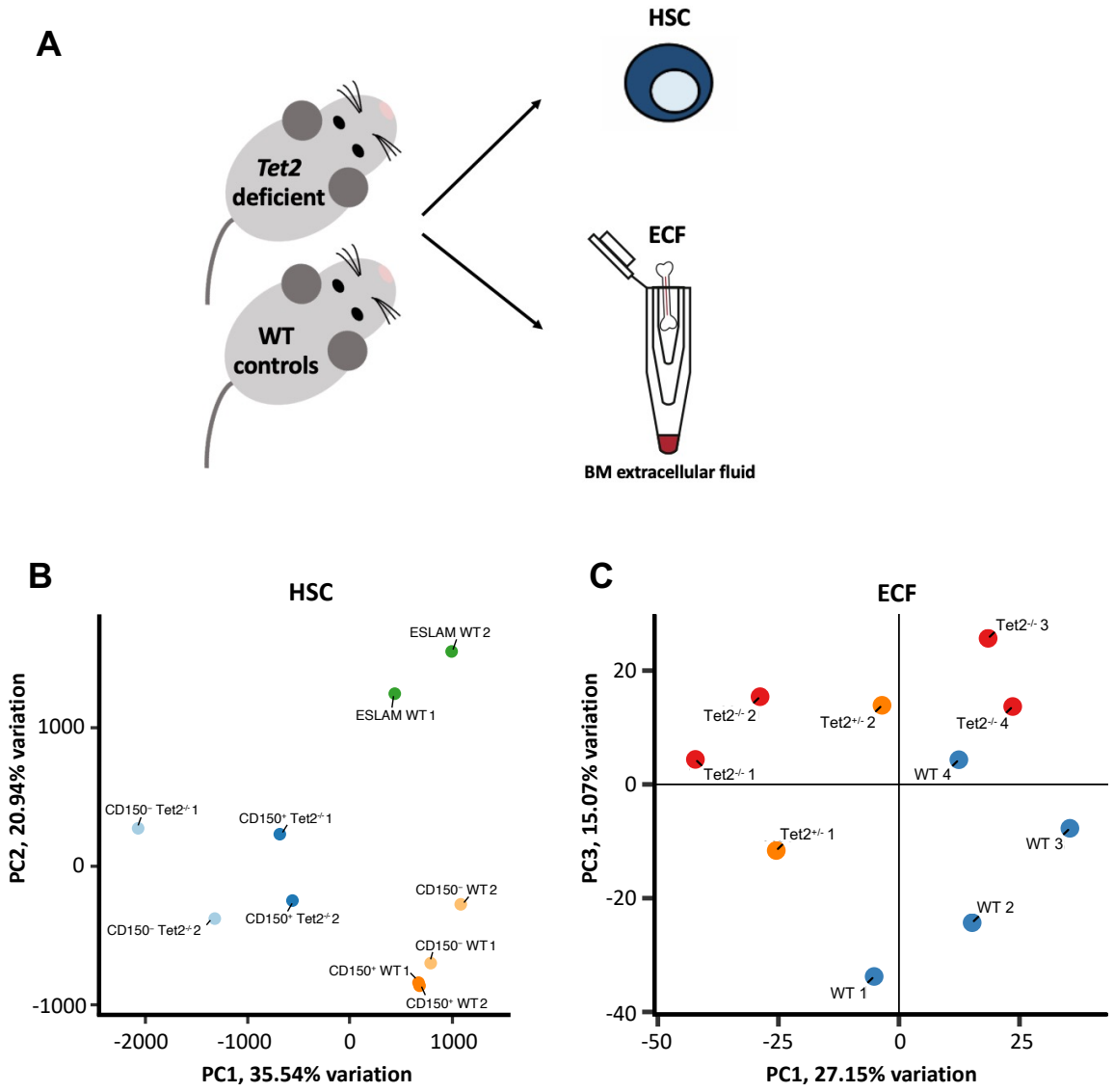


Figure 5.7 An overview of an experimental set up for the downstream proteome analyses*; A) Experimental design, proteome of freshly isolated HSC and BM extracellular fluid (ECF) isolated from Tet2-deficient and wild-type control mice was analysed, these were two independent, unpaired experiments; B) PCA[4] analysis of the HSC dataset; C) PCA[4] analysis of the BM ECF dataset. ESLAM = cells with a CD45^{+}EPCR^{+}CD150^{+}CD48^{-} phenotype, CD150^{+} = cells with a CD45^{+}CD150^{+}CD48^{-} phenotype, CD150^{-} = cells with a CD45^{+}CD150^{-}CD48^{-} phenotype. Plots made with ggplot2[260], stats[366], PCAtools[367].*

Further analysis of the BM ECF proteome detected 7652 proteins, indicating significant differences in protein abundances, with 128 proteins being more abundant

in *Tet2*$^{-/-}$ samples and 73 proteins more abundant in the WT samples (abs(log_2FC ≥ 0.5) & adj p-val < 0.1) (Figure 5.8).

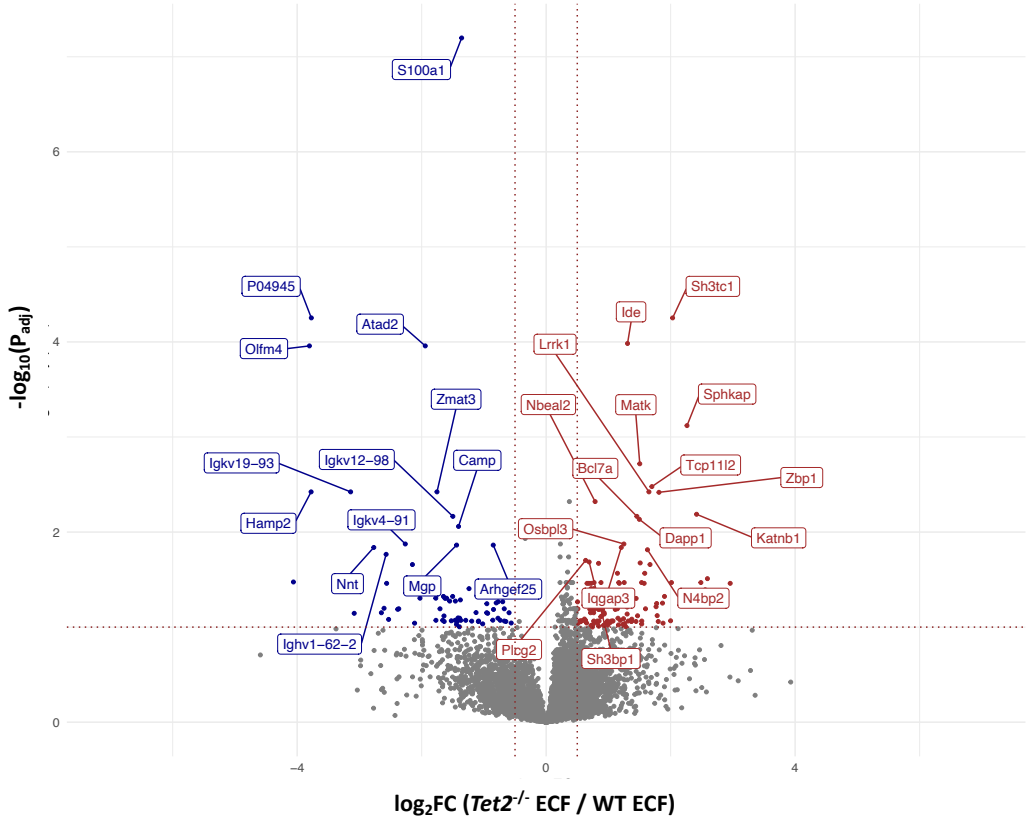


Figure 5.8 Proteins in the bone marrow extracellular fluid are upregulated in Tet2$^{-/-}$ mice in comparison to WT controls; *Volcano plot showing differentially abundant ECF proteins, labelled top 30 DA proteins. Plot made with ggplot2*[260].

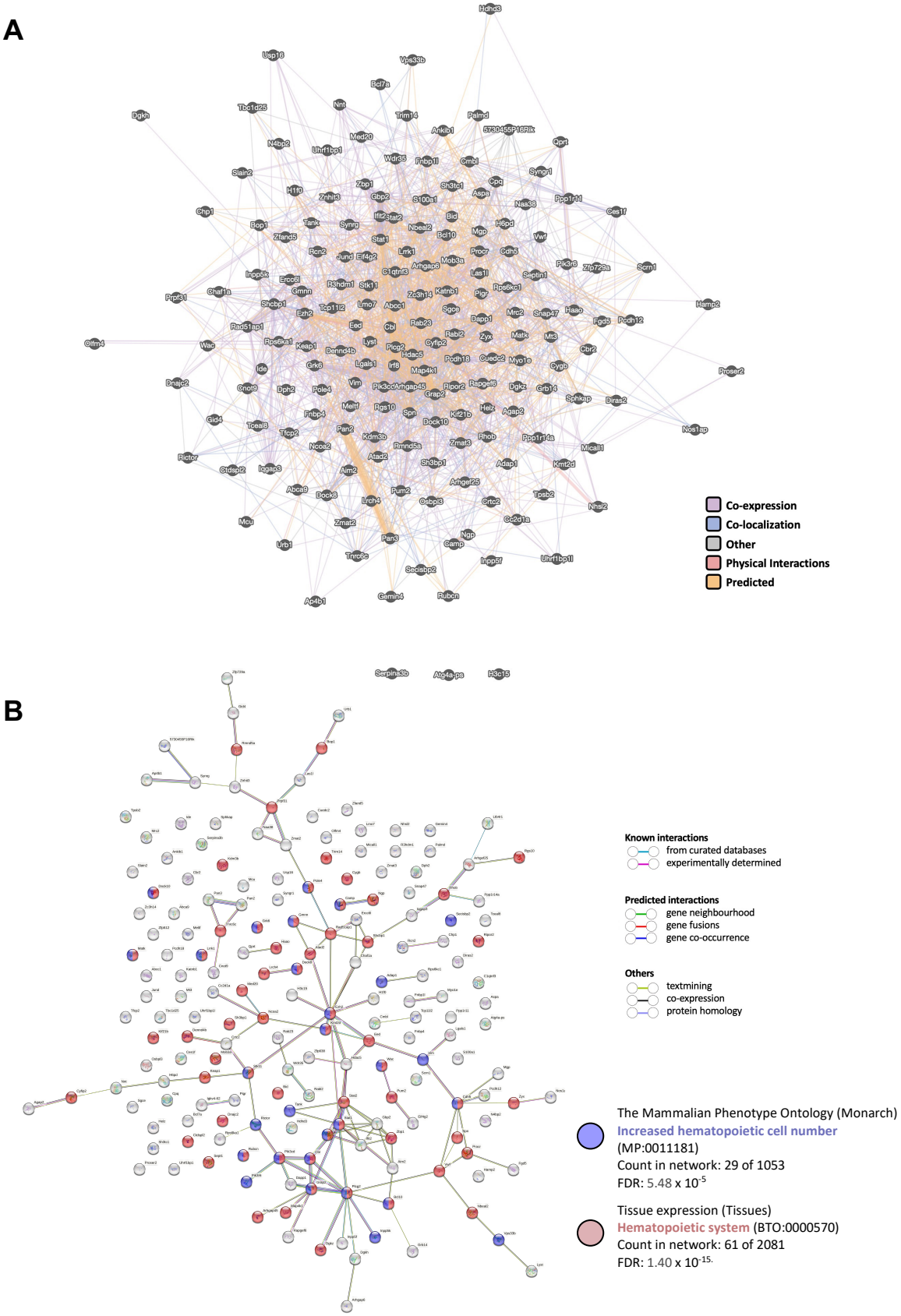


Figure 5.9 Analyses of the bone marrow extracellular fluid proteome of Tet2$^{-/-}$ mice identify dysregulation in abundances of proteins important in the haematopoietic system*; A) Interaction graph for all differentially abundant ECF proteins constructed using the Genemania database (https://genemania.org), 179 nodes, 2260*

links in total, 56% of links were classified as co-expression; B) Static picture of interactions found for all differentially abundant proteins in the String database (https://string-db.org), 183 nodes. Plots made with Genemania (https://genemania.org) and String (https://string-db.org).

Next, we undertook an analysis of protein interactions of all differentially abundant proteins using the Genemania database (https://genemania.org). This analysis showed high connectivity, with most interactions (56% of links) classified as co-expression, which suggested that identified proteins may function in the same regulatory pathways (Figure 5.9A). Further String analysis (https://string-db.org) confirmed that many of detected differentially abundant proteins are expressed in the haematopoietic tissue and are associated with an increased hematopoietic cell number (Figure 5.9B), which is understandable as loss of function mutations in the *Tet2* gene lead to increased HSC self-renewal.

To examine which BM cell subpopulations might secrete identified proteins, I used the Baccin et al.[274] dataset accessed via CellHub (Nygen analytics, database.nygen.io/) which contains a single-cell transcriptome analysis of BM resident cells. Mapping of identified up- and down- regulated proteins to the Baccin et al.[274] dataset suggested that these proteins might be secreted by neutrophils and dendritic cells, respectively (Figure 5.10). As neutrophils mainly secrete inflammatory cytokines this might suggest BM inflammation upon *Tet2* deficiency - a well-known fact in many haematological malignancies.

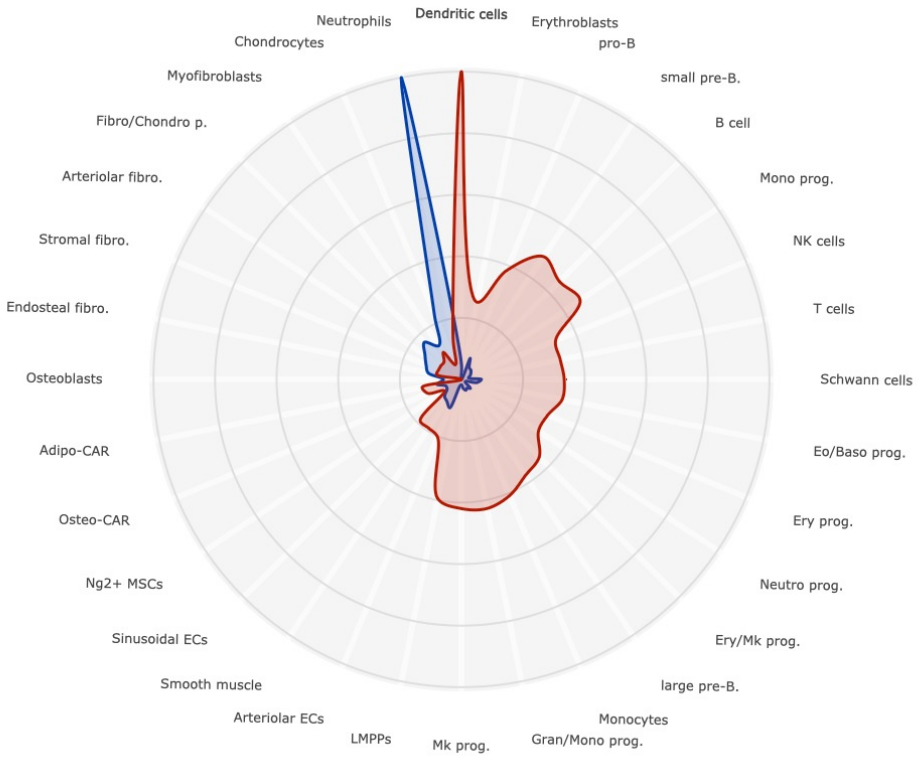


Figure 5.10 Differentially abundant proteins in the extracellular fluid bone marrow of Tet2$^{-/-}$ mice are mainly secreted by neutrofils and dentritic cells; *Radar plot with mapped all differentially abundant proteins in the ECF analysis onto the Baccin et al.[274] dataset. Plot made with Nygen analytics (database.nygen.io/).*

5.4. Ligand-receptor analyses of differentially abundant HSC and ECF proteins

To take advantage of having information about the proteome for both HSC and BM extracellular fluid, I aimed to study ligand-receptor interactions. For this aim I used the Omnipath's[261] ligrecextra database[264] which contains information on ligand-receptor interactions. Firstly, I identified direct interactions of differentially abundant HSC and BM ECF ligands and receptors with ligands and receptors found in the Omnipath's[261] ligrecextra database[264]. Next, I built protein interaction graphs aiming to investigate communication in the *Tet2*-mutated BM niche by looking into interactions of HSC ligands (Figure 5.11A) and receptors (Figure 5.11B) with receptors and ligands, respectively, found in the Omnipath's[261] ligrecextra database[264]. The same analysis was performed for ECF ligands (Figure 5.11C) and receptors (Figure 5.11D).

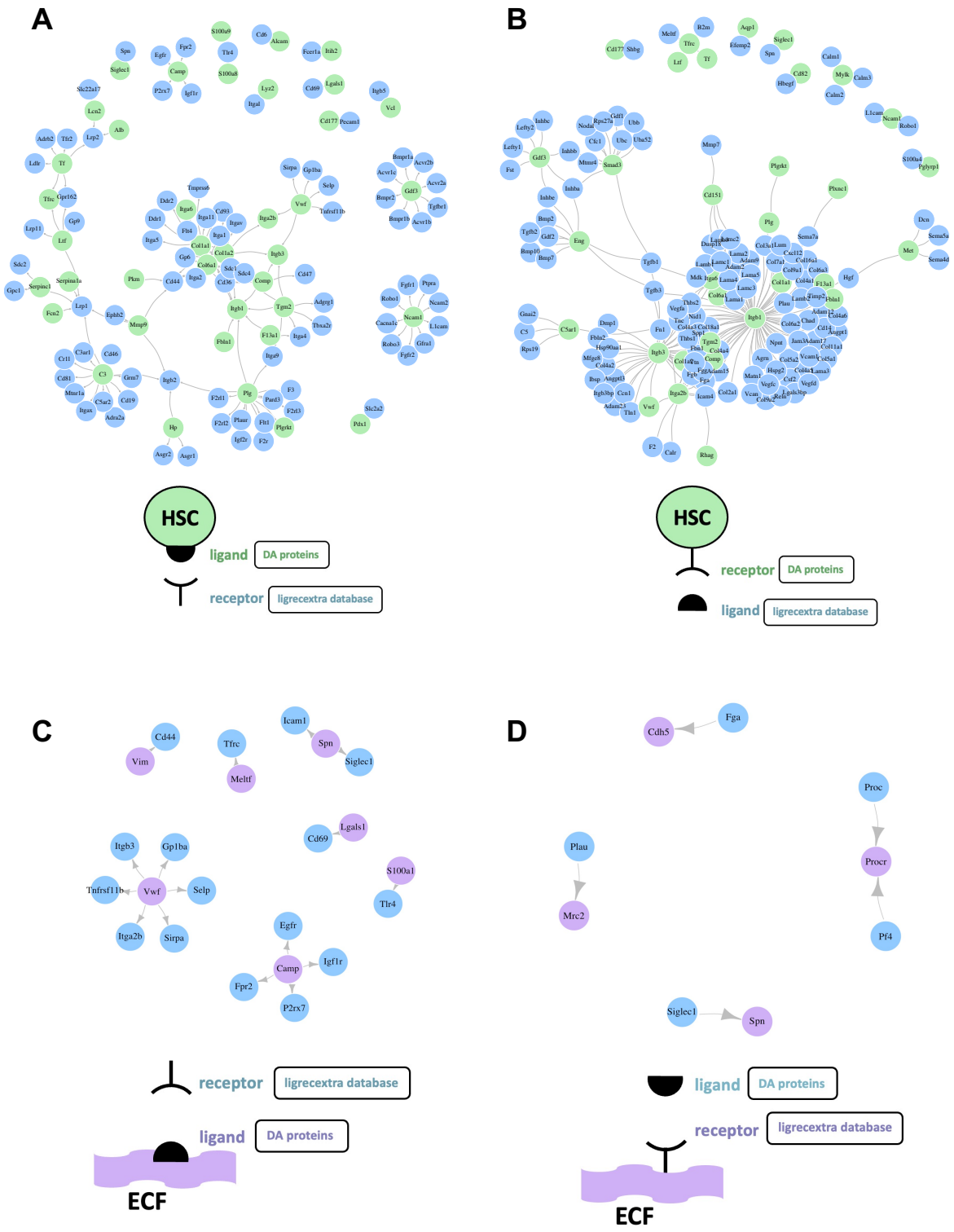
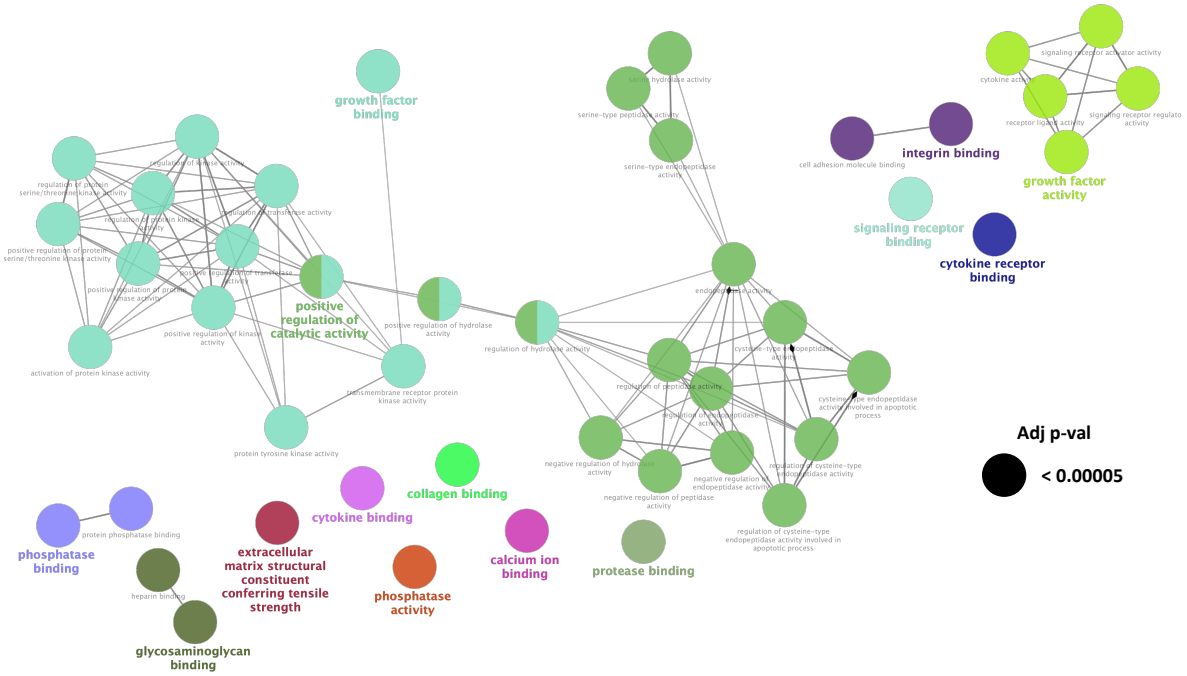


Figure 5.11 Ligand-receptor analyses of differentially abundant HSC and BM ECF proteins identify a set of interactions disturbed in the Tet2-deficient niche*; A-D) Interaction networks built with the ligrecextra database[264] accessed via Omnipath[261], A) ligands identified in HSC proteome analysis (green) and receptors found in the OmniPath[261] database (blue), B) receptors identified in HSC proteome analysis (green) and ligands found in the OmniPath[261] database (blue), C) ligands identified in ECF proteome analysis (violet), receptors found*

in the OmniPath[261] database (blue); D) receptors identified in ECF proteome analysis (violet), ligands found in the OmniPath[261] database (blue). DA = differentially abundant. Plots made with Omnipath[261], igraph[258], graphics[259].

A



B

Growth factor binding		Integrin binding			Cytokine receptor binding
ACVR1B,	ENG*,	ADAM15,	**FN1**,	JAM3,	ADAM17,
ACVR1C,	**FGFR1**,	ADAM17,	GFRA1,	L1CAM,	ANGPT1,
ACVR2A,	FGFR2,	ADAM9,	IBSP,	LAMA3,	CD44,
ACVR2B,	FLT1,	ANGPTL3,	**ICAM1**,	LAMA5,	CDH5*,
AGRN,	FLT4,	CALR,	ICAM4,	LAMB1,	CSF2,
BMPR2,	GPC1,	CCN1,	ITGA1,	LAMB2,	CXCL12,
CCN1,	IGF1R,	CD151*,	ITGA11,	LAMC1,	ENG*,
CD36,	**IGF2R**,	CD177*,	ITGA2*,	MFGE8,	**ITGB3***,
COL1A1*,	**ITGA6***,	CD81,	**ITGA2B**,	NPNT,	LEFTY1,
COL1A2*,	**ITGAV**,	COL16A1,	ITGA4,	**SEMA7A**,	LEFTY2,
COL2A1,	**ITGB3***,	**COL3A1**,	ITGA5,	SPP1,	PF4,
COL3A1,	LRP2,	COL4A3,	**ITGA6***,	THBS1,	SMAD3*,
COL4A1,	**RPS19**,	COMP*,	ITGA9,	TIMP2,	**TGFB1**,
COL5A1,	TGFB3,	CXCL12,	ITGAL,	TLN1,	TGFB2,
COL6A1*,	TGFBR1,	EGFR,	**ITGAV**,	VCAM1,	TGFB3,
EGFR,	THBS1,	**FBLN1***,	ITGAX,	VTN,	TGFBR1,
	VEGFA	**FBN1**,	**ITGB1***,	**VWF***	**VEGFA**,
			ITGB2,		VEGFC,
			ITGB3*,		VEGFD
			ITGB5,		

***Figure 5.12 Analysis of Tet2$^{-/-}$ HSC interactome with the BM extracellular fluid shows expected enrichment in growth factor binding, integrin binding and cytokine receptor binding**; A) Gene Ontology[195]*

analysis performed with ClueGO[254] for all ligands and receptors identified in Figure 5.11A-D graphs, node sizes reflect the statistical significance of the terms, force directed layout presented by the Kappa score, layout adjusted to minimise label overlap (see Methods for more details), Molecular Function, thresholds - min 15 analysed genes (default = 3) / each term, all remaining thresholds were set to the default settings; B) A table showing proteins contributing to the enrichement of growth factor binding, integrin binding and cytokine receptor binding identified in A), a red star marks DA HSC or ECF proteins, all proteins without a red start are ligands/receptors that were found in the ligrecextra database[264] in Figure 5.12, proteins highlighted in orange were already discussed by Dr Daniel Bode in his PhD thesis[1]. Plots made with Cytoscape[253], ClueGO[254].

To get a better understanding of the function of identified ligand-receptor interactions, Gene Ontology[195] analysis was performed. Among the expected to find GO[195] terms were growth factor binding, integrin binding and cytokine receptor binding (Figure 5.12A). Among the proteins contributing to these enrichments, we found well-known targets playing key roles in haematopoiesis, e.g. CXCL12, CALR, CD44, as well as proteins already discussed by Dr Daniel Bode in his PhD thesis, e.g. collagens (COL1A1, COL1A2, COL2A1, COL3A1, COL6A1), fibulin (FBLN1) or von Willebrand factor (VWF) (Figure 5.12B).

To further examine pathways connecting differentially abundant ligands/receptors in ECF to differentially abundant ligands/receptors on HSC, I performed the shortest path analysis. This analysis did not identify any paths from ligand/receptors on HSC to ligands/receptors in ECF, but we did identify pathways in the opposite direction (Figure 5.13). Interestingly, this analysis indicated four direct interactions between ECF-HSC ligand-receptors pairs (Figure 5.14A): Von Willebrand factor (VWF) - integrin alpha-IIb protein (ITGA2B), VWF - beta3 subunit of the integrin alphaIIb/beta3 (αIIbβ3) receptor (ITGB3), melanotransferrin (MELTF) - transferrin receptor (TFRC) and Sialophorin (SPN) – Sialoadhesin (SINGLEC1).

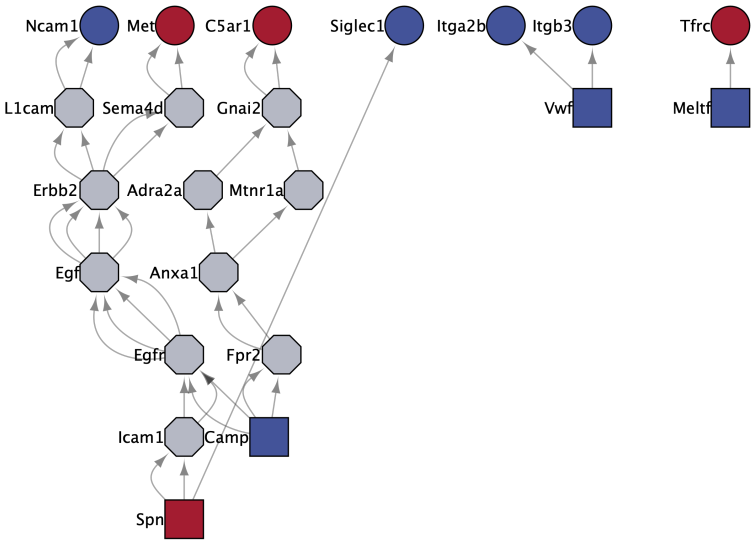


Figure 5.13 Shortest path analysis reveals potentially interesting direct interactions between HSC and ECF DA proteins; *Shortest path analysis performed using the Omnipath[261] database, ellipse node shape = DA proteins in the HSC datasets, square node share = DA proteins in the ECF dataset, octagon node shape = intermediate targets identified with Omnipath, blue/red/grey = genes showing reduced/increased/no change in (respectively) expression in Tet2$^{-/-}$ HSC in comparison to WT HSC (as identified with the proteome analysis), directed interaction graph showing path analysis from DA ECF ligands/receptors to DA HSC ligands receptors. Plots made with Omnipath[261], igraph[258], graphics[259], Cytoscape[253], ClueGO[254]. DA = differentially abundant protein.*

A query of the Genemania database (https://genemania.org) confirmed SIGLEC1-SPN interaction and marked as “predicted” ITGB3-VWF and ITGA2B-VWF interactions (Figure 5.14B). To determine connectivity of these ligands/receptors to TET2, shortest path analyses to and from TET2 were performed with OmniPath[261] (Figure 5.14C-F). These analyses revealed a set of proteins potentially playing a role in extrinsic interactions of *Tet2*-deficient HSCs. Interestingly, the analysis of paths from VWF, SPN and MELTF to Tet2 (Figure 5.14E) indicated many downregulated intermediate nodes, suggesting that these pathways might be potentially important for the HSC function upon *Tet2* mutations.

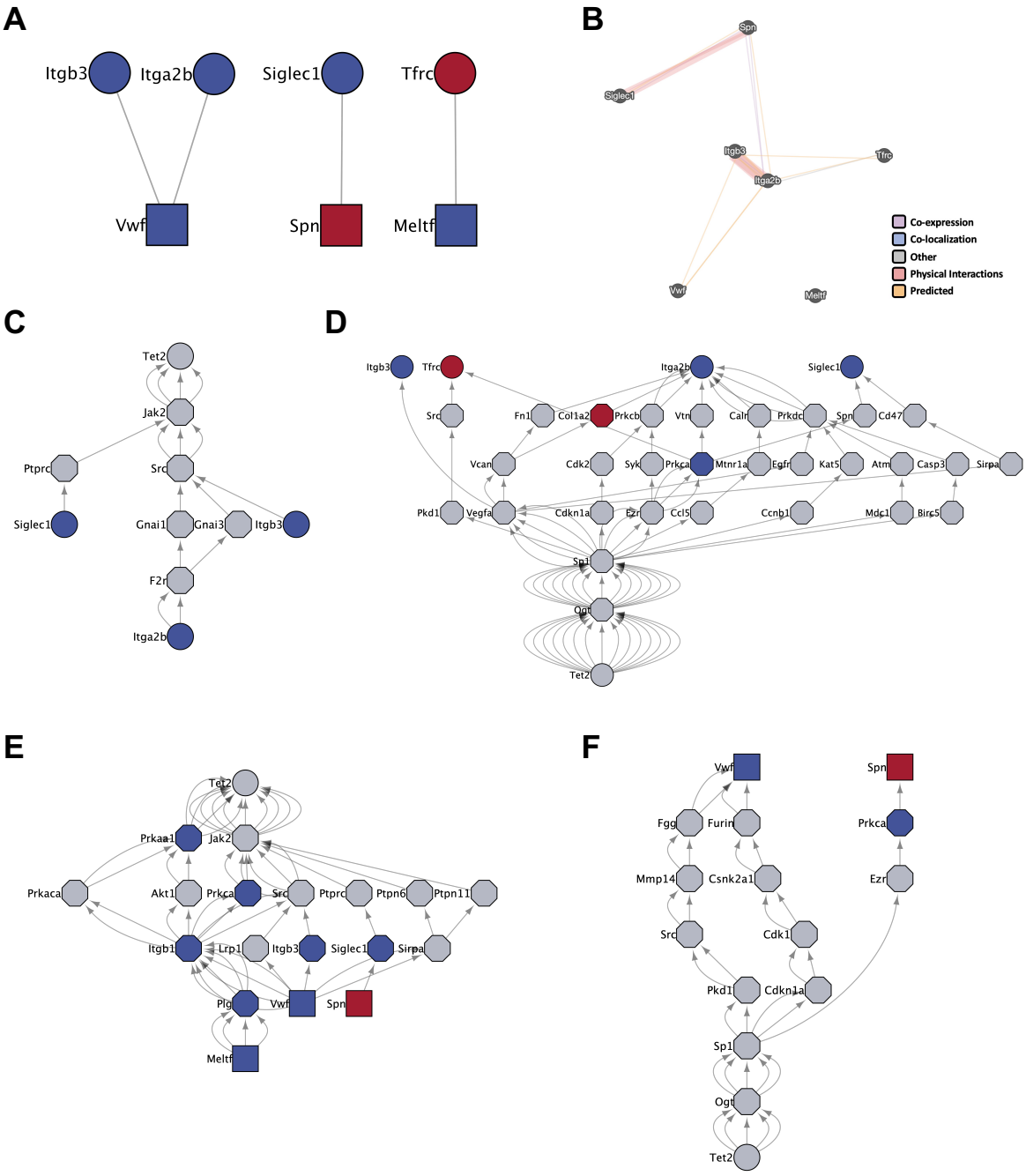


Figure 5.14 Direct ligand-receptor analyses of differentially abundant HSC and ECF proteins reveal a set of interactions potentially disturbed in the BM niche upon loss of function in Tet2*, A) Identified direct ligand-receptor pairs, analysis performed using the ligrecextra database[264] accessed via Omnipath[261]; B) Interaction graph for all ligands/receptors in A), constructed using the Genemania database (https://genemania.org), 7 input proteins, 20 links in total, 82% of links were classified as physical interactions; C-F) Shortest path analysis performed using the Omnipath[261] database, ellipse node shape = DA proteins in the HSC datasets, square node share = DA proteins in the ECF dataset, octagon node shape = intermediate targets identified with Omnipath, blue/red/grey = genes showing reduced/increased/no change in (respectively) expression in Tet2$^{-/-}$ HSC in comparison to WT HSC (as identified with the proteome analysis), C) directed interaction graph showing path analysis from DA HSC ligands/receptors to Tet2, D) directed interaction graph showing path analysis from Tet2 to DA HSC ligands/receptors, E) directed interaction graph showing path*

analysis from DA ECF ligands/receptors to Tet2, F) directed interaction graph showing path analysis from Tet2 to DA ECF ligands/receptors. Plots made with Omnipath[261], Cytoscape[253], Genemania (https://genemania.org).

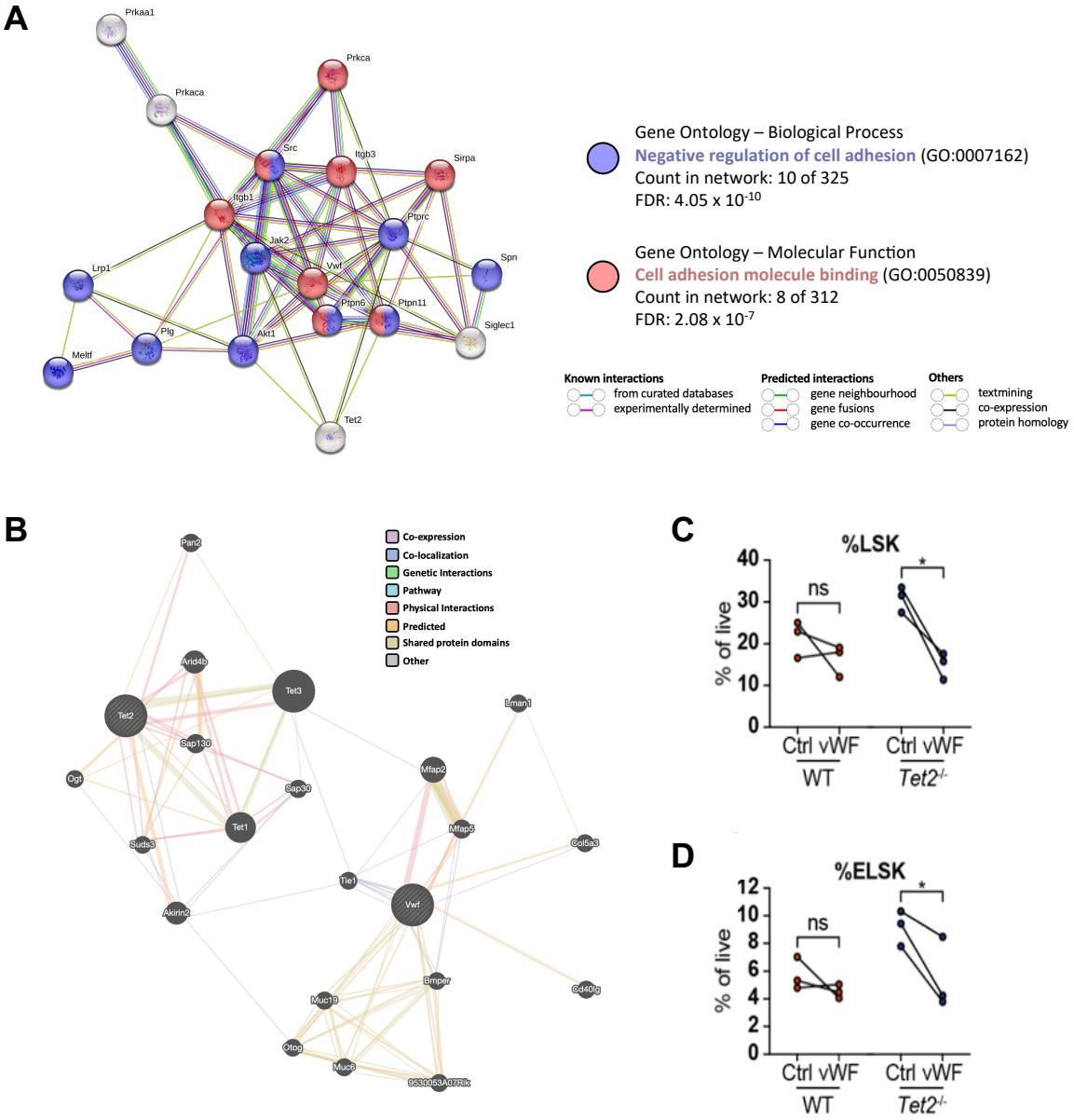


Figure 5.15 VWF and other proteins mediating cell adhesion might regulate Tet2$^{-/-}$ HSC self-renewal. *A) Static picture of interactions found for the nodes identifies in Figure 5.14E in the String database (https://string-db.org), 19 nodes; B) Interaction graph for Tet2 and VWF + 20 related genes constructed using the Genemania database (https://genemania.org), 22 nodes, 124 links in total, 45% of links were classified as physical interactions; C-D) Frequency of C) HSPCs (LSK) and D) HSCs (ELSK) cells following 28-day culture of WT and Tet2$^{-/-}$ ESLAM HSCs on hydrogels functionalized with VWF, two-tailed t-test, p-value < 0.05, ns = non-significant. Plots made with String (https://string-db.org), Genemania (https://genemania.org) and GraphPad Prism (Boston, Massachusetts USA, www.graphpad.com).*

To further investigate paths identified in Figure 5.14E, the String database (https://string-db.org) was surveyed (Figure 5.15A). Analysis indicated strong connectivity of all nodes found in Figure 5.14E. Interestingly, almost all these nodes were proteins involved in the regulation of cell adhesion (Figure 5.15A). Among the nodes with the highest betweenness centrality was Von Willebrand factor (VWF), a protein highlighted by Dr Daniel Bode as potential candidate for HSC self-renewal regulation[1]. VWF-TET2 connectivity analysis with Genemania (https://genemania.org) indicated little VWF-TET2 relatedness, however this might again be explained by relatively little published data on TET2 interactome (Figure 5.15B). Interestingly, ESLAM cell culture on hydrogels functionalised with VWF reduced self-renewal advantage of *Tet2*$^{-/-}$ ESLAM cells, as observed by the reduced frequency of LSK (Figure 5.15C) and ELSK (Figure 5.15D) cell fractions in *Tet2*$^{-/-}$, but not WT ESLAM cell cultures. Altogether, this further validates the hypothesis that VWF might regulate self-renewal capacity of *Tet2*$^{-/-}$ HSCs.

5.5. Discussion

Proteomic analyses enable exploration of extracellular interactions, which is essentially not possible with RNA sequencing. Analyses presented in this chapter provide insight into extrinsic interactions of HSCs with the bone marrow extracellular fluid in the *Tet2*-deficient niche.

Differential analyses of proteome of HSC isolated from *Tet2*-deficient mice indicated upregulation of calprotectin (S100A8 and S100A9 complex) in *Tet2*-deficient samples in comparison to WT control[1]. However, increased expression of S100A9 was not confirmed by flow cytometry analysis of *Tet2*$^{-/-}$ and WT HSC. This is interesting as analyses presented in Chapter 4 showed decreased expression of the RAGE ligands in human HSPC in patients diagnosed with MDS with *TET2* mutations. As noted earlier (see Chapter 4), Wang et al. reported an elevated level of S100A8/A9 in BM plasma in MDS patients, which correlated with poor prognosis for those patients[341]. Altogether, this might suggest that calprotectin expression is not directly regulated by *TET2*. The absence of functional TET2 protein might be not sufficient for its increased expression, and as *TET2* is usually mutated early in the

malignancy development this might imply that calprotectin is only upregulated at the later stages of malignancy progression. Further studies are needed to determine whether calprotectin could be used as a clinical target or whether it should “just” continue to be used as a useful disease marker.

Ligand-receptor analyses of HSC proteome and proteome of the BM ECF pointed to Von Willebrand factor (VWF) as a potential regulator of HSC self-renewal. Von Willebrand factor plays a role in the platelet adhesion following injury and in Factor VIII stabilization in the coagulation pathway and is critical for hemostasis[368], however its function in stem cells remains unknown. Interestingly, presented ligand-receptor analyses identified ITGB3-VWF and ITGA2B-VWF interactions, where VWF showed downregulated abundance in the BM ECF and ITGA2B and ITGB3 downregulated abundance in HSC. Moreover, pathway analyses demonstrated several downregulated intermediate protein nodes connecting TET2 and VWF, suggesting that these pathways might be potentially important for the HSC function upon *Tet2* mutations. Moreover, WT and *Tet2*$^{-/-}$ ESLAM cell culture on hydrogels functionalized with VWF reduced self-renewal advantage of *Tet2*$^{-/-}$ ESLAM cells. This drew our attention as the role of VWF in HSC is poorly understood. In 2013 Sanjuan-Pla et al. showed that VWF marks megakaryocyte-primed HSCs[369,370], moreover, it seems that megakaryocytes control proliferation of the HSC subpopulation expressing VWF[371]. The link between TET2 and VWF in HSCs was initially highlighted by Daniel Bode in his PhD thesis[1], but has not been thoroughly investigated. Presented ligand-receptor analyses of HSCs with the BM ECF, using an experimental protocol adapted from Jassinskaja et al.[244], constitute a novel approach in studying TET2 function in HSCs. The subsequent experimental validation offered a valuable observation given the selective effect in *Tet2*$^{-/-}$ HSCs, with no detected effects in WT HSCs. The observed response was specific to *Tet2*$^{-/-}$ HSCs which suggests a potential clinical relevance in designing therapeutic targets. Further studies of the function of VWF in *Tet2*-deficient HSCs, including spatial analyses of the bone marrow niche, will determine its applicability in clinical settings. Altogether, this suggests that VWF might regulate self-renewal capacity of *Tet2*$^{-/-}$ HSCs.

Studies of HSC communication with the BM niche are pivotal for discovering new therapeutic targets for various haematological malignancies. Moreover, emerging technologies, such as spatial proteome analyses, will provide unprecedented insights into BM niche interactome. Analyses presented in this chapter contribute to painting a picture of HSC extrinsic interactions with the niche. Further studies of identified differentially abundant ligands and receptors will help clarifying their molecular mechanisms in the BM.

6. Discussion

6.1. A summary of presented analyses

Haematological malignancies constitute a heterogenous group of severe diseases, with ~10,000 new cases of leukaemia diagnosed only in the UK each year where approximately half of them lead to death[372]. Understanding HSC biology, where the disease is often suspected to originate, will enable the discovery of new molecular targets to design successful therapies.

Following the discovery of a clonal background of blood cancers, numerous genes were found to be recurrently mutated in malignant haematopoietic cell clones isolated from patients[373]. One such gene is *TET2*, often mutated in haematological malignancies of the myeloid lineage[374]. Since a discovery of *TET2* mutations 16 years ago[81–85] scientists have been trying to determine how *TET2*-deficiency drives malignant cell expansion. Despite extensive research on TET2 function in blood cells, the molecular mechanism governing increased self-renewal of *TET2*-mutated HSC remain elusive. In this thesis I aimed to address this research gap on TET2.

In Chapter 3, I studied regulation of gene expression in mouse HSCs with *TET2*-deficiency using chromatin accessibility and transcriptome data. Next, in Chapter 4 I aimed to explore signalling pathways governing cell expansion in human HSPC isolated from patients diagnosed with MDS with *TET2* mutations. Finally, in Chapter 5 I looked into ligand-receptor interactions of mouse *TET2*-deficient HSC with the BM niche using proteome data. Together, analyses presented in this thesis cover a range of TET2 biology – from chromatin accessibility, transcriptome and proteome studies of HSCs from a *Tet2*-deficient mouse model to RNAseq analyses of HSPCs isolated from patients diagnosed with MDS with *TET2* mutations.

6.2. Gene regulation in *Tet2*-deficient mouse HSCs

In Chapter 3 of my thesis I aimed to study gene regulation in HSCs using a mouse model with deletion of exons 8, 9 and 10 in the catalytic double-stranded β-helix domain of the *Tet2* gene[2]. I have applied two alternative methods to integrate single-cell ATAC sequencing and single-cell RNA sequencing datasets – a closest gene analysis and a transcription factor regulatory network. Although gene expression and chromatin accessibility of *Tet2*-deficient HSC have been studied before[128,134,146], presented analyses contribute to our understanding on TET2 biology by providing additional resource on mouse *Tet2*$^{-/-}$ HSC gene expression state.

There are several limitations of the analyses presented in this chapter. Some of them have already been addressed in the discussion section of Chapter 3, such as e.g., limited information on transcription factors available in the online databases which hindered drawing conclusions from the regulatory network analysis. Naturally, that also limits conclusions from the performed String (https://string-db.org), Genemania (https://genemania.org) and OmniPath[261] analyses. Another concern is the use of genomic distance between open chromatin peaks and the gene of interest in the closest gene integrative analyses to study gene regulation. This assumption, made for many analyses in the field nowadays, is a broad generalization, as the closest accessible chromatin region does not necessary regulate expression of the gene of interest[225,375]. A potential solution would be to study the genome structure with Hi-C sequencing[225].

Importantly, the choice of integrative approaches to consolidate information on the chromatin accessibility and transcriptome should be discussed. Both chromatin accessibility and transcriptome datasets were generated at the single cell level, however the presented integrative analyses were performed on a pseudo-bulk level. A decision to not use single cell based integrative tools was mainly dictated by the fact that scATACseq and scRNAseq data were unpaired and so information did not come from the same cell. Moreover, scATACseq was generated using the 10xGenomics technology and scRNAseq was plate-based, resulting in substantial differences in the number of analysed cells in each assay. There are tools that

enable integration from separate cells[228] however as transcription is time-sensitive in any case, analyses of these datasets needed to be imputed nonetheless, therefore a more supervised approach was chosen to get more reliable insights into TET2-deficient HSC biology and search for dysregulated targets.

Altogether, these analyses pointed to a set of targets potentially mediating disturbed HSC behaviour upon the loss of function of TET2, such as *Dusp1*, *Egr1*, *Nr4a1*, *Rgs1* and *Smarcc1,* among others. Despite the fact that none of the genes selected for in the RT-qPCR assay were significantly differentially expressed between *Tet2*$^{-/-}$, *Tet2*$^{+/-}$ and WT HSCs, collectively, results presented in this chapter constitute a stockpile of information available to support further studies.

6.3. Molecular repercussions of *TET2* mutations in HSPCs in patients diagnosed with MDS

In Chapter 4 of my thesis, I aimed to examine molecular states of human *TET2*-deficient haematopoietic cells. We studied three cell subpopulations: HSPC-enriched fraction (CD34^{+}), STRC fraction (CD34^{+}CD38^{+}) and the HSPC-fraction (CD34^{+}CD38^{-}). These analyses revealed different transcriptional profiles of each cell subpopulation, with HSPCs displaying the biggest changes in terms of numbers of up- and downregulated genes. Moreover, the presented studies suggested several genes potentially important in *TET2*-mutated MDS development, such as RAGE ligands and *Mmrn1*.

The main limitations hindering the presented analyses is the fact that all samples within the comparison groups belonged to the same sex. Unfortunately, this implies that differences between groups could be potentially convoluted by sex differences. Another limitation is the number of analysed samples, preferably, more than 3 samples / condition should be studied.

The control samples used for comparative analyses in this chapter should also be discussed. Healthy control samples were used for this aim - these were three healthy commercially obtained BM samples (Lonza human BM, cat # 2M-125C).

Optimal controls in case of studying transcriptome of haematopoietic cells isolated from patients with *TET2* mutations would be cells isolated from MDS patients without *TET2* mutations. However, this comes with several complications due to patient heterogeneity in terms of molecular mechanisms of disease development. As disease might progress differently in different patients, comparisons of patients diagnosed with and without TET2 mutations might lead to incorrect conclusions if the remaining mutational profile were different. Moreover, other mutations might disrupt similar pathways as loss of function TET2 further complicating any comparisons. The best control would be normal, non-mutated cells isolated from the same patient, this however comes with additional difficulty of not being able to prospectively identify which haematopoietic cells carry the mutation prior to sequencing. Nevertheless, examination of gene expression in patients diagnosed with MDS with similar mutational profile, but without TET2 mutations might have provided more insights.

Collectively, the presented results provide a valuable resource for further studies. Performed analyses suggest potentially interesting targets (*Mmrn1* and RAGE ligands among others) possibly deregulated in HSPCs in patients diagnosed with MDS with *TET2*-mutations. Further studies could determine their role in the MDS progression and connectivity to TET2 function.

6.4. Disturbed signalling between *Tet2*-deficient HSCs and their bone marrow niche

In Chapter 5 of my thesis, I firstly investigated how targets deregulated in HSPCs isolated from patients diagnosed with *TET2*-mutated MDS identified in Chapter 4 map onto proteomics dataset of *Tet2*$^{-/-}$ mouse HSPC[1]. Next wet-lab experiments examining calprotectin function in HSC were performed. Finally, I looked into HSC interactions with the BM ECF and performed ligand-receptor analyses.

The mapping of RNA onto proteomic data was limited by several assumptions, the main one being that all RNA molecules will form functional proteins, which naturally is not the case. However, analyses of an overlap between

these two modalities might help understanding which of the RNA molecules end up forming proteins in the *Tet2*-deficient environment. This is of course limited due to an integration across species and by the presence of other mutations in HSPCs isolated from patients diagnosed with MDS, which might cause differences in cell behaviour. Nevertheless, these analyses might help to shed light on potentially important proteins playing a role in the MDS progression.

A limitation of the presented ligand-receptor analyses between HSC and BM ECF is that despite careful preparation of the extracellular fluid samples they might contain proteins that came from cells degraded during samples preparation. Moreover, the ligand-receptor analyses themselves might provide many false positive interactions due to limited information present in ligand-receptor databases. However, these analyses still provide a useful screen for potentially important protein interactions disturbed upon *Tet2*-deficiency which can help designing wet-lab validation experiments to test discovered hypotheses.

Altogether analyses of differentially abundant proteins in the BM extracellular fluid in *Tet2*$^{-/-}$ mice identified a set of proteins that might be contributing to the disturbed equilibrium in *Tet2*-deficient BM environment.

6.5. Current understanding of HSC biology following *TET2* mutations

It is a rapidly advancing time for research on TET2 role in haematopoietic cells. Following the discovery of *TET2* mutations 16 years ago, the collective work of researchers across the globe identified various TET2 partner proteins[128] and genes commonly co-mutated with TET2 in haematological malignancies[122], examined TET2 enzymatic independent activity[156] and TET2 function on RNA[146,242] and discovered hypermethylation of enhancers in haematopoietic cells following loss of TET2 function[70,128,132] - and these are only a few of areas of active TET2 research.

In general, the results presented in this thesis align with reported deregulation of HSC behaviour upon loss of TET2 function. A field of ongoing investigation is the exploration of DNA methylation, chromatin accessibility and transcriptome dynamics

in haematopoietic cells following TET2 mutations[70,125,128,132,135,160,376]. scATACseq and scRNAseq analyses described in Chapter 3 indicated a more active state of *Tet2*$^{-/-}$ HSCs compared to WT HSCs, as shown by a global increase in chromatin accessibility and gene expression. As discussed in Chapter 3, this is in agreement with a study by Zou et al.[146] published last year, which suggested that loss of TET2 function leads to globally decreased ubiquitination of histone H2A and an increase in chromatin accessibility and gene expression[146].

Other area of active research is the role of inflammation in haematological malignancies[152,154,155,377–379]. The analysis of transcriptome of HSPC isolated from patients diagnosed with MDS with TET2 mutations described in Chapter 4 suggested deregulated inflammatory pathways which is in line with reported role of TET2 in the regulation of inflammatory response in myeloid cells[87,156,324,325]. One of the proteins secreted by immune cells during inflammation and examined in this thesis is calprotectin – a complex formed by S100A8 and S100A9 proteins and a well-known biomarker in haematological malignancies[340]. Flow cytometry analysis in Chapter 5 suggested that mouse HSCs do not express S100A8 or S100A9 and no differences in numbers of *Tet2*$^{-/-}$ and WT HSCs were found following a treatment with recombinant S100A9 or paquinimod (a S100A9 inhibitor) in comparison to not-treated control cells. These *in vitro* experiments did not confirm bioinformatic analyses discussed in Chapters 4 and 5, however, it should be noted that addition of more biological replicates to the experimental design might have provided more comprehensive results on the potential link of calprotectin and TET2.

Another active field of research are studies of HSC microenvironment during malignancy development[49,164,165,360,380]. Recent advancements in spatial proteome analyses[211] open new doors for our understanding of the haematopoietic stem cell niche. When combined with single-cell proteomic[210,212] they might explain molecular mechanisms governing deregulated expression of targets identified in this thesis, such as MMRN1 and RAGE ligands among others. Hopefully, these cutting-edge technologies will help deciphering more protein interactions in *TET2*-deficient HSCs and therefore enable building more comprehensive protein signalling networks which would allow generation of novel hypotheses[226].

6.6. Future directions

Collectively, the analyses presented in this thesis provide insights into molecular mechanisms governing cell behaviour upon *Tet2*-deficiecy at chromatin, RNA and proteome levels across mice and humans. Several improvements could have been implemented to the experimental design. Firstly, we would measure chromatin accessibility and gene expression levels from the same cell, for example using SNARE-seq[381] or 10x Genomics multiomics[382,383], and not as a set of two independent experiments. Moreover, we would test all overlapping targets in the qPCR assay, not only selected ten. In Chapter 4, we would perform sequencing of more samples, not just three per group, and balance male/female samples equally across *TET2*-mutated / healthy conditions. In Chapter 5, more targets would have been tested *in vitro* to look for potential candidates regulating HSC fate following *Tet2* deficiency. Moreover, adding information about localisation of HSC in the BM niche with spatial proteomics assays would help decipher roles of distinct cell subpopulations in the increased self-renewal of *Tet2*-mutated cells. Furthermore, future studies building up on the analyses presented in this thesis should consist of functional *in vitro* and *in vivo* experiments aiming to validate roles of identified targets in *Tet2*-deficient HSCs.

Altogether, performed analyses provide insights into pathways and targets potentially playing a role in the regulation of HSC self-renewal upon *Tet2* deficiency. I hope that presented results will constitute a valuable resource for further studies on mechanisms of expansion of mutated cell clones in haematological malignancies and contribute to developing novel targeted therapies and helping patients.

Abbreviations

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxymethyluracil
5mC	5-methylcytosine
α-KG	α-ketoglutarate
abs(log_2FC)	Absolute value of the log base 2 fold change
AGPS	Alkylglycerone Phosphate Synthase
AI	Artificial Intelligence
AML	Acute Myeloid Leukaemia
ATACseq	Assay for Transposase-Accessible Chromatin
BER	Base Excision Repair
BM	Bone Marrow
BM ECF	Bone Marrow Extracellular Fluid
BSF	Biological Service Facility
CALM1	Calmodulin-1
CB	Cord Blood
CBFα	Core-Binding Factor α
CBFβ	Core-Binding Factor β
CDH1	Cadherin-1
CEBPα	CCAAT-enhancer Binding Protein α
CHIP, CH	Clonal Haematopoiesis (CH) of Indeterminate Potential
CLP(s)	Common Lymphoid Progenitor(s)
CMML	Chronic Myelomonocytic Leukaemia
CMP(s)	Common Myeloid Progenitor(s)
COL	Collagens
CORO7	Coronin-7
CR	Cysteine-Rich domain
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
CUT&Tag	Cleavage Under Targets and Tagmentation

ChIPseq	Chromatin Immunoprecipitation sequencing
DE	Differential Expression analysis
DEG	Differentially Expressed Genes
distToTSS	Distance to the Transcription Start Site
DNMT	DNA Methyltransferases
DNMT3A	DNA Methyltransferase 3A
DSBH	Double-Stranded β-Helix domain
ECF	Extracellular Fluid
ECM	Extracellular Matrix
EDTA	Ethylenediaminetetraacetic Acid
ELSK	$EPCR^{+}Lin^{-}Sca1^{+}cKit^{+}$
EPDR1	Ependymin-Related Protein 1
ESLAM	$CD45^{+}EPCR^{+}CD150^{+}CD48^{-}$
FACS	Fluorescence Activated Cell Sorting
FBLN1	Fibulin
FBS	Fetal Bovine Serum
FDR	False Discovery Rate
FLT3	FMS-like Tyrosine kinase 3
FLT3-ITD	FLT3 Internal Tandem Duplication mutations
FMO	Fluorescence Minus One
GLCNAC	N-acetylglucosamine
GO	Gene Ontology[195]
GSEA	Gene Set Enrichment Analysis[196,197]
HAT	Histone Acetyltransferases
HDAC	Histone Deacetylases
HIF1α	Hypoxia-Inducible Factor 1α
HMT	Histone Methyltransferases
HSC(s)	Haematopoietic Stem Cell(s)
HSPC(s)	Haematopoietic Stem and Progenitor Cell(s)
HuRi	The Human Reference Interactome[273]
IT-HSC(s)	Intermediate-term HSC(s)
ITGA2B	Integrin alpha-2b
ITGB3	Integrin beta-3

JAK2	Janus kinase 2
KDM	Histone Demethylases
KO	Knock Out
LCI	Low Complexity Insert
LC-MS	Liquid chromatography and mass spectrometry
lncRNAs	Long Noncoding RNAs
LSI	Latent Semantic Indexing
LSK	$Lin^{-}Sca\text{-}1^{+}cKit^{+}$
LT-HSC(s)	Long-Term Haematopoietic Stem Cell(s)
m6A	N6-methyladenosine
MDS	Myelodysplastic Syndrome
MELTF	Melanotransferrin
MFI	Mean Fluorescence Intensity
MPN	Myeloproliferative Neoplasm
MPP(s)	Multipotent Progenitor(s)
MS	Mass Spectrometry
MSigDB	Molecular Signatures Database[255,256]
Mmrn1, MMRN1	Multimerin-1
ncRNA(s)	Noncoding RNA(s)
NPM1	Nucleophosmin
OGT	O-linked N-acetylglucosamine (GLCNAC) Transferase
PBS	Phosphate-Buffered Saline
PCA	Principal Component Analysis[4]
PEF	PBS, 2% FBS, 2 mM EDTA
PRKAA1	Protein Kinase AMP-activated catalytic subunit alpha 1
Pdzd2	PDZ domain-containing protein 2
QC	Quality Control
rAAV6	Recombinant Adeno-Associated Virus 6
RNAseq	RNA sequencing
ROS	Reactive Oxygen Species
RPKM	Reads Per Kilobase per Million mapped reads
RT-qPCR	Reverse Transcription quantitative Polymerase Chain Reaction
scATACseq	Single-cell Assay for Transposase-Accessible Chromatin

scRNAseq	Single-cell RNA sequencing
SINGLEC1	Sialoadhesin
SLC14A1	Solute Carrier family 14 member 1
SPN	Sialophorin
ST-HSC(s)	Short-Term Haematopoietic Stem Cell(s)
SVD	Singular Value Decomposition
TDG	Thymine DNA Glycosylase
TET1	Ten-Eleven Translocation 1
TET2, Tet2	Ten-Eleven Translocation 2
TET3	Ten-Eleven Translocation 3
TF(s)	Transcription Factor(s)
TF-IDF	Term Frequency-Inverse Document Frequency
TFRC	Transferrin Receptor
TKD	Tyrosine Kinase Domain
TNFα	Tumor Necrosis Factor α
TPM	Transcripts Per Million
TSS	Transcription Start Site
UMAP	Uniform Manifold Approximation and Projection[186]
UMI	Unique Molecular Identifiers
VAF	Variant Allele Frequency
VWF	Von Willebrand Factor
WNN	Weighted Nearest Neighbour
WT	Wild Type
WT1	Wilms Tumor 1

References

1. Bode Daniel. Identification of novel drivers of haematopoietic stem cell fate using low cell number proteomics and single cell profiling, University of Cambridge, published at: https://www.repository.cam.ac.uk/items/a0ac8d47-ec68-4794-ab55-916bdbc11673. (2021).
2. Ko, M. *et al.* Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *PNAS* **108**, 14566–14571 (2011).
3. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171–181 (2014).
4. PCA, developed by Pearson (1901) and Hotelling (1933). PCA, Principal Component Analysis.
5. Lintel 24, a bloodletting ritual, British Museum. https://www.britishmuseum.org/collection/object/E_Am1923-Maud-4.
6. Mystery cults, bloody sacrifices and purification. *National Archaeological Museum* https://www.namuseum.gr/en/monthly_artefact/mystery-cults-bloody-sacrifices-and-purification/.
7. Blood | Science Museum. https://www.sciencemuseum.org.uk/objects-and-stories/medicine/blood.
8. Jacobson, L. O., Simmons, E. L., Marks, E. K. & Eldredge, J. H. Recovery from radiation injury. *Science* **113**, 510–511 (1951).
9. Till, J. E. & McCULLOCH, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res* **14**, 213–222 (1961).

10. Olson, O. C., Kang, Y.-A. & Passegué, E. Normal Hematopoiesis Is a Balancing Act of Self-Renewal and Regeneration. *Cold Spring Harb Perspect Med* **10**, a035519 (2020).
11. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: A Human Perspective. *Cell Stem Cell* **10**, 120–136 (2012).
12. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241**, 58–62 (1988).
13. Purton, L. E. Adult murine hematopoietic stem cells and progenitors: an update on their identities, functions, and assays. *Experimental Hematology* **116**, 1–14 (2022).
14. Rix, B., Maduro, A. H., Bridge, K. S. & Grey, W. Markers for human haematopoietic stem cells: The disconnect between an identification marker and its function. *Front. Physiol.* **13**, (2022).
15. Kent, D. G. *et al.* Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. *Blood* **113**, 6342–6350 (2009).
16. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218–221 (2011).
17. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).
18. Kucinski, I. *et al.* A time and single-cell resolved model of murine bone marrow hematopoiesis. *Cell Stem Cell* **31**, 244-259.e10 (2024).

19. Vaziri, H. *et al.* Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. *Proc Natl Acad Sci U S A* **91**, 9857–9860 (1994).
20. Bryder, D., Rossi, D. J. & Weissman, I. L. Hematopoietic Stem Cells: The Paradigmatic Tissue-Specific Stem Cell. *The American Journal of Pathology* **169**, 338–346 (2006).
21. Suda, J., Suda, T. & Ogawa, M. Analysis of Differentiation of Mouse Hemopoietic Stem Cells in Culture by Sequential Replating of Paired Progenitors. *Blood* **64**, 393–399 (1984).
22. Bernitz, J. M., Kim, H. S., MacArthur, B., Sieburg, H. & Moore, K. Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. *Cell* **167**, 1296-1309.e10 (2016).
23. Foudi, A. *et al.* Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nat Biotechnol* **27**, 84–90 (2009).
24. Wilson, A. *et al.* Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell* **135**, 1118–1129 (2008).
25. Reya, T. *et al.* A role for Wnt signalling in self-renewal of haematopoietic stem cells. *Nature* **423**, 409–414 (2003).
26. Nemeth, M. J. & Bodine, D. M. Regulation of hematopoiesis and the hematopoietic stem cell niche by Wnt signaling pathways. *Cell Res* **17**, 746–758 (2007).
27. Pajcini, K. V., Speck, N. A. & Pear, W. S. Notch signaling in mammalian hematopoietic stem cells. *Leukemia* **25**, 1525–1532 (2011).
28. Duncan, A. W. *et al.* Integration of Notch and Wnt signaling in hematopoietic stem cell maintenance. *Nat Immunol* **6**, 314–322 (2005).

29. Stier, S., Cheng, T., Dombkowski, D., Carlesso, N. & Scadden, D. T. Notch1 activation increases hematopoietic stem cell self-renewal in vivo and favors lymphoid over myeloid lineage outcome. *Blood* **99**, 2369–2378 (2002).
30. Lemos, T. & Merchant, A. The hedgehog pathway in hematopoiesis and hematological malignancy. *Front Oncol* **12**, 960943 (2022).
31. Trowbridge, J. J., Scott, M. P. & Bhatia, M. Hedgehog modulates cell cycle regulators in stem cells to control hematopoietic regeneration. *Proc Natl Acad Sci U S A* **103**, 14134–14139 (2006).
32. Sugiyama, T., Kohara, H., Noda, M. & Nagasawa, T. Maintenance of the Hematopoietic Stem Cell Pool by CXCL12-CXCR4 Chemokine Signaling in Bone Marrow Stromal Cell Niches. *Immunity* **25**, 977–988 (2006).
33. Broudy, V. C. Stem Cell Factor and Hematopoiesis. *Blood* **90**, 1345–1364 (1997).
34. Ohta, M., Greenberger, J. S., Anklesaria, P., Bassols, A. & Massagué, J. Two forms of transforming growth factor-beta distinguished by multipotential haematopoietic progenitor cells. *Nature* **329**, 539–541 (1987).
35. Hattori, K. *et al.* Vascular Endothelial Growth Factor and Angiopoietin-1 Stimulate Postnatal Hematopoiesis by Recruitment of Vasculogenic and Hematopoietic Stem Cells. *J Exp Med* **193**, 1005–1014 (2001).
36. Kimura, S., Roberts, A. W., Metcalf, D. & Alexander, W. S. Hematopoietic stem cell deficiencies in mice lacking c-Mpl, the receptor for thrombopoietin. *Proc Natl Acad Sci U S A* **95**, 1195–1200 (1998).
37. Jang, Y.-Y. & Sharkis, S. J. A low level of reactive oxygen species selects for primitive hematopoietic stem cells that may reside in the low-oxygenic niche. *Blood* **110**, 3056–3063 (2007).

38. Ushio-Fukai, M. & Rehman, J. Redox and metabolic regulation of stem/progenitor cells and their niche. *Antioxid Redox Signal* **21**, 1587–1590 (2014).
39. Cipolleschi, M. G., Sbarba, P. D. & Olivotto, M. The Role of Hypoxia in the Maintenance of Hematopoietic Stem Cells. *Blood* **82**, 2031–2037 (1993).
40. Cipolleschi, M. G. *et al.* The expansion of murine bone marrow cells preincubated in hypoxia as an in vitro indicator of their marrow-repopulating ability. *Leukemia* **14**, 735–739 (2000).
41. Simsek, T. *et al.* The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. *Cell Stem Cell* **7**, 380–390 (2010).
42. Ho, T. T. *et al.* Autophagy maintains the metabolism and function of young and old stem cells. *Nature* **543**, 205–210 (2017).
43. Schofield, R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells* **4**, 7–25 (1978).
44. Kiel, M. J. *et al.* SLAM Family Receptors Distinguish Hematopoietic Stem and Progenitor Cells and Reveal Endothelial Niches for Stem Cells. *Cell* **121**, 1109–1121 (2005).
45. Chen, J. Y. *et al.* Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature* **530**, 223–227 (2016).
46. Saçma, M. *et al.* Haematopoietic stem cells in perisinusoidal niches are protected from ageing. *Nat Cell Biol* **21**, 1309–1320 (2019).
47. Christodoulou, C. *et al.* Live-animal imaging of native haematopoietic stem and progenitor cells. *Nature* **578**, 278–283 (2020).
48. Comazzetto, S., Shen, B. & Morrison, S. J. Niches that regulate stem cells and hematopoiesis in adult bone marrow. *Developmental Cell* **56**, 1848–1860 (2021).

49. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nat Rev Mol Cell Biol* **20**, 303–320 (2019).

50. Moore, K. A. & Lemischka, I. R. Stem cells and their niches. *Science* **311**, 1880–1885 (2006).

51. Wilkinson, A. C., Igarashi, K. J. & Nakauchi, H. Haematopoietic stem cell self-renewal in vivo and ex vivo. *Nat Rev Genet* **21**, 541–554 (2020).

52. Zhao, A., Zhou, H., Yang, J., Li, M. & Niu, T. Epigenetic regulation in hematopoiesis and its implications in the targeted therapy of hematologic malignancies. *Signal Transduct Target Ther* **8**, 71 (2023).

53. Martire, S. & Banaszynski, L. A. The roles of histone variants in fine-tuning chromatin organization and function. *Nat Rev Mol Cell Biol* **21**, 522–541 (2020).

54. Sun, X.-J., Man, N., Tan, Y., Nimer, S. D. & Wang, L. The Role of Histone Acetyltransferases in Normal and Malignant Hematopoiesis. *Front Oncol* **5**, 108 (2015).

55. Wallace, L. & Obeng, E. A. Noncoding rules of survival: epigenetic regulation of normal and malignant hematopoiesis. *Front Mol Biosci* **10**, 1273046 (2023).

56. Park, S.-Y. & Kim, J.-S. A short guide to histone deacetylases including recent progress on class II enzymes. *Exp Mol Med* **52**, 204–212 (2020).

57. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet* **13**, 343–357 (2012).

58. Schwarzer, A. *et al.* The non-coding RNA landscape of human hematopoiesis and leukemia. *Nat Commun* **8**, 218 (2017).

59. Shang, R., Lee, S., Senavirathne, G. & Lai, E. C. microRNAs in action: biogenesis, function and regulation. *Nat Rev Genet* **24**, 816–833 (2023).

60. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* **24**, 430–447 (2023).

61. Johnnidis, J. B. *et al.* Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature* **451**, 1125–1129 (2008).

62. O'Connell, R. M. *et al.* Sustained expression of microRNA-155 in hematopoietic stem cells causes a myeloproliferative disorder. *J Exp Med* **205**, 585–594 (2008).

63. Lechman, E. R. *et al.* Attenuation of miR-126 activity expands HSC in vivo without exhaustion. *Cell Stem Cell* **11**, 799–811 (2012).

64. Petriv, O. I. *et al.* Comprehensive microRNA expression profiling of the hematopoietic hierarchy. *Proceedings of the National Academy of Sciences* **107**, 15443–15448 (2010).

65. Ortiz, G. G. R. *et al.* A state-of-the-art review on the MicroRNAs roles in hematopoietic stem cell aging and longevity. *Cell Communication and Signaling* **21**, 85 (2023).

66. Luo, M. *et al.* Long Non-coding RNAs Control Hematopoietic Stem Cell Function. *Cell Stem Cell* **16**, 426–438 (2015).

67. Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M. & Ballestar, E. Epigenetic control of myeloid cell differentiation, identity and function. *Nat Rev Immunol* **15**, 7–17 (2015).

68. Challen, G. A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* **44**, 23–31 (2011).

69. Kim, M. & Costello, J. DNA methylation: an epigenetic mark of cellular memory. *Exp Mol Med* **49**, e322–e322 (2017).

70. Tulstrup, M. *et al.* TET2 mutations are associated with hypermethylation at key regulatory enhancers in normal and malignant hematopoiesis. *Nat Commun* **12**, 6061 (2021).

71. Long, H. K., King, H. W., Patient, R. K., Odom, D. T. & Klose, R. J. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res* **44**, 6693–6706 (2016).

72. Cooper, D. N., Taggart, M. H. & Bird, A. P. Unmethylated domains in vertebrate DNA. *Nucleic Acids Res* **11**, 647–658 (1983).

73. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet* **17**, 551–565 (2016).

74. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).

75. Globisch, D. *et al.* Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. *PLOS ONE* **5**, e15367 (2010).

76. Li, W. *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res* **27**, 1243–1257 (2017).

77. Bowman, R. L. & Levine, R. L. TET2 in Normal and Malignant Hematopoiesis. *Cold Spring Harb Perspect Med* **7**, a026518 (2017).

78. Zhang, X., Zhang, Y., Wang, C. & Wang, X. TET (Ten-eleven translocation) family proteins: structure, biological functions and applications. *Sig Transduct Target Ther* **8**, 1–20 (2023).

79. Hu, L. *et al.* Structural insight into substrate preference for TET-mediated oxidation. *Nature* **527**, 118–122 (2015).

80. Yamagata, K. & Kobayashi, A. The cysteine-rich domain of TET2 binds preferentially to mono- and dimethylated histone H3K36. *J Biochem* **161**, 327–330 (2017).

81. Chiba, S. Dysregulation of TET2 in hematologic malignancies. *Int J Hematol* **105**, 17–22 (2017).

82. Delhommeau, F. *et al.* Mutation in TET2 in Myeloid Cancers. *New England Journal of Medicine* **360**, 2289–2301 (2009).

83. Langemeijer, S. M. C. *et al.* Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat Genet* **41**, 838–842 (2009).

84. Jankowska, A. M. *et al.* Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood* **113**, 6403–6410 (2009).

85. Tefferi, A. *et al.* TET2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. *Leukemia* **23**, 905–911 (2009).

86. Lv, H. *et al.* Characterization of global research hotspots and trends on ten-eleven translocation 2: visualization and bibliometric analysis. *Am J Transl Res* **14**, 8504–8522 (2022).

87. Ferrone, C. K., Blydt-Hansen, M. & Rauh, M. J. Age-Associated TET2 Mutations: Common Drivers of Myeloid Dysfunction, Cancer and Cardiovascular Disease. *Int J Mol Sci* **21**, 626 (2020).

88. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).

89. Fernández-Maestre, I., Cai, S. F. & Levine, R. L. A View of Myeloid Transformation through the Hallmarks of Cancer. *Blood Cancer Discov* **5**, 377–387 (2024).

90. Takahashi, K. & Tanaka, T. Clonal evolution and hierarchy in myeloid malignancies. *Trends Cancer* **9**, 707–715 (2023).

91. Tang, L., Huang, Z., Mei, H. & Hu, Y. Immunotherapy in hematologic malignancies: achievements, challenges and future prospects. *Sig Transduct Target Ther* **8**, 306 (2023).

92. Landau, D. A., Carter, S. L., Getz, G. & Wu, C. J. Clonal evolution in hematologic malignancies and therapeutic implications. *Leukemia* **28**, 34–43 (2014).

93. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing. *Nature* **481**, 506–510 (2012).

94. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).

95. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193–1203 (2016).

96. Young, A. L. *et al.* Quantifying ultra-rare pre-leukemic clones via targeted error-corrected sequencing. *Leukemia* **29**, 1608–1611 (2015).

97. Jan, M. & Majeti, R. Clonal evolution of acute leukemia genomes. *Oncogene* **32**, 135–140 (2013).

98. Corces, M. R., Chang, H. Y. & Majeti, R. Preleukemic Hematopoietic Stem Cells in Human Acute Myeloid Leukemia. *Front Oncol* **7**, 263 (2017).

99. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med* **4**, 149ra118 (2012).

100. Metzeler, K. H. *et al.* Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood* **128**, 686–698 (2016).

101. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472–1478 (2014).
102. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477–2487 (2014).
103. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488–2498 (2014).
104. Ayachi, S., Buscarlet, M. & Busque, L. 60 Years of clonal hematopoiesis research: From X-chromosome inactivation studies to the identification of driver mutations. *Experimental Hematology* **83**, 2–11 (2020).
105. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat Commun* **11**, 5327 (2020).
106. Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).
107. Bouligny, I. M., Maher, K. R. & Grant, S. Mechanisms of myeloid leukemogenesis: Current perspectives and therapeutic objectives. *Blood Rev* **57**, 100996 (2023).
108. Ntziachristos, P., Abdel-Wahab, O. & Aifantis, I. Emerging concepts of epigenetic dysregulation in hematological malignancies. *Nat Immunol* **17**, 1016–1024 (2016).
109. Scheuermann, J. C. *et al.* Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465**, 243–247 (2010).
110. Kühn, M. W. M., Pemmaraju, N. & Heidel, F. H. The evolving landscape of epigenetic target molecules and therapies in myeloid cancers: focus on acute myeloid leukemia and myeloproliferative neoplasms. *Leukemia* **39**, 1824–1837 (2025).

111. Nagase, R. *et al.* Expression of mutant Asxl1 perturbs hematopoiesis and promotes susceptibility to leukemic transformation. *J Exp Med* **215**, 1729–1747 (2018).
112. Inoue, D. *et al.* Myelodysplastic syndromes are induced by histone methylation–altering ASXL1 mutations. *J Clin Invest* **123**, 4627–4640 (2013).
113. Balasubramani, A. *et al.* Cancer-associated ASXL1 mutations may act as gain-of-function mutations of the ASXL1-BAP1 complex. *Nat Commun* **6**, 7307 (2015).
114. Meyer, C. *et al.* The MLL recombinome of acute leukemias in 2013. *Leukemia* **27**, 2165–2176 (2013).
115. Takahashi, S. & Yokoyama, A. The molecular functions of common and atypical MLL fusion protein complexes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**, 194548 (2020).
116. Joshi, K., Liu, S., Breslin S J, P. & Zhang, J. Mechanisms that regulate the activities of TET proteins. *Cell Mol Life Sci* **79**, 363 (2022).
117. Li, L. *et al.* Epigenetic inactivation of the CpG demethylase TET1 as a DNA methylation feedback loop in human cancers. *Sci Rep* **6**, 26591 (2016).
118. Dawlaty, M. M. *et al.* Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell* **9**, 166–175 (2011).
119. Joshi, K., Zhang, L., Breslin S.J., P., Kini, A. R. & Zhang, J. Role of TET dioxygenases in the regulation of both normal and pathological hematopoiesis. *Journal of Experimental & Clinical Cancer Research* **41**, 294 (2022).

120. Feng, Y., Li, X., Cassady, K., Zou, Z. & Zhang, X. TET2 Function in Hematopoietic Malignancies, Immune Regulation, and DNA Repair. *Front Oncol* **9**, 210 (2019).
121. Zhang, P. *et al.* TP53 mutations and TET2 deficiency cooperate to drive leukemogenesis and establish an immunosuppressive environment. *J Clin Invest* **135**, e184021 (2025).
122. Kunimoto, H. & Nakajima, H. TET2: A cornerstone in normal and malignant hematopoiesis. *Cancer Sci* **112**, 31–40 (2021).
123. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839–843 (2010).
124. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
125. Ito, K. *et al.* Non-catalytic Roles of Tet2 Are Essential to Regulate Hematopoietic Stem and Progenitor Cell Homeostasis. *Cell Rep* **28**, 2480-2490.e4 (2019).
126. Flores, J. C. *et al.* Comparative analysis of Tet2 catalytic-deficient and knockout bone marrow over time. *Exp Hematol* **124**, 45-55.e2 (2023).
127. Holmes, S. M. *et al.* Sex differences alter primitive progenitors in the C57BL/6 Tet2 knockout mouse model. *Exp Hematol* **145**, 104747 (2025).
128. Rasmussen, K. D. *et al.* TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. *Genome Res* **29**, 564–575 (2019).
129. Hon, G. C. *et al.* 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol Cell* **56**, 286–297 (2014).

130. Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. Role of Tet proteins in enhancer activity and telomere elongation. *Genes Dev* **28**, 2103–2119 (2014).
131. An, J. *et al.* Acute loss of TET function results in aggressive myeloid cancer in mice. *Nat Commun* **6**, 10071 (2015).
132. Rasmussen, K. D. *et al.* Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. *Genes Dev* **29**, 910–922 (2015).
133. Yamazaki, J. *et al.* TET2 Mutations Affect Non-CpG Island DNA Methylation at Enhancers and Transcription Factor-Binding Sites in Chronic Myelomonocytic Leukemia. *Cancer Res* **75**, 2833–2843 (2015).
134. Nakauchi, Y. *et al.* The Cell Type-Specific 5hmC Landscape and Dynamics of Healthy Human Hematopoiesis and TET2-Mutant Preleukemia. *Blood Cancer Discov* **3**, 346–367 (2022).
135. Tian, H., Luan, P., Liu, Y. & Li, G. Tet-mediated DNA methylation dynamics affect chromosome organization. *Nucleic Acids Res* **52**, 3654–3666 (2024).
136. Hong, T. *et al.* TET2 modulates spatial relocalization of heterochromatin in aged hematopoietic stem and progenitor cells. *Nat Aging* **3**, 1387–1400 (2023).
137. Costa, Y. *et al.* NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature* **495**, 370–374 (2013).
138. de la Rica, L. *et al.* PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome Biol* **14**, R99 (2013).
139. Suzuki, T. *et al.* RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. *Blood Adv* **1**, 1699–1711 (2017).

140. Fujiki, K. *et al.* PPARγ-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine. *Nat Commun* **4**, 2262 (2013).
141. Song, S. J. *et al.* MicroRNA-antagonism regulates breast cancer stemness and metastasis via TET-family-dependent chromatin remodeling. *Cell* **154**, 311–324 (2013).
142. Song, S. J. *et al.* The oncogenic microRNA miR-22 targets the TET2 tumor suppressor to promote hematopoietic stem cell self-renewal and transformation. *Cell Stem Cell* **13**, 87–101 (2013).
143. Islam, M. S., Leissing, T. M., Chowdhury, R., Hopkinson, R. J. & Schofield, C. J. 2-Oxoglutarate-Dependent Oxygenases. *Annu Rev Biochem* **87**, 585–620 (2018).
144. Matuleviciute, R., Cunha, P. P., Johnson, R. S. & Foskolou, I. P. Oxygen regulation of TET enzymes. *The FEBS Journal* **288**, 7143–7161 (2021).
145. Arguello, A. E. *et al.* Reactivity-dependent profiling of RNA 5-methylcytidine dioxygenases. *Nat Commun* **13**, 4176 (2022).
146. Zou, Z. *et al.* RNA m5C oxidation by TET2 regulates chromatin state and leukaemogenesis. *Nature* **634**, 986–994 (2024).
147. Li, Y. *et al.* TET2-mediated mRNA demethylation regulates leukemia stem cell homing and self-renewal. *Cell Stem Cell* **30**, 1072-1090.e10 (2023).
148. Lee, W.-H. *et al.* Distinct genetic landscapes and their clinical implications in younger and older patients with myelodysplastic syndromes. *Hematol Oncol* **41**, 463–473 (2023).
149. Gurnari, C. *et al.* TET2 mutations as a part of DNA dioxygenase deficiency in myelodysplastic syndromes. *Blood Adv* **6**, 100–107 (2022).

150. Huang, F. *et al.* TET2 deficiency promotes MDS-associated leukemogenesis. *Blood Cancer J* **12**, 141 (2022).
151. Sun, J. *et al.* SIRT1 Activation Disrupts Maintenance of Myelodysplastic Syndrome Stem and Progenitor Cells by Restoring TET2 Function. *Cell Stem Cell* **23**, 355-369.e9 (2018).
152. de Jong, M. M. E., Chen, L., Raaijmakers, M. H. G. P. & Cupedo, T. Bone marrow inflammation in haematological malignancies. *Nat Rev Immunol* **24**, 543–558 (2024).
153. Barreyro, L., Chlon, T. M. & Starczynowski, D. T. Chronic immune response dysregulation in MDS pathogenesis. *Blood* **132**, 1553–1560 (2018).
154. Mei, Y. *et al.* Bone marrow-confined IL-6 signaling mediates the progression of myelodysplastic syndromes to acute myeloid leukemia. *J Clin Invest* **132**, e152673 (2022).
155. Cai, Z. *et al.* Inhibition of Inflammatory Signaling in Tet2 Mutant Preleukemic Cells Mitigates Stress Induced Abnormalities and Clonal Hematopoiesis. *Cell Stem Cell* **23**, 833-849.e5 (2018).
156. Zhang, Q. *et al.* Tet2 is required to resolve inflammation by recruiting Hdac2 to specifically repress IL-6. *Nature* **525**, 389–393 (2015).
157. Zavras, P. D., Shastri, A., Goldfinger, M., Verma, A. K. & Sauntharajah, Y. Clinical Trials Assessing Hypomethylating Agents Combined with Other Therapies: Causes for Failure and Potential Solutions. *Clin Cancer Res* **27**, 6653–6661 (2021).
158. Gurion, R. *et al.* 5-azacitidine prolongs overall survival in patients with myelodysplastic syndrome - a systematic review and meta-analysis. *Haematologica* **95**, 303–310 (2010).

159. Clinical Trials register - Search for decitabine OR 2'-deoxy-5-azacytidine OR 5-aza-2'-deoxycytidine OR 5-azadeoxycytidine OR 5-deoxyazacytidine. https://www.clinicaltrialsregister.eu/ctr-search/search?query=decitabine+OR+2%27-deoxy-5-azacytidine+OR+5-aza-2%27-deoxycytidine+OR+5-azadeoxycytidine+OR+5-deoxyazacytidine.
160. Schiroli, G. *et al.* Cell of origin epigenetic priming determines susceptibility to Tet2 mutation. *Nat Commun* **15**, 4325 (2024).
161. Iyama, S. *et al.* Prognostic impact of TET2 mutations in patients with acute myeloid leukemia: HM-SCREEN-Japan 01 and 02 study. *Ann Hematol* **104**, 275–284 (2025).
162. Pan, X. *et al.* TET2 mutations contribute to adverse prognosis in acute myeloid leukemia (AML): results from a comprehensive analysis of 502 AML cases and the Beat AML public database. *Clin Exp Med* **24**, 35 (2024).
163. Jin, G.-B. *et al.* Neomorphic leukemia-derived mutations in the TET2 enzyme induce genome instability via a substrate shift from 5-methylcytosine to thymine. *Proc Natl Acad Sci U S A* **122**, e2418318122 (2025).
164. Mian, S. A., Ngo, S. & Bonnet, D. Is the bone marrow microenvironment the hidden catalyst in malignant haematopoiesis? *Leukemia* **39**, 1589–1592 (2025).
165. Fröbel, J. *et al.* The Hematopoietic Bone Marrow Niche Ecosystem. *Front Cell Dev Biol* **9**, 705410 (2021).
166. Colmone, A. *et al.* Leukemic cells create bone marrow niches that disrupt the behavior of normal hematopoietic progenitor cells. *Science* **322**, 1861–1865 (2008).

167. Waclawiczek, A. *et al.* Mesenchymal niche remodeling impairs hematopoiesis via stanniocalcin 1 in acute myeloid leukemia. *J Clin Invest* **130**, 3038–3050 (2020).

168. Krause, D. S. *et al.* Differential regulation of myeloid leukemias by the bone marrow microenvironment. *Nat Med* **19**, 1513–1517 (2013).

169. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

170. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

171. Patterson, S. D. & Aebersold, R. H. Proteomics: the first decade and beyond. *Nat Genet* **33 Suppl**, 311–323 (2003).

172. Chen, Y. *et al.* Chromatin accessibility: biological functions, molecular mechanisms and therapeutic application. *Sig Transduct Target Ther* **9**, 340 (2024).

173. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519–522 (2015).

174. Gopalan, S., Wang, Y., Harper, N. W., Garber, M. & Fazzio, T. G. Simultaneous profiling of multiple chromatin proteins in the same cells. *Mol Cell* **81**, 4736-4746.e5 (2021).

175. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**, 781 (2018).

176. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865–868 (2017).

177. Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol* **39**, 1246–1258 (2021).

178. Chen, A. F. *et al.* NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods* **19**, 547–553 (2022).

179. Jassinskaja, M., Gonka, M. & Kent, D. G. Resolving the hematopoietic stem cell state by linking functional and molecular assays. *Blood* **142**, 543–552 (2023).

180. Brenner, S., Jacob, F. & Meselson, M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* **190**, 576–581 (1961).

181. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).

182. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610–620 (2015).

183. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

184. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. 2021.05.05.442755 Preprint at https://doi.org/10.1101/2021.05.05.442755 (2021).

185. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

186. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).
187. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
188. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
189. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).
190. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
191. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
192. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
193. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* **52**, D672–D678 (2024).
194. Wu, G. & Haw, R. Functional Interaction Network Construction and Analysis for Disease Discovery. *Methods Mol Biol* **1558**, 235–253 (2017).
195. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
196. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267–273 (2003).

197. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

198. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).

199. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**, pdb.prot5384 (2010).

200. Cui, K. & Zhao, K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. *Methods Mol Biol* **833**, 413–419 (2012).

201. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

202. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).

203. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).

204. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

205. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).

206. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403–411 (2021).

207. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21**, 22 (2020).
208. Single-cell proteomics: challenges and prospects. *Nat Methods* **20**, 317–318 (2023).
209. Jassinskaja, M. *et al.* Selection of pre-leukemic hematopoietic stem cells driven by distinct extracellular matrix molecules. 2025.01.20.633881 Preprint at https://doi.org/10.1101/2025.01.20.633881 (2025).
210. Guo, T., Steen, J. A. & Mann, M. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature* **638**, 901–911 (2025).
211. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat Rev Mol Cell Biol* **20**, 285–302 (2019).
212. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology* **19**, 161 (2018).
213. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550–572 (2023).
214. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257–272 (2019).
215. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* **15**, 359–362 (2018).
216. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 111 (2020).
217. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

218. Schäfer, P. S. L., Dimitrov, D., Villablanca, E. J. & Saez-Rodriguez, J. Integrating single-cell multi-omics and prior biological knowledge for a functional characterization of the immune system. *Nat Immunol* **25**, 405–417 (2024).
219. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70–80 (2018).
220. Tarazona, S. *et al.* Harmonization of quality metrics and power calculation in multi-omic studies. *Nat Commun* **11**, 3092 (2020).
221. Le Novère, N. Quantitative and logic modelling of molecular and gene networks. *Nat Rev Genet* **16**, 146–158 (2015).
222. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology* **22**, 437–467 (1969).
223. Hall, B. A. & Niarakis, A. Data integration in logic-based models of biological mechanisms. *Current Opinion in Systems Biology* **28**, 100386 (2021).
224. Mazein, A. *et al.* A guide for developing comprehensive systems biology maps of disease mechanisms: planning, construction and maintenance. *Front Bioinform* **3**, 1197310 (2023).
225. Badia-I-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* **24**, 739–754 (2023).
226. Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S. & Saez-Rodriguez, J. Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol Syst Biol* **18**, e11036 (2022).
227. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
228. Jain, M. S. *et al.* MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biology* **22**, 346 (2021).

229. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
230. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* **20**, 1355–1367 (2023).
231. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**, 764–766 (2017).
232. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci U S A* **114**, E4914–E4923 (2017).
233. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187–200 (2021).
234. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* **13**, 966–967 (2016).
235. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* **17**, 159–162 (2020).
236. Hidalgo, M. R. *et al.* High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget* **8**, 5160–5178 (2016).
237. Martini, P., Sales, G., Massa, M. S., Chiogna, M. & Romualdi, C. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res* **41**, e19 (2013).

238. Babur, Ö. *et al.* Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns* **2**, 100257 (2021).
239. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
240. Li, Z. *et al.* Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. *Blood* **118**, 4509–4518 (2011).
241. Hawking, Z. L. & Allan, J. M. Landscape of TET2 Mutations: From Hematological Malignancies to Solid Tumors. *Cancer Med* **14**, e70792 (2025).
242. Fu, L. *et al.* Tet-Mediated Formation of 5-Hydroxymethylcytosine in RNA. *J Am Chem Soc* **136**, 11582–11585 (2014).
243. Méndez-Ferrer, S. *et al.* Bone marrow niches in haematological malignancies. *Nat Rev Cancer* **20**, 285–298 (2020).
244. Jassinskaja, M. *et al.* A complex interplay of intra- and extracellular factors regulates the outcome of fetal- and adult-derived MLL-rearranged leukemia. *Leukemia* **38**, 1115–1130 (2024).
245. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
246. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
247. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).

248. Cusanovich, D. A. *et al.* Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science* **348**, 910–914 (2015).
249. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).
250. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
251. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975–978 (2017).
252. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192 (2013).
253. Cytoscape: Network Data Integration, Analysis, and Visualization in a Box, https://cytoscape.org.
254. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
255. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
256. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
257. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363–1375 (2019).

258. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal* **Complex Systems**, 1695 (2005).
259. R Core Team and contributors worldwide. graphics, The R Graphics Package, R functions for base graphics.
260. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org. ggplot2.
261. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology* **17**, e9923 (2021).
262. Csabai, L. *et al.* SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Research* **50**, D701–D709 (2022).
263. Lo Surdo, P. *et al.* SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res* **51**, D631–D637 (2023).
264. ligrecextra, ligand-receptor interactions without literature reference, https://omnipathdb.org/interactions?datasets=ligrecextra.
265. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184–1191 (2009).
266. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
267. Fang, H. & Gough, J. The `dnet' approach promotes emerging research on cancer patient survival. *Genome Medicine* **6**, 64 (2014).
268. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* **36**, D646–D650 (2008).

269. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088 (2021).
270. Zhang, Y. *et al.* Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis. *Bioinformatics* btab036 (2021) doi:10.1093/bioinformatics/btab036.
271. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* **15**, 1484–1506 (2020).
272. Meldal, B. H. M. *et al.* The complex portal - an encyclopaedia of macromolecular complexes. *Nucleic Acids Res* **43**, D479–D484 (2015).
273. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
274. Baccin, C. *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat Cell Biol* **22**, 38–48 (2020).
275. Wilkinson, A. C., Ishida, R., Nakauchi, H. & Yamazaki, S. Long-term ex vivo expansion of mouse hematopoietic stem cells. *Nat Protoc* **15**, 628–648 (2020).
276. Becton, Dickinson and Company. FlowJo™ Software, 2023.
277. Tefferi, A. *et al.* Frequent TET2 mutations in systemic mastocytosis: clinical, KITD816V and FIP1L1-PDGFRA correlates. *Leukemia* **23**, 900–904 (2009).
278. Palomo, L. *et al.* Molecular landscape and clonal architecture of adult myelodysplastic/myeloproliferative neoplasms. *Blood* **136**, 1851–1862 (2020).
279. Bernard, E. *et al.* Molecular taxonomy of myelodysplastic syndromes and its clinical implications. *Blood* **144**, 1617–1632 (2024).

280. Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat Med* **24**, 103–112 (2018).
281. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209–2221 (2016).
282. Brady, S. W. *et al.* The genomic landscape of pediatric acute lymphoblastic leukemia. *Nat Genet* **54**, 1376–1389 (2022).
283. Smith, A. E. *et al.* Next-generation sequencing of the TET2 gene in 355 MDS and CMML patients reveals low-abundance mutant clones with early origins, but indicates no definite prognostic value. *Blood* **116**, 3923–3932 (2010).
284. Formattable, https://github.com/renkun-ken/formattable.
285. C, S. *et al.* Smarcc1/Baf155 couples self-renewal gene repression with changes in chromatin structure in mouse embryonic stem cells. *Stem cells (Dayton, Ohio)* **27**, (2009).
286. Tanaka, T. *et al.* An acute myeloid leukemia gene, AML1, regulates transcriptional activation and hemopoietic myeloid cell differentiation antagonistically by two alternative spliced forms. *Leukemia* **11 Suppl 3**, 299–302 (1997).
287. Wang, X. *et al.* Bach1 Induces Endothelial Cell Apoptosis and Cell-Cycle Arrest through ROS Generation. *Oxid Med Cell Longev* **2016**, 6234043 (2016).
288. Luchsinger, L. L. *et al.* Harnessing Hematopoietic Stem Cell Low Intracellular Calcium Improves Their Maintenance In Vitro. *Cell Stem Cell* **25**, 225-240.e7 (2019).

289. Wilson, N. K. *et al.* Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* **16**, 712–724 (2015).

290. Identification and characterization of in vitro expanded hematopoietic stem cells | EMBO reports. https://www.embopress.org/doi/full/10.15252/embr.202255502.

291. Guerrero Schimpf, M. L., Sparger, C. C., Huang, H.-T., Wade, D. & Figueroa, M. K. E. PDZD2 Is Essential for Steady-State Hematopoiesis and Its 37-Kda Secreted Product, sPDZD2, Functions As a Soluble Tumor Suppressor in AML. *Blood* **140**, 8696–8697 (2022).

292. VennDiagram, Hanbo Chen, https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf.

293. Land, R. *et al.* The Orphan Nuclear Receptor Nr4a1 Specifies Long-Term, Myeloid-Biased HSCs and Responds To PGE2. *Blood* **122**, 3675 (2013).

294. Ma, Z. *et al.* Hes1 deficiency causes hematopoietic stem cell exhaustion. *Stem Cells* **38**, 756–768 (2020).

295. Shang, F. & Rodewald, H.-R. Toward the dissection of hematopoietic stem cell fates and their determinants. *Current Opinion in Genetics & Development* **75**, 101945 (2022).

296. Pei, W. *et al.* Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding. *Cell Stem Cell* **27**, 383-395.e8 (2020).

297. Zhang, T. *et al.* Phosphorylation of TET2 by AMPK is indispensable in myogenic differentiation. *Epigenetics Chromatin* **12**, 32 (2019).

298. Chen, Q., Chen, Y., Bian, C., Fujiki, R. & Yu, X. TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature* **493**, 561–564 (2013).

299. de Pater, E. *et al.* Gata2 is required for HSC generation and survival. *J Exp Med* **210**, 2843–2850 (2013).
300. Frelin, C. *et al.* GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat Immunol* **14**, 1037–1044 (2013).
301. Ross, K. *et al.* Polycomb group ring finger 1 cooperates with Runx1 in regulating differentiation and self-renewal of hematopoietic cells. *Blood* **119**, 4152–4161 (2012).
302. Wang, Z.-Q. *et al.* Bone and haematopoietic defects in mice lacking c-fos. *Nature* **360**, 741–745 (1992).
303. Lacorazza, H. D. The reprogramming factor KLF4 in normal and malignant blood cells. *Front Immunol* **16**, 1584181 (2025).
304. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Sig Transduct Target Ther* **2**, 17023 (2017).
305. Ben-David, Y., Gajendran, B., Sample, K. M. & Zacksenhaus, E. Current insights into the role of Fli-1 in hematopoiesis and malignant transformation. *Cell Mol Life Sci* **79**, 163 (2022).
306. Chua, E. H. Z., Yasar, S. & Harmston, N. The importance of considering regulatory domains in genome-wide analyses - the nearest gene is often wrong! *Biol Open* **11**, bio059091 (2022).
307. Lea, A. J. *et al.* Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* **7**, e37513.
308. 10x Genomics Cell Ranger.
309. Asmar, F. *et al.* Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. *Haematologica* **98**, 1912–1920 (2013).

310. Ortmann, C. A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *N Engl J Med* **372**, 601–612 (2015).
311. Sekeres, M. A. & Taylor, J. Diagnosis and Treatment of Myelodysplastic Syndromes: A Review. *JAMA* **328**, 872–880 (2022).
312. Ogawa, S. Genetics of MDS. *Blood* **133**, 1049–1059 (2019).
313. Chen, T.-C. *et al.* Dynamics of ASXL1 mutation and other associated genetic alterations during disease progression in patients with primary myelodysplastic syndrome. *Blood Cancer Journal* **4**, e177–e177 (2014).
314. Thol, F. *et al.* IDH1 mutations in patients with myelodysplastic syndromes are associated with an unfavorable prognosis. *Haematologica* **95**, 1668–1674 (2010).
315. Daver, N. G. *et al.* TP53-Mutated Myelodysplastic Syndrome and Acute Myeloid Leukemia: Biology, Current Therapy, and Future Directions. *Cancer Discov* **12**, 2516–2529 (2022).
316. Maurya, N. *et al.* Comprehensive analysis of genetic factors predicting overall survival in Myelodysplastic syndromes. *Sci Rep* **12**, 5925 (2022).
317. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
318. Gene Set Enrichment Analysis (GSEA) User Guide, https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm.
319. Maifrede, S. *et al.* TET2 and DNMT3A Mutations Exert Divergent Effects on DNA Repair and Sensitivity of Leukemia Cells to PARP Inhibitors. *Cancer Res* **81**, 5089–5101 (2021).

320. Xie, J. *et al.* STING activation in TET2-mutated hematopoietic stem/progenitor cells contributes to the increased self-renewal and neoplastic transformation. *Leukemia* **37**, 2457–2467 (2023).
321. Kafer, G. R. *et al.* 5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes Genome Stability. *Cell Rep* **14**, 1283–1292 (2016).
322. Enrichment Map Gene Sets, BaderLab, University of Toronto, http://www.baderlab.org/GeneSets, Gene Set files downloaded from http://download.baderlab.org/EM_Genesets/, version: September 2024. Enrichment Map Gene Sets, BaderLab.
323. Caiado, F., Pietras, E. M. & Manz, M. G. Inflammation as a regulator of hematopoietic stem cell function in disease, aging, and clonal selection. *J Exp Med* **218**, e20201541 (2021).
324. Cull, A. H., Snetsinger, B., Buckstein, R., Wells, R. A. & Rauh, M. J. Tet2 restrains inflammatory gene expression in macrophages. *Exp Hematol* **55**, 56-70.e13 (2017).
325. Li, J. *et al.* Role of Tet2 in Regulating Adaptive and Innate Immunity. *Front. Cell Dev. Biol.* **9**, (2021).
326. Chen, B. *et al.* Cellular zinc metabolism and zinc signaling: from biological functions to diseases and therapeutic targets. *Signal Transduct Target Ther* **9**, 6 (2024).
327. Bonaventura, P., Benedetti, G., Albarède, F. & Miossec, P. Zinc and its role in immunity and inflammation. *Autoimmun Rev* **14**, 277–285 (2015).
328. Liu, W. *et al.* Lactate regulates cell cycle by remodelling the anaphase promoting complex. *Nature* **616**, 790–797 (2023).

329. Wang, L. *et al.* Co-implantation of magnesium and zinc ions into titanium regulates the behaviors of human gingival fibroblasts. *Bioact Mater* **6**, 64–74 (2021).
330. Xiao, W. *et al.* Therapeutic targeting of the USP2-E2F4 axis inhibits autophagic machinery essential for zinc homeostasis in cancer progression. *Autophagy* **18**, 2615–2635 (2022).
331. Moazed, V. *et al.* Copper and Zinc Levels in Myelodysplastic Syndrome Patients versus Healthy Subjects. *Asian Pac J Cancer Prev* **21**, 239–241 (2020).
332. Hu, L. *et al.* Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. *Cell* **155**, 1545–1555 (2013).
333. Katayama, N., Nishikawa, M., Komada, F., Minami, N. & Shirakawa, S. A role for calmodulin in the growth of human hematopoietic progenitor cells. *Blood* **75**, 1446–1454 (1990).
334. Ewerth, D. *et al.* APC/CCdh1 regulates the balance between maintenance and differentiation of hematopoietic stem and progenitor cells. *Cell Mol Life Sci* **76**, 369–380 (2019).
335. Grey, W. *et al.* Activation of the receptor tyrosine kinase RET improves long-term hematopoietic stem cell outgrowth and potency. *Blood* **136**, 2535–2547 (2020).
336. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).
337. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

338. Aljoufi, A. *et al.* Physioxia-induced downregulation of Tet2 in hematopoietic stem cells contributes to enhanced self-renewal. *Blood* **140**, 1263–1277 (2022).
339. Parmar, K., Mauch, P., Vergilio, J.-A., Sackstein, R. & Down, J. D. Distribution of hematopoietic stem cells in the bone marrow according to regional hypoxia. *Proceedings of the National Academy of Sciences* **104**, 5431–5436 (2007).
340. Razmkhah, F., Kim, S., Lim, S., Dania, A.-J. & Choi, J. S100A8 and S100A9 in Hematologic Malignancies: From Development to Therapy. *Int J Mol Sci* **24**, 13382 (2023).
341. Wang, Y.-H. *et al.* High BM plasma S100A8/A9 is associated with a perturbed microenvironmentand poor prognosis in myelodysplastic syndromes. *Blood Adv* **7**, 2528–2533 (2023).
342. Krečak, I. *et al.* Serum calprotectin: A circulating biomarker of the inflammatory state in Philadelphia-negative myeloproliferative neoplasms. *Blood Cells Mol Dis* **79**, 102344 (2019).
343. Song, R. & Struhl, K. S100A8/S100A9 cytokine acts as a transcriptional coactivator during breast cellular transformation. *Sci Adv* **7**, eabe5357 (2021).
344. Keyvani Chahi, A. *et al.* PLAG1 dampens protein synthesis to promote human hematopoietic stem cell self-renewal. *Blood* **140**, 992–1008 (2022).
345. Raffel, G. D. *et al.* Ott1(Rbm15) has pleiotropic roles in hematopoietic development. *Proceedings of the National Academy of Sciences* **104**, 6001–6006 (2007).
346. Gu, Z. ComplexHeatmap. (2016).
347. Anjos-Afonso, F. & Bonnet, D. Human CD34+ hematopoietic stem cell hierarchy: how far are we with its delineation at the most primitive level? *Blood* **142**, 509–518 (2023).

348. Hayward, C. P. *et al.* The cDNA sequence of human endothelial cell multimerin. A unique protein with RGDS, coiled-coil, and epidermal growth factor-like domains and a carboxyl terminus similar to the globular domain of complement C1q and collagens type VIII and X. *J Biol Chem* **270**, 18246–18251 (1995).
349. Hayward, C. P. *et al.* Multimerin is found in the alpha-granules of resting platelets and is synthesized by a megakaryocytic cell line. *J Clin Invest* **91**, 2630–2639 (1993).
350. Broxmeyer, H. E. Hematopoietic Stem Cell Intracellular Levels of *Ca*2+ to the Rescue! What Next? *Cell Stem Cell* **25**, 171–173 (2019).
351. Xia, P. *et al.* Roles of S100A8, S100A9 and S100A12 in infection, inflammation and immunity. *Immunology* **171**, 365–376 (2024).
352. Chen, X. *et al.* Induction of myelodysplasia by myeloid-derived suppressor cells. *J Clin Invest* **123**, 4595–4611 (2013).
353. Sallman, D. A., Cluzeau, T., Basiorka, A. A. & List, A. Unraveling the Pathogenesis of MDS: The NLRP3 Inflammasome and Pyroptosis Drive the MDS Phenotype. *Front Oncol* **6**, 151 (2016).
354. Simard, J.-C. *et al.* S100A8 and S100A9 induce cytokine expression and regulate the NLRP3 inflammasome via ROS-dependent activation of NF-κB(1.). *PLoS One* **8**, e72138 (2013).
355. Cheng, P. *et al.* S100A9-induced overexpression of PD-1/PD-L1 contributes to ineffective hematopoiesis in myelodysplastic syndromes. *Leukemia* **33**, 2034–2046 (2019).
356. Cluzeau, T. *et al.* Pro-inflammatory proteins S100A9 and tumor necrosis factor-α suppress erythropoietin elaboration in myelodysplastic syndromes. *Haematologica* **102**, 2015–2020 (2017).

357. Zambetti, N. A. *et al.* Mesenchymal Inflammation Drives Genotoxic Stress in Hematopoietic Stem Cells and Predicts Disease Evolution in Human Pre-leukemia. *Cell Stem Cell* **19**, 613–627 (2016).

358. Umemoto, T., Hashimoto, M., Matsumura, T., Nakamura-Ishizu, A. & Suda, T. Ca2+-mitochondria axis drives cell division in hematopoietic stem cells. *J Exp Med* **215**, 2097–2113 (2018).

359. Lee-Thedieck, C., Schertl, P. & Klein, G. The extracellular matrix of hematopoietic stem cell niches. *Adv Drug Deliv Rev* **181**, 114069 (2022).

360. Man, Y., Yao, X., Yang, T. & Wang, Y. Hematopoietic Stem Cell Niche During Homeostasis, Malignancy, and Bone Marrow Transplantation. *Front Cell Dev Biol* **9**, 621214 (2021).

361. Boyd, A. L. *et al.* Niche displacement of human leukemic stem cells uniquely allows their competitive replacement with healthy HSPCs. *J Exp Med* **211**, 1925–1935 (2014).

362. Kuiperij, H. B., van Pel, M., de Rooij, K. E., Hoeben, R. C. & Fibbe, W. E. Serpina1 (alpha1-AT) is synthesized in the osteoblastic stem cell niche. *Exp Hematol* **37**, 641–647 (2009).

363. van Pel, M. *et al.* Serpina1 is a potent inhibitor of IL-8-induced hematopoietic stem cell mobilization. *Proc Natl Acad Sci U S A* **103**, 1469–1474 (2006).

364. Yi, W. *et al.* Ferritin-mediated mitochondrial iron homeostasis is essential for the survival of hematopoietic stem cells and leukemic stem cells. *Leukemia* **38**, 1003–1018 (2024).

365. Balazs, A. B., Fabian, A. J., Esmon, C. T. & Mulligan, R. C. Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow. *Blood* **107**, 2317–2321 (2006).

366. K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **Volume 58**, Pages 453-467 (1971).

367. Blighe K, Lun A. PCAtools: PCAtools: Everything Principal Components Analysis. (2024).

368. Cortes, G. A., Moore, M. J. & El-Nakeep, S. Physiology, Von Willebrand Factor. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

369. Tong, J. *et al.* Hematopoietic Stem Cell Heterogeneity Is Linked to the Initiation and Therapeutic Response of Myeloproliferative Neoplasms. *Cell Stem Cell* **28**, 780 (2021).

370. Sanjuan-Pla, A. *et al.* Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature* **502**, 232–236 (2013).

371. Pinho, S. *et al.* Lineage-biased hematopoietic stem cells are regulated by distinct niches. *Dev Cell* **44**, 634-641.e4 (2018).

372. Leukaemia (all subtypes combined) statistics. *Cancer Research UK* https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia (2015).

373. Dunn, W. G., McLoughlin, M. A. & Vassiliou, G. S. Clonal hematopoiesis and hematological malignancy. *J Clin Invest* **134**, e180065.

374. Shih, A. H., Abdel-Wahab, O., Patel, J. P. & Levine, R. L. The role of mutations in epigenetic regulators in myeloid malignancies. *Nat Rev Cancer* **12**, 599–612 (2012).

375. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515–520 (2018).

376. Izzo, F. *et al.* DNA methylation disruption reshapes the hematopoietic differentiation landscape. *Nat Genet* **52**, 378–387 (2020).

377. Brück, O. *et al.* Immune profiles in acute myeloid leukemia bone marrow associate with patient age, T-cell receptor clonality, and survival. *Blood Adv* **4**, 274–286 (2020).

378. Chen, L. *et al.* A Single-Cell Taxonomy Predicts Inflammatory Niche Remodeling to Drive Tissue Failure and Outcome in Human AML. *Blood Cancer Discov* **4**, 394–417 (2023).

379. Balandrán, J. C., Lasry, A. & Aifantis, I. The Role of Inflammation in the Initiation and Progression of Myeloid Neoplasms. *Blood Cancer Discov* **4**, 254–266 (2023).

380. Bandyopadhyay, S. *et al.* Mapping the cellular biogeography of human bone marrow niches using single-cell transcriptomics and proteomic imaging. *Cell* **187**, 3120-3140.e29 (2024).

381. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**, 1452–1457 (2019).

382. Belhocine, K., DeMare, L. & Habern, O. Single-Cell Multiomics: Simultaneous Epigenetic and Transcriptional Profiling. *Genetic Engineering & Biotechnology News* **41**, 66–68 (2021).

383. Gill, J., Dasgupta, A., Manry, B. & Markuzon, N. Combining single-cell ATAC and RNA sequencing for supervised cell annotation. *BMC Bioinformatics* **26**, 67 (2025).

Bioinformatic software packages used by Monika Gońka in this thesis:

Software	Version	Reference
annotate	1.72.0	Gentry J, 2025, annotate: Annotation for microarrays. doi:10.18129/B9.bioc.annotate, R package, https://bioconductor.org/packages/annotate
AnnotationDbi	1.56.2	Pages H, Carlson M, Falcon S, Li N, 2022, AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor, https://bioconductor.org/packages/AnnotationDbi
AnnotationHub	3.2.2	Bioconductor Package Maintainer, M Morgan, M Carlson, D Tenenbaum, S Arora, V Oberchain, K Morrell, L Shepherd
ape	5.6.2	ape: Analyses of Phylogenetics and Evolution, Paradis E, Schliep K, 2019, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, Bioinformatics, doi:10.1093/bioinformatics/bty633
aplot	0.1.6	G Yu, 2022, aplot: Decorate a ggplot with Associated Information, https://CRAN.R-project.org/package=aplot
ArchR	1.0.2	Granja JM, Corces MR, Pierce SE, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis, Nat Genet. 2021, doi:10.1038/s41588-021-00790-6
beeswarm	0.4.0	A Eklund, J Trimble, 2021, beeswarm: The Bee Swarm Plot, an Alternative to Stripchart, R package, https://CRAN.R-project.org/package=beeswarm
Biobase	2.58.0, 2.54.0	Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oles AK, Pages H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M, 2015, Orchestrating high-throughput genomic analysis with Bioconductor, Nature

		Methods, http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html
BiocManager	1.30.22, 1.30.18	M Morgan, 2022, BiocManager: Access the Bioconductor Project Package Repository; R package, https://CRAN.R-project.org/package=BiocManager
Bioconductor	3.14	https://bioconductor.org
BiocSingular	1.10.0	Lun A, 2025, BiocSingular: Singular Value Decomposition for Bioconductor Packages, doi:10.18129/B9.bioc.BiocSingular, R package, https://bioconductor.org/packages/BiocSingular
biomaRt	2.50.3	Durinck S, Spellman P, Birney E, Huber W, 2009, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt, Nature Protocols
BioModel Analyser	1.14.0001	https://biomodelanalyzer.org, B Hall et al.
Biostrings	2.66.0, 2.62.0	Pages H, Aboyoun P, Gentleman R, DebRoy S, 2023, Biostrings: Efficient manipulation of biological strings, doi:10.18129/B9.bioc.Biostrings, R package, https://bioconductor.org/packages/Biostrings
BSgenome	1.66.3	Pages H, 2023, BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs, doi:10.18129/B9.bioc.BSgenome, R package, https://bioconductor.org/packages/BSgenome
BSgenome.Mmusculus.UCSC.mm10	1.4.3	Team TBD, 2021, BSgenome.Mmusculus.UCSC.mm10: Full genome sequences for Mus musculus (UCSC version mm10, based on GRCm38.p6), R package
CellRanger	2.0.0	10x Genomics Cell Ranger, 10xgenomics.com
CellRanger-ATAC	2.0.0	10x Genomics Cell Ranger, 10xgenomics.com
ChatGPT	3, 4	ChatGPT, OpenAI, https://chat.openai.com/
circlize	0.4.15	circlize implements and enhances circular visualization in

		R, Z Gu, L Gu, R Eils, M Schlesner, B Brors, Bioinformatics, 2014
ClueGO	2.5.10	ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, G Bindea, B Mlecnik, H Hackl, P Charoentong, M Tosolini, A Kirilovsky, W-H Fridman, F Pages, Z Trajanoski, J Galon
cluster	2.1.3	Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K, 2022, cluster: Cluster Analysis Basics and Extensions, R package
clusterProfiler	4.2.2	Yu G, Wang L, Han Y, He Q, 2012, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS: A Journal of Integrative Biology, doi: 10.1089/omi.2011.0118
colorspace	2.0.3	Zeileis A, Fisher JC, Hornik K, Ihaka R, McWhite CD, Murrell P, Stauffer R, Wilke CO, 2020, colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes, Journal of Statistical Software, doi: 10.18637/jss.v096.i01
Complex Heatmap	2.10.0	Gu, Z, 2016, Complex heatmaps reveal patterns and correlations in multidimensional genomic data, Bioinformatics
conda	4.9.2	https://docs.conda.io/en/latest/, https://github.com/conda
cowplot	1.1.1	wilkelab.org/cowplot, C O Wilke
Cytoscape	3.10.1	Cytoscape: Network Data Integration, Analysis, and Visualization in a Box, https://cytoscape.org
dbplyr	2.2.0	dbplyr: A dplyr Back End for Databases, H Wickham, M Girlich, E Ruiz, 2023, https://github.com/tidyverse/dbplyr, https://dbplyr.tidyverse.org/
DESeq2	1.34.0	Love M I, Huber W, Anders S, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology, 2014
devtools	2.4.3	H Wickham, J Hester, W Chang, J Bryan, 2021, devtools: Tools to Make Developing R Packages Easier,

		https://CRAN.R-project.org/package=devtools
dnet	1.1.7	Fang H, Gough J, 2014, The dnet approach promotes emerging research on cancer patient survival, Genome Medicine, doi:10.1186/s13073-014-0064-8, http://dx.doi.org/10.1186/s13073-014-0064-8.
DO.db	2.9	J Li, 2015, DO.db: A set of annotation maps describing the entire Disease Ontology, R package
DOSE	3.20.1	G Yu, L-G Wang, G-R Yan, Q-Y He, DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis, Bioinformatics, 2015
dplyr	1.1.3, 1.0.9	H Wickham, R Francois, L Henry, K Muller, 2022, dplyr: A Grammar of Data Manipulation, https://CRAN.R-project.org/package=dplyr
Enhanced Volcano	1.12.0	Blighe K, Rana S, Lewis M, 2022, EnhancedVolcano: Publication- ready volcano plots with enhanced colouring and labeling, https://github.com/kevinblighe/EnhancedVolcano
enrichplot	1.14.2	Yu G, 2022, enrichplot: Visualization of Functional Enrichment Result, https://yulab-smu.top/biomedical-knowledge-mining-book/
fgsea	1.20.0	G Korotkevich, V Sukhov, A Sergushichev, Fast gene set enrichment analysis, bioRxiv, 2019, doi:10.1101/060012
formattable	0.2.1	K Ren, K Russell, 2021, formattable: Create 'Formattable' Data Structures, R package, https://CRAN.R-project.org/package=formattable, https://github.com/renkun-ken/formattable
genefilter	1.76.0	R Gentleman, V J Carey, W Huber, F Hahne, 2021, genefilter: methods for filtering genes from high-throughput experiments, R package
geneplotter	1.72.0	R Gentleman, 2025, geneplotter: Graphics related functions for Bioconductor, doi:10.18129/B9.bioc.geneplotter, R package, https://bioconductor.org/packages/geneplotter

GenomeInfoDb	1.34.9, 1.30.1	Arora S, Morgan M, Carlson M, Pages H, 2022, GenomeInfoDb: Utilities for manipulating chromosome names, including modifying them to follow a particular naming style, https://bioconductor.org/packages/GenomeInfoDb
GenomicRanges	1.50.2, 1.46.1	Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V, 2013, Software for Computing and Annotating Genomic Ranges, PLoS Computational Biology, doi:10.1371/journal.pcbi.1003118
ggbeeswarm	0.6.0	E Clarke, S Sherrill-Mix, 2017, ggbeeswarm: Categorical Scatter (Violin Point) Plots, R package, https://CRAN.R-project.org/package=ggbeeswarm
ggforce	0.3.3	T L Pedersen, 2021, ggforce: Accelerating ggplot2, R package, https://CRAN.R-project.org/package=ggforce
ggfun	0.0.6	G Yu, S Xu, 2022, ggfun: Miscellaneous Functions for ggplot2, R package, https://CRAN.R-project.org/package=ggfun
ggnewscale	0.4.7	E Campitelli, 2022, ggnewscale: Multiple Fill and Colour Scales in ggplot2, https://CRAN.R-project.org/package=ggnewscale
ggplot2	3.4.4, 3.3.6	Wickham H, 2016, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org
ggplotify	0.1.0	G Yu, 2021, ggplotify: Convert Plot to grob or ggplot Object, R package, https://CRAN.R-project.org/package=ggplotify
ggrastr	1.0.1	V Petukhov, T van den Brand, E Biederstedt, 2021, ggrastr: Rasterize Layers for ggplot2, R package, https://CRAN.R-project.org/package=ggrastr
ggrepel	0.9.1	ggrepel.slowkow.com, K Slowikowski
ggridges	0.5.3	C O Wilke, 2021, ggridges: Ridgeline Plots in ggplot2, R package, https://CRAN.R-project.org/package=ggridges
ggtree	3.2.1	G Yu, Using ggtree to visualize data on tree-like

		structures, Current Protocols in Bioinformatics, 2020, doi: 10.1002/cpbi.96; G Yu, T Tsan-Yuk Lam, H Zhu, Y Guan, Two methods for mapping and visualizing associated data on phylogeny using ggtree, Molecular Biology and Evolution 2018, doi: 10.1093/molbev/msy194; G Yu, D Smith, H Zhu, Y Guan, T Tsan-Yuk Lam, ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data, Methods in Ecology and Evolution, 2017
ggupset	0.3.0	github.com/const-ae/ggupset, Constantin Ahlmann-Eltze
GO.db	3.14.0	M Carlson, 2021, GO.db: A set of annotation maps describing the entire Gene Ontology, R package
GOSemSim	2.20.0	Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S, 2010, GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics, doi: 10.1093/bioinformatics/btq064
gplots	3.1.3	G R Warnes, B Bolker, L Bonebakker, R Gentleman, W Huber, A Liaw, T Lumley, M Maechler, A Magnusson, S Moeller, M Schwartz, B Venables, 2022, gplots: Various R Programming Tools for Plotting Data, https://CRAN.R-project.org/package=gplots
gprofiler2	0.2.2	Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H, 2020, gprofiler2 - an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler, F1000Research, R package
graph	1.72.0	R Gentleman, E Whalen, W Huber and S Falcon, 2021, graph: graph: A package to handle graph data structures, R package
htmltools	0.5.2	J Cheng, C Sievert, B Schloerke, W Chang, Y Xie, J Allen, Posit Software, https://cran.r-project.org/web/packages/htmltools/index.html
htmlwidgets	1.5.4	R Vaidyanathan, Y Xie, JJ Allaire, J Cheng, C Sievert, K Russell, E Hughes, RStudio, https://cran.r-

		project.org/web/packages/htmlwidgets/index.html
igraph	1.3.1	Csardi G, Nepusz T, 2006, The igraph software package for complex network research, InterJournal, Complex Systems, https://igraph.org
IGV	2.16.1	J Robinson, H Thorvaldsdottir, W Winckler, M Guttman, E Lander, G Getz, J Mesirov, Integrative Genomics Viewer, Nature Biotechnology, 2011, https://software.broadinstitute.org/software/igv/
ipython	8.13.2	IPython Kernel for Jupyter, F Perez, B E Granger, IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, 2007, doi:10.1109/MCSE.2007.53, URL: https://ipython.org
IRanges	2.32.0, 2.28.0	Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V, 2013, Software for Computing and Annotating Genomic Ranges, PLoS Computational Biology, doi:10.1371/journal.pcbi.1003118
irlba	2.3.5	J Baglama, L Reichel, B W Lewis, 2021, irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices, R package, https://CRAN.R-project.org/package=irlba
Java	17.0.5	Java programming language, https://www.java.com/en/
jinja2	3.1.2	A Ronacher, https://jinja.palletsprojects.com/en/3.1.x/
jpeg	0.1.9	S Urbanek, https://github.com/s-u/jpeg
JupiterLab	5.1	Jupyter Notebooks - a publishing format for reproducible computational workflows, T Kluyver, B Ragan-Kelley, F Perez, B Granger, M Bussonnier, J Frederic, K Kelley, J Hamrick, J Grout, S Corlay, P Ivanov, D Avila, S Abdalla, C Willing, Jupyter Development Team, DOI:10.3233/978-1-61499-649-1-87, Ebook: Positioning and Power in Academic Publishing: Players, Agents and Agendas
knitr	1.39	Y Xie, 2022, knitr: A General-Purpose Package for Dynamic Report Generation in R; Y Xie, 2015, Dynamic

		Documents with R and knitr, 2nd edition, Chapman and Hall/CRC, ISBN 978-1498716963; Y Xie, 2014, knitr: A Comprehensive Tool for Reproducible Research in R. In V Stodden, F Leisch and R D Peng, editors, Implementing Reproducible Computational Research, Chapman and Hall/CRC, ISBN 978-1466561595
leiden	0.4.2	S T Kelly, 2020, leiden: R implementation of the Leiden algorithm, R package, https://github.com/TomKellyGenetics/leiden
limma	3.50.3	Ritchie M E, Phipson B, Wu D, Hu Y, Law C W, Shi W, Smyth G K, 2015, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Research
lmtest	0.9.40	A Zeileis, T Hothorn, 2002, Diagnostic Checking in Regression Relationships, R News, https://CRAN.R-project.org/doc/Rnews/
locfit	1.5.9.7	C Loader, 2023, locfit: Local Regression, Likelihood and Density Estimation, R package, https://CRAN.R-project.org/package=locfit
matplotlib	3.7.1	Matplotlib: A 2D Graphics Environment, J D Hunter, Computing in Science & Engineering, 2007
mgcv	1.8.40	Wood S N, 2011, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, Journal of the Royal Statistical Society; Wood S N, N Pya and B Saefken, 2016, Smoothing parameter and model selection for general smooth models (with discussion), Journal of the American Statistical Association; Wood S N, 2004, Stable and efficient multiple smoothing parameter estimation for generalized additive models, Journal of the American Statistical Association; Wood S N, 2017, Generalized Additive Models: An Introduction with R (2nd edition), Chapman and Hall/CRC; Wood S N, 2003, Thin-

		plate regression splines, Journal of the Royal Statistical Society
modelr	0.1.8	H Wickham, 2020, modelr: Modelling Functions that Work with the Pipe, R package, https://CRAN.R-project.org/package=modelr
nlme	3.1.157	J Pinheiro, D Bates and R Core Team, 2022, nlme: Linear and Nonlinear Mixed Effects Models, R package, https://CRAN.R-project.org/package=nlme; Pinheiro JC, Bates DM, 2000, Mixed-Effects Models in S and S-PLUS, Springer, New York, doi: 10.1007/b98882, https://doi.org/10.1007/b98882
numpy	1.24.3	Array programming with NumPy, Harris CR, Millman KJ, van der Walt SJ et al., Nature, 2020, DOI: 10.1038/s41586-020-2649-2
OmnipathR	3.2.8	Integrated intra- and intercellular signaling knowledge for multicellular omics analysis, D Turei, A Valdeolivas, L Gul, N Palacio-Escat, M Klein, O Ivanova, M Olbei, A Gabor, F Theis, D Modos, T Korcsmaros, J Saez-Rodriguez, Molecular Systems Biology, 2021, DOI = 10.15252/msb.20209923
org.Mm.eg.db	3.14.0	Carlson M, 2019, org.Mm.eg.db: Genome wide annotation for Mouse
pandas	2.0.1	Data Structures for Statistical Computing in Python, W McKinney, Python in Science Conf., 2010
patchwork	1.1.1	patchwork.data-imaginist.com/index.html, T L Pedersen
PCAtools	2.6.0	K Blighe, A Lun, 2021, PCAtools: PCAtools: Everything Principal Components Analysis, R package, https://github.com/kevinblighe/PCAtools
Perplexity	2025	Perplexity AI, Inc., https://www.perplexity.ai
pip	23.1.2	pip: the package installer for Python, https://pip.pypa.io/en/stable/
plotly	4.10.0	Plotly, open source https://dash.plotly.com
plyr	1.8.9,	Wickham H, 2011, The Split-Apply-Combine Strategy for

	1.8.7	Data Analysis, Journal of Statistical Software, https://www.jstatsoft.org/v40/i01/
png	0.1.7	Simon Urbanek, https://cran.r-project.org/web/packages/png/index.html
purrr	1.0.2, 0.3.4	Wickham H, Henry L, 2023, purrr: Functional Programming Tools, R package, https://github.com/tidyverse/purrr, https://purrr.tidyverse.org/
Python	3.11.3	Python, programming language, https://www.python.org
qvalue	2.26.0	J D Storey, A J Bass, A Dabney and D Robinson, 2021, qvalue: Q-value estimation for false discovery rate control, R package, http://github.com/jdstorey/qvalue
R	4.1.2	R Core Team, 2021, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
RANN	2.6.1	S Arya, D Mount, S E Kemp and G Jefferis, 2019, RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric, R package, https://CRAN.R-project.org/package=RANN
RColorBrewer	1.1.3	E Neuwirth, 2022, RColorBrewer: ColorBrewer Palettes, R package, https://CRAN.R-project.org/package=RColorBrewer
RCurl	1.98.1.6	D T Lang, 2022, RCurl: General Network (HTTP/FTP/...) Client Interface for R, R package, https://CRAN.R-project.org/package=RCurl
RCy3	2.14.2	Gustavsen A J, Pai S, Isserlin R, Demchak B, Pico R A, 2019, RCy3: Network Biology using Cytoscape from within R, F1000Research, doi:10.12688/f1000research.20887.3
readr	2.1.4, 2.1.2	H Wickham, J Hester, R Francois, J Bryan, S Bearrows, Posit, https://github.com/mandreyel/, J Jylanki, M Jorgensen, https://cran.r-project.org/web/packages/readr/index.html
readxl	1.4.0	H Wickham and J Bryan, 2022, readxl: Read Excel Files,

		https://CRAN.R-project.org/package=readxl
reshape2	1.4.4	H Wickham, 2007, Reshaping Data with the reshape Package, Journal of Statistical Software, http://www.jstatsoft.org/v21/i12/
reticulate	1.28	Ushey K, Allaire J, Tang Y, 2023, reticulate: Interface to Python, https://rstudio.github.io/reticulate/, https://github.com/rstudio/reticulate
Rgraphviz	2.38.0	K D Hansen, J Gentry, L Long, R Gentleman, S Falcon, F Hahne and D Sarkar, 2021, Rgraphviz: Provides plotting capabilities for R graph objects, R package
rjson	0.2.21	A Couture-Beil, 2022, rjson: JSON for R, R package, https://CRAN.R-project.org/package=rjson
RJSONIO	1.3.1.8	D Temple Lang, J Wallace, 2023, RJSONIO: Serialize R Objects to JSON, JavaScript Object Notation, R package, https://CRAN.R-project.org/package=RJSONIO
rmarkdown	2.14	JJ Allaire, Y Xie, J McPherson, J Luraschi, K Ushey, A Atkins, H Wickham, J Cheng, W Chang, R Iannone, 2022, rmarkdown: Dynamic Documents for R, R package, URL https://rmarkdown.rstudio.com; Y Xie, J J Allaire, G Grolemund, 2018, R Markdown: The Definitive Guide. Chapman and Hall/CRC, ISBN 9781138359338, https://bookdown.org/yihui/rmarkdown; Y Xie, C Dervieux, E Riederer, 2020, R Markdown Cookbook, Chapman and Hall/CRC, ISBN 9780367563837, https://bookdown.org/yihui/rmarkdown-cookbook
ROCR	1.0.11	Sing T, Sander O, Beerenwinkel N, Lengauer T, 2005, ROCR: visualizing classifier performance in R, Bioinformatics, http://rocr.bioinf.mpi-sb.mpg.de
rpart	4.1.16	T Therneau, B Atkinson, 2022, rpart: Recursive Partitioning and Regression Trees, R package, https://CRAN.R-project.org/package=rpart
RSQLite	2.2.14	K Muler, H Wickham, D A James, S Falcon, 2022, RSQLite: SQLite Interface for R, R package,

		https://CRAN.R-project.org/package=RSQLite
RStudio	1.4.1717	RStudio Team, 2020, RStudio: Integrated Development for R. RStudio, PBC, Boston, http://www.rstudio.com/; Posit Team (2025). RStudio: Integrated Development Environment for R, Posit Software, PBC, Boston, MA. URL http://www.posit.co/
rsvd	1.0.5	Erichson NB, Voronin S, Brunton SL, Kutz JN, 2019, Randomized Matrix Decompositions Using R, Journal of Statistical Software, doi: 10.18637/jss.v089.i11, https://doi.org/10.18637/jss.v089.i11
Rtsne	0.16	L J P van der Maaten, G E Hinton, Visualizing High-Dimensional Data Using t-SNE, Journal of Machine Learning Research, 2008; L J P van der Maaten, Accelerating t-SNE using Tree-Based Algorithms, Journal of Machine Learning Research, 2014; J H Krijthe, 2015, Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation, https://github.com/jkrijthe/Rtsne
rvest	1.0.2	H Wickham, 2021, rvest: Easily Harvest (Scrape) Web Pages, R package, https://CRAN.R-project.org/package=rvest
S4Vectors	0.36.2, 0.32.4	Pages H, Lawrence M, Aboyoun P, 2023, S4Vectors: Foundation of vector-like and list-like containers in Bioconductor, doi:10.18129/B9.bioc.S4Vectors, R package, https://bioconductor.org/packages/S4Vectors
scales	1.2.1	H Wickham, D Seidel, 2022, scales: Scale Functions for Visualization, R package, https://CRAN.R-project.org/package=scales
scatterpie	0.1.7	G Yu, 2021, scatterpie: Scatter Pie Plot, R package, https://CRAN.R-project.org/package=scatterpie
scipy	1.10.1	https://docs.scipy.org/doc/scipy/
sctransform	0.3.3	Hafemeister C & Satija R, Normalization and variance stabilization of single-cell RNA-seq data using regularized

		negative binomial regression, Genome Biol, 2019
seaborn	0.12.2	seaborn: statistical data visualization, M Waskom, Journal of Open Source Software, 2021, doi = 10.21105/joss.03021
Seurat	4.4.0, 4.1.1	satijalab.org/seurat/index.html, Hao et al. Integrated analysis of multimodal single-cell data, Cell, 2021, [Seurat V4]
SeuratObject	5.0.0, 4.1.0	R Satija, A Butler, P Hoffman, T Stuart, 2022, SeuratObject: Data Structures for Single Cell Data, R package, https://CRAN.R-project.org/package=SeuratObject
splines	4.1.2	R Core Team, 2021, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
STAR	2.7	Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNAseq aligner, Bioinformatics, 2013, doi:10.1093/bioinformatics/bts635
stats	4.1.2	R Core Team, 2021, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
stats4	4.1.2	R Core Team, 2021, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
stringi	1.7.6	Gagolewski M, 2021, stringi: Fast and portable character string processing in R, Journal of Statistical Software, R package, https://stringi.gagolewski.com/
stringr	1.5.0	Wickham H, 2023, stringr: Simple, Consistent Wrappers for Common String Operations, https://stringr.tidyverse.org, https://github.com/tidyverse/stringr
Summarized Experiment	1.28.0, 1.24.0	Morgan M, Obenchain V, Hester J, Pages H, 2023, SummarizedExperiment: SummarizedExperiment container, doi:10.18129/B9.bioc.SummarizedExperiment,

		R package, https://bioconductor.org/packages/SummarizedExperiment
tibble	3.2.1, 3.1.7	Muller K, Wickham H, 2023, tibble: Simple Data Frames, https://tibble.tidyverse.org/, https://github.com/tidyverse/tibble
tidygraph	1.2.1	T L Pedersen, 2022, tidygraph: A Tidy API for Graph Manipulation, R package, https://CRAN.R-project.org/package=tidygraph
tidyr	1.3.0, 1.2.0	Wickham H, Vaughan D, Girlich M, 2023, tidyr: Tidy Messy Data, R package, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org
tidytree	0.3.9	G Yu, 2022, tidytree: A Tidy Tool for Phylogenetic Tree Data Manipulation, R package, https://CRAN.R-project.org/package=tidytree
tidyverse	1.3.1	Wickham H, Averick M, Bryan J, Chang W, McGowan LD, Francois R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Muller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H, 2019, Welcome to the tidyverse, Journal of Open Source Software, doi: 10.21105/joss.01686
treeio	1.18.1	LG Wang, TTY Lam, S Xu, Z Dai, L Zhou, T Feng, P Guo, CW Dunn, BR Jones, T Bradley, H Zhu, Y Guan, Y Jiang, G Yu, treeio: an R package for phylogenetic tree input and output with richly annotated and associated data, Molecular Biology and Evolution, 2020, doi: 10.1093/molbev/msz240
UpSetR	1.4.0	N Gehlenborg, 2019, UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets, R package, https://CRAN.R-project.org/package=UpSetR
uwot	0.1.11	J Melville, 2021, uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for

		Dimensionality Reduction, R package, https://CRAN.R-project.org/package=uwot
VennDiagram	1.7.3	H Chen, https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf
vipor	0.4.5	S Sherrill-Mix, E Clarke, 2017, vipor: Plot Categorical Data Using Quasirandom Noise and Density Estimates, R package, https://CRAN.R-project.org/package=vipor
viridis	0.6.2	S Garnier, N Ross, R Rudis, A P Camargo, M Sciaini, C Scherer, 2021, Rvision - Colorblind-Friendly Color Maps for R, R package
VSCode	1.67.2	VisualCodeStudio, Microsoft, https://code.visualstudio.com
werkzeug	2.3.4	https://werkzeug.palletsprojects.com/en/3.0.x/
xfun	0.31	Y Xie, 2022, xfun: Supporting Functions for Packages Maintained by Y Xie, R package, https://CRAN.R-project.org/package=xfun
xlrd	2.0.1	J Machin, https://pypi.org/project/xlrd/1.1.0/
xlsxwriter	3.1.2	J McNamara, https://github.com/jmcnamara/XlsxWriter
XML	3.99.0.9	D T Lang, 2022, XML: Tools for Parsing and Generating XML Within R and S-Plus, R package, https://CRAN.R-project.org/package=XML
xml2	1.3.3	H Wickham, J Hester and J Ooms, 2021, xml2: Parse XML, R package, https://CRAN.R-project.org/package=xml2
xtable	1.8.4	D B Dahl, D Scott, C Roosen, A Magnusson, J Swinton, 2019, xtable: Export Tables to LaTeX or HTML, R package, https://CRAN.R-project.org/package=xtable
yaml	2.3.5, 0.2.5	S P Garbett, J Stephens, K Simonov, Y Xie, Z Dong, H Wickham, J Horner, Reikoch, W Beasley, B OConnor, G R Warnes, M Quinn,Z N Kamvar,github.com/vubiostat/r-yaml