Characterising The Functional Role of GATA1short Isoform in Normal and Malignant Haematopoiesis



David Cruz Hernandez

Green Templeton College

Weatherall Institute of Molecular Medicine

University of Oxford

A Thesis Submitted for The Degree of

Doctor of Philosophy

in

Medical Sciences

Michaelmas 2021

Acknowledgments

First and foremost, I would like to express my gratitude to all patients and their families who donated samples for this project, without whom none of this research would have been possible. I would like to thank the Lady Tata Memorial Trust for providing funding for this DPhil project. I also want to thank my supervisor Paresh Vyas not only for welcoming me in his lab at a crucial time in my life but also for providing mentorship on how to be a better scientist. I want to thank him for nurturing my creativity and scientific curiosity. I wish to acknowledge the technical help provided by all members of the Vyas group in this research project. Specially to Marlen Metzner who directly helped me with next generation sequencing and analysis. I wish to extend my special thanks to Mirian Angulo Salazar and Batchimeg Usukhbayar for organizing, processing, and bio-banking patient samples. I would also like to thank my co-supervisor Irene Roberts for her scientific guidance and support on this project. Finally, I would like to thank my family for their unconditional support, whom I owe my deepest gratitude.

In memory of all those lives tragically lost to the COVID-19 pandemic.

Abstract

ML-DS and its precursor disease TAM are genetically simple models of acute myeloid leukaemia to study how two or more genetic changes synergize to induce maturation arrest and leukaemic transformation. Both TAM and ML-DS are driven by mutations in the gene encoding the X-linked erythroid-megakaryocyte transcription factor GATA1 that result in truncation of 83 N-terminal amino acids to produce GATA1short (GATA1s). Through single cell RNA sequencing of trisomy 21 haematopoietic primitive, progenitor and precursor cells derived from cord blood, I identified 5 cell clusters with robust *GATA1* expression that were named: EMMPPs, cycling-EMMPPs, megakaryocyte progenitor, erythroid progenitor, and basophil/eosinophil/mast cell progenitor. I showed that in TAM, *GATA1s* mutations resulted in aberrant gene expression in all GATA1 expressing cell compartments and led to maturation defects of the erythroid and megakaryocytic lineages. These defects were most likely explained by sustained expression of *GATA2*, *RUNX1* and *NFE2* among other genes. I optimized and employed a CRISPR-Cas9 system to recreate *GATA1s* mutation and showed that the erythroid maturation defect and sustained GATA2 expression in TAM can be recapitulated *in vitro*. This experimental approach prioritised cell compartments where trisomy 21 and GATA1s have the greatest synergy and would allow me to determine the contribution of trisomy 21 to TAM and ML-DS. Moreover, I optimized CUT&TAG to specifically profile genome wide GATA1 biding sites in low cell number. I showed biding of GATA1s at the -77kb enhancer of GATA2 in an ML-DS derived cell line. We hypothesize that sustained *GATA2* expression in TAM and in ML-DS might be driven by GATA1s biding at the -77kb GATA2 enhancer. Finally, I developed a highly sensitive flow cytometric assay to detect wild type cells expressing both GATA1 isoforms and mutant TAM and ML-DS cells with exclusive expression of GATA1s. I envision this assay can substitute for next generation sequencing of *GATA1* mutational profiling when not available to help diagnose TAM and ML-DS in clinical settings. Word count: 36599

Table of Contents

TABLE OF CONTENTS **5**

INTRODUCTION **8**

1.1 GATA1 IN NORMAL HAEMATOPOIESIS 15

1.1.1 GATA1 Does Not Have an Intrinsic Role in Haematopoietic Stem Cells Maintenance *15*

1.1.2 Lineage Specification Occurs Early and Generates a GATA1+ and GATA1- Primed Multipotential Progenitor *16*

1.1.3 GATA1 Maintains Immature Cells of The Erythroid and Megakaryocyte Lineage In Sufficient But Not Excessive Numbers *20*

1.1 GATA1 STRUCTURE AND FUNCTION 21

1.1.4 Alternative splicing of GATA1 Locus *21*

1.1.5 GATA1 Functional Domains *22*

1.2 GATA1SHORT IN MALIGNANT HAEMATOPOIESIS 25

1.2.1 Trisomy 21 Alters Haematopoietic Stem Cell Homeostasis *27*

1.2.2 GATA1s in TAM and ML-DS *28*

1.2.3 Transforming mutation in TAM to ML-DS *31*

1.3 DPHIL PROJECT OBJECTIVES 32

MATERIALS AND METHODS **33**

1.4 CELL LINE CULTURE 33

1.5 IMMUNOPHENOTYPING OF TRISOMY 21 CORD BLOOD, DS NEONATES, TAM AND ML-DS SAMPLES 33

1.5.1 Sample Processing *33*

1.5.2 Flow cytometry and Intracellular Flow Cytometry *34*

1.6 CRISPR-CAS9 OF PRIMARY CD34+ CELLS AND K562 CELLS 35

1.6.1 CD34+ Cell Enrichment and Colony Forming Assay *35*

1.6.2 Sample Processing *36*

1.6.3 DNA Isolation, Genotyping of Insertions and Deletions and Sanger Sequencing *37*

1.7 GENOTYPING FOR GATA1S MUTATIONS IN TAM AND ML-DS AND TRISOMY 21 CORD BLOOD SAMPLES 38

1.7.1 DNA Isolation *38*

1.7.2 Library Preparation and Next Generation Sequencing *38*

1.8 Cleavage Under Targets and Tagmentation 38

1.8.1 Sample Processing *38*

1.8.2 DNA Isolation *40*

1.8.3 Initial PCR Amplification of DNA Fragments *40*

1.8.4 PCR Cycle Number Determination Based on qPCR *40*

1.8.5 Library Preparation and Next Generation Sequencing *41*

1.9 Single Cell RNA sequencing 41

1.9.1 Sample Processing *41*

1.9.2 Library Preparation *42*

1.9.3 Next Generation Sequencing *42*

1.10 Bioinformatic Analysis 42

1.10.1 GATA1 Locus Genotyping DATA Processing *42*

1.10.2 CUT&TAG Data Processing *43*

1.10.3 Single Cell RNA Sequencing: Quality Control *43*

1.10.4 Single Cell RNA Sequencing: Low-Dimension Embedding *44*

1.10.5 Single Cell RNA Sequencing: Cluster Annotation *44*

1.10.6 Single Cell RNA Sequencing: Creating a Reference and Query Mapping *45*

DIAGNOSIS OF TAM AND ML-DS BY INTRACELLULAR DETECTION OF GATA1S 45

1.11 Dual Immuno-Staining of GATA1 N' and C' Termini Can Identify Cells with Exclusive GATA1s Expression 47

1.12 Intracellular Detection of GATA1s is Comparable to NGS Detection of *GATA1s* Mutations 50

1.13 Cell Surface Expression Alone is Insufficient to Detect TAM and ML-DS Cells 58

1.14 Intracellular Detection of GATA1s as a MRD Tool: A Proof of Principle 61

1.15 Intracellular Detection of GATA1s in TAM and ML-DS is Highly Sensitive and Specific 64

SINGLE CELL TRANSCRIPTIONAL REFERENCE MAP OF T21 HAEMATOPOIESIS 69

1.16 Integration of 4 T21 Transcriptomes of Haematopoietic Samples 70

1.17 Single Cell Cluster Identification 75

1.18 Query Mapping of Transient Abnormal Myelopoiesis onto Trisomy 21 Haematopoiesis 91

1.19 Identification of *GATA1s* Mutations with VarTrix in a Cell Mix Experiment 95

1.20 Variant Calling of *GATA1s* mutations with VarTrix in TAM cells 101

TRANSCRIPTIONAL AND PHENOTYPIC CONSEQUENCES OF GATA1SHORT EXPRESSION 104

1.21 Differential Gene Expression Between Trisomy 21 *GATA1* Wild Type Cells and Trisomy 21 *GATA1s* TAM Cells 106

1.21.1 GATA1s TAM Cells are Marked by a Defective Megakaryocytes 110

1.21.2 GATA1s TAM Cells are Marked by a Defective Erythropoiesis 111

1.21.3 Functional Consequences of GATA1s Mutations in Terminal Erythropoiesis 114

GENOME-WIDE IDENTIFICATION OF GATA1 BIDING SITES BY CUT&TAG 123

1.22 Optimization of CUT&TAG For GATA1 N' and C' Terminus In Low Cell Numbers 124

1.23 C&T Detect: A qPCR Based Method to Quantify Known Transcription Factor Binding Sites 126

1.24 CUT&TAG Bioinformatic Analysis and Quality Control 131

1.25 GATA1s Binds at the -77 GATA2 Enhancer in an ML-DS Cell line 137

DISCUSSION 140

CONCLUSION 145

REFERENCES 146

APPENDIX I. SUPPLEMENTARY FIGURES 155

APPENDIX II. SUPPLEMENTARY TABLES 161

APPENDIX III. PUBLICATION OBTAINED DURING MY DPHIL PROJECT 168

List of Figures

FIGURE 1.1 SCHEMATIC REPRESENTATION OF GATA1 EXPRESSION IN THE HAEMATOPOIETIC HIERARCHY 19

FIGURE 1.2 *GATA1* TRANSCRIPT PROCESSING AND STRUCTURE. 22

FIGURE 1.3 GATA1 AMINO ACID SEQUENCE CONSERVATION 23

FIGURE 1.4 STRUCTURE OF GATA1 N' TERMINAL TAD DOMAIN. 25

FIGURE 1.5 MULTISTEP MODEL OF MYELOID PRE-LEUKAEMIA AND LEUKAEMIA IN DOWN SYNDROME 27

FIGURE 3.1 SCHEMATIC REPRESENTATION OF DUAL IMMUNOSTAINING OF GATA1 N' AND C' TERMINUS 48

FIGURE 3.2 DUAL IMMUNOSTAINING AGAINST GATA1 IN A CELL MIX EXPERIMENT 49

FIGURE 3.3 SCHEMATIC REPRESENTATION OF THE EXPERIMENTAL APPROACH 51

FIGURE 3.4 MULTIPARAMETER FLOW CYTOMETRIC PANEL 55

FIGURE 3.5 SIMPLE LINEAR REGRESSION BETWEEN NGS, IFC AND FC 57

FIGURE 3.6 GATA1S^{ONLY} CELLS IN TAM AND ML-DS CO-LOCATE IN THE CD45LOWCD117+ GATE 59

FIGURE 3.7 CD7 IS EXPRESSION IN TAM AND ML-DS AS WELL AS GATA1^{WT} CELLS. 60

FIGURE 3.8 IMMUNOPHENOTYPING OF ML-DS CELLS BEFORE AND AFTER CHEMOTHERAPY 64

FIGURE 3.9 RECEIVER OPERATING CHARACTERISTIC ANALYSIS OF CONVENTIONAL FC AND IFC METHODS 66

FIGURE 3.10 VALIDATION OF THE IFC SENSITIVITY. 68

FIGURE 4.1 QUALITY CONTROL OF T21 SINGLE CELL RNA SEQUENCING 71

FIGURE 4.2 HARMONY INTEGRATION OF 4 T21 DONOR SAMPLES. 73

FIGURE 4.3 A UNIFIED TRANSCRIPTIONAL LANDSCAPE OF T21 SINGLE CELLS COMPOSED OF 20 CELL CLUSTERS VISUALIZED IN UMAP SPACE 74

FIGURE 4.4 IDENTIFICATION OF MATURE T, B, NK CELLS AS WELL AS MONOCYTE AND DENDRITIC CELLS BY CANONICAL MARKERS 75

FIGURE 4.5 HEATMAP GENERATED USING THE PHEATMAP LIBRARY PACKAGE FOR CLUSTER IDENTIFICATION 77

FIGURE 4.6 UCELL GENE SET SCORING FOR CLUSTER ANNOTATION 78

FIGURE 4.7 IDENTIFICATION OF PROGENITOR, PRECURSOR, AND MATURE CELL BY EXPRESSION DENSITY OF CD34 AND KIT 79

FIGURE 4.8 HSC/MPPS ARE MARKED BY *KLF2, FLT3, HLF* AND *AVP* EXPRESSION 80

FIGURE 4.9 HEATMAP SHOWING THE TOP 30 GENES WITH THE HIGHEST LOG2 FOLD CHANGE IN THE MOST IMMATURE CELL COMPARTMENTS.....82

FIGURE 4.10 HEATMAP SHOWING THE TOP 30 GENES WITH THE HIGHEST LOG2 FOLD CHANGE IN THE GATA1+ IMMATURE CELL COMPARTMENTS.....84

FIGURE 4.11 GSEA ANALYSIS TO DETERMINE THE IDENTITY OF CLUSTER 6.....86

FIGURE 4.12 HEATMAP SHOWING THE TOP 30 GENES WITH THE HIGHEST LOG2 FOLD CHANGE DURING ERYTHROPOIESIS.....88

FIGURE 4.13 GENE AND PROTEIN EXPRESSION OF GATA1 T21 MYELOPOIESIS.....90

FIGURE 4.14 *GATA1* AND *GATA2* EXPRESSION IN MEGAKARYOPOIESIS, EO/BA/ MAST CELL MATURATION, AND ERYTHROPOIESIS.....91

FIGURE 4.15 QUALITY CONTROL OF A TAM SAMPLE 92

FIGURE 4.16 QUERY MAPPING OF TAM CELLS ONTO SINGLE CELL TRISOMY 21 HAEMATOPOIESIS.....94

FIGURE 4.17 A CELL MIX EXPERIMENT TO TEST VARTRIX 96

FIGURE 4.18 VARTRIX CAN IDENTIFY *GATA1S* MUTATIONS IN SINGLE CMKS CELLS 98

FIGURE 4.19 VARTRIX ASSIGNS AMBIGUOUS GENOTYPES TO A FRACTION OF MUTANT CMKS 100

FIGURE 4.20 VARTRIX CAN IDENTIFY GATA1S MUTATION IN TAM AND FIND CELLULAR HETEROGENEITY OF TAM 102

FIGURE 4.21 IDENTIFICATION OF GATA*1S* MUTATIONS IN SINGLE CELLS DEPEND ON THE LEVEL OF *GATA1* EXPRESSION 104

FIGURE 5.1 TRISOMY 21 AND GATA1S MUTANT TAM CELLS 105

FIGURE 5.2 DIFFERENTIAL GENE EXPRESSION.....106

FIGURE 5.3 UPREGULATED GENES IN *GATA1S* MUTANT TAM CELLS.....109

FIGURE 5.4 *HPSE2* AND *TGB1* OVER EXPRESSION IN TRISOMY 21 HAEMATOPOIESIS AND IN TAM 110

FIGURE 5.5 DOWNREGULATED GENES IN TAM *GATA1S* MUTANT CELLS.....112

FIGURE 5.6 ERYTHROID MATURATION ARREST – DOWNREGULATED GENE EXPRESSION 113

FIGURE 5.7 ERYTHROID MATURATION ARREST – UPREGULATED GENE EXPRESSION 114

FIGURE 5.8 *GATA1, GATA2* AND *RUNX1* EXPRESSION IN T21 AND TAM ERYTHROPOIESIS 117

FIGURE 5.9 *GATA1, GATA2* AND *RUNX1* EXPRESSION IN T21 AND TAM MEGAKARYOPOIESIS 118

FIGURE 5.10 GENE ENGINEERING OF GATA1S MUTATIONS 120

FIGURE 5.11 ERYTHROID MATURATION ARREST AND SUSTAINED GATA2 EXPRESSION IN GATA1^{WT} TRISOMY 21 AND GATA1S TAM CELLS 121

FIGURE 5.12 HYPERPROLIFERATIVE BLAST COLONIES DUE TO GATA1S MUTATIONS IN DISOMIC CELLS 123

FIGURE 6.1 CUT&TAG OPTIMIZATION 128

FIGURE 6.2 PCR CYCLE OPTIMIZATION OF CUT&TAG DNA YIELD BY QPCR 130

FIGURE 6.3 CUT&TAG PROFILES GENOME WIDE H3K27ME3 EPIGENETIC MARKS IN LOW CELL NUMBERS 134

FIGURE 6.4 CUT&TAG PROFILES GENOME WIDE GATA1 BINDING SITES IN LOW CELL NUMBERS. 135

FIGURE 6.5 TOP 10 MOTIFS IDENTIFIED BY MEME IN GATA1 CUT&TAG 136

FIGURE 6.6. GATA1 BINDING AT THE GATA1 LOCUS IDENTIFIED BY CUT&TAG 138

FIGURE 6.7 GATA1S BIDING AT A DISTAL GATA2 ENHACER SITE IN CMKS 139

List of Tables

TABLE 1. FLOW CYTOMETRY AND INTRACELLULAR FLOW CYTOMETRY ANTIBODIES FOR TAM AND ML-DS 161

TABLE 2. NGS GATA1 PRIMER SETS FOR GENOTYPING TAM AND ML-DS SAMPLES 161

TABLE 3. CRISPR CAS9 SGRNAS FOR GATA1 EXON 2 DELETION AND PCR AMPLIFICATION 161

TABLE 4. C&T DETECT AND CUT&TAG PRIMER SEQUENCES 162

TABLE 5 HAEMATOPOIETIC SPECIFIC LINEAGE GENE SIGNATURES FROM ROY A. ET AL[12] 163

TABLE 6. PATIENT SAMPLES, GENOTYPING AND FLOW CYTOMETRIC FEATURES 164

TABLE 7. RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS 165

TABLE 8. FACS SORTING PANEL FOR SINGLE CELL RNA SEQUENCING 166

TABLE 9. FACS CELL SORTING FOR SINGLE CELL RNA SEQUENCING 166

TABLE 10. GENOTYPING FREQUENTING PER CELL CLUSTER TAM 167

List of Abbreviations

AA: Amino Acid....21

AF647: Alexa Flour 647....47

AMKL: Acute Megakaryoblastic Leukaemia....26

AML: Ccute Myeloid lLeukaemia....13

AUC: Under The Curve....44

BM: Bone Marrow....33

C': Carboxyl Terminus....22

CMP : Common Myeloid Progenitor....18

CUT&Tag: Cleavage Under Targets and Tagmentation....33

DMSO: Dimethyl Sulfoxide....33

DS: Down Syndrome....13

EMPP : Erythroid-Megakaryocyte Primed Multipotential Progenitor....18

FACS: Fluorescent Activated Cell Sorting....35

FBS: Foetal Bovine Serum....33

FL:full-length....21

FMOs: Fluorescent -minus-one....35

GMP: Granulocyte-Macrophage Progenitor....18

HS: Heparan Suphate....107

HSC: Hematopoietic Stem Cell....14

HSPCs: Hematopoietic Stem and Progenitor Cells....29

iFC: Intracellular Flow Cytometry....47

IMDM:Iscove's Modified Dulbecco's Medium....33

iPSC: Induced Pluripotent Stem Cell....28

LMPP: Lymphoid-Primed Multi-Potential Progenitor....18

MEP: : Erythroid-Megakaryocyte Progenitor....17

mL: Milli Liters....34

ML-DS: Myeloid Leukaemia of Down Syndrome....13

MNCs: Mono Nuclear Cells....34
MPP: Multipotent Potent Progenitor....15
MRD: Measurable Residual Disease....46
mRNAs: messenger Ribo Nucleic Acid....21
N': AminoTerminus....22
ng: nanograms....36
NGS: Next-Generation DNA sequencing....46
PB: Peripheral Blood....33
PBS: Phosphate Buffer Saline....33
PCA: Principal Components Analysis....44
PCR: Polymerase Chain Reaction....37
PE: : R-Phycoerythrin....47
PPV: Positive Predictive Value....65
pRB: Retinoblastoma Protein....25
qPCR: Quantitative PCR....48
RNP: RiboNucleo Protein....36
T21: Trisomy 21....13
TAD: Transactivation Domain....22
TAM: Transient Abnormal Myelopoiesis....13
TFs: Transcription Factors....14
ug: Microgram....33
ul: microliter....34
UMAP: Uniform Manifold Approximation and Projection....44
VAF: Variant Allele Frequency....56
ZF: Zinc Finger DNA Binding Domain....23

1 Introduction

Each cell of the 50 trillion cells in the human body contains 2 meters of DNA. If you were to add up all the DNA in your body, it comes to 100 trillion meters of DNA, more than enough to knit a jumper. Good luck with the tangling! To avoid the tangling, our 2 meters of DNA are condensed into chromosomes, a series of colourful yarns that are correctly packed inside the nucleus of each cell. There are 23 pairs of chromosomes per cell to be exact. Chromosome 21 is the smallest of the DNA yarns, yet it contains more than 300 genes. In rare circumstances, instead of the usual two yarns for chromosome 21, babies are born with an extra one. We refer to this phenomenon as trisomy 21 (T21) or Down Syndrome, a genetic disorder characterized by intellectual disability, heart disease and a staggering propensity to develop blood cancer.

Young children with Down Syndrome (DS) have a 20-fold increased risk of developing acute lymphoblastic leukaemia and 150-fold increased risk of developing acute myeloid leukaemia (AML) compared to children without Down Syndrome[1,2] . This type of leukaemia with erythroid and megakaryoblastic features is known as myeloid leukaemia of Down Syndrome (ML-DS) and it is often preceded by a temporally restricted, preleukaemic phase referred as Transient Abnormal Myelopoiesis or TAM. In the early 2000's, it became clear that TAM and ML-DS cells were linked by acquired mutations in the haematopoietic transcription factor GATA1[3-6]. Although, the altered DNA sequence in the *GATA1* gene varied, virtually all mutations gave rise to exclusive expression of a N'-terminal truncated GATA1 protein known as GATA1short or GATA1s. Now we know that during fetal life, up to 30% of babies with Down Syndrome will acquire *GATA1s* mutations[7] that will cooperate with trisomy 21 to trigger the first steps in the development of acute myeloid leukaemia.

Acute myeloid leukaemia, like most blood cancers arises from multiple genetic changes that are gradually acquired in the blood cells during one's lifetime. Most of these

genetic changes have unnoticeable effects. Occasionally, some of them will disrupt normal patterns of gene expression to alter cellular maturation, proliferation and self-renewal to induce advantageous pre-leukaemic phenotypes[8]. This expanded clonal population of pre-leukaemic cells have an increased probability of acquiring secondary driver mutations. This in turn, will make them more fit than their normal counterparts. Eventually in adult AML, 5 recurrent genetic changes[9] (range: 2-10 mutations) are required to transform blood progenitor cells into acute leukaemia. In paediatric ML-DS, only three recurrent genetic changes are required: constitutional Trisomy 21, a *GATA1s* mutation and most often a single allele mutation in genes encoding the cohesin complex[10]. It is the sum of all these genetic changes that results in absence of normal blood production and an excessive proliferation of immature myeloid cells blocked in cellular maturation.

A halt in normal cell maturation due to acquired genetic changes is one important hallmark of acute myeloid leukaemia and probably of most blood cancers[11]. The maturation of blood cells is a result of a progressive restriction of a haematopoietic stem cell's ability to self-renew. This progressive restriction of self-renewal is accompanied by an attainment of lineage specific patterns of gene expression. Therefore, haematopoietic stem cells (HSC) give rise to a series of committed progenitor cells with limited self-renewal that can expand and further differentiate into an eventual non-dividing cell type of our blood system, such as red blood cells or platelets[12]. It has been estimated that the turnover of the blood system is about 1 trillion cells per day[13]. This remarkable continuous demand of blood cells can only be achieved by a highly regulated process of cell maturation. With their unique ability to read the DNA code and interpret the genome, as well as their ability to respond to environmental stimuli, transcription factors are key determinants of this process[14]. Consequently, the dysregulation of these transcription factor networks is emerging as the culprit of leukaemogenesis[14,15].

1.1 GATA1 in Normal Haematopoiesis

Cellular maturation from haematopoietic stem cells (HSC) is characterized by progressive loss of self-renewal, a constraint of multi lineage potential and an attainment of lineage specific patterns of gene expression. This phenomenon has been attributed in part to the concerted activity of tissue specific transcription factors (TFs) and general transcription factors, which respond to environmental cues such as cytokines[16], to ultimately read the DNA code to specify patterns of gene expression. Transcription factors can be classified in a simplified way, according to their function: i) those that are involved in HSC development and maintenance such as SCL[17,18], RUNX1[19] and GATA2[20]; ii) those that coordinate specification of blood lineages such as KLF1 and FLI1 for the erythroid and megakaryocytic lineages respectively[21]; iii) those that maintain immature already committed precursor cells in sufficient numbers such as Gfi1 for the neutrophil lineage[22]; and iv) those that determine terminal differentiation or mature cell function such as NF-E2 for pro-platelet formation in megakaryocytes. This seemingly simple classification of transcription factor function is not absolute as a single transcription factor can play different and fundamental roles at any time during blood cell ontogeny and function. Exemplified by GATA3, whose expression is high in haematopoietic stem cells, where in part, it determines self-renewal[23] and in naïve CD4+ T cells, where it determines differentiation to T helper type 2 cells[24]. With this classification of transcription factor function in mind, I now ask what is the functional role of the transcription factor GATA1 in haematopoiesis?

1.1.1 GATA1 Does Not Have an Intrinsic Role in Haematopoietic Stem Cells Maintenance

Human haematopoietic stem cells reside within the flow cytometric population of mononuclear cells in either bone marrow or cord blood, defined by a pattern of expression of cell surface proteins: $CD34^+$ $CD38^{-/low}CD10^-CD45RA^-$ $CD90^+$ and absence of lymphoid and myeloid mature markers[25,26]. GATA1 is not expressed within this compartment at the RNA level. Instead, its expression starts at $CD34^+CD38^{low}$

multipotent potent progenitor (MPP) compartment marked by CD71 expression[27], a population of cells downstream of HSCs with limited self-renewal[26]. This simple observation of absent *GATA1* expression in haematopoietic stem cells rules out an intrinsic involvement of GATA1 in the regulation of HSC homeostasis.

1.1.2 Lineage Specification Occurs Early and Generates a *GATA1*+ and *GATA1*- Primed Multipotential Progenitor

The molecular mechanism by which TFs coordinate specification of blood lineages from haematopoietic stem cells remains an intensive research topic. GATA1 expression can be observed in erythrocytes and megakaryocytes[28], as well as mast cells[29], eosinophils and basophils[29,30]. How HSCs differentiate into GATA1 expressing cells of these lineages through a series of intermediate stages is still not understood. There are at least 4 reasons why it has been difficult to study human haematopoietic lineage specification:

1. Human haematopoietic lineage specification has been primarily based on functional assays that demonstrate what a progenitor cell can give rise to, rather than what it actually gives rise to in unperturbed condition during *in vivo* haematopoiesis.
2. Intermediate, multipotent or bipotent progenitor cells have been isolated based on cell surface proteins that do not necessarily isolate cells to functional homogeneity.
3. It is not known if such cell surface markers used are absolute or stable or whether they change across different stages of human development or across different sources of haematopoietic stem and progenitor cells such as mobilized peripheral blood vs bone marrow.

4. The user defined condition of an *in vitro* culture assay, such as the concentration and type of cytokines used and the time when the output is assessed, might not be optimal and can accidentally skew the readout of these assays.

For example, haematopoietic stem cells are defined by two key properties, the ability to self-renew and the ability to form all blood cell types. These two properties are functional in nature and can readily be tested in transplantation settings. CD34 cell surface expression has been used as a positive marker to enrich for human HSC activity in clinical and experimental settings[31]. The addition of negative and positive markers such as CD38, CD45RA and CD49f, has allowed for further enrichment to the extent that HSC activity can be isolated at the single cell level[25]. At the same time, human HSCs can reside in the CD34 negative fraction of cord blood mononuclear cells[32], but at much lower frequencies. Similarly, the human erythroid-megakaryocyte progenitor (MEP)[33], an intermediate bipotent progenitor with both erythroid and platelet producing megakaryocytic lineage potential, is not functionally pure. MEPs are immunophenotypically defined as CD34+CD38+IL3RA-CD45RA. When assessed at the single cell level, MEPs are only 10% bipotent, the rest are already committed unipotent erythroid cells (70%) or unipotent megakaryocytic cells (5%) (See Juliana Xavier-Ferrucio, Diane S. Krause concise review detailing the output of MEPs[34]). Therefore, it has become difficult to study the molecular process by which human HSCs differentiates into MEPs, which in turn gives rise to erythroid and megakaryocyte progenitors, because human HSC as well as MEPs are poorly defined. Nonetheless limited by these observations, in the classical view of human lineage specification, human HSCs generate a series of progenitor cells with restricted lineage potential and limited self-renewal.

In this model, human HSCs ($CD34^+$ $CD38^{-/low}CD10^-CD45RA^-$ $CD90^+$) generate multipotent progenitor cells (MPPs, $CD34^+$ $CD38^{-/low}CD10^-CD45RA^-$ $CD90^-$) marked by reduced CD90 expression with limited self-renewal but absence of lineage specification.

MPPs can then differentiate into lymphoid-primed multi-potential progenitors (LMPPs, $CD34^+$ $CD38^{-/low}CD10^-$ $CD45RA^+$ $CD90^-$) marked by gain expression of CD45RA, with granulocytic, monocytic, B and T cell potential but cannot generate erythroid or megakaryocytic cells[35]. While the erythroid/megakaryocytic potential is retained in the CD45RA negative fraction marked by gain of CD38 expression, in a population of cells referred as the common myeloid progenitor[33] (CMP, $CD34^+$ $CD38^+$ $CD10^-$ $CD45RA^-$ $CD123^+$). With the advent of optimized single cell culture conditions as well as single cell transcriptomic assays, it has been shown that the cell-surface defined cellular compartment of human haematopoietic intermediate bipotent/multipotent stages such as the erythroid-megakaryocyte progenitor (MEP), and the granulocyte-macrophage progenitor (GMP), or the common myeloid progenitor (CMP), once thought to be homogenous, are now shown to be functionally and transcriptionally heterogenous at the single cell level[36-41]. Drissen R, et al[36] showed that single cells downstream of the human MPP compartment, with the CD34+ CD38+ CD123+ CD45RA- immunophenotype and marked by CD131 expression ($CD131^+$ CMPs), generated mast cells, basophils, and eosinophils as well as erythroid and megakaryocytic cells *in vitro* but no other myeloid or lymphoid lineage. While single cells within the same compartment, but marked by CD114 expression instead, generated neutrophils, monocytes, as well as T, B and NK cells but no other myeloid cells. Consequently, they proposed a model in which HSC give rise to MPPs,[26] which bifurcate into the either a *GATA1* expressing erythroid-megakaryocyte primed multipotential progenitor (Erythroid/megakaryocyte-primed MPP or EMPP) marked by CD131 expression[36] with erythroid, megakaryocytic, basophil, eosinophil and mast cell potential, and a *GATA1* negative lymphoid-primed multi-potential progenitor (LMPP)[35] with granulocytic, monocytic, Natural Killer, B and T cell potential[38] (). It is the EMPP[36] then that can differentiate into a megakaryocyte-erythroid bipotent progenitor[41], a mast cell-basophil bipotent progenitor[36], and an eosinophil progenitor[39]. The existence of a $GATA1^+$ EMPP progenitor cell has been reinforced by single cell RNA sequencing of haematopoietic cells. Whereby inferred trajectories of cellular differentiation supported

the presence of a shared megakaryocyte-erythroid-mast cell progenitor downstream of the haematopoietic stem cell[42-45]. Single cell RNA sequencing provides a sampling or snapshots of single cells as they move through a differentiation landscape, and therefore might represent unperturbed haematopoiesis.

This model of cell specification has two implications, first cell fate decisions occur earlier than anticipated, at the level of the MPP[27]. Second, one of the earliest restrictions in cell potential generates a *GATA1* expressing multipotent progenitor (EMPP). Yet, it is not known if induction of GATA1 expression is what determines lineage specification or simply expression of GATA1 reinforces a cell fate decision that has already occurred. As ablation of the *Gata1* locus in murine cells does not selectively abrogates erythroid, megakaryocyte and mast cells (discussed in the next sub section), it is most likely that GATA1 expression in EMPPs only reinforces or coincides with cell fates decisions that have already taken place.

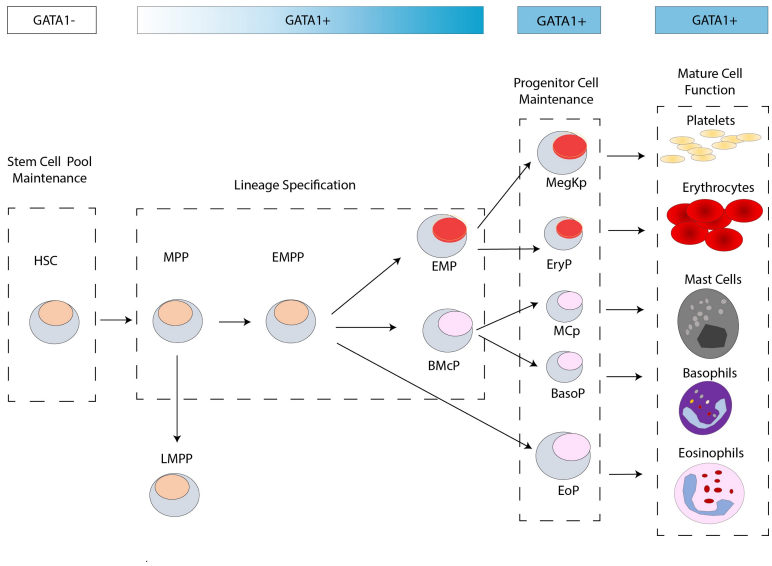


Figure 1.1 Schematic representation of GATA1 expression in the haematopoietic hierarchy

Haematopoietic Stem Cells (HSCs) do not express GATA1 but give rise to a Multipotent Progenitors (MPPs), which in turn differentiate into either a GATA1 expressing Erythroid-Megakaryocyte Multipotential Progenitors (EMPPs) or a GATA1 non-expressing lymphoid-primed multipotential progenitors (LMPPs). EMPPs give rise to an Erythroid-Megakaryocyte Progenitors (MEPs), a Basophil-Mast Cell Progenitors (BMcPs) and an Eosinophil Progenitors (EoPs). Adapted from Drissen R, et al[36]. Dotted squares classify transcription factor function in haematopoietic cell maturation.

1.1.3 GATA1 Maintains Immature Cells of The Erythroid and Megakaryocyte Lineage In Sufficient But Not Excessive Numbers

A classical experiment of forced expression of GATA1 in chicken myeloblast that do not express GATA1 resulted in a cellular conversion: monocytic cells transformed into to red blood cells, platelet producing thromboblasts or eosinophils[46]. This seemingly simple cellular conversion attributed an important role of GATA1 in orchestrating or instructing lineage specific patterns of gene expression. The role of GATA1 in erythropoiesis[47], through the generation of the Gata1-null mice[48], showed that erythroid precursors formed in the absence of GATA1. However, Gata1-null erythroid cells were arrested at an early proerythroblast-like state characterized by high levels of Gata2 and c-myb expression. These two factors are normally downregulated during erythroid maturation. Similarly, megakaryocyte progenitor cells form in the albescence of Gata1 expression but are hyperproliferative and produce abnormal platelets with compromised function[49-51]. As such, GATA1 ablation does not affect mast cell numbers[52]; yet low GATA1 expression induces abnormal mast cell morphology and maturation[52,53].

An exception can be made, contrary to the erythroid, megakaryocytic and mast cell lineages, targeted deletion of a high-affinity GATA1 biding site in the *Gata1* promoter leads to selective loss of the eosinophil lineage[54] and a drastic reduction of mature basophils and basophils progenitors[55]. Moreover, GATA1 expression is necessary for orchestrating terminal differentiation of these cells by evidence of GATA1 biding to promoters of cells-specific genes, such as *PRG2* (eosinophil granule major basic protein), *Ms4a2* (the basophil and mast cell membrane spanning 4-domains 2a), *GYPA* (erythroid specific glycophorin A), and *GP1BB* (platelet specific glycoprotein Ib Beta Chain). From these experiments I can conclude that the primary role of GATA1 is to maintain immature already committed cells of the erythroid, megakaryocyte, basophil, and mast cell lineage in sufficient numbers and to orchestrate or instruct terminal cell maturation of these cells.

1.1 GATA1 Structure and Function

Transcription factors have the unique ability to bind DNA in a sequence specific manner. Because of this ability, they are responsible for interpreting the genome to specify patterns of gene expression. Thus, transcription factors play a fundamental role in determining cell identity and function and their dysregulation is often a driver of cancers. In 1980s, owing to the understanding of the regulation of haemoglobin expression, a new transcription factor containing a dual zinc finger DNA biding domain was purified[56,57] and cloned[58]. This transcription factor was termed GATA1 based on the nucleotide composition of its DNA biding motif, (A/T) GATA (A/G).

1.1.4 Alternative splicing of *GATA1* Locus

The gene encoding the GATA1 protein resides in the X-chromosome. It is composed of 6 exons and two promoters, one of which is specific for Sertoli cells of the testis[59], while the other one is specific to haematopoietic tissue. Once transcribed, the human *GATA1* transcript is alternatively spliced to produced two mRNAs[60]. One of the mRNAs contains all exons and it is referred as the full-length (fl) transcript. The fl-transcript can be translated into two protein isoforms: a 413-amino acid (AA) protein referred as GATA1full-length, and a N' terminally truncated, 330-AA protein isoform referred as GATA1 short or GATA1s[61]. GATA1s results from the alternative use of a translation initiation codon AUG in exon 3 and lacks the first 83 amino acids. The naturally occurring GATA1s isoform is the same as the N-terminal truncated GATA1 isoform generated in TAM and ML-DS by *GATA1s* mutations. The second mRNA transcript lacks exon 2 but retains exon 1 and exon 3 to 6. This mRNA is referred as the short transcript and it can only be translated into the GATA1s protein (Figure 1.2). The phenomenon of alternative splicing of exon 2 of *GATA1* could be unique to humans as *Gata1* short transcripts have not been reported in murine tissue[60]. Whereas, alternative splicing of transcripts encoding transcription factors is a common phenomenon[62], no regulatory role has been attributed to the alternative splicing of exon 2 of human *GATA1*.

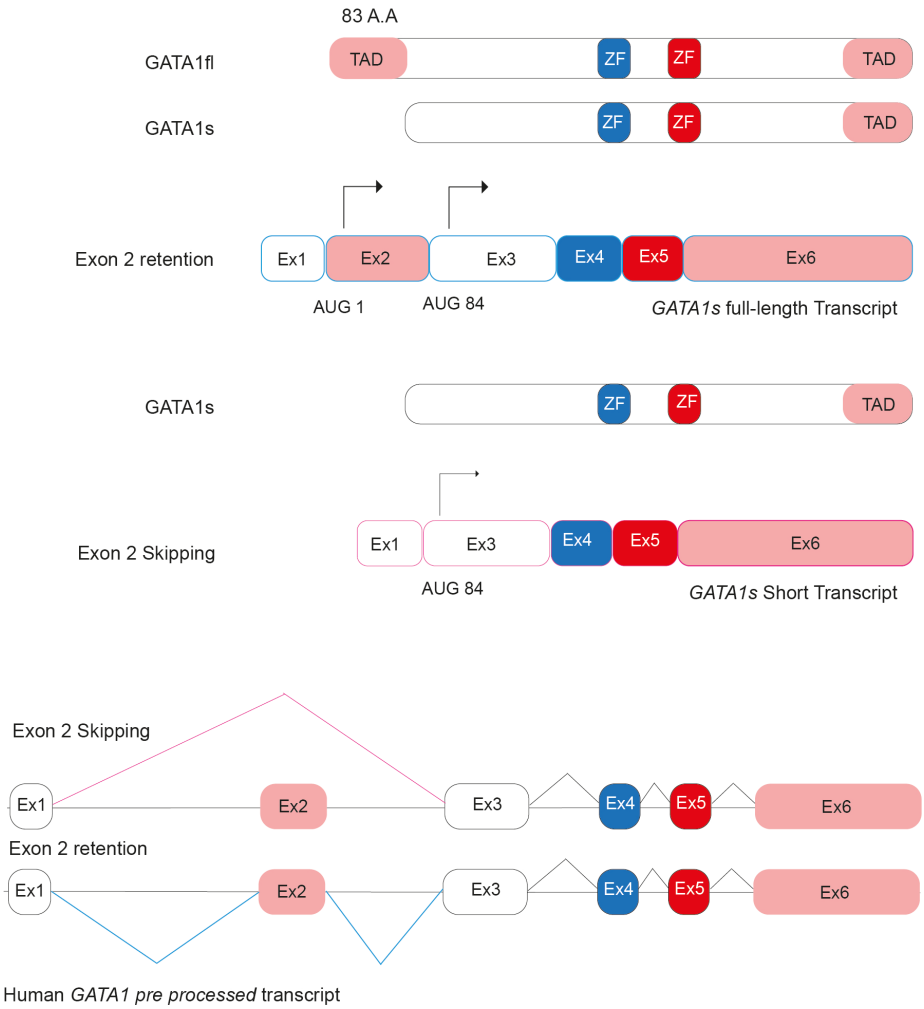


Figure 1.2 *GATA1* transcript processing and structure.

The *GATA1* pre-messenger RNA is alternative spliced to produce two transcripts through exon skipping. When translated, they produce two protein isoforms termed GATA1fl and GATA1s. Both isoforms consist of two zinc finger DNA biding domains and a C' terminal TAD domain. It is only GATA1fl that contains an N' terminal TAD domain.

1.1.5 GATA1 Functional Domains

The structure of GATA1fl isoform is composed of two transactivation domains (TAD)[63,64] at the amino (N') and carboxyl (C') termini that are important for activation of gene expression. These two TAD domains flank two zinc-finger DNA biding domains encoded in exon 4 and 5 respectively (Figure 1.2). Amino acid sequence conservation among 106 homologs using the ConSurf[65] algorithm demonstrates that both zinc-finger DNA biding domains and the C' terminal TAD domain are slowly evolving and implies that these domains are important for GATA1 function (Figure 1.3 A). Yet, with

approximately 60% amino acid sequence similarity, the two zinc-finger DNA biding domains (ZF) have acquired distinct functions[66] (Figure 1.3 B). The C' terminal ZF determines DNA recognition and biding[67,68] while the N' terminal ZF confers binding stability[69] and coordinates protein-protein interactions[70]. This differential function of the two ZFs can be observed at the genetic and phenotypic level. Genetic changes that affect the amino acid sequence of the ZF have only been described for the N' terminal ZF[71]. These changes are expected to disrupt protein-protein interactions and are responsible for haematological disorders of the erythroid and megakaryocyte lineages[71,72]. Genetic changes that affect the amino acid sequence of the C' terminal ZF would most likely disrupt DNA recognition and biding. This could be why these changes are not observed in nature as they would eliminate GATA1 ability to interpret the genome.

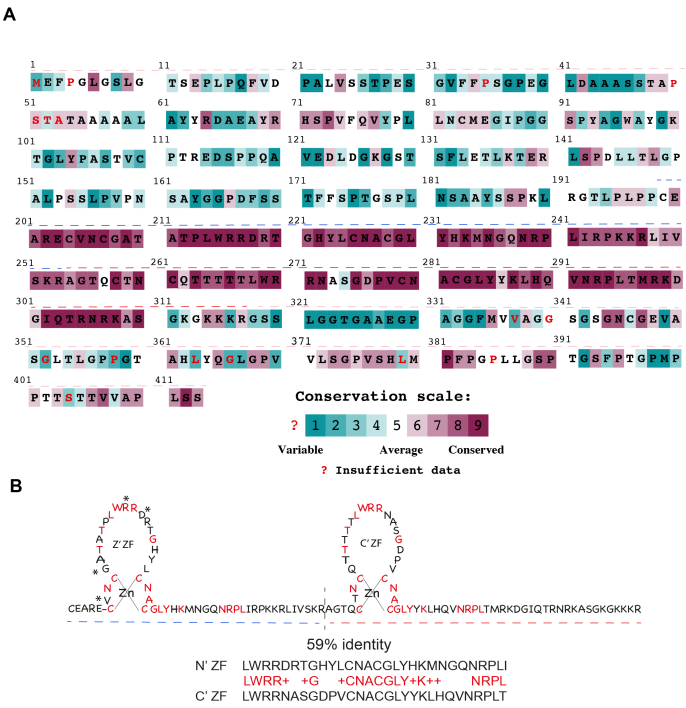


Figure 1.3 GATA1 amino acid sequence conservation

A) GATA1 amino acid sequence conservation with ConSurf algorithm. Dotted lines represent the beginning and end of each domain. N' and C' terminal TAD domain (pink), N' terminal ZF domain (blue), C' terminal ZF domain (red). B) Amino Acid sequence similarity between the N' (blue dotted lines) and C' (red dotted lines) terminal ZF domains, where red letters denote conserved amino acids. Asterisks denote amino acids that are often mutated and lead to haematological disorders. Adapted from Trainor CD, et al [69].

GATA1short on the other hand, retains the DNA biding ability[61] but completely lacks the N' terminal TAD domain. Unlike the ZFs domains, the N' terminal transactivation domain exhibits limited amino acid sequence conservation. TAD domains are intrinsically disordered amino acid sequences with limited secondary or tertiary structure[73]. When TAD domains are fused to heterologous DNA-binding domains, they can activate transcription[63]. Traditionally, TAD domains were classified based on their amino acid composition and historically they were referred to as "acid blobs and negative noodles", because within a TAD domain there is a cluster of negatively charged amino acids[74]. Consequently, the high acidity within a TAD domain was considered of importance. This prevailing model has been challenged, and in some cases through directed site mutagenesis, it has been shown that hydrophobic residues are more important than acidic ones in determining gene activation[75]. Recently, a functional screen of more than 6000 rationally designed variants of the yeast Gcn4 TAD domain performed by Staller et al[76], prioritized the ability of disordered acidic region to keep short, linear, hydrophobic motifs exposed for interactions with coactivators. Gcn4 TAD variants with the highest activity of gene expression, were those with aromatic residues exposed to solvent. This model of TAD function referred as Acid Exposure Model[76] predicts that a TAD domain will be composed of an exquisite balance between acidic, intrinsically disordered residues that forces and maintains hydrophobic amino acids exposed to solvent where they are available to bind coactivators[77].

Similarly, the N' terminal TAD domain of GATA1 is characterized by residues with high probability of being part of a disordered region as predicted by IUPred3[78] (Figure 1.4). Within this TAD domain, two short linear motifs with predicted buried hydrophobic amino acids can be observed. Interestingly, these two short linear motifs place high in the ConSurf [65] conservation scale and have extremely low probability of being part of a disordered region. These hydrophobic residues might play an important role in determining protein-protein interaction with coactivators. One of these two short motifs

was recently identified as a retinoblastoma protein(pRb) interacting site; whereby, Gata1 forms a complex with pRb and E2F2 to stall proliferation of maturing red blood cells[79]. Until this day, the proteins that interact with the GATA1 N' terminal transactivation domain remain largely undefined. How the lack of GATA1 N-terminal TAD domain affects haematopoiesis in the context of trisomy 21 to trigger pre-leukaemia and myeloid leukaemia remains unknown.

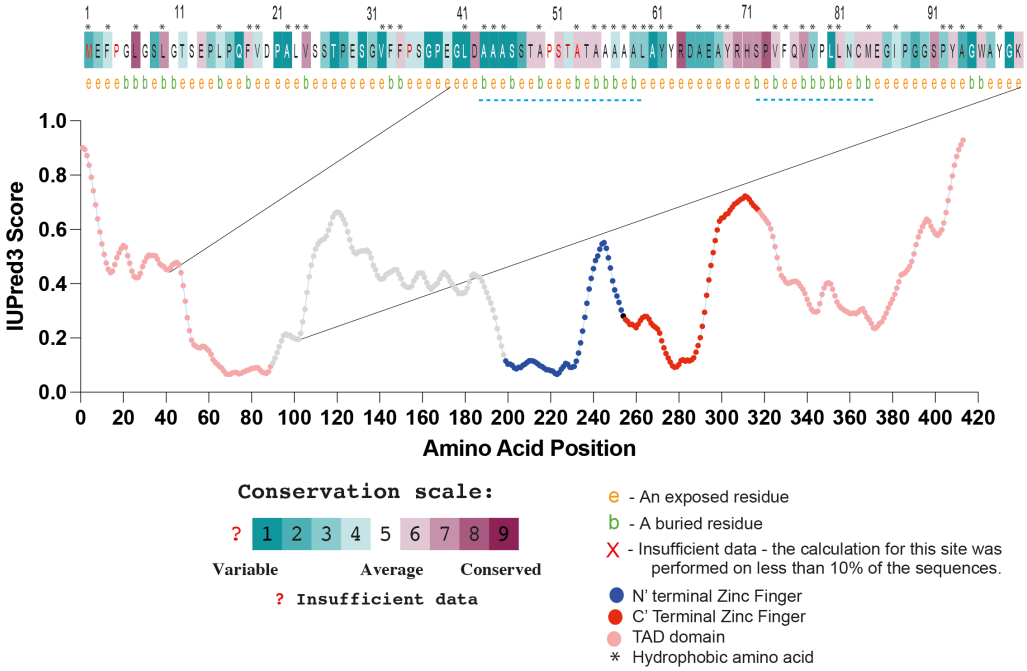


Figure 1.4 Structure of GATA1 N' terminal TAD domain.

Top, amino acid conservation of the N' terminal TAD domain with ConSurf algorithm demonstrates two short linear motifs of buried residues (dotted blue lines) which corresponds to low probability of being part of a disordered region as calculated by IUPred3, bottom. Colour coding of the amino acid sequence position is as in figure 1.2.

1.2 GATA1short in Malignant Haematopoiesis

For the past years multiple groups have confirmed the increased susceptibility of Down syndrome children to ML-DS. Recently, a new study defined the risk of ML-DS in a cohort of 3.9 million children with and without Down syndrome. This study demonstrated that DS children have 400-fold increased risk to develop ML-DS compared to non-Down syndrome children within the first 5 years of life[80]. To this day, we have

limited understanding of the molecular mechanism underpinning the staggering propensity for ML-DS in Down syndrome. Unlike adult AML, which is genetically complex, ML-DS arises from only three distinct genetic perturbations: an extra copy of chromosome 21, a single allele mutation in the gene encoding the transcription factor GATA1 acquired during fetal life, and post-natal acquisition of single allele mutation in any of the genes encoding either the cohesin complex, the *JAK* family kinases, or epigenetic regulators (Figure 1.5)[10]. ML-DS shares many features with acute megakaryoblastic leukaemia (AMKL). That is, leukaemic blast cells are characterized by aberrant and variable expression of megakaryocytic markers such as CD36[81,82,83], MPL[81] CD41[82,83], CD42b[81,82], CD61[82,83]. Unlike AMKL, TAM and ML-DS cells aberrantly expresses erythroid markers such as CD71[81,82], CD235a[82], as well as markers of stem and progenitor cells such as CD33[81,82,83], CD34[81,82,83], CD117[81,82] as well as CD38[82,83] and CD45low[82,83]. How an extra copy of chromosome 21, a truncating *GATA1* mutations and secondary mutations in the cohesin complex for example, cooperate to alter stem cell homeostasis, erythroid and megakaryocytic cellular maturation and lead to leukaemia transformation is explored in the next section.

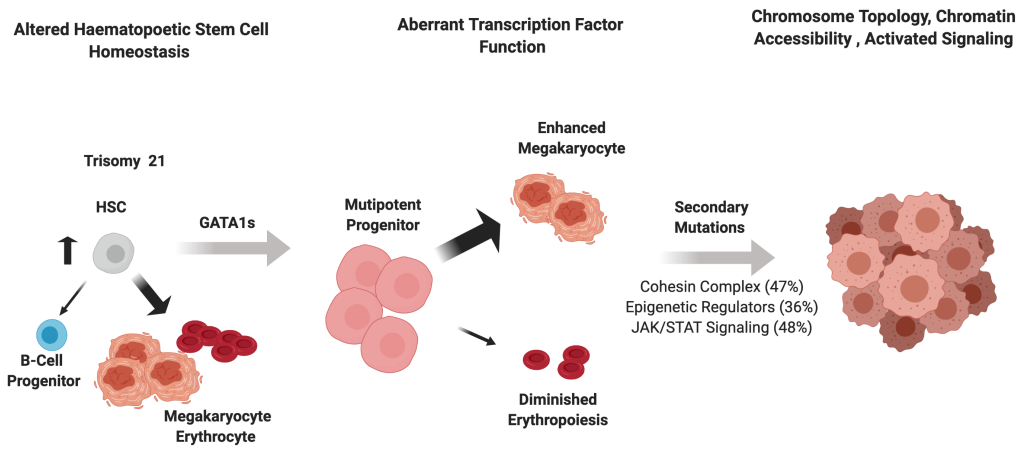


Figure 1.5 Multistep model of myeloid pre-leukaemia and leukaemia in Down syndrome

Trisomy 21 is the initial genetic change in ML-DS leukaemogenesis. It leads to expansion of haematopoietic stem cells and the erythroid/megakaryocytic compartment but impairment of lymphopoiesis. It is expected that, close to 30% of Down syndrome children acquire mutations in the GATA1 locus during fetal life. These mutations abrogate expression of the full-length GATA1 isoform but leave expression of the short GATA1s intact inducing a myeloproliferative disorder known as TAM. 10% of TAM children develop ML-DS before the age of 4 after acquiring additional secondary mutations.

1.2.1 Trisomy 21 Alters Haematopoietic Stem Cell Homeostasis

Trisomy 21 is the first hit in the multi-step model of leukaemogenesis in Down syndrome. Trisomy 21 or partial T21 in most cases results from nondisjunction of chromosome 21 during meiosis. Because of this, all cells of an individual with Down syndrome have an extra copy of chromosome 21, including the haematopoietic cells. There are about 233 protein coding genes as well as 423 non-protein coding genes in chromosome 21, of which 48% has not been completely annotated[84]. A cell-autonomous gene imbalance due to T21 might be responsible for perturbation of fetal haematopoiesis. In the few cases where ML-DS develops from partial T21 suggest that the leukaemic risk is confined to an 8.5Mb region on chromosome 21 where key haematopoietic genes lie, such as *RUNX1*, *ETS2*, *ERG*, *DYRK1A*[85].

Some of the haematopoietic abnormalities in Down syndrome neonates include a reduced count of neutrophils, and platelets and an increase concentration of haemoglobin in the peripheral blood[86,87]. Recently, single cell RNA sequencing of human fetal bone marrow samples taken at 12-13 weeks of gestation corroborated some of these findings[88]. Here, T21 fetal bone marrow samples, when compared to normal disomic samples of similar gestational age, had a significantly expanded erythroid and dendritic cell lineage compartment. They also had a diminished megakaryocyte and lymphoid lineage compartment. In the same study, it was reported that the erythroid lineage in T21 had a 5-fold increase in immature erythroid progenitor cells. This finding aligns with observation from *in vitro* hemopoietic differentiation assays of trisomy 21 induced pluripotent stem cells (iPSC) that have shown skewed haematopoiesis towards the production of erythroid cells[89,90].

Furthermore, immunophenotypic profiling of haematopoietic stem and progenitor cells in fetal liver showed that T21 causes an expansion of the cell-surface defined HSCs and MEPs, as well as a reduced GMP compartment and a severed impairment of B-cell lymphopoiesis[91]. Moreover, through in-vitro clonogenic assays, it was shown that T21 HSCs, CMPs as well as MEPs formed more colonies in response to cytokines than their disomic counterparts. In these experiments, the clonogenic output of HSC, MPPs, CMPs and MEPs was skewed towards the megakaryocytic lineage despite both T21 and disomic compartments having equivalent expression of the *TPO* receptor[91]. It is most likely that trisomy 21 causes genome-wide perturbation of gene expression and no single gene will be able to explain the phenotypes observed.

1.2.2 GATA1s in TAM and ML-DS

An N'-terminal truncating mutation in *GATA1* gene (*GATA1s*) is the second hit in the multi-step model of leukaemogenesis in Down syndrome. Up to 30 % of neonates with Down syndrome acquire *GATA1s* mutations during fetal life and develop TAM.

Whole genome sequencing of single cell derived clones of haematopoietic stem and progenitor cells (HSPCs) determined that during fetal life, HSPCs acquired approximately 100 base substitution per year[92]. This mutation rate is 5.8 times higher than that observed in post-infant HSPCs cells. Furthermore, T21 fetal HSPCs had in average 34 extra somatic mutations compared to disomic cells of similar gestational age. Although, the molecular mechanism behind this higher mutation rate during T21 fetal liver haematopoiesis remains to be deciphered, it can be attributed in part to a diminished ability of T21 cells to repair DNA damage. Morawiec Z., et al. identified that peripheral blood lymphocytes from DS patients are characterised by a significant higher level of endogenous DNA damage as well as a diminished ability to repair DNA damage when compared to healthy controls[93]. By analysing the spectrum of *GATA1s* mutations in Down Syndrome, Cabelof D., et al provided evidence that the base excision DNA repair pathway is compromised in DS individuals[94]. Consequently, the high mutational rate in the *GATA1* locus during fetal liver haematopoiesis can be attributed in part to a diminished ability of T21 cells to repair them.

GATA1s mutation in T21 are stereotypical, they either occur in exon 2 or the beginning of exon 3 but all of them abrogate production of GATA1 full-length and lead to exclusive expression of GATA1s. As the *GATA1* locus lies in the X-chromosome, in TAM and ML-DS, single allele mutations lead to the production of only the GATA1s isoform. Outside of trisomy 21, genetic germline variants in the *GATA1* locus that also abrogated GATA1 full-length have been described[93] but here they are not leukaemogenic[94-96]. Instead, patients present clinical haematological abnormalities related to Diamond-Blackfan anaemia mainly characterized by ineffective erythropoiesis. Once acquired *in utero*, *GATA1s* mutations will cooperate with trisomy 21 to result in TAM. TAM has a variable clinical presentation. In some cases, it is clinically silent and resolves within 3 months post birth and in others it is life threatening. Approximately 20% of clinically diagnosed TAM patients die within 6 months of life.

Experimental evidence from transgenic murine models[97-99] as well as patient-derived induced pluripotent cells[100] have demonstrated that exclusive GATA1s expression promotes expansion of megakaryocytes at the expense of erythropoiesis. This cellular phenotypes of exclusive GATA1s expression were corroborated by gene expression analysis which showed that GATA1s^{only} erythroid progenitor cells expressed low levels of genes whose product participate in haemoglobin synthesis such as *HBZ, HBG1, HBE1, AHSP/ERAF, and ALAS2*. At the same time, these cells expressed high levels of megakaryocytic affiliated genes. Chromatin immunoprecipitation followed by sequencing (ChIP-seq), Cleavage Under Targets and Release using Nuclease (CUT&RUN) and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) have been recently utilized to study the mechanism by which exclusive GATA1s expression leads to aberrant erythroid maturation[99,100]. Although, the majority of Gata1 biding sites are equivalently bound by Gata1s and Gata1 wild type (WT), binding to some DNA cis-elements specific for erythroid gene expression may be reduced in absence of the N-terminus of Gata1. Gata1s failed to repress *Gata2* and *Runx1* transcription, despite both Gata1 and Gata1s having similar chromatin occupancy at both loci[99]. A marked decreased in trimethylation at histone H3 lysine 27 (H3K27me3), brought by EZH2 a component of the PRC2 complex, was observed at the *Runx1* promoter. This explained in part the sustained expressing of Runx1[100]. Yet, at GATA2 locus the difference in H3K27me3 was minimal. Consequently, the mechanism by which GATA1s fails to repress GATA2 still needs to be clarified. Other groups have proposed a GATA1 repressive mechanism at the GATA2 locus through histone H4 at lysine 20 methyl marks imparted by SETD8 (KMT5A)[101,102]. Elucidating the mechanism by which GATA1s fails to repress GATA2 is important as overexpression of *GATA2* is associated with a poor prognosis in AML in general[103] .

GATA2 and RUNX1 expression are required for haematopoietic stem and progenitor cells (HSPCs) to balance self-renewal and differentiation. Their expression

is downregulated during terminal differentiation of the erythroid but not of the megakaryocyte lineage[104-106]. Moreover, RUNX1 has been shown to play a fundamental role during lineage fate decisions at the erythroid/megakaryocyte branching point by repressing KLF1 transcription, an erythroid specific transcription factor and promoting a megakaryocytic, FLI1-driven, gene expression program[106]. It is still not clear if GATA1s inability to repress at least two key transcription factors, GATA2 and RUNX1, cooperates with trisomy 21 to predispose blood progenitor cells to leukaemia transformation. It is also not known if there are other key transcription factors that are dysregulated upon exclusive expression of GATA1s.

1.2.3 Transforming mutation in TAM to ML-DS

The progression of TAM to ML-DS requires a third genetic change that can cooperate with exclusive GATA1s expression and trisomy 21. The use of next generation sequencing technologies in the transformation of TAM to ML-DS have identified several recurrent genetic alterations in ML-DS. For example, mutations in members of the cohesin complex (*RAD21, STAG1/2, SMC1/3, CTCF*), epigenetic regulators (e.g., *KANSL1, EZH2*, and *SUZ12*) and signalling molecules (e.g., *JAK* family, *MPL, KIT*)[10,107,108]. In a recent large-scale sequencing analysis of 111 TAM and 141 ML-DS samples, our group found that 1.6 additional somatic variants were present in ML-DS samples compared with 0.4 variants per sample in TAM[10]. This suggest that transformation of TAM to ML-DS is most likely due to a single mutation in genes encoding any member of cohesin complex (47%), the *JAK* family kinases, *MPL*, *KIT* (48%) or epigenetic regulators such as *KANSL1, EZH2*, and *SUZ12* (36%). This mutation frequency in ML-DS samples suggest that transformation occurs by cooperation between aberrant transcription factor function due to exclusive GATA1s expression and activated signalling and/or deregulation of epigenetic process.

With the advent of gene editing technologies such as CRISPR/Cas9, it becomes possible to functionally annotate the genetic variants during transformation of TAM to ML-DS. A murine model that closely resembles Gata1s driven pre-leukaemia served as a platform on which to interrogate predicted loss-of-function variants required to induce frank leukaemia[10]. In this murine model, a similar frequency of variants inducing activated signalling (32%) were observed. There was also an over representation of mutations in epigenetic regulators (40%). Surprisingly, there was a reduced mutational frequency in cohesin complex (5%), a four-protein ring-shaped structure necessary to bring distant regulatory units near promoters through DNA-loop formation. This tells us that cohesin mutations might not be easily studied in murine tissue because of the difference in chromosome architecture between species. Moreover, it suggests that trisomy 21 background might be required to produce the high cohesin mutation frequency observed in ML-DS.

1.3 DPhil Project objectives

The progressive nature of AML implies that a comprehensive understanding of the pre-leukaemic phase is required for i) early diagnosis, ii) monitoring of disease progression, and for iii) understanding aberrant transcription factor-mediated cell maturation. In Down Syndrome, TAM is considered a pre-leukaemic phase where *GATA1s* mutations and trisomy 21 cooperate to provide a clonal advantage[109]. When this pre-leukaemic clone acquires secondary mutations in genes encoding the cohesin complex for example, it transforms into ML-DS[10]. Transformation of TAM to ML-DS occurs in 13-33%[110] of the cases. Consequently, a comprehensive understanding of TAM offers opportunities for i) early diagnosis of young infants at risk of developing ML-DS, ii) monitoring expansion or shrinkage of the pre-leukaemic TAM clone and for iii) understanding the molecular mechanism by which exclusive GATA1s expression alters cellular maturation.

The aim of my DPhil project is to develop a same-day diagnostic test for TAM and ML-DS that can substitute next generation sequencing when not available. This test can serve to monitor TAM and ML-DS clones during the early months after birth. Second, it aims at generating a comprehensive single cell transcriptional reference landscape of T21 haematopoietic stem, progenitor, and precursor cells. Third, it aims at using this transcriptional reference landscape to identify cellular heterogeneity in TAM clones, and to defined maturation arrest and cell dysfunction of the erythroid and megakaryocytic lineage. Fourth, through gene editing technologies, it explores the functional consequences of exclusive GATA1s expression in the context of erythroid differentiation. Finally, during this DPhil project, I optimized Cleavage Under Targets and Tagmentation (CUT&TAG)[111], a new technology that allows to profile genome-wide GATA1 biding sites in as low as 50,000 cells. The results of my DPhil project are presented in this thesis.

2 Materials and Methods

2.1 Cell Line Culture

K562, THP1 and CMK cell lines were stored in 10% dimethyl sulfoxide (DMSO) in fetal bovine serum (FBS) at -80 °C. Cell vials were thawed before used at 37 °C in a water bath and washed twice with 2mL of warm 10% FBS in PBS (GIBCO) with 20ug/mL of DNase (FACS buffer) before culture. Cell lines were cultured at 37 °C with 5% CO_2, at a density of 1E6 cells per mL with 10% FBS in Iscove's Modified Dulbecco's Medium (IMDM). In the case of CMK cells, 20% of FBS was used.

2.2 Immunophenotyping of Trisomy 21 Cord Blood, DS Neonates, TAM and ML-DS Samples

2.2.1 Sample Processing

Down syndrome cord blood, DS neonates peripheral blood (PB), TAM PB and ML-DS bone marrow (BM) samples were obtained with informed consent from parents according to the Declaration of Helsinki and approved by the Thames Valley Research

Ethics Committee (06MRE12-10; NIHR portfolio no. 6362). Frozen cord blood samples from healthy donors were purchased from the National Health Services Cord Blood Bank, United Kingdom. Bags of cord blood or cryovials of mononuclear cells (MNCs) were stored in liquid nitrogen and thawed out at 37 °C in a water bath when needed. To isolate mononuclear cells, cord blood was diluted 1 in 5 with warm FACS buffer, and peripheral blood or bone marrow was diluted 1 in 2 with FACS buffer. Cord blood samples were spun down at 300g for 10 minutes and the supernatant was removed. The cell pellet was then resuspended in 35 mL of FACS buffer and layered on 15ml Ficoll-Paque™ Plus density gradient media. For bone marrow and peripheral blood pellets were resuspended in 15 mL of FACs buffer and samples were layered on 5 ml of Ficoll-Paque™ Plus density gradient media. Samples were spun down for 25 minutes at 400 g, with no brake. The interface of MNCs was collected and washed twice with 2 mL of FACS buffer before use. A cell wash is defined by mixing the cell suspension with 2mL of FACs buffer, followed by centrifugation at 300 g for 5 minutes at 4 °C.

2.2.2 Flow cytometry and Intracellular Flow Cytometry

For cell surface flow cytometry and intracellular flow cytometry, MNCs were first stained for dead cells by incubating each sample in 1 mL of ice-cold PBS with 0.5 ul of fixable viability dye-near-IR (Thermo Fisher Scientific) for 15 minutes in ice. Cells were then washed with 2 mL of FACS buffer, and the cell pellet was resuspended in 100 ul of FACS buffer with the appropriate amounts of cell surface antibodies (see Table 1). Cell suspension was incubated on ice for 30 minutes in the dark and then washed with 2mL of FACS buffer. For cell surface flow cytometry only, cells were filtered and resuspended in 300 ul of FACS buffer before data acquisition.

For intracellular flow cytometry, the Transcription Factor Buffer Set (BD biosciences) was used as specified. Briefly, cell pellet was fixed with 1ml of 1X Perm/Fix buffer and incubated on ice for 55 minutes. Fixation was quenched by directly adding 1

mL of 1X Perm/Wash buffer and spun down for 6 minutes at 400g. Samples were washed one more time with 2mL of Perm/Wash buffer. Cell pellet was resuspended in 100 ul of 1X Perm/Wash buffer with rabbit anti-human GATA1 phycoerythrin (PE)-conjugated (clone D52H6, Cell Signalling) and rabbit anti-human GATA1 (clone D24E4, Cell Signalling). Before use, antibody clone D24E4 was conjugated to Alexa Fluor 647 with the Zenon Alexa Fluor-647 Rabbit IgG Labeling Kit (Thermo Fisher Scientific) as specified by the vendor. Samples were incubated on ice for 55 minutes before being washed two times with 2mL of 1X Perm/Wash buffer. Cells were filtered and resuspended in 300 ul of FACS buffer before data acquisition.

Anti-Mouse Ig control compensation beads (BD Biosciences) were used to set up flow cytometry panels. For cell surface flow cytometry, fluorescent-minus-one (FMOs) technical controls were used to check the flow cytometry compensation parameters and to set the gating of positive signal. For intracellular flow cytometry, FMOs as well as IgG controls or cell lines known to be negative for GATA1 were used to set the gating of positive signal. Samples were run on a BD FACSymphony flow cytometer (Becton Dickinson) and analysed in FlowJo-10.6.0 (Becton Dickinson). A population was defined as containing ≥10 events, unless otherwise stated. Statistical analysis was performed with Prism V10 (GraphPad San Diego USA).

2.3 CRISPR-Cas9 of Primary CD34+ cells and K562 cells

2.3.1 CD34+ Cell Enrichment and Colony Forming Assay

CD34+ cell enrichment from cord blood was performed by staining MNCs for flow cytometry with fixable viability dye-near-IR as before followed by 2 ul of anti-human CD34-APC antibody (clone 581- BioLegend). 1,000 CD34+ (> 94% purity) cells were sorted by fluorescent activated cell sorting (FACS) into 4 ml of MethoCult SFH4436 (Stem Cell Technologies). MethoCult cell suspension was vortexed for 1 minute and 1.1 mL were dispensed in 25mm dishes and incubated at 37C for 14 days. Colonies were

inspected morphology and catalogued as either erythroid or myeloid colonies. If they were red because of haemoglobin synthesis they were catalogued as erythroid, otherwise they were catalogued as myeloid. The phenotype of erythroid and myeloid colonies was corroborated by flow cytometry and classified as either erythroid or myeloid if they expressed CD71 and Glycophorin A/CD235A or CD14/CD15 respectively.

2.3.2 Sample Processing

CD34+ cells were FACS sorted from cryopreserved MNCs as previously mentioned into 100 ul of stem cell expansion media containing the following cytokines: SCF 100ng/ml, FLT3-lig, 100ng/mL, TPO 50ng/ml, IL6 20ng/ml, UM171 35nM, SR-1 500nM in SFEM II basal media. Cells were washed once, and the cell pellet was resuspended in stem cell expansion media at 1E5 cells per mL. Cells were expanded for 48 hours at 37 °C with 5% CO2. After expansion, 50-100K cells were washed with 2 mL of PBS and the cell pellet was resuspended in 20 ul of P3 solution (Lonza). To form the ribonucleoprotein (RNP), 3ug of Alt-R® S.p. Cas9 Nuclease V3 (Integrated DNA Technologies) were mixed with 1.6 ug of target specific sgRNA (see Table 2 for sgRNA sequences) resuspended in nuclease free water. The RNP solution was incubated at 25 °C for 10 minutes in a thermocycler. After the incubation, the RNP solution was added to the 20 ul cell suspension and mixed up and down 5 times. All the volume was then transfer to an electroporation cassette (Lonza) and cells were electroporated with program EZ100 in a 4-D nucleofector (Lonza). In the case of K562 cells, 50-200K cells were washed and resuspended in 20 ul of SF solution (Lonza) and mixed with the RNP solution as before. K562 were electroporated with program FF120 in a 4-D nucleofector (Lonza). The voltage, the pulse time and length for EZ100 and FF120 are proprietary to Lonza. After electroporation, K562 cells were transfer to 1 mL of cell culture media and culture for 72 hours before DNA isolation. For primary CD34+ cells, cells were recovered overnight in stem cell retention media before FACS sorting CD34+ cells for a colony forming assay. For an erythroid differentiation assay, targeted cells were recovered

overnight before transferring to erythroid promoting media: SCF (10ug/ml), EPO (4Units/ul), IL3(10ug/ml), IL6 (100ug/ml) in SFEM II basal media.

2.3.3 DNA Isolation, Genotyping of Insertions and Deletions and Sanger Sequencing

DNA isolation was performed with the QuickExtrac DNA extraction solution (Lucigen) by resuspended a cell pellet or a cell colony in 50 ul of QuickExtract solution. Cell lysis suspension was incubated at 65°C for 10 minutes follow by 5 minutes of incubation at 95 °C. DNA was not quantified and 0.5 ul of DNA suspension were used for polymerase chain reaction (PCR) amplification only.

To genotype for insertions and deletions after CRISPR Cas9 targeting, first PCR amplification of *GATA1* exon 2 was done by mixing 0.5ul of DNA-QuickExtract solution with 1 ul of forward primer, 1 ul of reverse primer and 25 ul of DreamTaq Green Master Mix (2x) (Thermo Scientific), the volume was increased to 50 ul with nuclease free water. Amplification was done in a thermocycler with the following program:

Cycle 1: 95 °C for 1 minute
Cycle 2: 95 °C for 30 seconds
Cycle 3: 60 °C for 30 seconds
Cycle 4: 72 °C for 1 minute
Repeat Cycles 2-4, 30 times
Fina extension: 72°C for 5 min

PCR reactions were run on 1% agarose gel for 30 minutes. The correct DNA size (745 base pairs) was extracted from the agarose gel with the QIAquick Gel extraction kit (Qiagen). DNA was quantified with the NanoDrop 2000 spectrophotometer (Thermo Scientific). Sanger DNA sequencing of PCR amplicons was performed at the Weatherall Institute of Molecular Medicine sequencing facility in an ABI-3730. A reaction was prepared by mixing 3.2 pico mols of forward or reverse DNA primer with 30 ng of PCR amplified DNA. Sanger Sequencing traces were used as input for the ICE Synthego webserver tool (https://ice.synthego.com) to quantify insertion and deletions.

2.4 Genotyping for GATA1s mutations in TAM and ML-DS and Trisomy 21 Cord Blood Samples

2.4.1 DNA Isolation

For TAM, ML-DS and T21 cord blood samples, frozen cell pellets were used for DNA isolation with the Qiagen Kit QIAamp DNA Blood Mini Kit as specified. DNA was eluted in 20-50ul of water and quantify with a NanoDrop 2000 spectrophotometer (Thermo Scientific).

2.4.2 Library Preparation and Next Generation Sequencing

50 ng of genomic DNA was used for PCR amplification using FastStart High Fidelity PCR System (Roche, Burgess Hill, UK). Five *GATA1* specific primer pairs with CS1 and CS2 tags (supplemental Table 2) were used. Patient specific PCR products were pooled and diluted 1:50 before barcoding with the Access Array Barcode library (Fluidigm, Cam- bridge, UK; PN 100-4876) as recommended. Barcoded products were pooled proportionately and purified using Beckman Coulter Agencourt AMPure XP beads (Beckman Coulter, High Wycombe, UK). Products were diluted to 4 nM and pair-end sequenced on an Illumina MiSeq using 300 bp-phycoerythrin sequencing.

2.5 Cleavage Under Targets and Tagmentation

2.5.1 Sample Processing

K562, THP1 and CMK cell lines were cultured for up to 3 passages before use. The viability of each cell line was above 90% in all cases. Cell count was performed with the NucleoCounter NC-300 (ChemoMetec), solution 13.

To perform CUT&TAG, between 2,000 to 50,000 cells were initially washed with 500 ul of Washing Buffer (20mM HEPES pH7.6, 150mM NaCl, 0.5mM Spermidine, 1% BSA), spun down at 300g for 5 minutes. To isolate nuclei, cell pellet was resuspended in 100 ul of Nuclear Isolation Buffer (20mM HEPES pH7.6, 150mM NaCl, 2mM EDTA, 0.5mM Spermidine, 1% BSA, 0.01 % Digitonin, 0.1% NP40, 0.1% Tween 20, 1%

Protease Inhibitor) and incubated in ice for 10 minutes. After the incubation, each sample was spun down at 300 g for 5 minutes and the supernatant was carefully removed. Nuclei pellet was resuspended in 100 ul of ice-cold Nuclear Isolation Buffer. Each 100 ul sample was transferred to a PCR tube (STAR Labs), then 10 ul of activated Concanavalin A, BioMag@ Plus beads (Bangs Laboratories) were added. To allow concanavalin beads to bind nuclei glycoproteins, each sample was mixed by pipetting up and down slowly for 3 times and incubated at room temperature for 30 minutes. After the incubation, 2 ul (204ng) of H3K27me3 antibody (C36B11, Cell Signalling) were added to each positive control sample. 5 ul (500ng) of rabbit IgG (66362S, Cell Signalling) were added to each negative control sample. 3ul (453 ng) of rabbit anti GATA1 N' (3535T cell Signalling) or 5 ul (80ng) of rabbit anti-GATA1 C' (4589S cell Signalling) were added to each test sample. Samples were incubated with antibodies at 4 °C for 3 hours in a nutator. After incubation, samples were placed in a PCR magnetic holder for 5 minutes before the supernatant was removed. Beads/nuclei were resuspended in 200 ul of Dig-Wash Buffer (20mM HEPES pH7.6, 150mM NaCl, 0.5mM Spermidine, 1% BSA, 0.01 % Digitonin, 0.01% NP40, 1% Protease Inhibitor). Samples were placed back in the magnetic holder for 5 minutes before supernatant was removed. Nuclei/beads were resuspended in 50 ul of Dig-Wash Buffer with 5 ul of (600ng) of 1/10 diluted Guinea Pig anti rabbit secondary antibody (NBP1-72763, Novousbio). Samples were incubated for 30 minutes in a nutator at room temperature. After the incubation, samples were placed on the magnetic holder for 5 minutes before the supernatant was removed. To completely remove unbound secondary antibody and minimize background, nuclei/beads were washed twice with 200 ul of Dig-Wash buffer while on the magnetic holder. Nuclei/beads were resuspended in 50ul Di-300 buffer (Dig-Wash buffer + 300mM NaCl) with 2.5 ul of pA/G-Tn5 protein (EpiCypher). Samples were incubated for 1 hour at room temperature in a nutator. Samples were placed on the magnetic holder for 5 minutes and wash two times with Dig-300 buffer while on the magnet to remove unbound pA/G-Tn5. Bead/nuclei were

resuspended in 50 ul of cold Tagmentation Buffer (10mM MgCl2 in Dig 300 Washing Buffer) and incubated for 1 hour at 37°C in a thermocycler.

2.5.2 DNA Isolation

After tagmentation, proteinA/G-Tn5 and fragmented DNA were released by resuspending each sample in 600uL of Biding Buffer (Monarch PCR and DNA Clean-up kit). Samples were vortexed vigorously and transfer to a column. Columns were spun down for 1 minute at 16,000 g. The flowthrough was removed, and the column was washed with 200uL of Wash Buffer (monarch PCR and DNA Clean-up kit). The previous step was repeated one more time. DNA was eluted with 22 ul of warm elution buffer (Qiagen) by spinning the column for 1 minute at 16,000 g.

2.5.3 Initial PCR Amplification of DNA Fragments

22 uL of eluted DNA of each sample were mixed with 2 ul of P5 (10uM) primer, 2uL indexed P7 primer (10uM) and 25 uL NEBnext Hi Fi master mix – non hot start (M0541L, NEB). See Table 4 for primer sequences. The following PCR program was run for each sample:

- Cycle 1: 58 °C for 5 min (gap filling)
- Cycle 2: 72 °C for 5 min (gap filling)
- Cycle 3: 98 °C for 45 sec
- Cycle 4: 98 °C for 15 sec
- Cycle 5: 60 °C for 10 sec
- Repeat Cycles 4-5, X times
- 72°C for 1 min and hold at 4 °C

For experiments with 50,000 cells 8 initial cycles of PCR amplification were performed. For samples between 2000 to 50,000 cells 10 initial cycles of PCR were performed.

2.5.4 PCR Cycle Number Determination Based on qPCR

After the initial PCR amplification, 5 ul of sample were taken and the rest was stored on iced until needed. A secondary PCR was set by adding 0.25 ul of P5 and 0.25 ul of indexed P7 primer, 4.5 ul of H2O and 5 ul of Luna Universal qPCR Master Mix

(NEB). A qPCR reaction was run on the QuantStudio 7 Flex Real Time PCR system (ThermoFisher) for 40 cycles with the follow settings:

- Cycle 2: 95 °C for 60 sec
- Cycle 3: 95 °C for 15 sec
- Cycle 5: 60 °C for 30 sec
- Repeat Cycles 4-5, 40 times

Using the following options:

- Experiment type: Standard curve
- Reagent: SYBRTM Green reagents
- Reporter: SYBRTM
- Quencher: None
- Passive reference dye: ROXTM
- Ramp speed: Fast

After the qPCR has finished, the delta Rn (normalized fluorescence) vs Cycle number were plotted in prism and a sigmoidal curve was fitted. Then the cycle number where the Rn value plateaus was determined, and it was divided by 5 and its corresponding cycle number was used. The stored 45 ul of initial PCR amplification were further amplified with the same settings for the number of cycles determined by qPCR.

2.5.5 Library Preparation and Next Generation Sequencing

CUT&Tag libraries were pooled equimolar and sequenced in a NexSeq500 (Illumina) with a NextSeq 500/550 High Output Kit v2.5 (150 Cycles).

2.6 Single Cell RNA sequencing

2.6.1 Sample Processing

4 wild type trisomy 21 cord blood and 1 TAM cord blood with a known *GATA1s* mutation were processed as before. Mononuclear cells were resuspended in 100 uL of FACS buffer and stained for flow cytometry with cell surface antibodies listed in Table 8. For the TAM sample live, lineage negative, $CD45^{low}$ cells were FACS sorted. For wild type trisomy 21 samples EMPPs (defined as Lin- CD131+ CMP)[36], MegKP (defined as lin- CMP CD41+)[40], MK-MEP (lin- MEP CD71+CD41+)[37] were enriched by FACS sorting, as well as primitive (lin- CD34+CD38-), progenitor (lin- CD34+ CD38+) , HSPCs (lin-

CD34+), and precursor cells (lin- CD34- CD117+), where lineage (lin) is defined as CD2+ CD3+ CD4+ CD8a+ CD11b+ CD14+ CD15+ CD19+ CD20+ and CMP as CD34+CD38+CD123+CD45RA- and MEP as CD34+CD38+CD123-CD45RA-. All compartments were pooled together and a 30% cell loss was assume, then the cell pellet was resuspended at a concentration of 1200 cells / ul in FACS buffer. 16,516 cells per sample were loaded in a Chromium 10X Chip K (10X Genomics) with the goal of capturing 10,000 cells per sample. Reverse transcription was performed in a Bio-Rad C100 Touch thermocycler as specified in the Chromium Next GEM Single Cell 5' Reagent Kits v2 (Dual Index), CG000330 Rev A protocol.

2.6.2 Library Preparation

For each sample, the cDNA was amplified for 13 cycles for low RNA content cells assuming a target recovery of 10,000 cells. 50 ng of amplified cDNA for each T21 sample were dual indexed. For T21 samples with *GATA1s* mutations only 20 ul of amplified cDNA library were dual indexed. Dual indexing of each library was done as specified on step 5 (5' Gene Expression (GEX) Library Construction) of the CG000330 Rev A protocol.

2.6.3 Next Generation Sequencing

Dual indexed libraries were quantified with 2100 Bioanalyzer high Sensitivity DNA Kit (Agilent). The libraries were pooled equimolar for a total of 20nM library to achieve 100,000 reads per cell in 4 lanes of a NovaSeq (GeneWiz).

2.7 Bioinformatic Analysis

2.7.1 GATA1 Locus Genotyping DATA Processing

Mapping was done by the Burrows-Wheeler[16] algorithm and variants were called using Varscan.[17] Called variants were inspected in the Integrative Genomics Viewer. Variant allele frequency (VAF%) was determined using the UNIX command "grep" on fastq files. For female samples the VAF% was multiplied by 2.

2.7.2 CUT&TAG Data Processing

For each sample, lane-specific fastq files were merged with the UNIX command *cat*. For each merged fastq file, the Nextera adaptor sequences were removed with Cutadapt/2.10 (https://cutadapt.readthedocs.io/en/stable/). A quality control of fastq files was obtained with the FastQC/0.11.9 (Braham Bioinformatics). MultiQC/1.9 was used to aggregate FasQC[112] results into a single report. Fastq files with inserts between 10 to 700 nucleotides long were mapped to the human genome version hg38 with the help of Bowtie2[113]. The 9th column from Bowtie2 output sam files was extracted as this column contains the fragment distribution. The fragment distribution for each sample was plotted with Prism (GraphPad). Samtools/1.10[114] was used to sort sam files by coordinate. Samtools/1.10 was used to convert a sorted sam file to a bam file. Then bedtools/2.29.2[115] was used to convert the bam file to a bed file. The bed files for each sample as well as the IgG control bed files were used to call peaks with web server SEACR (http://seacr.fredhutch.org)[116]. The analysis of CUT&TAG data sets is described by Zheng Y et al (2020). Protocol.io. The CUT-RUNtools2.0[117] pipeline was also used, which performs the same steps as described previously in one single pipeline.

2.7.3 Single Cell RNA Sequencing: Quality Control

For T21 normal and TAM samples, demultiplexed fastq files for each sample were aligned to the human genome (version hg 38) with CellRanger Single Cell Software (10X Genomics, Inc)[118] to generate filtered gene expression matrixes. The Seurat/4.0[119] R package was used to for quality control and analysis. Cells with fewer than 7500 and more than 350 genes were kept for analysis. Genes expressed in fewer than 10 cells were also excluded. To exclude doublets derived from transcriptionally distinct cells, the DoubletFinder[120] package was used with an expected 5% doublet rate for trisomy 21 and 0.8% for TAM. Low quality cells defined as having a high percent of counts mapped to mitochondrial genes, were removed with miQC[121] with an posterior.cutoff = 0.95. After

the initial quality control filtering, 22388 T21 single cells were recovered in total, and 3877 TAM cells were recovered.

2.7.4 Single Cell RNA Sequencing: Low-Dimension Embedding

Single cells for each of the T21 samples were merged into a single gene expression matrix, then gene counts were normalized with the "Log Normalization" method and scaled using a scale factor of 10000. After finding 2000 variable genes, principal components analysis (PCA) was calculated with *RunPCA* function of Seurat. After this, T21 samples were integrated with the *RunHarmony* function of the Harmony[122] R package. From here, downstream analysis of T21 samples, such as clustering and non-linear dimensional reduction, used corrected Harmony embeddings instead of principal components. The number of harmony embeddings used was determined by the *ElbowPlot* function of Seurat. All the T21 cells were visualized in two dimensions with Uniform Manifold Approximation and Projection (UMAP)[123]. UMAP was calculated with the *RunUMAP2* function of the Symphony package, specifying 30 Harmony embeddings instead of PCAs. Cells were cluster into 20 groups by running two Seurat commands, *FindNeighbors* using 30 *harmony* embeddings and *FindClusters* with a resolution = 0.94.

2.7.5 Single Cell RNA Sequencing: Cluster Annotation

For T21 samples, annotation of clusters was facilitated by first identifying differentially expressed genes per cluster, based on the Wilcoxon Rank Sum test, using the command *wilcoxauc* of the Presto[124] R library. *Wilcoxauc* generates a list of statistic per cluster, which include log2 fold change of gene expression relative to other clusters, area under the curve (AUC) and adjusted p value. The AUC score for each gene for a given cell cluster was ranked and used as the input for gene set enrichment analysis (GSEA) with the FGSEA[125] R package with the function *fgsea*, against a curated list of 75 published haematopoietic lineage signatures[126]. The -log10 of the resulting adjusted p-values of the gene set enrichment score for each cluster was used to perform

hierarchical cluster. The results were plotted as heatmap with the Phemap R package (https://github.com/raivokolde/pheatmap). Some cells clusters were readily identified by canonical markers of T-cell (*CD3*, *CD4*), B-cell (*PAX5*, *CD19*), Dendritic Cells (*CD123*), Monocytes (*CD14*) and NK (*CD56*) cells. To verify the cell cluster annotation, a gene module scoring was calculated for each single cell using the AUcell[127], an R package. AUcell takes asset of curated gene list and calculates a gene score for each cell based on the Mann-Whitney U statistic. The gene lists were curated and taken from Roy et al, 2021[126]. UCell gene scores were visualized with *featurePlot* command of the Seurat package.

2.7.6 Single Cell RNA Sequencing: Creating a Reference and Query Mapping

After quality control, sample integration, low-dimensional reduction, cluster identification and annotation, an accurate reference map of T21 normal samples was built with the R package Symphony[128]. The reference was built with the Symphony command *buildReferenceFromSeurat* with the options: *save_umap = 'TRUE', save_uwot_path = 'cache_symphony.uwot'*. To generate a Query data set from TAM, and map this data set to the T21 reference, the Symphony command *mapQuery* was used. The T21 reference clusters were predicted on the Query data set with the command *knnPredict.Seurat.*

3 Diagnosis of TAM and ML-DS by Intracellular Detection of GATA1s

Around the world, 1 in 800 babies are born with Down Syndrome[129], of these 30% have acquired mutations in the *GATA1* locus that result in exclusive expression of short isoform of GATA1 (GATA1s) and are affected by TAM[7]. Clinical TAM has a mortality rate between 11% to 23%. Most patients with TAM go into spontaneous resolution within the first 3 months of life without chemotherapy intervention[130]. However, about 13% to 33 % of them progress to ML-DS before the age of 4 years (See Grimm J, et al. for a detailed

review[110]). Consequently, it is important to monitor the pre-leukaemic TAM clone in the first months after being born to identify DS children who are at risk of developing ML-DS.

Due to the natural history of TAM and ML-DS, it has been difficult to define a clear diagnostic criterion for these diseases. Historically, the diagnosis of TAM and ML-DS has been based on non-specific clinical and haematological parameters. For example, abnormal blood counts, a non-specified high percentage of circulating blast cells in the peripheral blood and conventional flow cytometric analysis of blast cells with erythroid and megakaryocytic markers[81-83,131] have been used. Unfortunately, an accurate diagnosis of TAM and ML-DS is complicated by the fact that all Down Syndrome neonates without a detectable *GATA1s* mutation have blast cells in blood smears ranging from 1% to 15% [7]. Moreover, the morphology and immunophenotypic profile of TAM and ML-DS blasts cells are shared with normal blast cells in neonates with Down Syndrome[87,132].

TAM and ML-DS cells are clonally linked by *GATA1s* mutations[3-5], and exclusive GATA1s expression is a disease driver. Consequently, detection of *GATA1s* mutations by highly sensitive next-generation DNA sequencing (NGS) has become the gold standard assay for a specific diagnosis of TAM and ML-DS[132,133]. However, NGS is time consuming, and it is not available in all hospitals. NGS also requires expensive equipment, and technical expertise in bioinformatics. These issues among others, make it challenging to incorporate NGS into routine clinical practice for the diagnosis of myeloid malignancies[134]. Most importantly, NGS is not a same day test. To overcome these issues and to offer an alternative method to NGS when not available, we developed a highly sensitive and specific intracellular flow cytometric test for the identification of cells with exclusive GATA1s expression. This test can be used to diagnose TAM, and to track and document the spontaneous resolution of TAM and measure the impact of chemotherapy in ML-DS. One day it could be used as measurable residual disease (MRD) tool in clinical settings.

3.1 Dual Immuno-Staining of GATA1 N' and C' Termini Can Identify Cells with Exclusive GATA1s Expression

The intracellular flow cytometry (iFC) test combines cell surface flow cytometry with intracellular detection of GATA1 isoforms. It was inspired by recent publications showing paraformaldehyde base intracellular detection of GATA1 and GATA2 in K562 cells[135] as well as immunohistochemical detection of GATA1 isoforms in primary patient samples[136]. It is based on three premises:

1. Normal cells express the two GATA1 isoforms while TAM and ML-DS cells express only the short GATA1 isoform.
2. The two GATA1 isoforms substantially differ in tertiary structure for them to be detected by independent antibodies.
3. Most importantly, the number of GATA1s molecules in the cell is higher than the limit of detection of conventional flow cytometers.

The iFC detection of GATA1 uses two commercially available rabbit anti GATA1 antibodies, D52H6 and D24E4. D52H6 is produced by immunizing animals with a synthetic peptide composed of residues surrounding Glutamate 13. These residues correspond to the N-terminal domain only present in GATA1 full-length. D24E4 is produced by immunizing animals with a synthetic peptide corresponding to residues surrounding alanine 184 of human GATA1. D24E4 should recognize the C-terminal part of GATA1 which is present in both GATA1 isoforms. To ensures that fluorescent molecules are above the level of detection for most flow cytometers which is 100 molecules per cell[137], D52H6 antibody is conjugated to R-Phycoerythrin (PE) and D24E4 to Alexa Flour 647 (AF647). These two fluorochromes have high quantum yields[137] and are readily excited by the 561nm and 640nm, two laser lines available in most flow cytometry analysers. It has been reported that the absolute number of GATA1 molecules ranges from 2,000 and 40, 000 proteins per nucleus in human haematopoietic stem and progenitor cells (HSPCs) and erythroid progenitors respectively[138]. Therefore, dual

immunostaining for flow cytometry with both anti GATA1 antibodies would identify normal cells (GATA1^{WT}) expressing both GATA1 isoforms as double positive events (positive for PE and AF647) and TAM or ML-DS cells as single positive events (positive for AF647 only) expressing only GATA1s (GATA1s^{only}). Consequently, it is the lack of the GATA1 N' terminus signal that marks TAM or ML-DS cells (Figure 3.1).

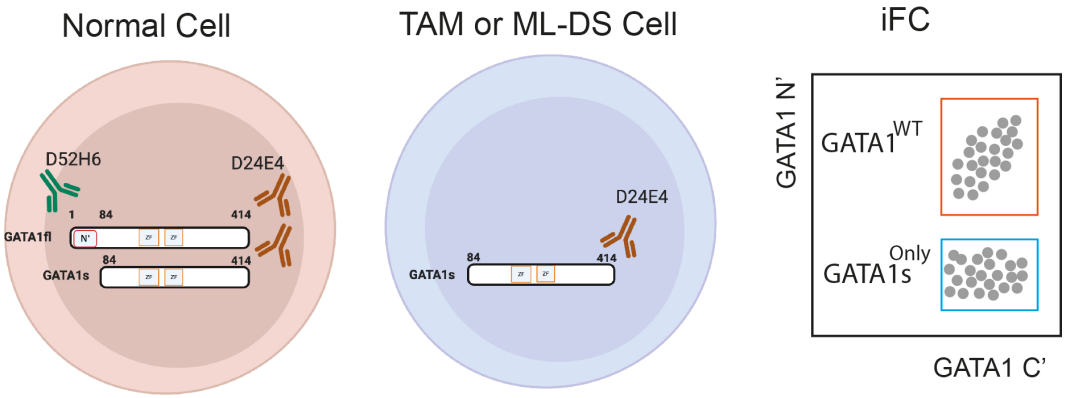


Figure 3.1 Schematic representation of dual immunostaining of GATA1 N' and C' terminus

D52H6 antibody against GATA1 N' terminus (green) detects only GATA1full-length (fl). D24E4 antibody directed against GATA1 C' terminus (brown) detects both GATA1fl and GATA1short. Left, a normal cell (red circle) expresses both GATA1full-length and GATA1short. Center, TAM or ML-DS cells (blue circle) lack GATA1full-length expression and only express GATA1short. Right, schematic representation of expected flow cytometric plot where cells (grey circles) are double positive for GATA1N' (Y-axis) and GATA1 C' (X-axis) (red square - GATA1^{WT}) or wild type for GATA1 and TAM or ML-DS cells are single positive for GATA1 C' only (blue squared -GATA1s^{only}) or mutant for GATA1.

To test the discriminatory power of the iFC method, first iFC was performed in a cell mix experiment where each cell line was mixed at equal ratios. I utilized THP1 cells that do not express GATA1 (negative control), K562 cells that express the two isoforms of *GATA1* (positive control) and CMK cells, a cell line derived from a patient with ML-DS with a mutation in exon 2 of the *GATA1* locus. GATA1 expression was corroborated by quantitative PCR (qPCR), with specific probes against either *Exon1-Exon2* or *Exon1-*

Exon3 boundaries, which detect both *GATA1full-length* or *GATA1short* transcripts respectively. As expected, K562 cells expressed both transcripts and CMK cells expressed only the short transcript. THP1 cells expressed neither of the transcripts (Figure 3.2B). Consequently, THP1 cells were used to set the background signal for both GATA1 N'-PE and GATA1 C'-AF647 antibodies. As predicted, K562 cells are double positive GATA1 N'-PE and GATA1 C'-AF647. A mutation in exon 2 of the *GATA1* locus in CMK cells abrogates GATA1 full-length production and consequently the PE signal, marking CMKs cells as single positive for the GATA1 C' terminus (Figure 3.2 A).

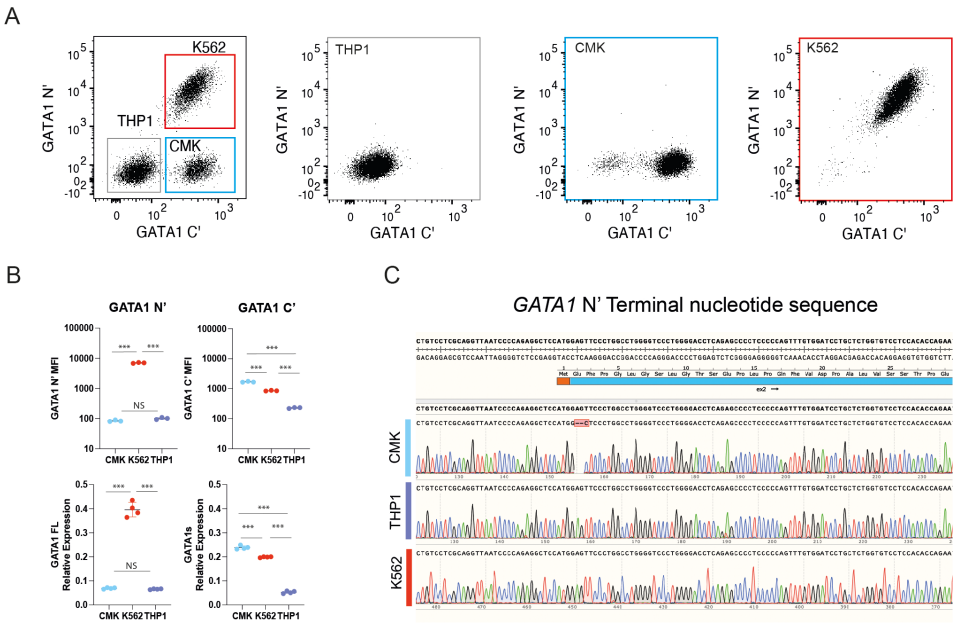


Figure 3.2 Dual immunostaining against GATA1 in a cell mix experiment

A) Left, a representative flow cytometric plot of dual immunostaining against GATA1 in a cell mix experiment. Right, a representative plot of dual immunostaining against GATA1 N' and C' terminus for each cell line used. Y-axis anti GATA1-N'-PE fluorescence, X-axis anti GATA1-C'-AF674 fluorescence. B) MFI - Mean fluorescence intensity, (top) and relative gene expression (bottom) of GATA1 in CMK cells (red), K562 cells (blue) and THP1cells (purple) of three technical replicates. C) Chromatogram traces of Sanger sequencing of the GATA1 N' terminal nucleotide sequence for each cell line used. CMK cells are marked by a frameshift at Glutamate 2 that disrupts the codon sequence, coloured coded as in B. Unpaired two-tail student T-test with p value style APA: 0.12 (ns), 0.033(*), 0.002 (**), <0.01 (***)

3.2 Intracellular Detection of GATA1s is Comparable to NGS Detection of *GATA1s* Mutations

To determine if the iFC method can be used as an alternative method to NGS when not available, a series of samples (n=42) with and without *GATA1s* mutations were subjected to both mutational profiling of the *GATA1* locus as well as intracellular flow cytometric detection of GATA1 (Figure 3.3). Due to the majority of *GATA1s* mutations in TAM or ML-DS occur in exon 2 or the beginning of exon 3[7], for each sample these two exons were PCR amplified with 5 sets of PCR primers (Table 2). Amplicons were pooled and deep sequenced by NGS, and the variant allele frequency (VAF%) was calculated. Second, a multiparameter flow cytometry panel was designed which included intracellular detection of GATA1 N' and C' terminus, as well as cell surface detection of CD7, CD117, CD45, CD36, Glycophorin A, and CD41 (Figure 3.4). Isotype controls and fluorescent-minus-one (FMOs) controls were used to set a positive gate for each antibody. Third, the samples included 7 peripheral blood (PB) samples from 7 DS neonates with no detectable mutations (control samples), 26 PB samples from 24 TAM patients with *GATA1s* mutations and 8 bone marrow (BM) samples from 5 ML-DS patients, three of which were taken post chemotherapy (control samples), whose *GATA1s* mutations were no longer detected by NGS. Subject characteristics are Table 6.

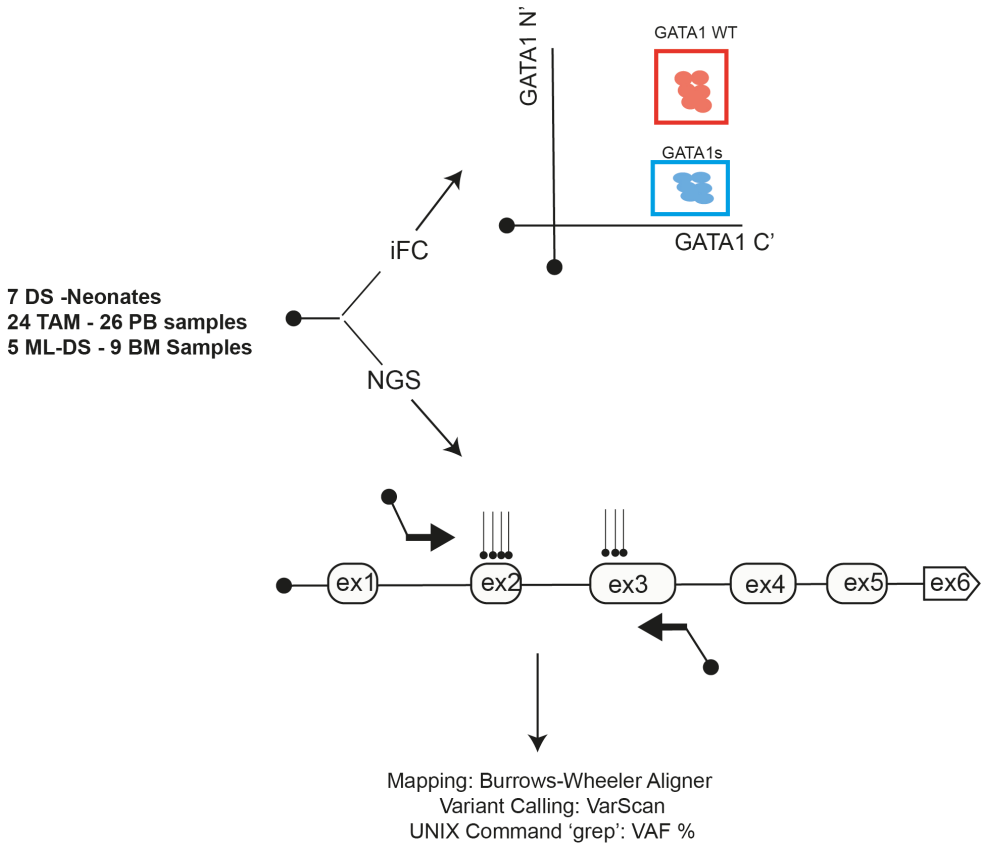
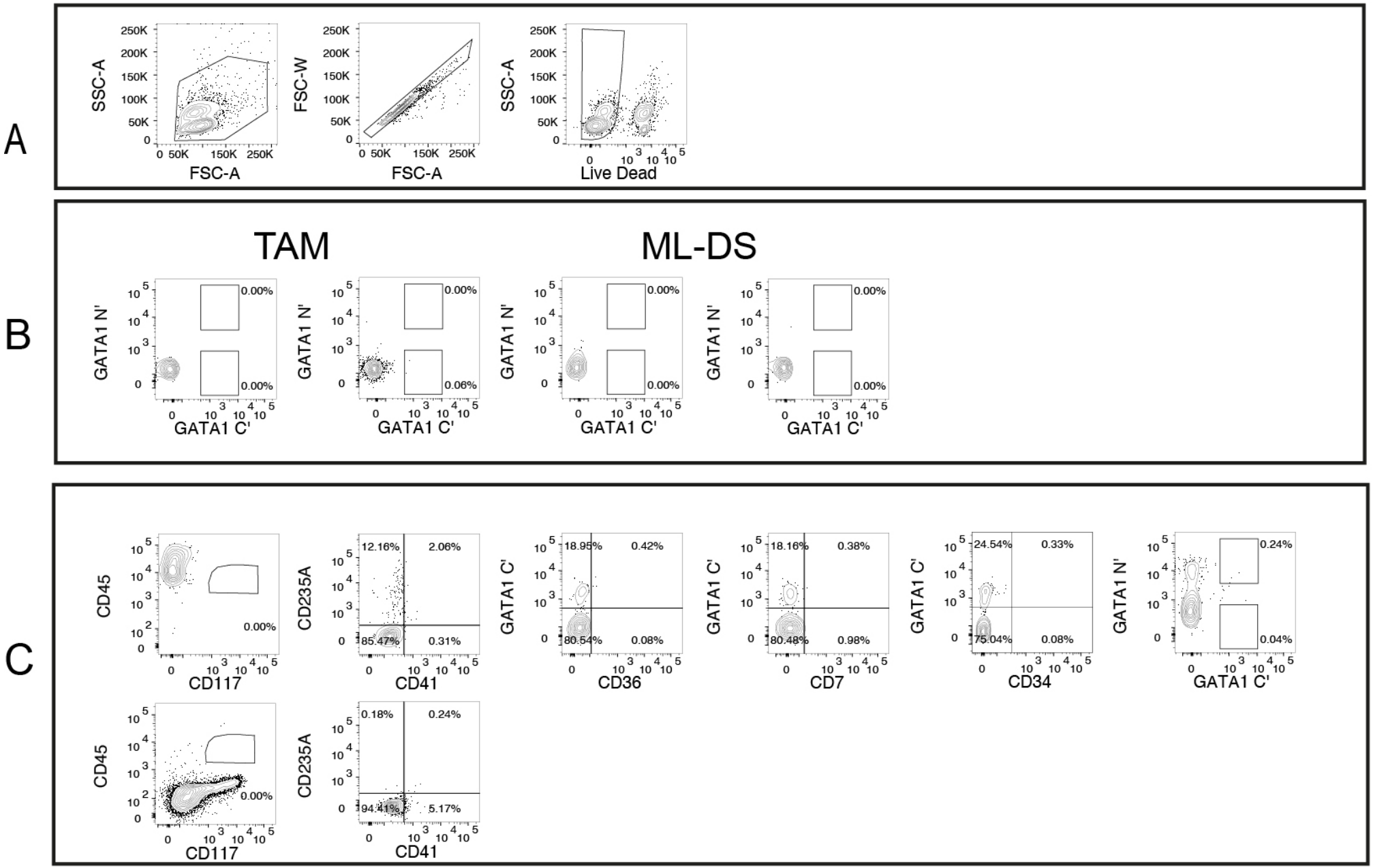
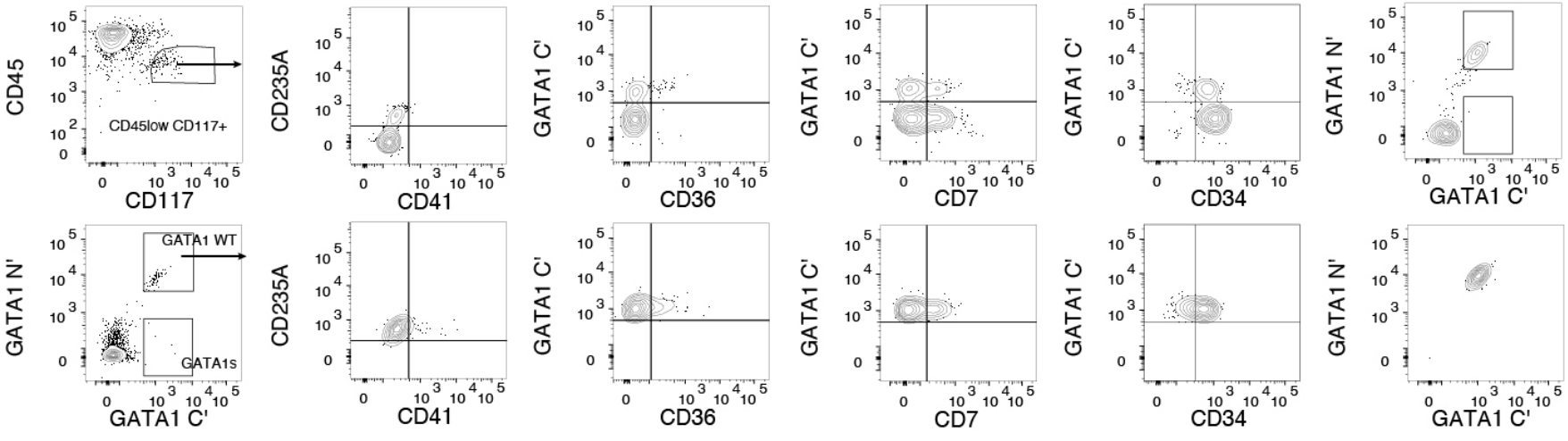


Figure 3.3 Schematic representation of the experimental approach

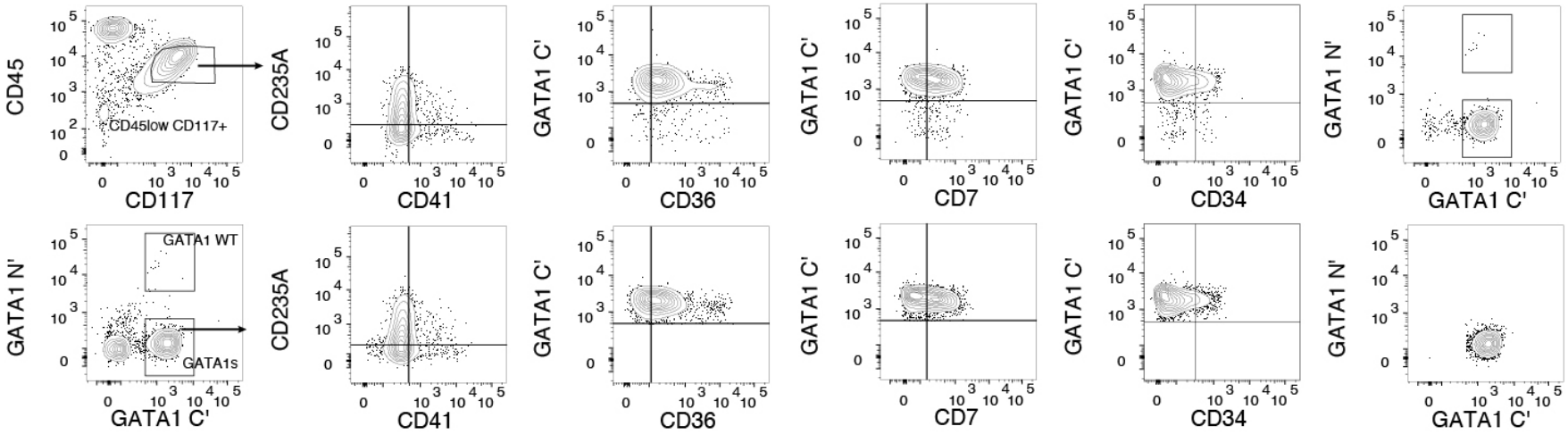
Peripheral Blood (PB) and Bone Marrow (BM) samples were subjected to both iFC (top) and NGS sequencing of *GATA1* exon 2 and 3 (bottom). Top, schematic of iFC plot of wild-type cells (red gate) and GATA1s-only expressing cells (blue gate) immunostained with anti GATA1 N terminus (y-axis) and anti-GATA1 C terminus (x-axis). Bottom, arrows indicate PCR amplification of exon 2 and exon 3 where most of *GATA1s* mutations (inverted lollipops) are found.


A
SSC-A
FSC-A
FSC-W
Live Dead
B
TAM
ML-DS
GATA1 N'
GATA1 C'
0.00%
0.06%
C
CD45
CD117
CD235A
CD41
CD36
CD7
CD34
GATA1 C'
GATA1 N'
12.16%
2.06%
85.47%
0.31%
18.96%
0.42%
80.54%
0.08%
18.16%
0.38%
80.48%
0.98%
24.54%
0.33%
75.04%
0.08%
0.24%
0.04%
0.18%
0.24%
94.41%
5.17%

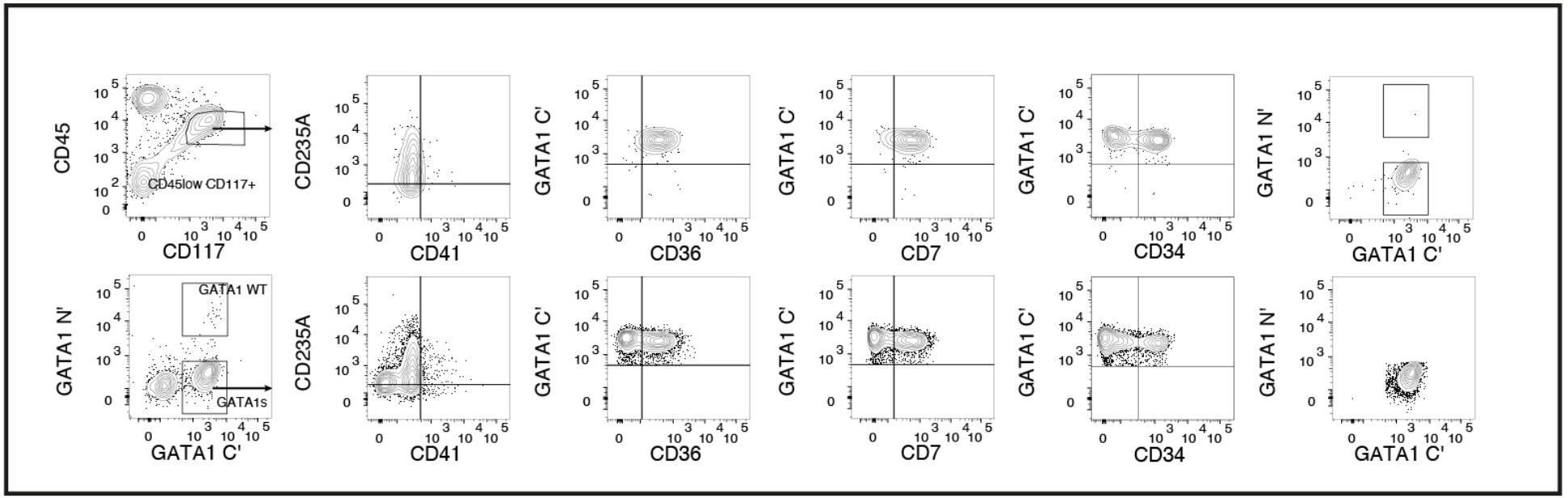
D


CD45
CD45low CD117+
CD117
CD235A
CD41
GATA1 C'
CD36
CD7
CD34
GATA1 N'
GATA1 WT
GATA1s

E


CD45
CD45low CD117+
CD117
CD235A
CD41
GATA1 C'
CD36
CD7
CD34
GATA1 N'
GATA1 WT
GATA1s

F


CD45
CD45low CD117+
CD117
CD235A
CD41
GATA1 C'
CD36
CD7
CD34
GATA1 N'
GATA1 WT
GATA1s

G

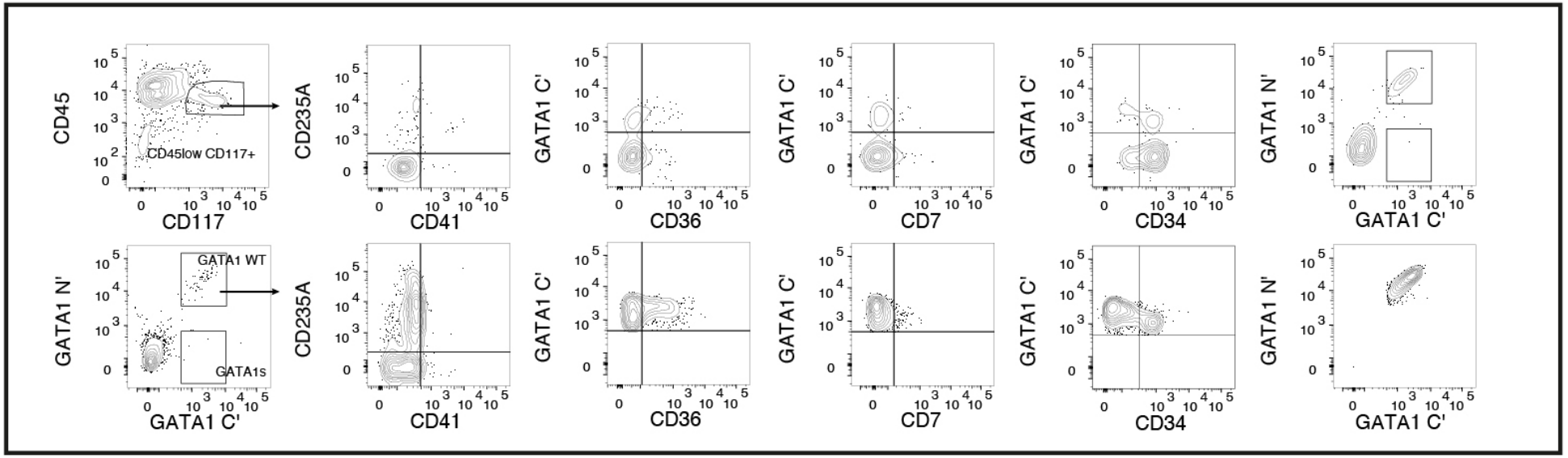

CD45
CD45low CD117+
CD117
CD235A
CD41
GATA1 C'
CD36
CD7
CD34
GATA1 N'
GATA1 WT
GATA1s

Figure 3.4 Multiparameter flow cytometric panel

A) Cells were first identified by their forward and side scatter properties, then doublet cells were gated out and live cells were defined as Near-IR negative (x-axis). After excluding doublets and dead cells, live cells were displayed as CD45 vs CD117 (top) or GATA1N' vs GATA1 C' (bottom) for panels D, E, F, G. B) Positive gates for $GATA1^{WT}$ and $GATA1s^{only}$ were first defined by staining two TAM samples (left) and two ML-DS samples (right) with rabbit PE-IgG and AF647-IgG isotype controls. C) ML-DS samples were used as fluorescent-minus-one (FMOs) controls to verify the positive gates for C117 (left top), CD45 (left bottom), as well as CD41, CD235A, CD36, CD7, CD34 and GATA1 C terminus. Representative flow cytometric plots of Down syndrome (DS) neonates without *GATA1s* mutations by NGS (D), DS children with TAM(E), DS young infants with ML-DS before chemotherapy (F) or after chemotherapy (G). For D, E F and G, live cells were first displayed in the CD45 vs CD117 (top) axis or GATA1N' vs GATA1 C' axis (bottom).

The variant allele frequency is defined as the number of mutant reads divided by the number of total reads in a NGS assay. The *GATA1* gene lies on the X-chromosome, if the sex of the patient is female there are two alleles and if the sex is male, there is only one. As *GATA1s* mutations tend to be heterozygous, for female patients the NGS VAF obtained underestimates the number of TAM or ML-DS cells by a factor of 2. For example, a VAF of 5% in a female TAM patient corresponds to 10% of cells being mutant, while a VAF of 5% in a male TAM patient, corresponds to 5% of cells being mutant. Once adjusted for sex, that is doubling the VAF % for female patients, the *GATA1s* VAF for each sample is proportional to the number of TAM or ML-DS cells. I hypothesize that if the iFC method can identify mutant cells it would be highly corelated to VAF % determined by NGS. Thus, for each sample the VAF of the *GATA1s* mutation was plotted against the frequency of live, GATA1s^{only} cells (see Supplementary Figure 1 for representative iFC plots of TAM samples). Linear regression analysis revealed a significant concordance (R^2 = 0.75; p-value=0.0001) between iFC and NGS methods (Figure 3.5 A). As expected, samples with a high VAF had a high number of GATA1s^{only} cells. More importantly, samples with no detectable *GATA1s* mutations by NGS (7 PB-DS samples and 3 post chemotherapy BM- ML-DS samples) had virtually no detectable GATA1s^{only} cells.

From this observation I hypothesized that there would be less concordance if conventional cell surface flow cytometry (FC) was used to identify TAM and ML-DS cells. To test this, first I identified aberrantly expressed cell surface markers that are positive on TAM and ML-DS cells and have previously been used to diagnose these diseases[81,82]. Of these, CD7 often described as a lymphoid lineage marker, CD117 the receptor for the stem-cell-factor cytokine and CD45, if lowly expressed marks immature cells, have been used. Once again, for the same set of samples the *GATA1s* VAF% was plotted against the frequency of live, CD45^{low}, CD7^{+} and CD117^{+} cells. As expected, there was less concordance (R^2 = 0.61; P =0.0001) between the *GATA1s* VAF and percentage of live

cells with CD45^{low} CD117^{+} CD7^{+} expression (Figure 3.5 B). Samples with no detectable *GATA1s* mutations had also expression of these markers. Therefore, I hypothesize that CD7, CD117 and CD45 expression would also be found in normal GATA1 wild type or GATA1 negative cells, which would confound their use to identify TAM and ML-DS cells during diagnosis. This hypothesis is explored in the following section.

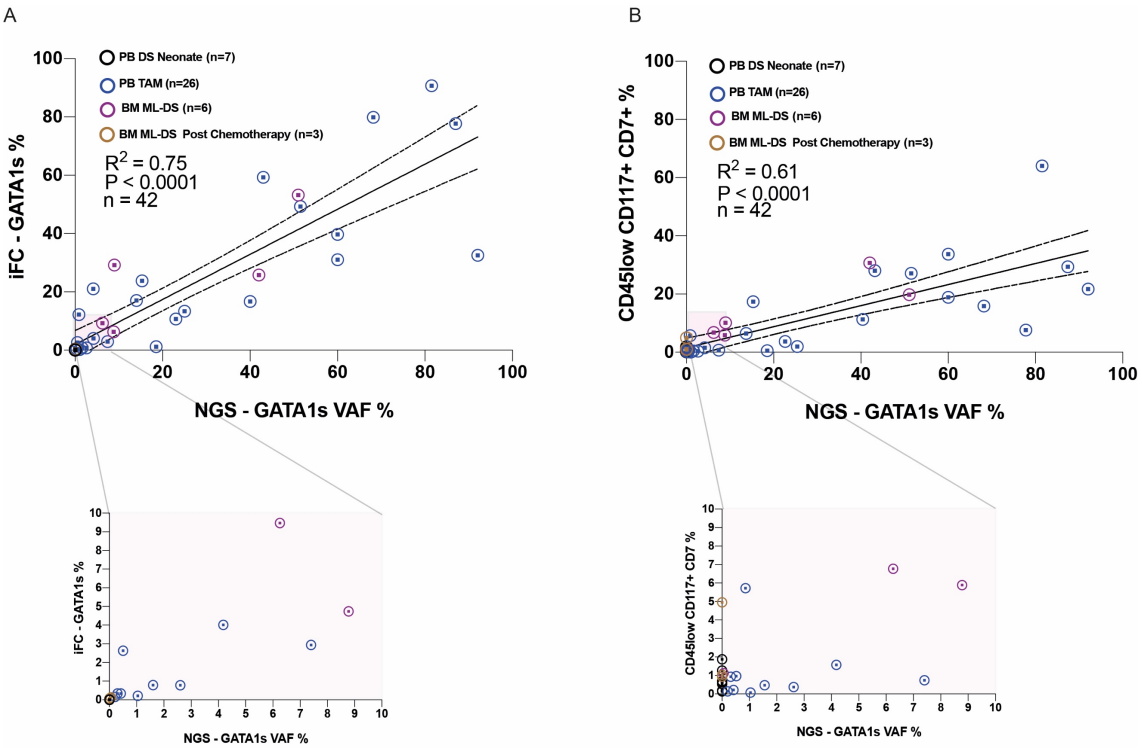


Figure 3.5 Simple linear regression between NGS, iFC and FC

A) Linear regression analysis showing correlation between *GATA1s* VAF (x-axis) by NGS and percentage of GATA1s^{only} cells detected by iFC, as a percentage of live cells (y-axis). Samples tested: PB DS neonates screened by NGS and found to have no GATA1 mutations (red open circles), PB TAM (blue open circles), ML-DS BM (purple open circles), and ML-DS post chemotherapy treatment BM (brown open circles). Dotted lines show 95% confidence intervals. R^2 = 0.75; P < 0.0001; n = 42. For samples in which more than 1 mutation was identified, the biggest VAF value was used. Bottom, samples with 1% to 10% VAF and 1% to 10% GATA1s^{only} cells are magnified to illustrate data clearly. (B) Linear regression analysis showing correlation between *GATA1s* VAF (x-axis) by NGS and percentage of cells in the live, CD45^{low}CD117^{+}CD7^{+} gate (y-axis) by flow cytometry. Dotted lines show 95% confidence intervals. R2 = 0.61; P < 0.0001; n = 42. The rest of the panel is as in panel A.

3.3 Cell Surface Expression Alone is Insufficient to Detect TAM and ML-DS Cells

Although immunophenotyping of AML samples by multiparameter flow cytometry is a rapid and efficient technique to detect and catalogue leukaemic blast, it suffers from some limitations. Cell surface markers are not always stable through disease and these markers are not exclusive, that is they can be shared with normal cells[139]. CD117 and $CD45^{low}$ are markers that are shared with normal stem and progenitor cells. CD7 is regarded as a lymphoid marker while glycophorin A is an immature erythroid specific marker. CD41 is a megakaryocytic and platelet marker and the expression of glycophorin A and CD41 are regulated by GATA1. It is not known how Glycophorin A and CD41 expression changes in the context of exclusive GATA1s expression in trisomy 21 cells. Therefore, patterns of expression of CD45, CD7 and CD117 were explored in TAM and ML-DS samples. First, live $GATA1s^{only}$ events in TAM and ML-DS samples were identified and the expression of CD45 and CD117 was evaluated (Figure 3.6 A). The majority of $GATA1s^{only}$ cells in TAM and ML-DS co-located in the $CD45^{low}$ $CD117^{+}$ gate. Linear regression analysis showed a high concordance between the percent of live $GATA1s^{only}$ and the percent of live cells with $CD45^{low}$ $CD117^{+}$ expression (R^2 = 0.82; p-value =0 .0001) (Figure 3.6 B). However, it is in this compartment that normal bone marrow stem and progenitor cells also reside (1.7+/-0.4% of total nucleated BM cells[140]), some of which express the two GATA1 isoforms in TAM, ML-DS and normal DS neonates.

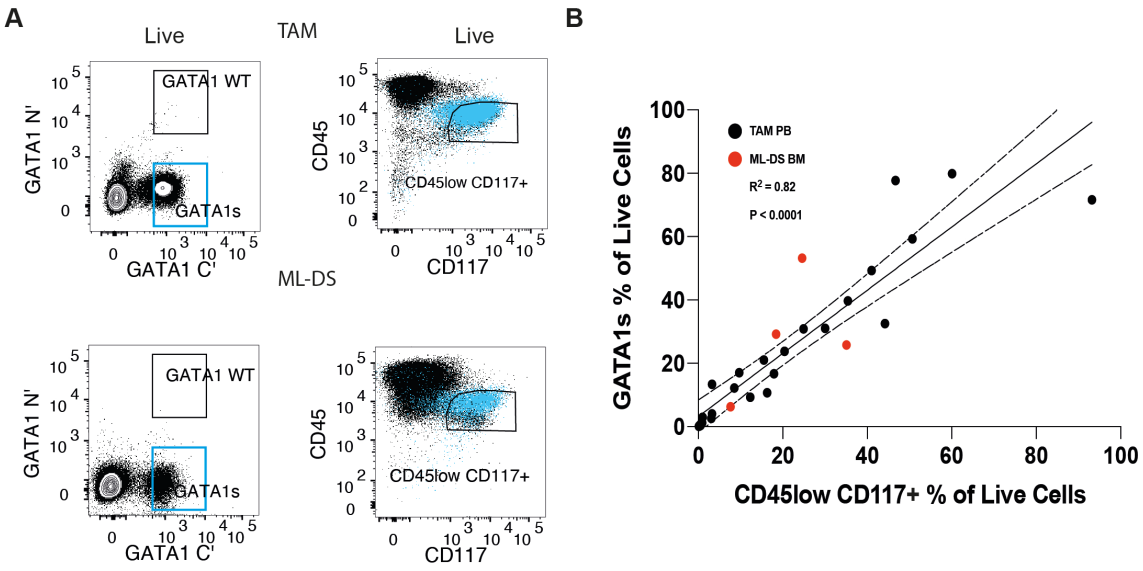


Figure 3.6 GATA1s^{only} cells in TAM and ML-DS co-locate in the CD45lowCD117+ gate

A) Representative flow cytometry plots overlaying GATA1s^{only} cells (blue gate) on the CD45^{low} CD117^{+} blast gate in TAM (top) and ML-DS (bottom). B) Linear regression analysis between live CD45^{low} CD117^{+} cells (x-axis) and the live GATA1s^{only} cells (y-axis) in peripheral blood TAM (black circles) and bone marrow ML-DS (red circles). R2 = 0.82, P<0.0001, n=32 samples (26 TAM and 5 ML-DS).

CD7 is often expressed in immature B, T and Natural killer cells. Detailed analysis of fetal liver haematopoietic stem and progenitor cells in normal and Down syndrome samples by Dr. Anindita Roy determined that CD7 expression ubiquitously expressed and is not limited to the CD45RA^{+} lymphoid compartment (A. Roy PhD thesis). To assess expression of CD7 in control TAM and ML-DS samples, first cells positive for CD117, CD45^{low} and CD7 were identified for each sample. Although the percent of live cells positive for CD45^{low} CD117+ CD7+ was higher in TAM (12.3% +/-10.9%) and ML-DS cells (14.6% +/- 10.5%) than DS neonates with no *GATA1s* mutations (1.8%+/- 3.1%), or post chemotherapy ML-DS (2.3%+/- 2.2%), it varied highly. The reason for this is that all samples show a population of CD45^{low}CD117^{+}cells, with CD7 expression irrespective of *GATA1s* mutation status. GATA1 negative, GATA1s^{only} and GATA1^{WT} cells, all show expression of CD7 (Figure 3.7). Due to the presence of a small population of CD45^{low}CD117^{+} CD7^{+} cells (1.8+/- 3.1%) in DS neonates without *GATA1s* mutation

and in ML-DS post chemotherapy (16.3 +/- 12.5%) confounds the use of surface CD7 expression to identify *GATA1s* mutant cells in TAM and ML-DS.

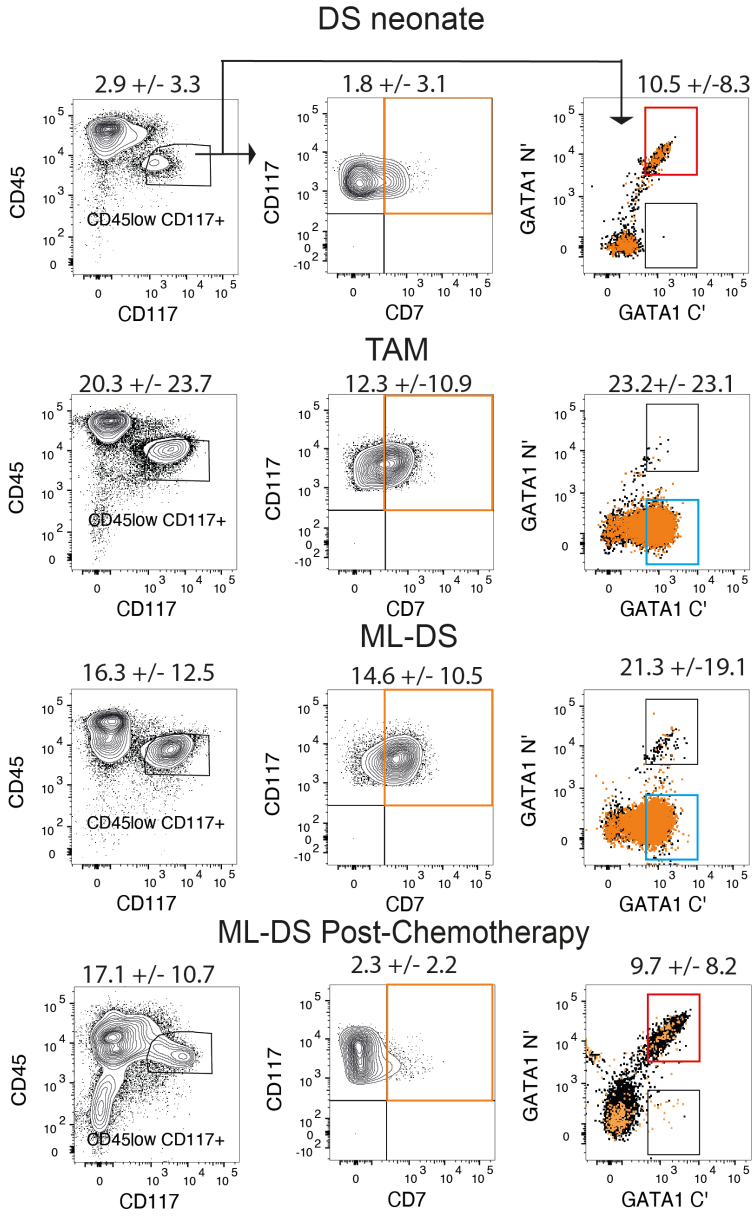


Figure 3.7 CD7 is expression in TAM and ML-DS as well as GATA1^{WT} cells.

(E) Representative flow cytometry plots of PB DS neonate, PB TAM, and BM ML-DS before and after chemotherapy. Cells within the CD45^{low} CD117^{+} (left) were plotted in the CD117 vs CD7 (middle) or in the GATA1 N terminus (GATA1 N') vs GATA1 C terminus (GATA1 C') (left). CD117^{+} CD7^{+} cells (orange box) were overlaid (as orange dots) on the GATA1 N' and GATA1 C' axis. Figures indicate the mean percentage +/- 1 standard deviation (SD) of cells as percent of all live cells for comparison. For the GATA1 N' vs GATA1 C' plots, the percentages referred to either GATA1 WT (red box) for DS neonates and ML-DS post chemotherapy samples or GATA1s (blue box) for TAM and ML-DS samples. Number of samples tested: 7 PB DS neonate samples, 26 PB TAM samples, and 6 and 3 BM ML-DS samples before and after chemotherapy, respectively.

3.4 Intracellular Detection of GATA1s as a MRD Tool: A Proof of Principle

While ML-DS patients respond well to chemotherapy[141], mainly due to ML-DS cells having an enhanced sensitivity to chemotherapy agents[142], there are still approximately 10% of patients who fail to enter complete remission or who relapsed[142-144]. For AML, complete remission (CR) criteria following treatment include less than 5% of blast cells in bone marrow as well as recovery of peripheral blood counts. Morphology based assessment of leukaemia blast cells is often used to defined CR, but it has a limit of detection of 1 leukaemic cell in 20 normal cells (5%)[145]. MRD, or measurable residual disease, is a newly emerging set of tools that includes multiparameter flow cytometry (sensitivity: 0.1% or 1 leukaemic cell in 1,000 cells), qPCR (0.01% or 1 leukaemic cell in 10,000 cells) and conventional NGS (0.3% to 1% or 1 leukaemic cell in 100 normal cells)[7] that can identify leukaemic cells far below the morphology-based detection of 5%[145]. Multiparameter flow cytometry based MRD requires leukaemic associated immunophenotypes, which have not yet been defined for ML-DS and might differ before and after treatment. qPCR MRD based methods, while being highly sensitive, require specific, well characterized DNA variations such as hotspots mutations in the gene encoding NPM1, chromosomal translocations such as *RUNX1-RUNX1T1*, *CBFB-MYH11*, and *PML-RARA*, among others. Due to the lack of suitable markers in TAM and ML-DS, MRD detection of leukaemic cells at complete remission or during diagnosis of TAM or ML-DS have been limited to NGS[146-148].

Why is MRD important in TAM and ML-DS? First, for the past 20 years it has been established that presence of MRD positive cells after treatment is associated with an increased risk of AML relapse and short patient survival in paediatric or adult AML[145]. Second, a highly sensitive and specific MRD tool could help determine Down syndrome patients at risk of developing leukaemia and it can also help identify the ~10% of ML-DS patients who will relapse after chemotherapy treatment.

I anticipate that iFC detection of GATA1s^{only} cells in TAM and ML-DS coupled to conventional cell surface flow cytometry (FC) could be used as an MRD tool. However, it needs to be clarified that the iFC detection of GATA1s^{only} cells as MRD tool will required extensive testing before it is accepted as a standardized assay. First, a binary cut-off value should be defined for calling positive and negative in a training set of samples and validated in a second confirmation set. Moreover, the MRD assay would also have to be applied to a larger ML-DS cohort following chemotherapy treatment with appropriate follow up to determine if there is a high probability for MRD^{+} patients to relapse. Nonetheless, as a proof of principle I applied iFC detection of GATA1s^{only} cells in 5 ML-DS patients, three of which had no detectable *GATA1s* mutations by NGS after chemotherapy treatment. First, I immunophenotyped the GATA1 positive compartment in ML-DS samples before and after treatment with the goal of measuring bone marrow reconstitution (Figure 3.8). The immature compartment (CD45^{low} CD117^{+}) in ML-DS before treatment was composed mainly of GATA1s^{only} mutant cells expressing high levels of CD36 (87.6% +/- 7.4%) and CD7(86.32% +/- 7.2%), and variable expression of glycophorin A (48.2 +/- 28%) and CD34 (34% +/- 25%) with low levels of CD41(12% +/- 21.0%) (Figure 3.8 B C). After chemotherapy treatment, cells within the CD45^{low}CD117^{+} compartment were mainly GATA1^{WT} cells by evidence of expression of both GATA1 N' and C' terminus. GATA1^{WT} cells expressed reduced levels of CD36 (38.4% +/- 18.3%) and CD7 (20.7% +/- 21%), as well as high levels of CD34 (50% +/- 18%). This data shows that chemotherapy can eradicate GATA1s^{only} cells, and residual GATA1^{WT} cells can then expand with an immature phenotype. These two aspects, the quantification of GATA1s^{only} cells and reconstitution of a GATA1^{WT} compartment, make iFC an ideal MRD tool in TAM and ML-DS. A binary cut-off value to determine MRD^{+} and MRD^{-} as well as the sensitivity and specificity of the iFC method is explored in the next section.

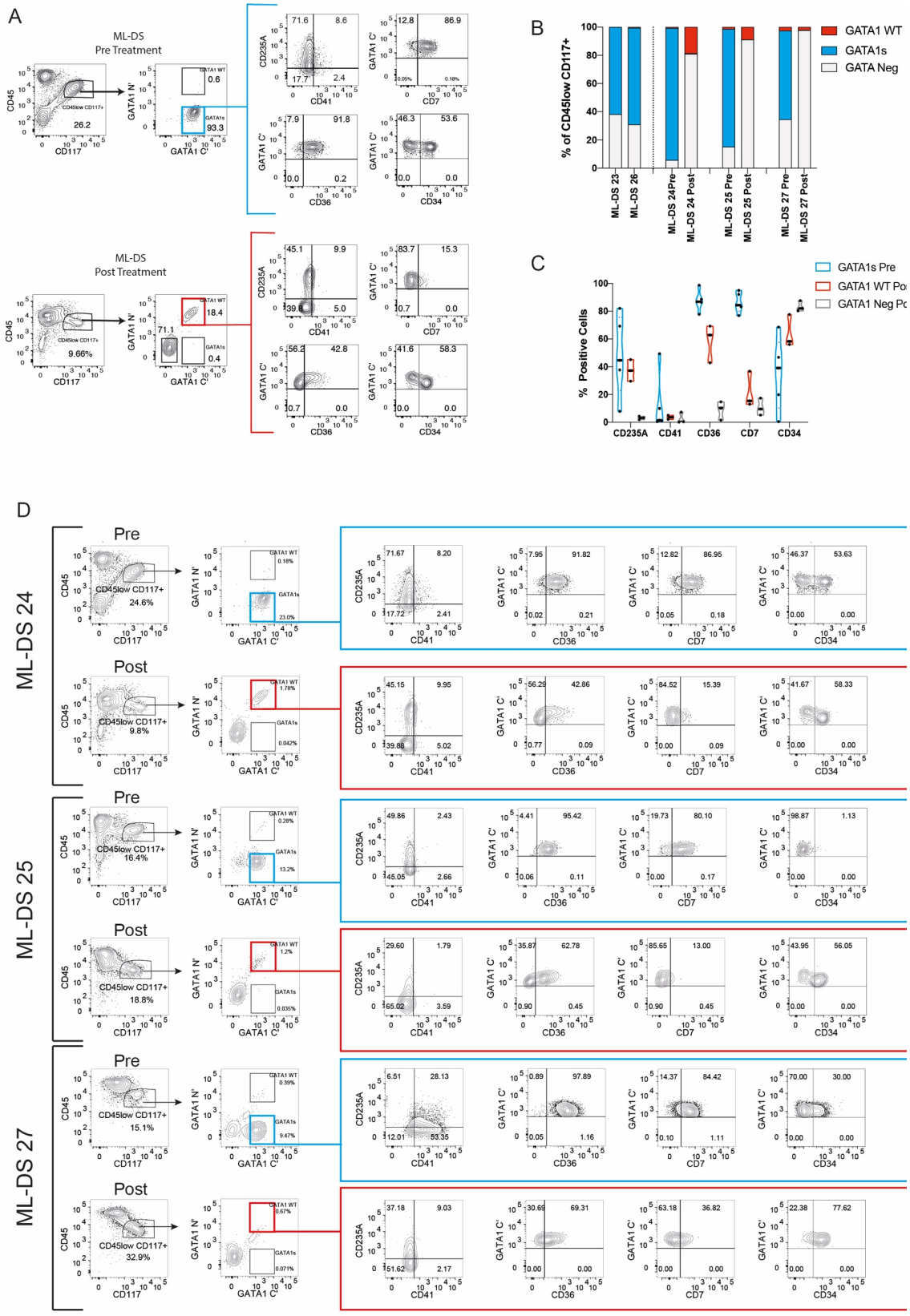

A
ML-DS
Pre Treatment
ML-DS
Post Treatment
B
% of CD45low CD117+
GATA1 WT
GATA1s
GATA Neg
ML-DS 23
ML-DS 26
ML-DS 24Pre
ML-DS 24 Post
ML-DS 25 Pre
ML-DS 25 Post
ML-DS 27 Pre
ML-DS 27 Post
C
% Positive Cells
GATA1s Pre
GATA1 WT Po
GATA1 Neg Po
CD235A
CD41
CD36
CD7
CD34
D
ML-DS 24
ML-DS 25
ML-DS 27
Pre
Post

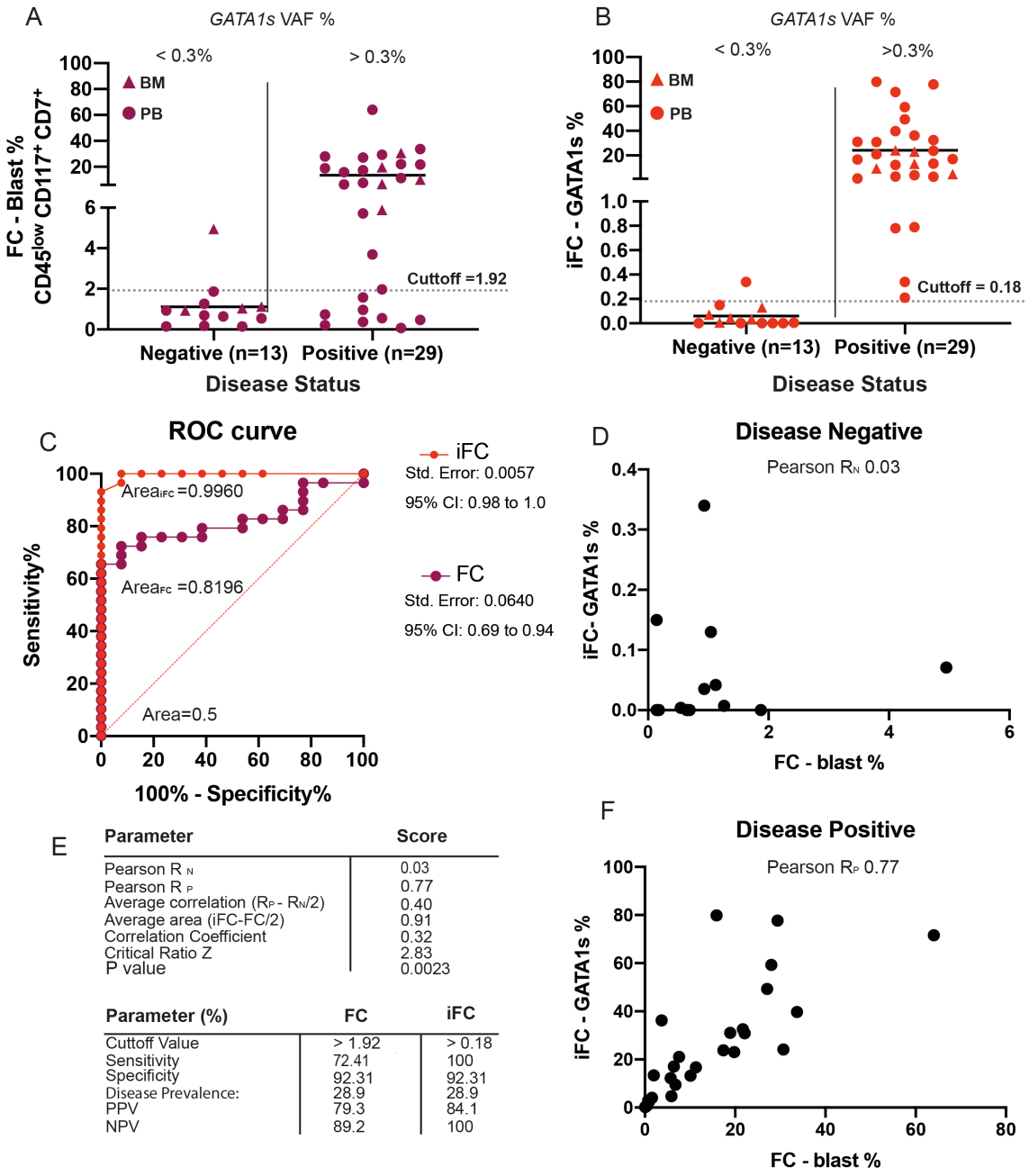
Figure 3.8 Immunophenotyping of ML-DS cells before and after chemotherapy

A) Representative flow cytometry panels of ML-DS samples before (top) and after chemotherapy (bottom) in which additional markers previously described as aberrant (CD235A, CD41, CD7, CD36, and CD34) were examined in GATA1s^{only} cells (blue gate) and GATA1^{WT} cells (red gate). (B) Bar graph showing the proportion of GATA1s^{only} (blue), GATA1^{WT} (red) and GATA1^{negative} (grey) cells within the CD45^{low} CD117^{+} gate in 5 pre-chemotherapy and 3 post-chemotherapy ML-DS samples. (C) Violin plots showing the percent of cells positive for CD235A, CD41, CD36, CD34, and CD7 in GATA1s^{only} cells before chemotherapy (blue), GATA1^{WT} after chemotherapy (red), and GATA1^{negative} cells after chemotherapy (grey). D) Flow cytometry analysis of CD45^{low}CD117^{+} in paired ML-DS samples before (top) and after chemotherapy (bottom) as in A. Cell frequencies are shown as percent of live cells for CD45^{low} CD117^{+} gate, the GATA1^{WT} gate and the GATA1s gate for comparison. Expression of CD41, CD235a, CD36, CD7 CD34 is shown as percent of GATA1s^{+} cells (blue gating) or GATA1^{WT} cells (red gating) of 3 ML-DS paired samples before chemotherapy (top) and after chemotherapy (bottom)

3.5 Intracellular Detection of GATA1s in TAM and ML-DS is Highly Sensitive and Specific

If the iFC method were to be used as an MRD tool in ML-DS, it would require a sensitivity similar to the current gold standard and better than the conventional cell surface flow cytometry. NGS, the gold standard has a sensitivity of 1.0% to 0.3%[7]. Because the purpose of the MRD tool is to determine the presence of MRD^{+} cells according to a cut-off value, first I classified the same set of samples previously used (7 DS neonates, 26 TAM and 9 ML-DS) as MRD^{+} or MRD^{-} if the *GATA1s* VAF was above or below 0.3% respectively. This classification determined that 13 samples were MRD^{-} and 29 samples were MRD^{+}. For each sample the frequency of TAM or ML-DS cell by conventional cell surface markers (CD45^{low} CD117^{+} CD7^{+}) was determined (Figure 3.9 A). Similarly, the frequency of TAM and ML-DS cells by intracellular detection of GATA1s^{only} cells was also determined (Figure 3.9 B). Then, receiver operating characteristic (ROC) analysis was employed to determine the ideal cut-off value for each method[149]. ROC analysis identifies the sensitivity and specificity for a range of cut-off values for each method. In simple terms, sensitivity is defined as the fraction of MRD^{+} cells (as determined by NGS) that the iFC test correctly identifies as positive, and specificity is the fraction of MRD^{-} cells that that the iFC test correctly identifies as negative. In practice, one should aim to maximize both sensitivity and specificity, but it

is not always possible. It is most likely that for a given cut-off value there will be a trade-off between specificity and sensitivity. If the cut-off value is too high for either the conventional flow cytometry or iFC, we run the risk of missing samples with low levels of TAM or ML-DS cells, although we wouldn't mistakenly identify disease cells in patients who don't have it. If the cut-off value is too low, we would be able to identify rare TAM and ML-Ds cells but at the same time we run the risk of classifying young infants as having TAM or ML-DS cells when truly they don't have any. ROC analysis identified an optimal cut-off value for the flow cytometry cell surface-based identification of TAM and ML-DS cells as 1.92% or 1 leukaemic cells per 50 normal cells. At this value, the test was able to identify 19 out of 29 samples as MRD+ (66%) and 12 out of 13 as MRD^- (92%). With an area under the ROC curve of 0.81, an 81% chance for the for the conventional flow cytometric method to identify TAM and ML-DS samples as MRD^+ (Figure 3.9). The optimal cut-off value for the iFC method was 0.18% or ~1 leukaemic cell in 1,000 normal cells. At this cut-off value, the iFC method identified 29 out of 29 (100%) samples as MRD^+ and 12 out of 13 as MRD^- (92%), with an area under the ROC curve of 0.99, a 99% chance for iFC method to identify TAM and ML-DS samples as MRD^+(Figure 3.9 C). The difference in areas under the curve between iFC or FC was statistically significant (Z=2.83, p-value =0.0023). Moreover, the ROC analysis can also help determine the positive predictive value (PPV) for a given diagnostic method if the prevalence of the disease is known. In the context of MRD, a PPV is the probability that a patient has leukaemic cells if the MRD assay is positive adjusted by the prevalence of TAM. It has been documented that *GATA1s* mutations occur in up to 28.9 % of young infants with Down syndrome[7] and are affected by TAM, with this disease prevalence a positive predictive value of 84.1% was determined for the iFC and 79.3% for conventional cell surface flow cytometry (Figure 3.9 E).



Parameter	Score
Pearson R N	0.03
Pearson R P	0.77
Average correlation (RP - RN/2)	0.40
Average area (iFC-FC/2)	0.91
Correlation Coefficient	0.32
Critical Ratio Z	2.83
P value	0.0023

Parameter (%)	FC	iFC
Cuttoff Value	> 1.92	> 0.18
Sensitivity	72.41	100
Specificity	92.31	92.31
Disease Prevalence:	28.9	28.9
PPV	79.3	84.1
NPV	89.2	100

Figure 3.9 Receiver operating characteristic analysis of conventional FC and iFC methods

A) Dot plots shows the % of malignant cells (y-axis) defined by surface flow cytometry (FC-$CD45^{low}$, $CD7^{+}$, $CD117^{+}$) in 7 DS neonates, 26 TAM and 9 ML-DS samples categorized as disease negative (*GATA1s* VAF < 0.3%, n=13) or disease positive (*GATA1s* VAF >0.3%, n=29) according to the GATA1 variant allele frequency identified by next generation sequencing. Dots mark peripheral blood samples (PB) and triangles mark bone marrow samples (BM), dotted grey lines shows the cut-off value calculated based on the ROC analysis. B) Dot plots showing the % of malignant cells (y-axis) defined by intracellular flow cytometry identification of $GATA1s^{+}$ cells (iFC-GATA1s) in the same set of samples, the rest is as in B. (**C**) ROC curves for the iFC method (red dots) or FC method (purple dots) were calculated by plotting sensitivity % (y-axis) vs 100%-Specificity % (x-axis) at different cut-off values. The area under the curve

is shown below each curve. Std error, standard error; CI, confidence interval. D) Pearson correlation between the percent of malignant cells identified by the FC and iFC methods among the disease negative samples. Pearson R_N, correlation coefficient. E) Top, parameters calculated according to Hanley and McNeil[150] to assess if the difference in areas under the curve between iFC or FC is statistically significant (Z=2.83, P value =0.0023). Bottom, parameters calculated from the ROC analysis. PPV, positive predicted value; NPV negative predicted value.

The previous ROC analysis determined that the iFC method is superior to conventional cell surface flow cytometry and performs similar to NGS. To confirm the cut-off value of 0.18% as determined by ROC analysis, I serially diluted a GATA1s^{only} cell line (CMK) into a mixture of GATA1 negative (THP1) and wild type GATA1-expressing K562 cell (Figure 3.10 A, B). I then, tried to define the sensitivity of the iFC method by recording 500,000 flow cytometric events and determining the frequency of GATA1s^{only} cells. I identified GATA1s^{only} cells at a frequency greater than ~0.1% (0.78% - 0.11%) (Figure 3.10C). Similarly, at this frequency, GATA1s^{only} cells were not detected in three ML-DS samples taken post-chemotherapy with no detectable *GATA1s* mutations by NGS (Figure 3.10 D). The three ML-DS patients did not relapse 4 years after the sample was taken.

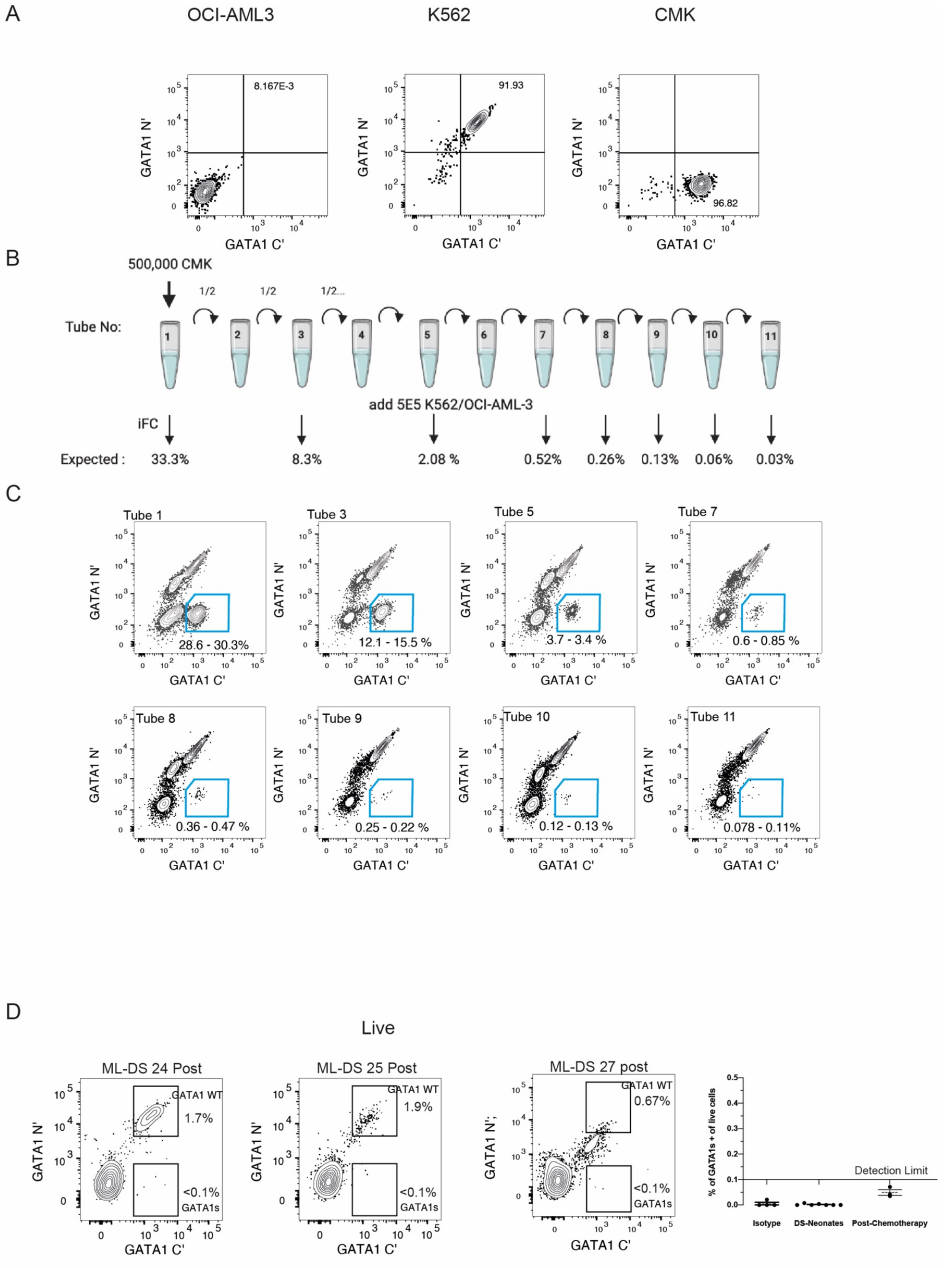


Figure 3.10 Validation of the iFC sensitivity.

A) Flow cytometry plots of dual immunostaining with antibodies against the N' and C' termini of GATA1 in: OCI-AML3 cells (left), a GATA1 negative cell line; K562 cells (middle), a GATA1^{WT} cell line; and CMK (right), a GATA1s^{only} cell line derived from a ML-DS patient. (B) Schematic representation of a serial dilution of CMK cells. A total of 500,000 CMK cells were serially diluted 11 times into an OCI-AML3/K562 cells mix and iFC performed on tubes 1, 3, 5, 7, 8, 9, 10, and 11 Expected frequency of CMK cell is shown below each tube. (C) Representative flow cytometry panels of serially diluted GATA1s CMK cells (blue gate) in a cell mixture of CMK/ OCI-AML3/K562. A total of 500,000 live events were recorded per tube but only a fraction of events is shown for clarity. Each gate contained more than 300 events. Percent of cells in the blue gate from 2 replicates is shown. (D) Top, 3 flow cytometry plots showing the percentage of GATA1^{WT} and GATA1s^{+} cells in the live CD45^{low} CD117^{+} cell compartment in 3 post-chemotherapy ML-DS samples in which no *GATA1s* mutations were identified by NGS. Below, graph of the percentage of GATA1s^{+} cells live CD45^{low} CD117^{+} gate in the isotype controls (4 samples), DS neonates (7 samples), and 3 ML-DS post-chemotherapy samples.

In this chapter I have presented a novel multiparameter flow cytometric technique that combines cell surface expression of Glycophorin A, CD41, CD34, CD36, 117, CD34, CD7 as well as dual intracellular detection of GATA1 isoforms with the goal of identifying leukaemic cells with exclusive GATA1s expression in TAM and ML-DS patients. I showed that this method is highly sensitive and specific and can identify 1 leukaemic cells in 1000 normal cells. Although, the intracellular flow cytometry protocol requires up to 3 hours, it can easily be adapted to clinal haematology laboratories that routinely run flow cytometric assays. Moreover, flow cytometry is a single cell assay, consequently the method develop here is agnostic to the number of GATA1s clones. One observation from this analysis is that most TAM and ML-DS samples expressed mid-levels of Glycophorin A, an erythroid marker while they expressed low levels of CD41, a megakaryocytic maker. This seems to be counterproductive as ML-DS and TAM are categorised as megakaryoblastic leukaemia and pre-leukaemia. The reason for this discrepancy might have to do with the anti CD41 antibody used in this analysis. The clone of the anti CD41 antibody used here is HIP8 while the cloned used in other publications to profile CD41 expression in TAM and ML-DS is P2. Nonetheless, this multiparameter flow cytometry can be easily customized to add other megakaryocytic markers such as CD61, and CD42. Finally, I hope that this method one day can be used as an MRD assay in clinical settings.

4 Single Cell Transcriptional Reference Map of T21 Haematopoiesis

Recent advances in microfluidic capture of a high number of single cells has allowed construction of comprehensive single cell transcriptional landscapes of human tissue samples. With the advent of new integration algorithms, it is now possible to construct such single cell transcriptional landscapes from multiple donors across ontogeny or disease states in hematopoesis[119,122,126]. Here, I aim to generate an accurate

single cell transcriptional landscape of trisomy 21 haematopoiesis to study cell heterogeneity in TAM as well as aberrant cell maturation due to *GATA1s* mutations.

Uniform manifold approximation and projection (UMAP) is being employed for linear dimensionality reduction of the large number of single cells. UMAP provides the most meaningful organization of cell clusters and preserve the local and large-scale structure of the data[151] and allows for visualization of thousands of single cells in two dimensions. However, the novo integration of both trisomy 21 and mutant TAM single cells can distort both the local and large-scale structure of the data despite batch or sample correction. This approach can skew the representation of all single cells in two dimensions, which can hinder the interpretation of such. Moreover, as new single RNA sequencing is performed of new TAM or ML-DS samples, a new analysis would have to be performed again and again which might or might not lead to same conclusions. Therefore, it is preferable to generate an accurate single cell reference data set of trisomy 21 single cells and freeze it. The reference map can then be use at any time to chart single cell RNA sequencing data from TAM or ML-DS samples. If coupled to variant calling *GATA1s* mutations in single cells, mapping of query cells onto a reference map would have three advantages. First, it allows to identify the cellular heterogeneity of TAM. Second, after inferred cellular differentiation, it permits to map exactly where during haematopoiesis cell maturation arrest has occurred. Third, and most importantly, it allows to accurately compare mutant GATA1s cells to their closest normal counterparts.

4.1 Integration of 4 T21 Transcriptomes of Haematopoietic Samples

Single cells from 4 trisomy 21 cord blood donors with no *GATA1s* mutation detected by NGS were captured with the 10X Genomics platform (see Table 9 for cell sorting frequencies). I chose cord blood as it is a readily available source of haematopoietic stem and progenitor cells. Three of the cord blood donors are female while only one is male. For each T21 sample, I captured 5,888, 8,482, 5,647 and 6,492 single cells respectively, with 3,400, 2,583, 2,579, and 3,200 median genes per cell for

each sample (Supplementary Figure 2, Appendix I). First, to filter out compromised cells of low quality, a new data-driven method named miQC[121] was used. This method models both the high percent of read counts mapping to mitochondrial genes and the number of detected genes per single cells. In this way, miQC can provide a single cell scoring to identify low quality cells (red dots) marked by high percent of mitochondrial read and a low gene count (Figure 4.1 A). Then, a second quality control metric was applied. DoubletFidner[120], identifies “doublet cells” (red dots) originating from more than one cell being co encapsulated in oil droplets during sample preparation (Figure 4.1 B). It primarily identifies doublets coming from cell types that substantially differ from each other. This metric identified 294, 424, 282 and 325 doublets for each T21 sample, respectively. Subsequently, outlier cells were excluded with less than 350 and more than 7500 genes (for sample T21-3, cells with more than 6000 genes were excluded). All cells were then merged into a single gene expression matrix composed of 22,310 cells in total. Of all cells in the merged data set, 22.7% (5084 cells) of all cells were derived from sample 1, 31.1% (6949 cells) of cells from sample 2, 20.1 % (4500 cells) of cells from sample 3, and 25.8% (5777 cells) of cells from sample 4.

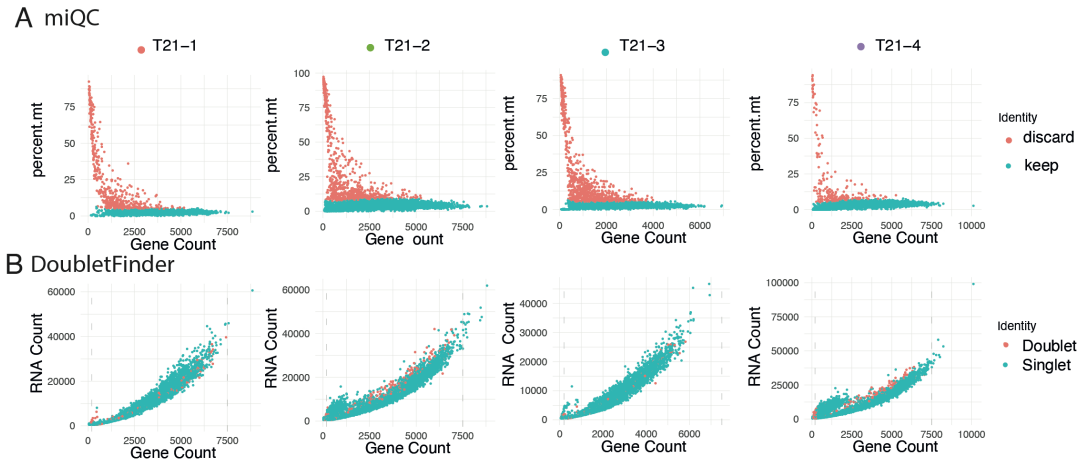


Figure 4.1 Quality control of T21 single cell RNA sequencing

Single cell RNA sequencing was performed in 4 T21 samples (marked 1 to 4) with no *GATA1*s mutations. A) The percentage of reads mapping to mitochondrial genes (percent.mt) (Y-axis) was plotted against the number of genes (Gene Count) (X-axis) for each cell where each dot represents a cell. MiQC was used to determine the cut-off percentage of mitochondrial reads

counts that separates high quality cells (green dots) from low quality cells (red dots). Low quality cells were discarded, and good quality cells were kept. B) RNA counts (Y-axis) were plotted against gene counts (X-axis) and doublets as determined by DoubletFinder were coloured red and were discarded, while singlets were coloured green and were kept for downstream analysis. Single cells with more than 350 or less than 7500 genes (dotted lines) were kept for further analysis.

Seurat[119], a user friendly platformed designed for the analysis of single cell RNA sequencing was used for data normalization and scaling, as well as to perform linear dimensional reduction. Seurat applies a global-scaling normalization method to account mainly for the difference in sequencing depth. I then used Harmony[122] to integrate the 4 trisomy 21 donor samples in an unsupervised manner. Uniform Manifold Approximation and Projection (UMAP) was used to visualize 22,310 single cells in two dimensions. In Figure 4.2 A, we can observe that the 4 T21 samples can be uniquely identified and thus integration to remove donor effects is required. Harmony was able filter out donor effects (Figure 4.2C), sex effects (Figure 4.2D) and cell cycle effects (Figure 4.2B) to integrate the 4 donors into a unifying transcriptional landscape (Figure 4.3).

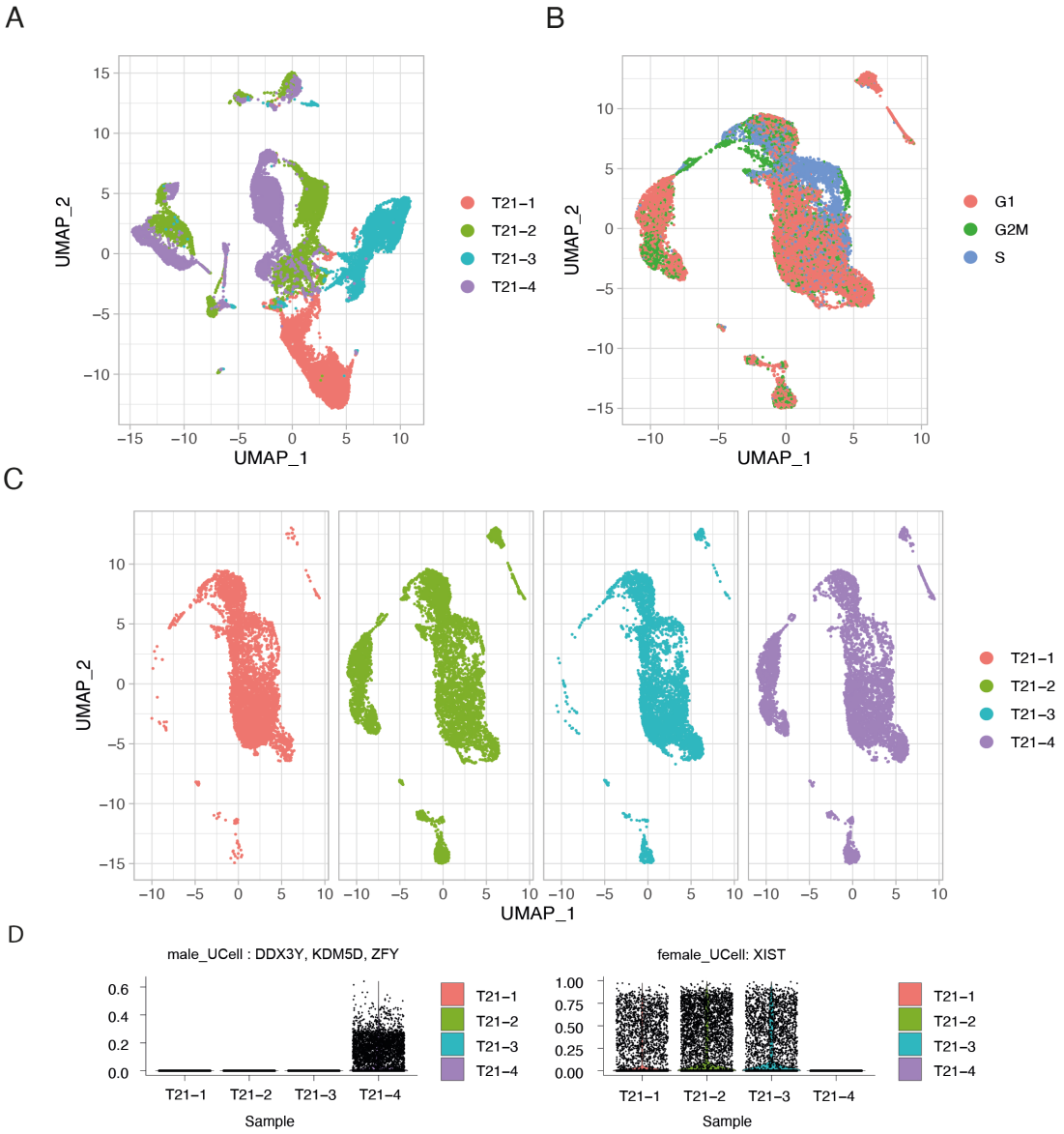


Figure 4.2 Harmony integration of 4 T21 donor samples.

A) linear dimensionality reduction of 4 T21samples in UMAP space before Harmony integration (A) and after harmony integration (B) and (C). B) Single cells are coloured coded according to predicted cell cycle phase. C) Single T21 samples are plotted in UMAP space after harmony integration. D) UCell is used to calculate gene signature scoring for genes in the Y-chromosome (left) for male samples and in the X-chromosome (right) for female samples.

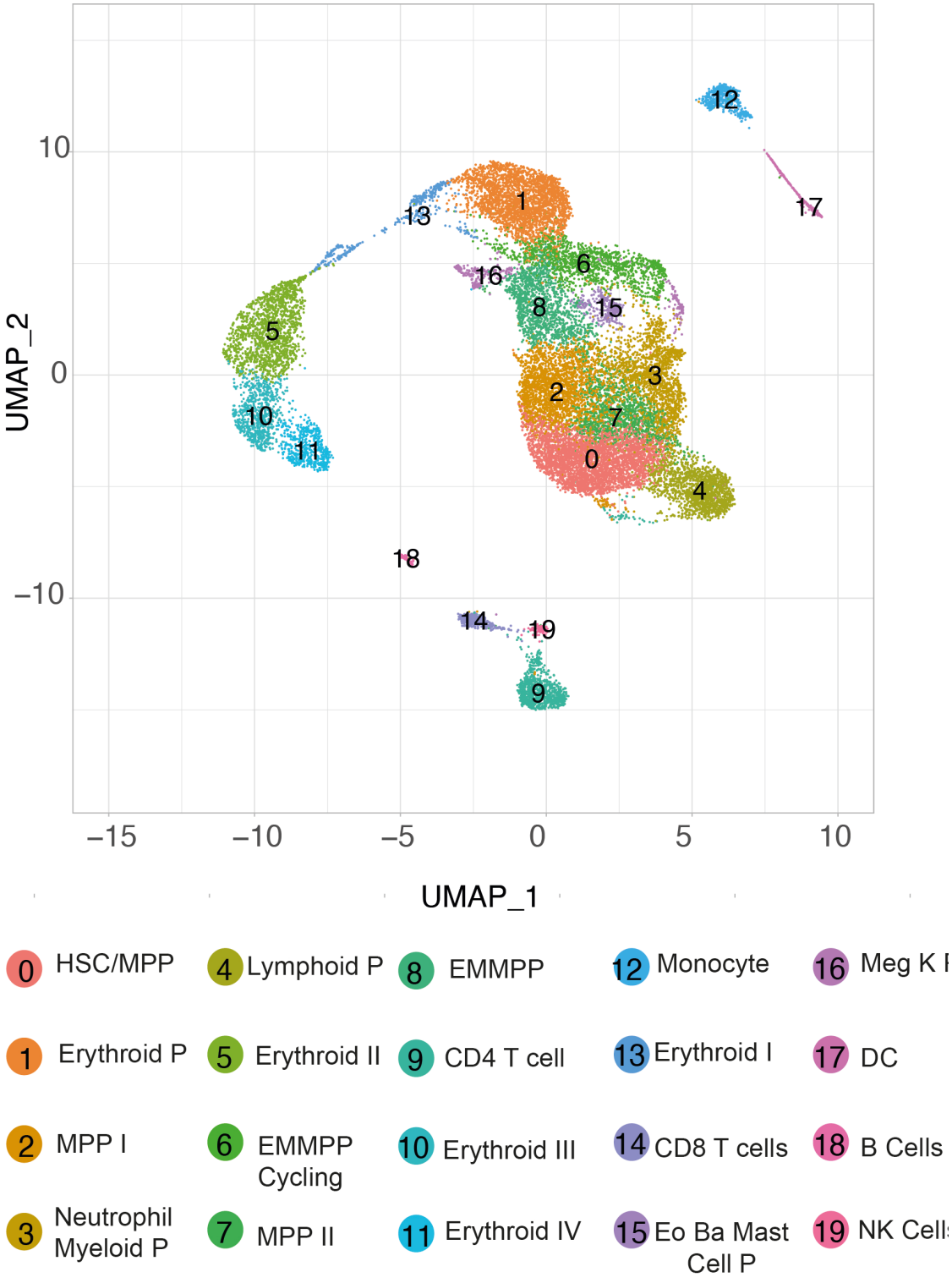


Figure 4.3 A unified transcriptional landscape of T21 single cells composed of 20 cell clusters visualized in UMAP space

4.2 Single Cell Cluster Identification

The *FindClusters* command of the Seurat package was used to identify 20 unique clusters label from 0 to 19. Seurat applies a graph-based clustering approach based on principal components analysis. Annotation of the 20 clusters was initially performed by exploration of known canonical markers. Cluster 9 and 14 were made up of mature *CD4*+ and *CD8*+ T-cells, cluster 18 was made up of B-cells (*PAX5*+, *CD19*+), cluster 19 was made up of Natural Killer cells (*CD56/NCAM1*+), cluster 12 is composed of monocytes (*CD14*+) and cluster 17 is composed of dendritic cells (*CD123/IL3RA*+) (Figure 4.4).

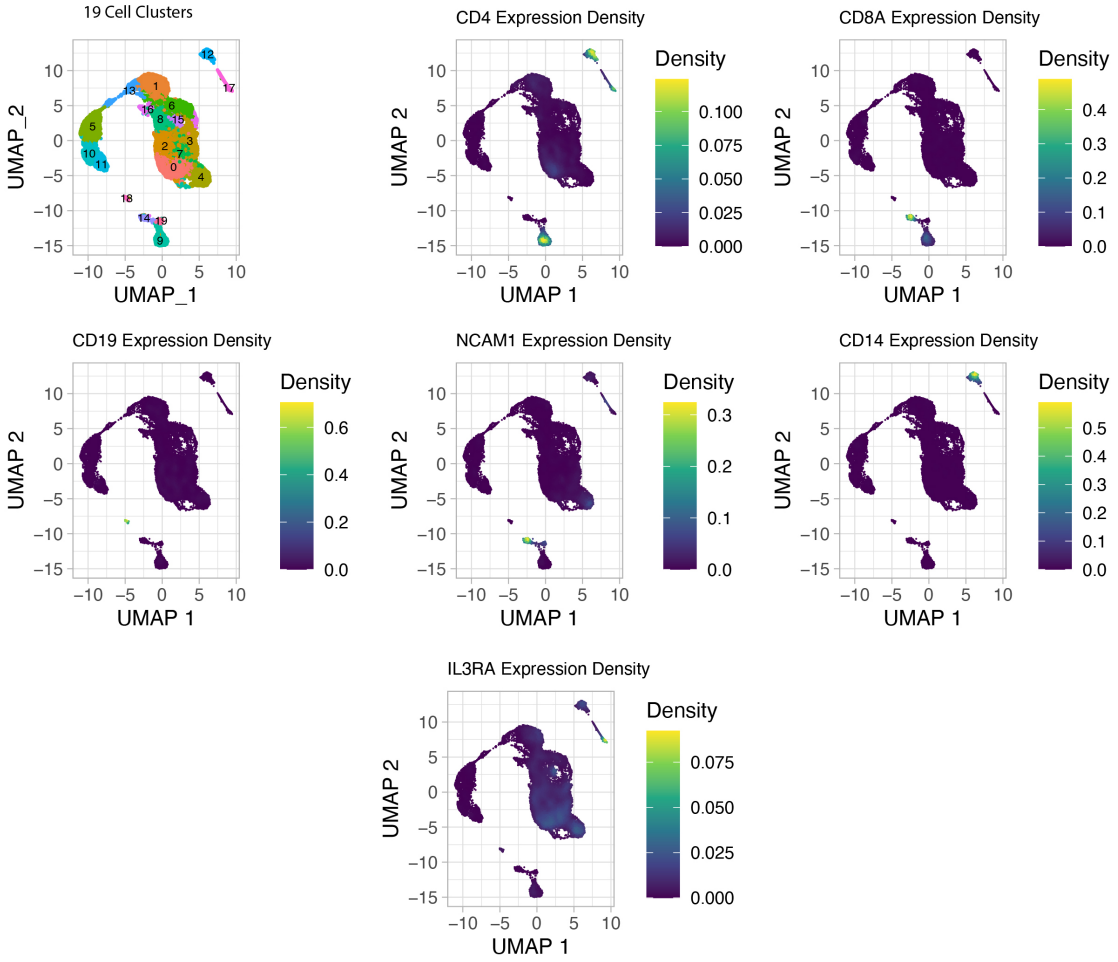


Figure 4.4 Identification of mature T, B, NK cells as well as monocyte and dendritic cells by canonical markers

Kernel density estimation of canonical gene expression by Nebulosa. Top left, UMAP space of 20 clusters, the rest represents density expression of CD19, CD4, CD8A, NCAM encoding CD56, and CD123 encoded by IL3RA.

To facilitate annotation of the rest of the clusters, a set of signature genes specific for haematopoietic stem and progenitor cells from 5 single-cell RNA sequencing publications collected and curated by Roy A, et al. 2020[126] were used. These set of genes were used for gene set enrichment analysis for each cell cluster identified. The -log10 of the resulting adjusted p-value of gene set enrichment scores for each cell group was used to perform hierarchical clustering. This analysis demonstrated that clusters 5, 10, and 11 are mainly of the erythroid lineage, while cluster 13 and 16 are of the erythroid/megakaryocyte lineages. Cluster 1, 6, 8, 15, are made up of cells from the erythroid/megakaryocyte and myeloid lineages, most likely of the basophil/eosinophil/mast cell lineages. Finally, clusters 0, 2, 3, 4, 7 represented HSC/MPP and progenitor cells (Figure 4.5).



Figure 4.5 Heatmap generated using the pheatmap library package for cluster identification

First ranking all the genes for each cluster (x-axis) was achieved with the presto library, then gene set enrichment analysis was performed with a curated haematopoietic gene set (y-axis) using the fgsea library package. The negative log 10 of the adjusted p-value of positive gene set enrichment scores was used for hierarchical clustering and plotting. The -log10 significant p value of 0.05 is equal to 1.3. Values greater than 1.3 are significant and correspond to green colour.

The annotation of each cluster was further refined with the help of a curated set of genes[126] specific for erythroid (*EPOR, KLF1*), myeloid (*MPO, AZU1, ELANE*), lymphoid (*VPREB1, VPREB2*), megakaryocyte (*GP9, VWF*), HSC/MPP (*AVP, HOPX*, *MLLT3*), and eosinophil/basophil/mast cell (*IL5RA, LMO4, EPX*), see Table 5 for a complete list of genes. UCell[127], a bioinformatic package based on the Mann-Whitney U statistic was used to calculate gene signature scores (Figure 4.6).

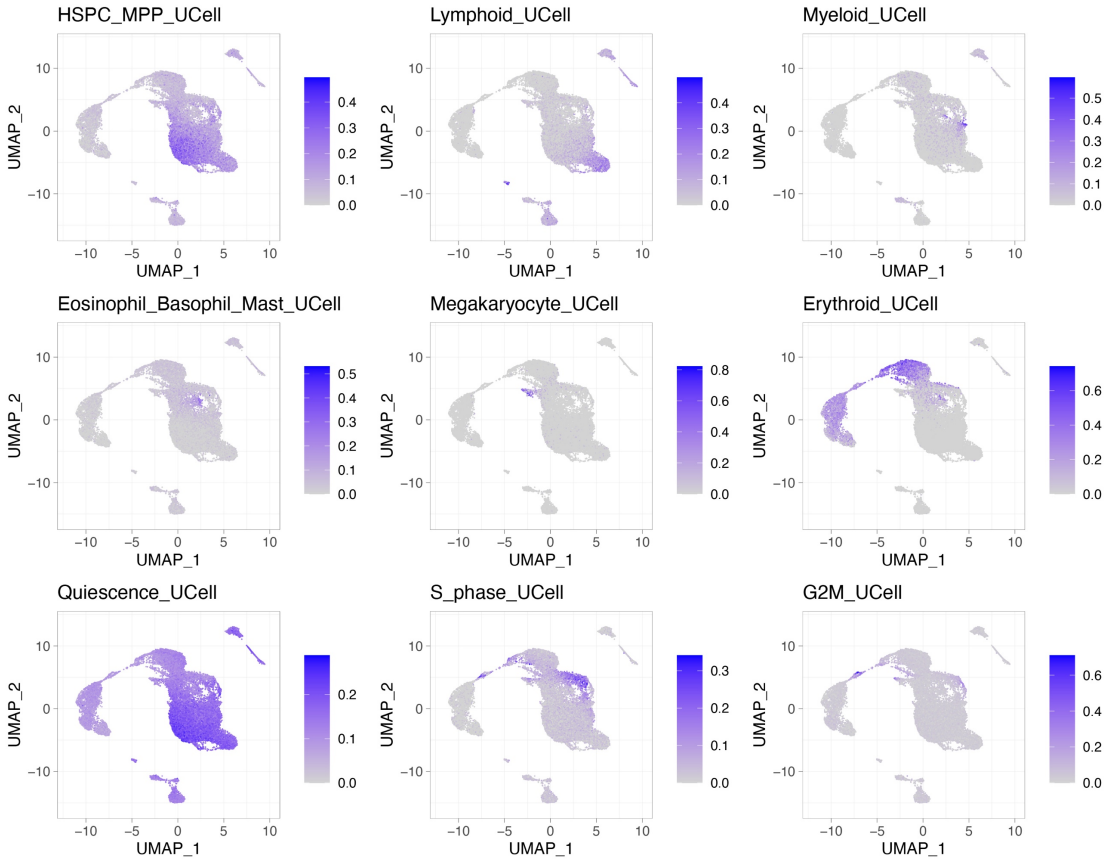


Figure 4.6 UCell gene set scoring for cluster annotation

UCell scoring based on the Mann-Whitney U statistic was used to evaluate the strength of a curated gene set for each haematopoietic lineage shown. Uniform manifold approximation and projection (UMAP) of 22,310 single T21 haematopoietic stem and progenitor cells where 20 cell clusters are identified by a graph-based clustering approach and optimized based on the Louvain algorithm. Haematopoietic stem and multipotent progenitor (HSC/MPP), Multipotent Progenitor I (MPP I), Multipotent Progenitor II (MPP II), Erythroid/Megakaryocyte Multi-Potent Progenitor (EMMPP), Natural Killer Cell (NK Cell), Dendritic Cell (DC), Megakaryocyte Progenitor (Meg K P).

Human haematopoietic cells can be classified in a simplified way as stem and progenitor cells marked by cell surface expression of CD34, precursor cells marked by KIT expression, and mature cells with no KIT or CD34 expression. Due to the sparsity of single cell RNA expression, the visualization of gene expression is not always optimal. Nebulosa[152] , a bioinformatic tool based on a kernel density estimation, can help to visualize gene expression. In simple words, kernel density estimation is a statistical method that replaces a discrete histogram for a continues one, useful when one needs to visualize only the shape of the data. Inspection of *CD34* and *KIT* density expression suggest that clusters 0, 2, 3, 4 and 7 are stem and progenitor cells while clusters 1, 6, 8, 15, 16 are precursor cells marked by gain of KIT expression. Clusters 13, 5, 10, and 11 are $CD34^{-}$ KIT^{-} mature cells (Figure 4.7)

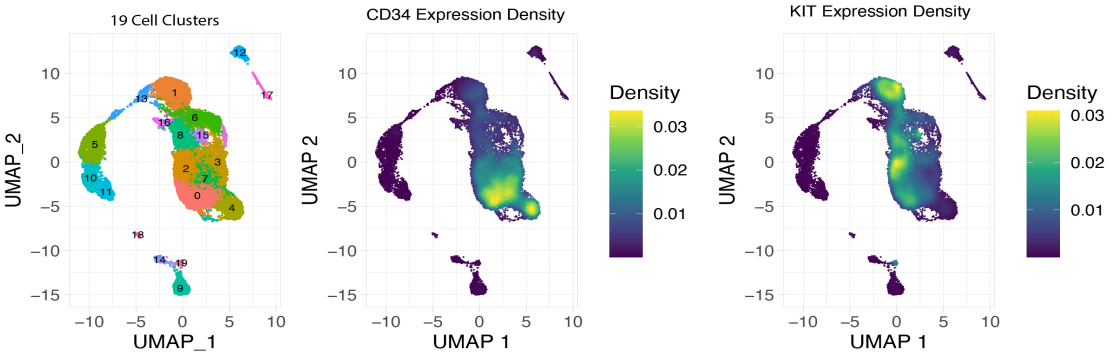


Figure 4.7 Identification of progenitor, precursor, and mature cell by expression density of CD34 and KIT

Kernel density estimation with Nebulosa is used to visualize CD34 (middle) and KIT (right) expression density to identify stem and progenitor, precursor, and mature haematopoietic cell clusters.

The overall architecture of the transcriptional reference map of trisomy 21, consisted of cluster 0 representing the most primitive cells marked by a series of genes that have been associated with self-renewal such as *KLF2*[153] or have been used to mark human HSCs such as *FLT3* [154] *HLF*[155], and *AVP*[156]. Consequently, this cluster was named HSC/MPP for haematopoietic stem and multipotential progenitors (Figure 4.8).

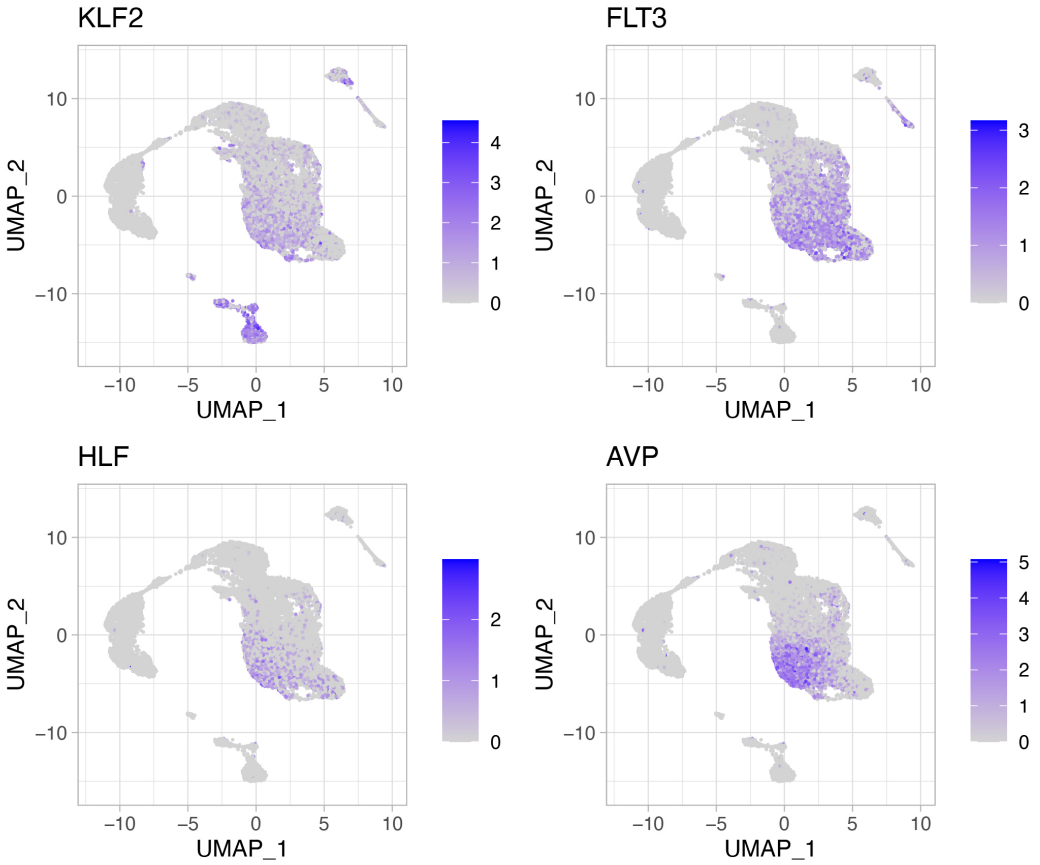


Figure 4.8 HSC/MPPs are marked by *KLF2, FLT3, HLF* and *AVP* expression

Normalized, log transformed expression of *KLF2, FLT3, HLF* and *AVP* visualized in UMAP space

Differential gene expression analysis among the *CD34+* stem and progenitor cell clusters (0,2,3,4,7), based on the Wilcoxon Rank-Sum test, showed that cluster 2, and 7, which surround HSC/MPP are primarily marked by downregulation of HSC/MPP associated genes (Figure 4.9) and were named MPP I and MPP II respectively. Cluster 2 or MPPI shares genes with cluster 8 (*KIT*, *RNF130, FCER1A)* and thus it might represent an intermediate state between HSC/MPP and cluster 8. Cluster 3, 4 and 8 are already marked by lymphoid (*CD7*, *IGHM*, *LTB*, *IRF8*), myeloid (*MPO*, *MIF*) and erythroid/megakaryocyte/basophil (*GATA1*, *ITGA2B*, *CPA3*, *CSF1*) associated genes respectively (Figure 4.9). Therefore, they were named Neutrophil/Myeloid Progenitor (cluster 3), Lymphoid progenitor (cluster 4), and Erythroid-Megakaryocyte Multi-Potential Progenitor (EMMPP) (cluster8).

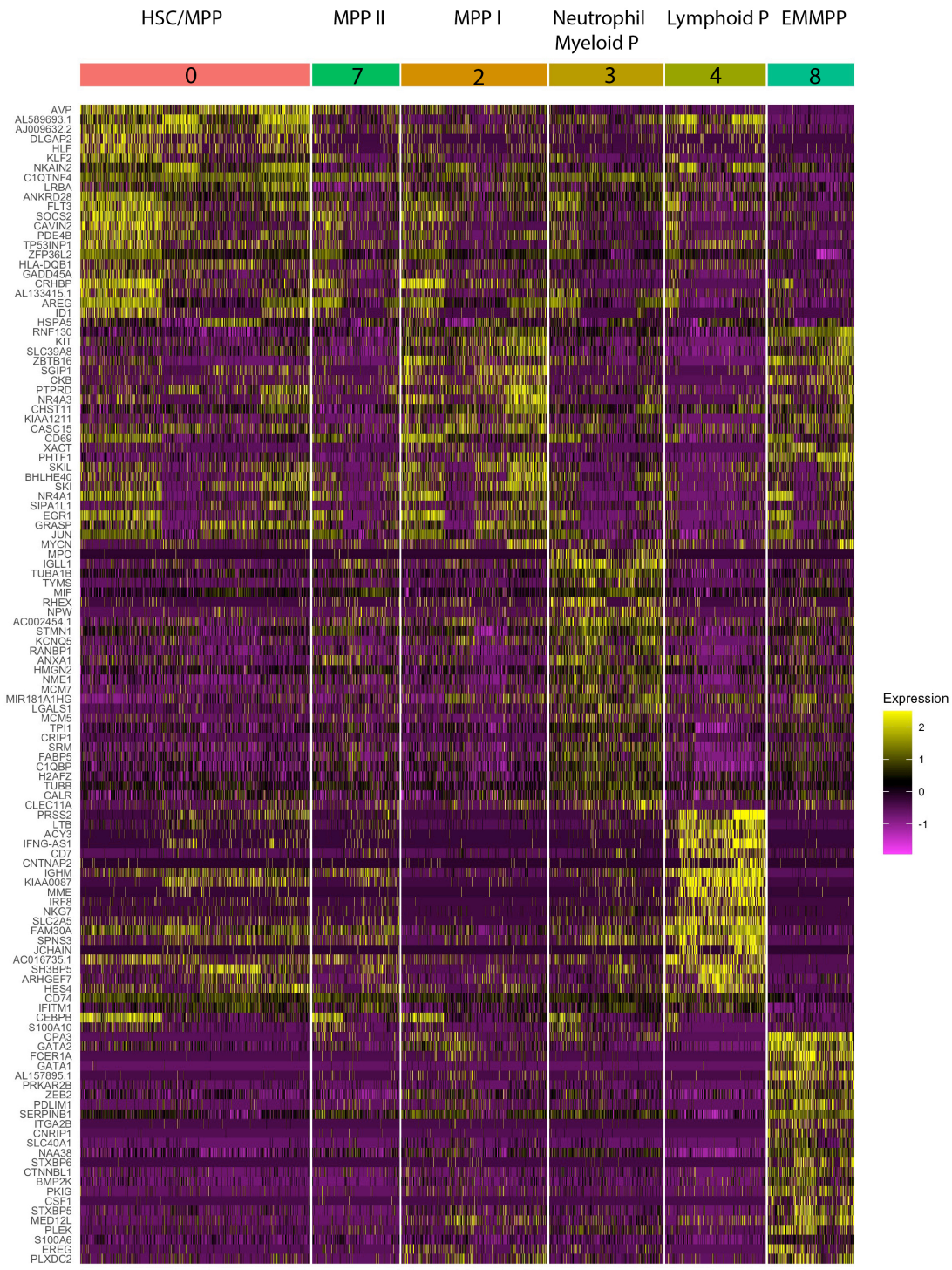


Figure 4.9 Heatmap showing the top 30 genes with the highest log2 fold change in the most immature cell compartments

Heatmap showing the top 30 most differentially expressed genes in the immature compartments determined by Wilcoxon Rank-Sum test. Rows represent genes and columns represent single cells. The colour coding of cell compartments and naming follows as Figure 4.6.

Clusters surrounding the EMMPP cluster, correspond to megakaryocyte progenitors (Meg K P) (cluster 16), erythroid progenitors (Erythroid P) (cluster 1) and eosinophil/basophil/mast cell progenitors (Eo/Ba/Mast Cell P) (cluster 15) as well as a cluster 6, mainly marked by upregulation of genes specific for the S-phase of the cell cycle, such as MCM2-7 whose product forms the minichromosomal maintenance protein complex, a DNA helicase necessary for DNA replication (Figure 4.10).

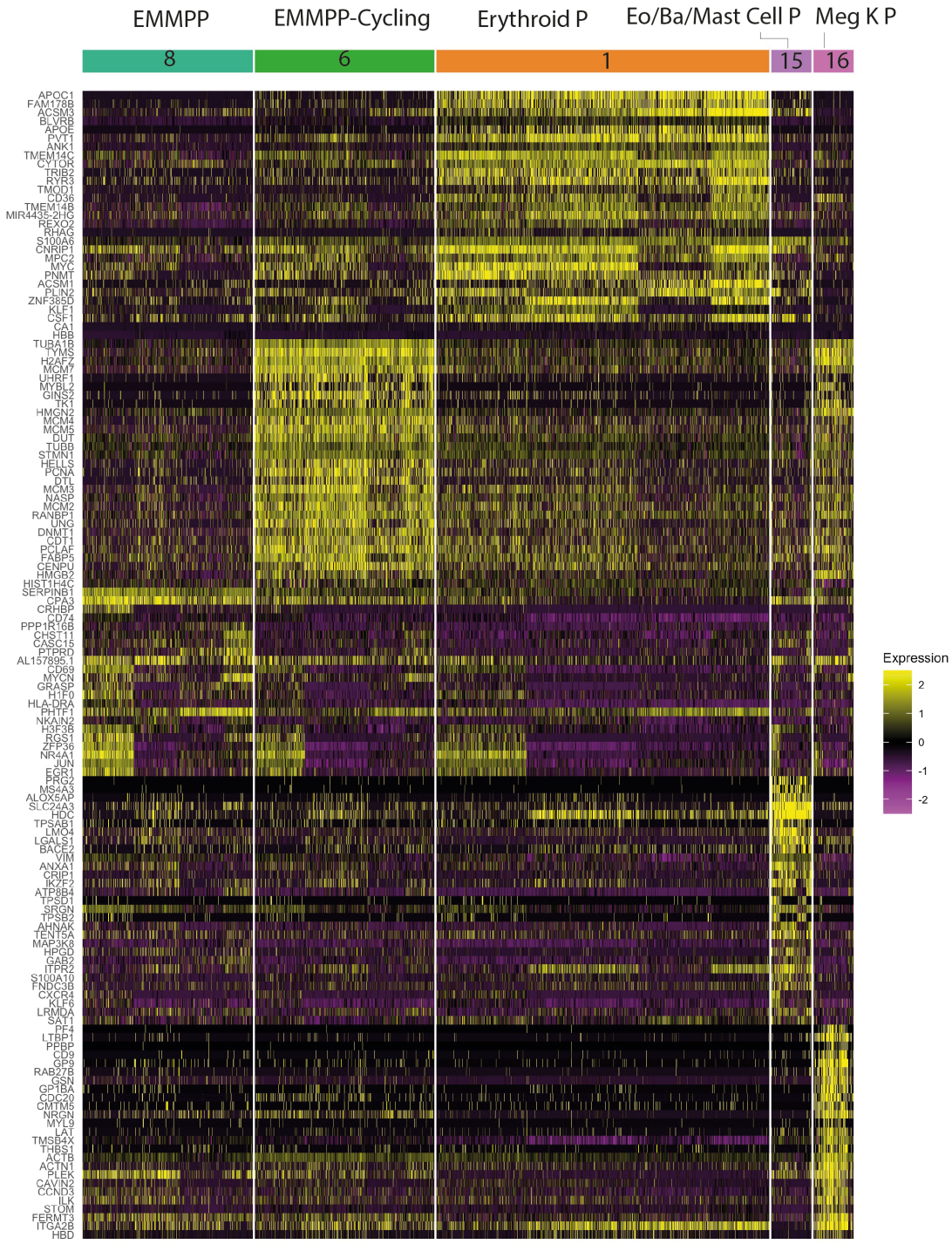


Figure 4.10 Heatmap showing the top 30 genes with the highest log2 fold change in the GATA1+ immature cell compartments.

Heatmap showing the top 30 most differentially expressed genes in the GATA1+ cell compartments determined by Wilcoxon Rank-Sum test. Rows represent genes and columns represent single cells. The colour coding of cell compartments and naming follows as Figure 4.6.

To identify which cell type makeup cluster 6, I performed gene set enrichment analysis. First by ranking all genes belonging to cluster 6 according to its AUC score. Where a gene with an AUC score of 1 can perfectly classify the cell cluster relative to the other clusters, and an AUC score of 0.5 has no predictive value. Then I obtained representative genes from each cluster with an AUC value greater than 0.8. I applied gene set enrichment analysis (GSEA) to test if our a priori defined cluster specific gene sets were significantly enriched in the top genes that define cluster 6 (Figure 4.11). This analysis demonstrated that cluster 6 is composed of cells expressing genes corresponding to the erythroid, megakaryocytic, and eosinophil/basophil/mast cell lineages. A closer inspection of the leading-edge gene subsets for each cluster, the core of genes that accounts for the gene set enrichment signal, showed that *GATA2, FCER1A* and *CPA3 (mast cell carboxypeptidase A)* expression is shared between cluster 6 and EMMPPs. Similarly, *GATA2* and *LMO4 and IKZF2* expression was shared between cluster 6 and eo/ba/mast cells, *ITGA2B* was shared between megakaryocytes progenitor and cluster 6, while *GATA1* and *TFR2* are shared between erythroid progenitors and cluster 6 among others. I named cluster 6 cycling-EMMPPs.

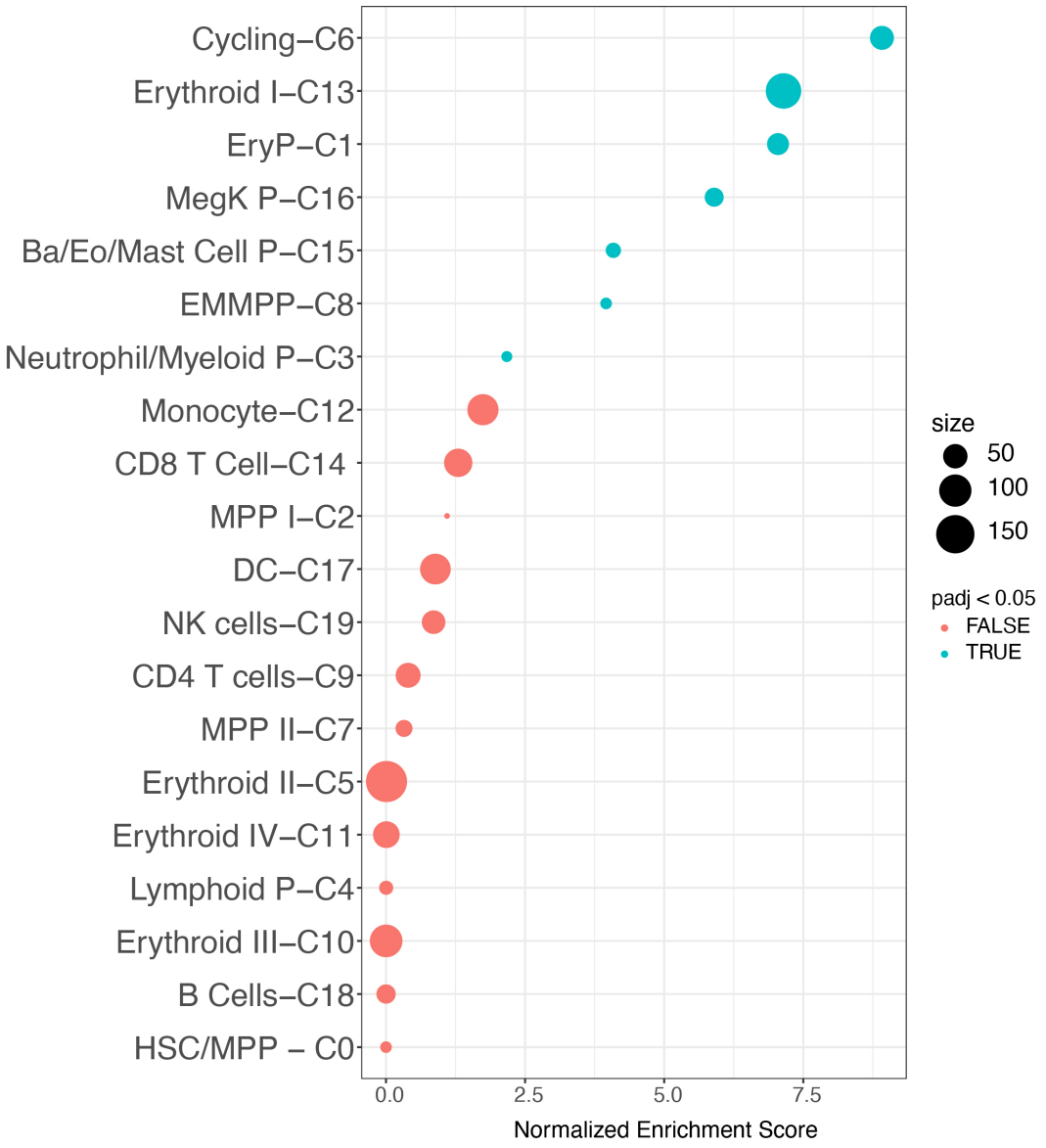


Figure 4.11 GSEA analysis to determine the identity of cluster 6.

Gene set enrichment analysis (GSEA) was used to determine if an a priori defined, cluster specific gene set (y-axis) is statistically enriched (x-axis) in top of a list of representative genes of cluster 6. The size of the dot corresponds to the number of genes with an AUC score greater than 0.8. Red means normalized enrichment scores are not statistically significant and green means they are statistically significant.

Finally, Cluster 1, 13, 5, 10 and 11 represents different stages of erythroid cell maturation and were named Erythroid P, Erythroid I, Erythroid II, Erythroid III, Erythroid IV, respectively. Cluster 1 is the most erythroid immature cell type marked by *GATA2, KIT* as well as *CD36, CD105 and CD71(TFRC)* expression but no Glycophorin A

expression. Glycophorin A, as well as other genes encoding erythroid glycoproteins (*GYPE*, *GYPB*) start being express in cluster 13, and continue throughout clusters 5, 10 and 11. Haemoglobin genes (*HBB, HBA1/A2, HBG1/HBG2*) are also start being expressed at cluster 13 and continue throughout maturation. Cluster 13 represent an intermediate stage of cycling cells marked by down regulation of *GATA2* and increase expression of *MKI67*, *CDC20, CCNB1, CCNB2*, as well as *EPOR* (EPO receptor). Cluster 11 is the most mature erythroid cell type identified in this analysis and is marked by down regulation of GATA1, KLF1 and SOX6 among other genes (Figure 4.12).

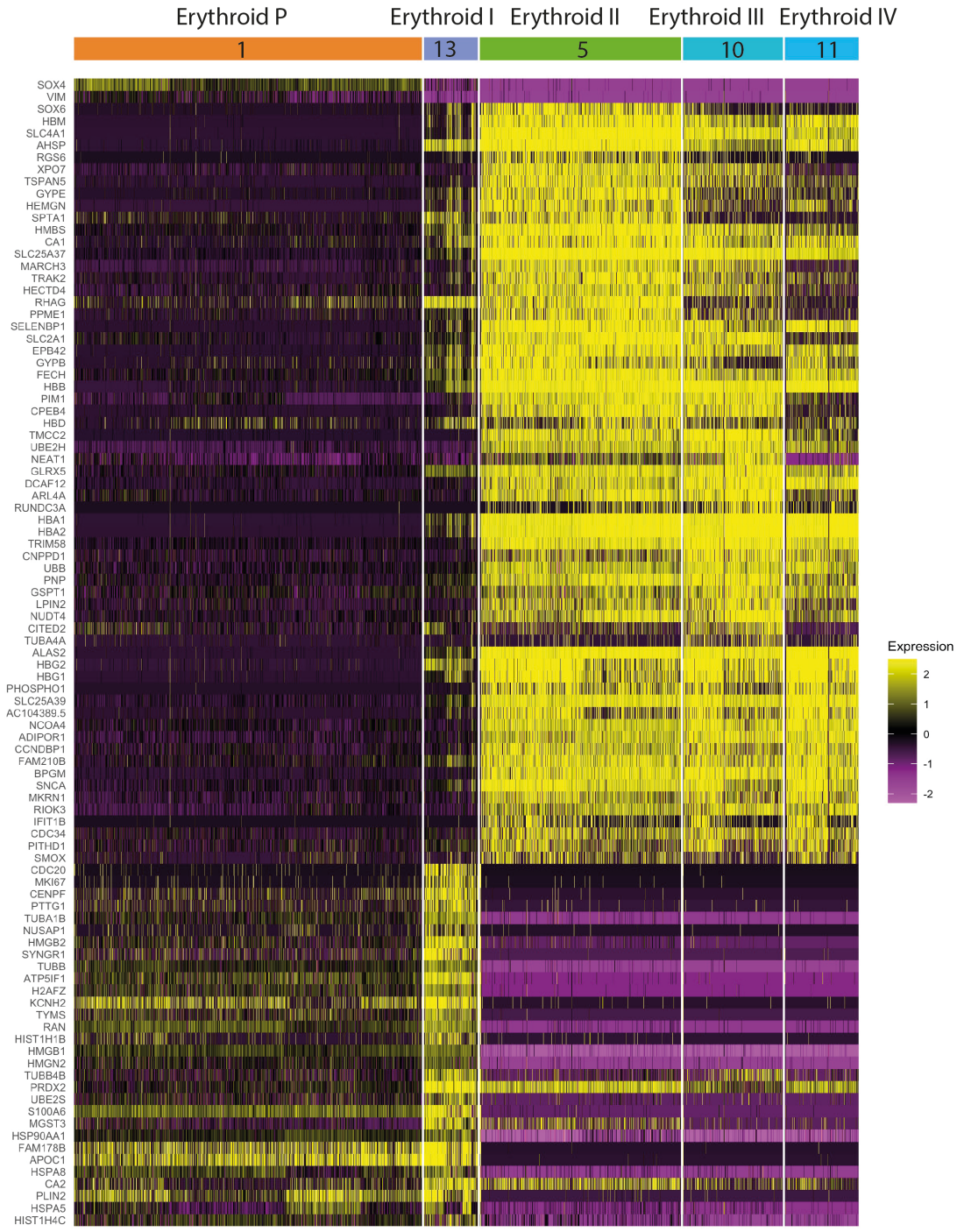


Figure 4.12 Heatmap showing the top 30 genes with the highest log2 fold change during erythropoiesis.

Heatmap showing the top 30 most differentially expressed during erythropoiesis determined by Wilcoxon Rank-Sum test. Rows represent genes and columns represent single cells. The colour coding of cell compartments and naming follows as Figure 4.6

Haematopoietic stem cell differentiation is phenotypically marked by gained of CD38 expression, whereby CD38- HSC/MPP give rise to CD38+ CMPs that further differentiate into erythroid/megakaryocyte bipotent progenitor cells among others. GATA1 expression can be observed in mature and immature erythrocytes, megakaryocytes, mast cells, eosinophils, and basophils at the RNA and protein level. To define the earliest compartment where GATA1 expression can be detected at the protein level in the T21 haematopoietic stem and progenitor cells, I applied the intracellular flow detection of GATA1 C' and N' terminus in a multiparameter flow cytometry that included detection of CD34, CD38, CD45RA, CD123 and CD10 and GATA2 (Supplementary Figure 3, appendix I), in 20 trisomy 21 cord blood samples. This analysis demonstrated that although GATA1 protein expression could be detected as early as in the most primitive compartment (primitive: lineage$^-$ CD34$^+$ CD38$^-$) (mean = 2.2 +/- 4.53 s.d), the Common Myeloid Progenitor (CMP: lineage$^-$ CD34$^+$ CD38$^+$ CD10$^-$ CD123+ CD45RA$^-$) was marked with robust GATA1 expression and was surprisingly split into GATA1- (mean = 75.8% +/- 12.2% s.d) and GATA1+ (mean = 21.9% +/- 11.78% s.d.) cells. Similarly, the Megakaryocyte Erythroid Progenitor (MEP: lineage$^-$ CD34$^+$ CD38$^+$ CD10$^-$ CD123- CD45RA-) was also split into GATA1- (mean = 38.6.8% +/- 14.5% s.d.) and GATA1+ (mean = 57.6% +/- 14.8 s.d.) cells (Figure 4.13). While the lymphoid/myeloid compartments marked by CD45RA expression such as the GMP, and B/NK progenitor, were GATA1 negative. Consequently, gain of CD38 expression which marks CMPs correlates with GATA1 expression and cellular commitment. In our single cell trisomy 21 landscape, *CD38* gene expression starts from cluster 2 or MPP1, and cluster 8 or EMMPP, of which EMMPPs are characterized by robust *GATA1* expression.

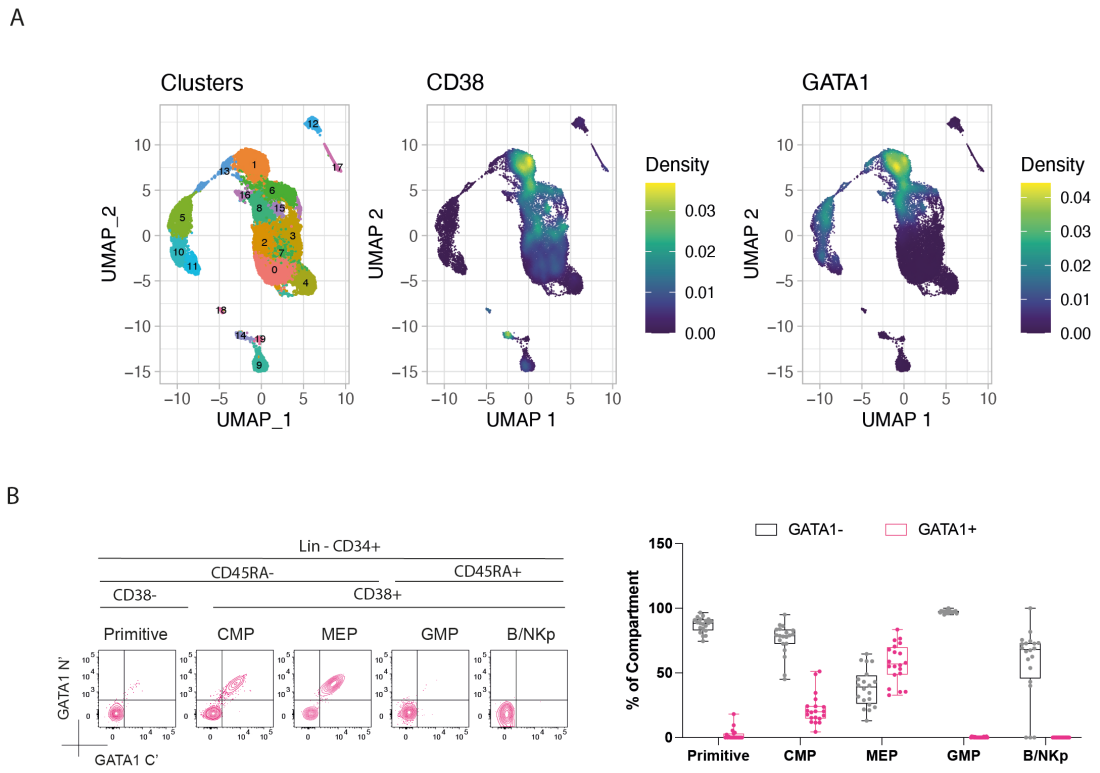


Figure 4.13 Gene and protein expression of GATA1 T21 myelopoiesis.

A) UMAP visualization of 20 T21 haematopoietic cell clusters (left), expression density of *CD38* (middle) and *GATA1* (right). B) Intracellular flow cytometry detection of GATA1 in 20 T21-samples for each haematopoietic compartment: primitive (lineage-, CD34+, CD38-, CD45RA-, CD10-; mean 2.2 +/- 4.53 standard deviation), CMP(lineage-, CD34+, CD38+, CD123+, CD45RA-, CD10-; mean21.9 +/- 11.78 s.d.), MEP(lineage-, CD34+, CD38+,CD123- CD45RA-, CD10-; mean 57.6 +/- 14.8s.d) , GMP (lineage-, CD34+, CD38+,CD123+ CD45RA+, CD10-; mean 0.2 +/- 0.34 s.d), B/NKp (lineage-, CD34+, CD38+, CD45RA+, CD10+; mean 0.0 +/- 0.0 s.d)

From this analysis I was able to identify the main inferred cellular trajectories for each of the three lineages that express *GATA1* and *GATA2*. Megakaryocyte, erythroid and eo/ba/mast cell development most likely occurs as cells progress from the HSC/MPP (cluster 0) to MPP1 (cluster 2) and EMMPPs (cluster 8) (Figure 4.14). EMMPPs, who express high levels of GATA1 most likely differentiate into megakaryocyte progenitors (cluster 16), erythroid progenitors (cluster 1) and eosinophil/basophil/mast cell progenitors (cluster 15) (Figure 4.14).

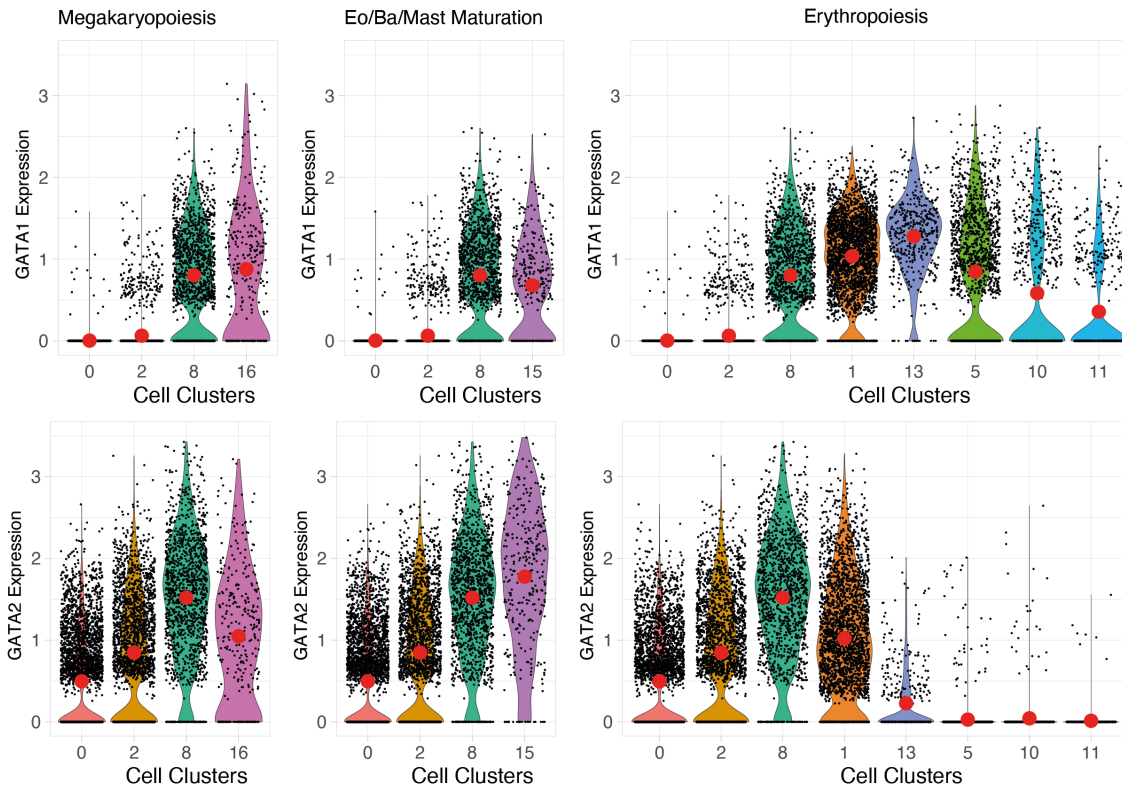


Figure 4.14 *GATA1* and *GATA2* expression in megakaryopoiesis, eo/ba/ mast cell maturation, and erythropoiesis.

Violin plots of normalized, log transformed *GATA1*(top) and *GATA2(bottom)* expression values per single cell (Y-axis) for each cell cluster (X-axis) in megakaryopoiesis (left), eo/ba/mast cell maturation (middle), and erythropoiesis (right). Red dots denote mean expression.

4.3 Query Mapping of Transient Abnormal Myelopoiesis onto Trisomy 21 Haematopoiesis

Following the construction of a single cell transcriptional reference landscape of trisomy 21 haematopoiesis, I can now map single cells from a cord blood of a TAM sample to study cell heterogeneity in TAM as well as aberrant cell maturation due to *GATA1s* mutations. Using intracellular flow cytometry, I have previously identified that cells with exclusive GATA1s expression reside or co-located in the $CD45^{low}$ KIT^{+} compartment (Figure 3.6). Consequently, single cell RNA sequencing was performed on lineage negative, $CD45^{low}$ cells from a TAM sample (Figure 4.15 C). Lineage markers included lymphoid (CD2, CD3, CD4, CD8a, CD19, CD20), and myeloid (CD11b, CD14, CD15 CD19 CD20) cell surface proteins. After removal of low-quality cells by miQC and

doublets with DoubletFinder as described before, I obtained 1750 single cells (Figure 4.15 A). This TAM sample was identified as male based on *DDX3Y, KDM5D,* and *ZFY* gene expression and lack of *XIST* expression (Figure 4.15 B). Consequently, in this sample there is only one GATA1 allele. Mutational profile of the TAM sample identified a G to T nucleotide change within exon 2 with a VAF of 12.6 % of cord blood MNCs. This G to T nucleotide change is predicted to transform a glutamate residue to a stop codon which would lead abrogated production of GATA1 full-length (Figure 4.15 C). iFC detection of GATA1s^{only} cells with in the CD45 low, lineage negative gate identified ~60% of mutant cells and only 3% of GATA1^{WT} cells, expressing both isoforms (Figure 4.15 C).

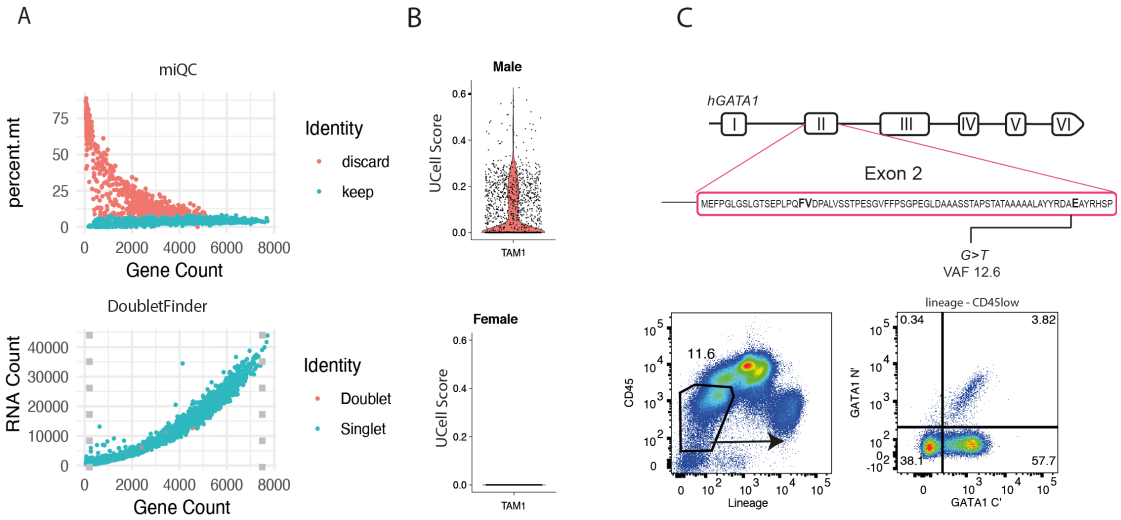


Figure 4.15 Quality control of a TAM sample

A) miQC and Dobletfinder identified low quality cells of high mitochondrial read counts (red dots top) and doublets (red dots-bottom) in one TAM sample. B) UCell scoring of male specific gene expression: *DDX3Y, KDM5D, ZFY* or female specific gene expression: *XIST* identifies the TAM sample as male. C) schematic representation of the mutation in exon 2 of GATA1(top) that leads to exclusive expression of GATA1s. Bottom, exclusive GATA1s expression was determined by flow cytometry dual immunostaining of GATA1 N' (y-axis) and GATA1 C' terminus (x-axis) after gating out doublet, dead and lineage positive cells.

I applied Symphony[128] for query mapping of the TAM single cells onto the previously defined T21 single cell reference map (Figure 4.16 B). Symphony is a novel algorithm that corrects for batch or sample effects and maps query cells to a precise

location in the reference UMAP embedding. Symphony allows for cluster annotation of query TAM cells, by assigning cells to already defined cell clusters in the T21 reference map. While all cells were assigned to already defined cell clusters, most TAM cells were mapped to clusters 8, 6, 16, 15 and 1, which correspond to EMMPPs, EMMPP-cycling, megakaryocyte progenitors, eosinophil/basophil/mast cell progenitors, and erythroid progenitor cells. I can confirm the accuracy of the mapping of TAM cells by verifying that most of these cells express high levels of *GATA1*. The effects of having *GATA1s* mutation in T21 haemopoiesis are studied in the next section.

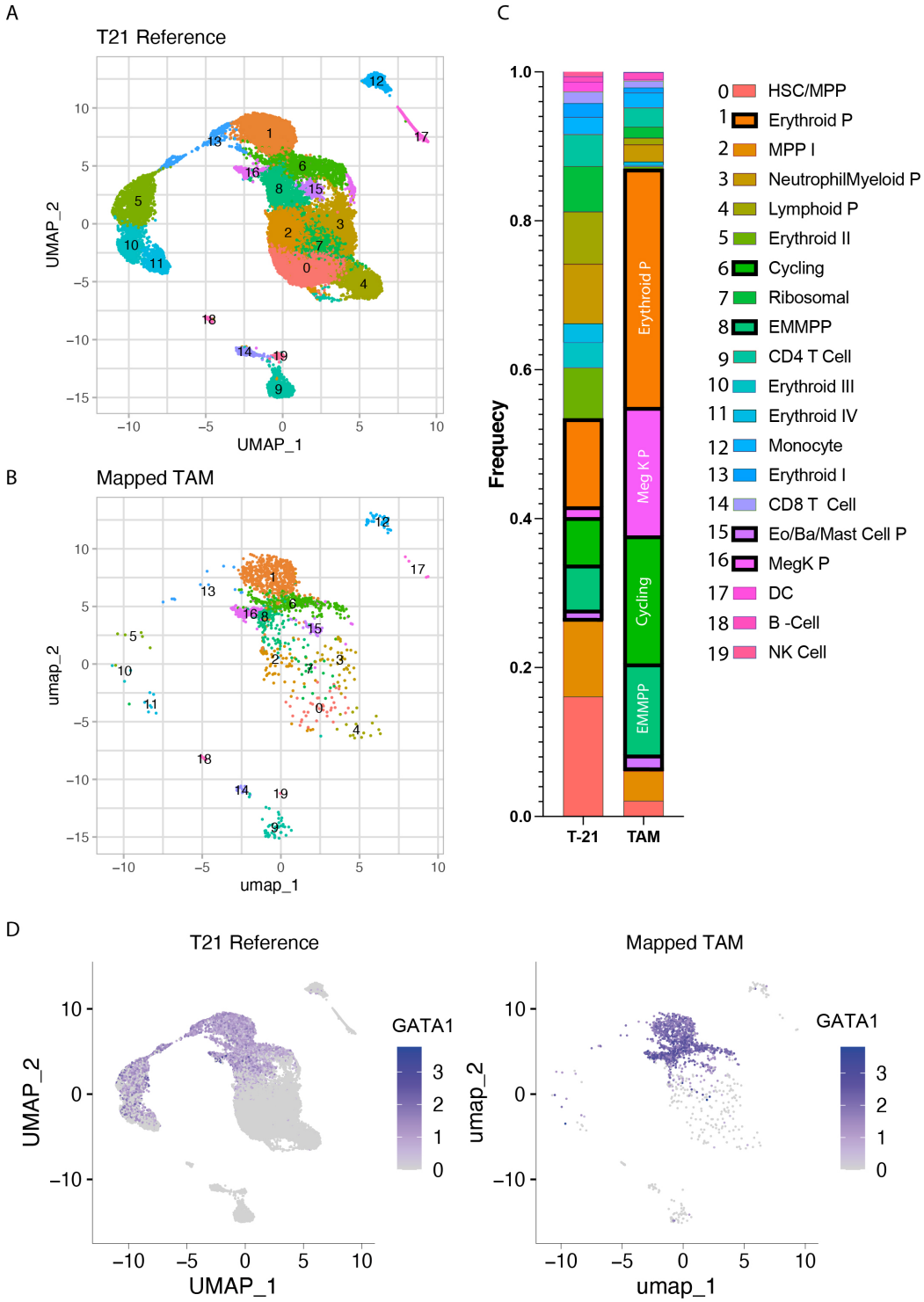


Figure 4.16 Query mapping of TAM cells onto single cell trisomy 21 haematopoiesis

A) Single cell transcriptional landscape of T21 haematopoiesis in UMAP space with 20 cell clusters (0-19). B) Query mapping of TAM ells after quality control. C) Cell frequencies of T21 reference cell clusters as well as of TAM cell clusters. D) Normalized, log transformed *GATA1* expression in T21 reference map (left), and in query TAM cells (right).

4.4 Identification of *GATA1s* Mutations with VarTrix in a Cell Mix Experiment

While single cell RNA sequencing methods have been optimized and the throughput is constantly increasing, detection of genetic variants from single cell RNA counts has been challenging. Detection of genetic variants at the single cell level permits the association of genotype and phenotype. In the context of leukaemia, identification of driver genetic changes at the single cell level can identify transcriptional difference between mutant and wild type cells. Unfortunately, sequencing coverage in single cell assays is limited to either the 5' or 3' end of the gene, especially in short-read based sequencing of the 10X Genomics platform. This makes it difficult to call genetic variants at the single cell level. Here I employed VarTrix to identify known *GATA1s* variants from single cells capture with the 10X Genomics platform. First, I tested VarTrix ability to identify a known mutant cell line in a cell line mix experiment. Then, I used VarTrix to identify *GATA1s* mutations in primary TAM single cells. VarTrix can extract known single cell variant information, but it does not perform variant calling. After cDNA mapping to the human genome, VarTrix only scans mapped DNA reads for already known genetic changes.

To assess the ability of VarTrix in identifying *GATA1s* variants I performed single cell RNA sequencing on a cell mix experiment, where K562, THP1s and CMKs, were mixed at equal ratios. I have previously identified that K562 cells express both isoforms of GATA1 and it is wild type for the *GATA1* locus. While CMK cells only express the short isoform due to a short indel in exon 2 of *GATA1*. THP1s, although wild type for the *GATA1* locus, they do not express *GATA1*. After removal of low-quality cells with miQC and doublets with DoubletFinder, I obtained 6567 single cells (Figure 4.17 C). Sex assignment to each cell line was performed as previously and it determined that CMKs and THP1s are male, while K562s are female (Figure 4.17 D). Moreover, K562s, THP1s and CMKs can be identified by differential levels of GATA1 C', GATA2 and PU.1 expression at the protein and RNA level respectively (Figure 4.17 A, B). Based on the

expression of *SPI1*, *GATA1*, and *GATA2*, I assigned each single cell to its corresponding cell type (Figure 4.17 D).

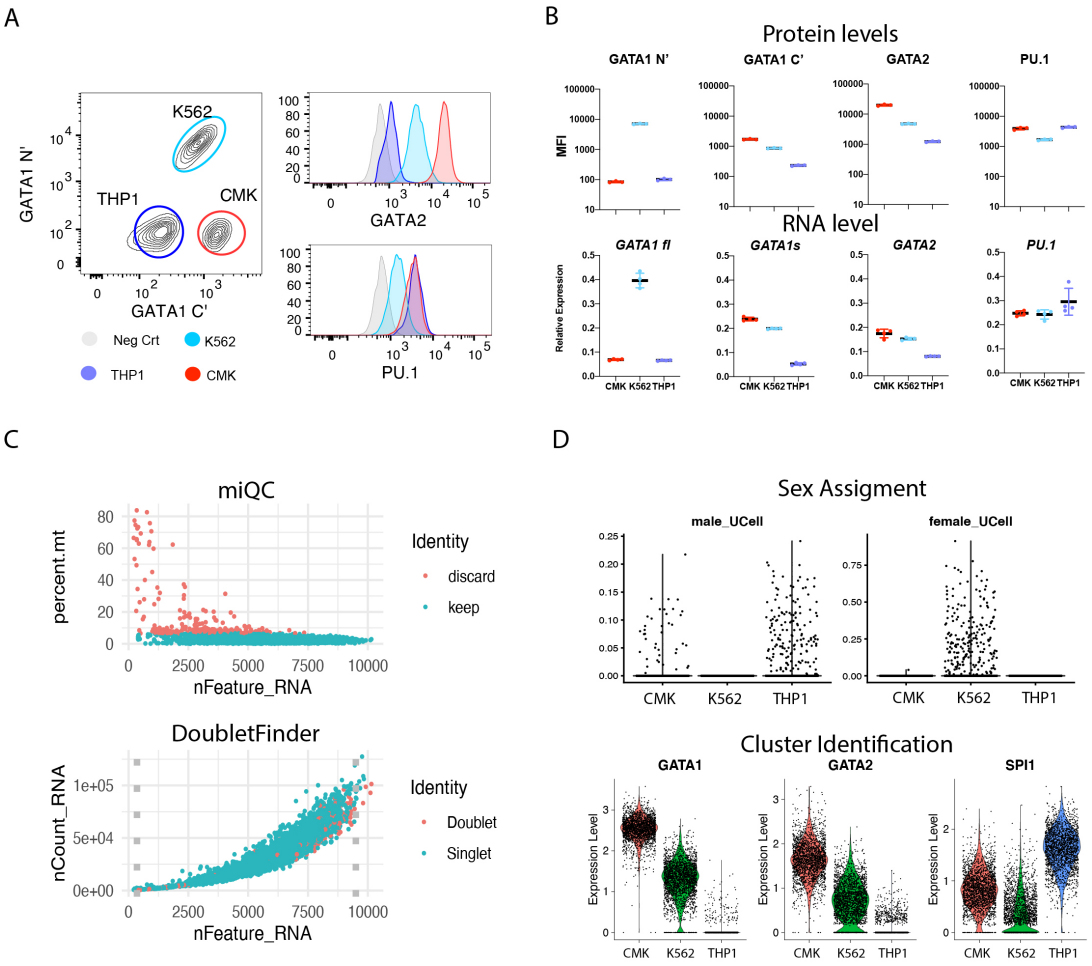


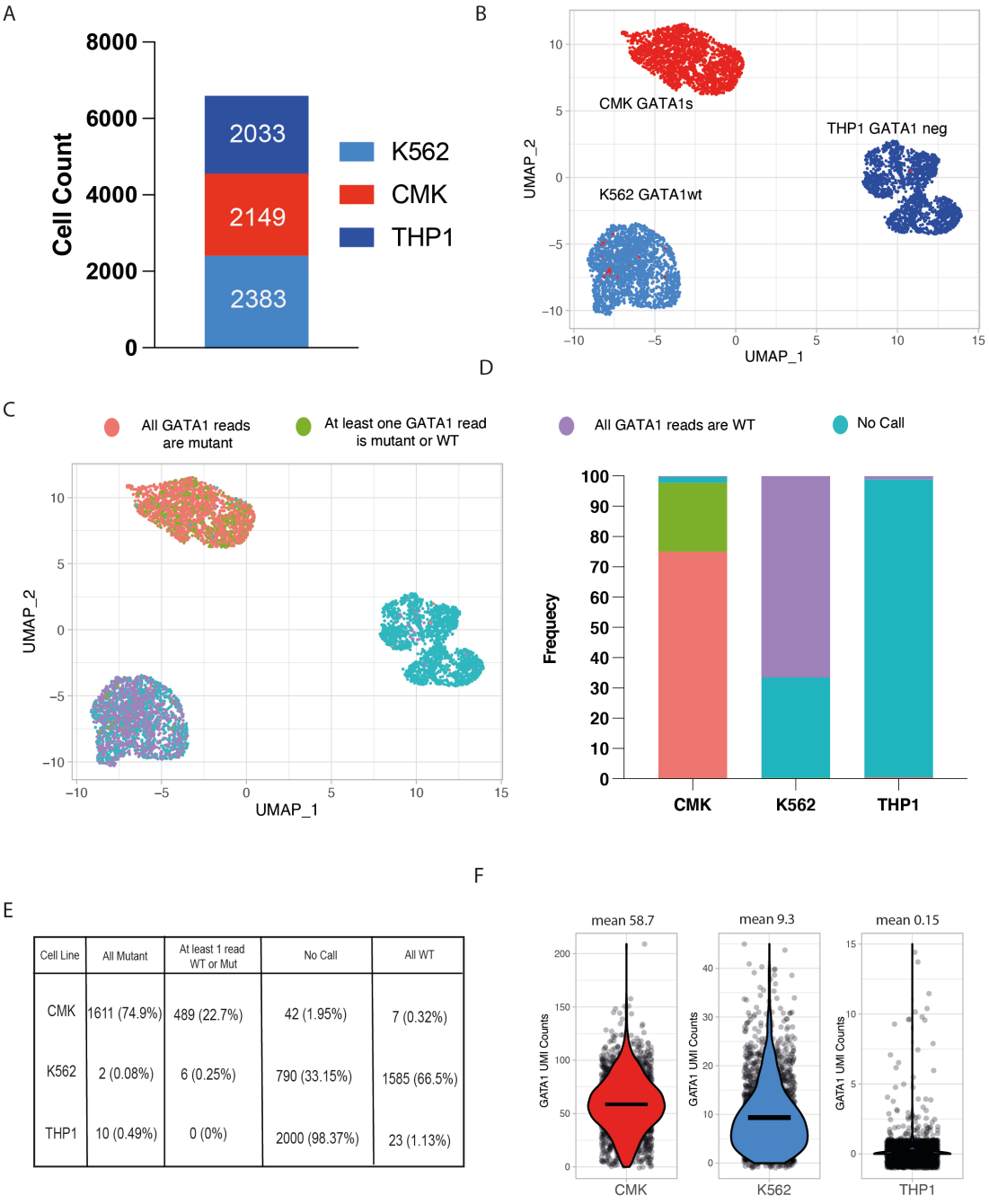
Figure 4.17 A cell mix experiment to test VarTrix

A) intracellular flow cytometry of GATA1 isoforms, GATA2, and PU.1 for K562, CMK, and THP1 mixed at equal ratios. B) Expression of GATA1 GATA1 isoforms, GATA2, and PU.1 at the protein level (top) based on iFC or RNA level (bottom) based on qPCR. MFI, mean fluorescent intensity. Relative Expression, inversion of the delta-delta CT method. C) Quality control of sequenced single cell from the mixed experiment with miQC (top) and DoubletFinder (bottom), red dots denote low quality or doublets respectively. D) Normalized, log transformed *GATA1*, *GATA2* and *SPI1* expression permits single cell phenotype assignment.

6,567 single cells with similar frequencies across the three cell types (36.8 % K562, 32.1% CMK, 32.1% THP1) were visualized in UMAP space (Figure 4.18 A, B). 3 distinct cell clusters were identified. I used VarTrix to identify mutant CMK cells with a known AG deletion in exon 2 of *GATA1* (Figure 3.2). VarTrix has four outputs per single cell. 1) *no*

call, if no reads contain the reference (wild type-WT) or alternative allele (mutant). 2) *all reads are WT*, if all reads belong to the reference allele. 3) *all reads are mutant*, if all reads belong to the alternative allele; 4) *at least one read is mutant or WT*; if of all the reads that map to this locus at least one is mutant or WT.

As expected, in most cells (98.67 % *no call*) no WT or mutant alleles were identified in THP1s cells as they do not express *GATA1*. While 66.5% of K562 cells were identified as having all WT reads. 74.9% of CMK cells were identified as having all mutant reads (Figure 4.18 C, E). In 33.1 % of K652 cells (790 cells out of 2383), and 1.9 % of CMK cells (42 cells out of 2149), VarTrix did not identify the reference or alternative allele respectively and cells were assigned as *no call*. VarTrix in order to identify the reference or alternative alleles, it not only requires the locus to be expressed but also how well a read aligns to the locus of interest. As we can see from Figure 4.18 F, CMKs cells have a GATA1 UMI mean count of 58, K562 cells of 9 and THP1 cells of 0 UMI per cell. The lower the expression of *GATA1* per single cell, the most likely it is for VarTrix to assign each cell as *no call*.



Cell Line	All Mutant	At least 1 read WT or Mut	No Call	All WT
CMK	1611 (74.9%)	489 (22.7%)	42 (1.95%)	7 (0.32%)
K562	2 (0.08%)	6 (0.25%)	790 (33.15%)	1585 (66.5%)
THP1	10 (0.49%)	0 (0%)	2000 (98.37%)	23 (1.13%)

Figure 4.18 VarTrix can identify *GATA1s* mutations in single CMKs cells

A) Cell frequency for each cell line: CMK (red), K562(light blue), THP1(dark blue). Cluster analysis and visualization of the three cell lines in UMAP space. Genotyping assignment by VarTrix and visualization in UMAP space. E) cell frequencies and absolute number of single cells for each genotype. F) GATA1 UMI counts (y-axis) for each cell line, the respective mean GATA1 UMI count is written on top.

Interestingly, 22. 7% of CMKs cells (489 out of 2149) were called as having at least one read mutant or wild type. This genotype can be explained if the cell line has a heterozygous mutation. Often gene variants are either homozygous or heterozygous. In a single cell, a heterozygous genotype can become homozygous after losing the wild type allele. It can be speculated that CMKs cells are heterozygous (22. 7% cells with mutant and wild type reads) and have become homozygous after losing the wild type allele (74.9% cells with only mutant reads). This implies that within our CMK cell line, some cells have a mutant and a wild type allele of GATA1 and that the full-length GATA1 isoform is express from the wild type allele. Intracellular detection of both the C' and N' terminus of GATA1 fails to detect any GATA1 N' terminus signal in CMKs cells (Figure 3.2). Moreover, the CMKs cell line was derived from a male paediatric patient (X, Y). Knowing that *GATA1* allele lies on the X chromosome, it means that CMKs cells only have one *GATA1* allele and are hemizygous for the mutations.

A more careful analysis of the genotyping assignment of CMKs cells shows that of 489 cells that were assigned a genotype of having *at least one read mutant or WT*, have at most only one read belonging to the WT allele (Ref-allele) (Figure 4.19 A,B). I refer to this genotype as ambiguous as it cannot be explained based on GATA1 wild type expression by intracellular flow cytometry. Moreover, principal component analysis between mutant CMK cells and CMK cells with the ambiguous genotype failed to identified gene expression variance that can separate both genotypes (Figure 4.19 C). Differential gene expression analysis between both genotypes in CMK cells only identified two differentially express genes, *MTRNR2L12* and *HBG1* with 0.6 and 0.3 log2 fold change but they are not statistically significant (Figure 4.19 D). It is most likely that the ambiguous genotype of having at least one wild type allele is a technical artifact generated from the sequence quality of read and the mapping quality of the read. It is most likely that CMK cells with the ambiguous phenotype are all mutant cells.

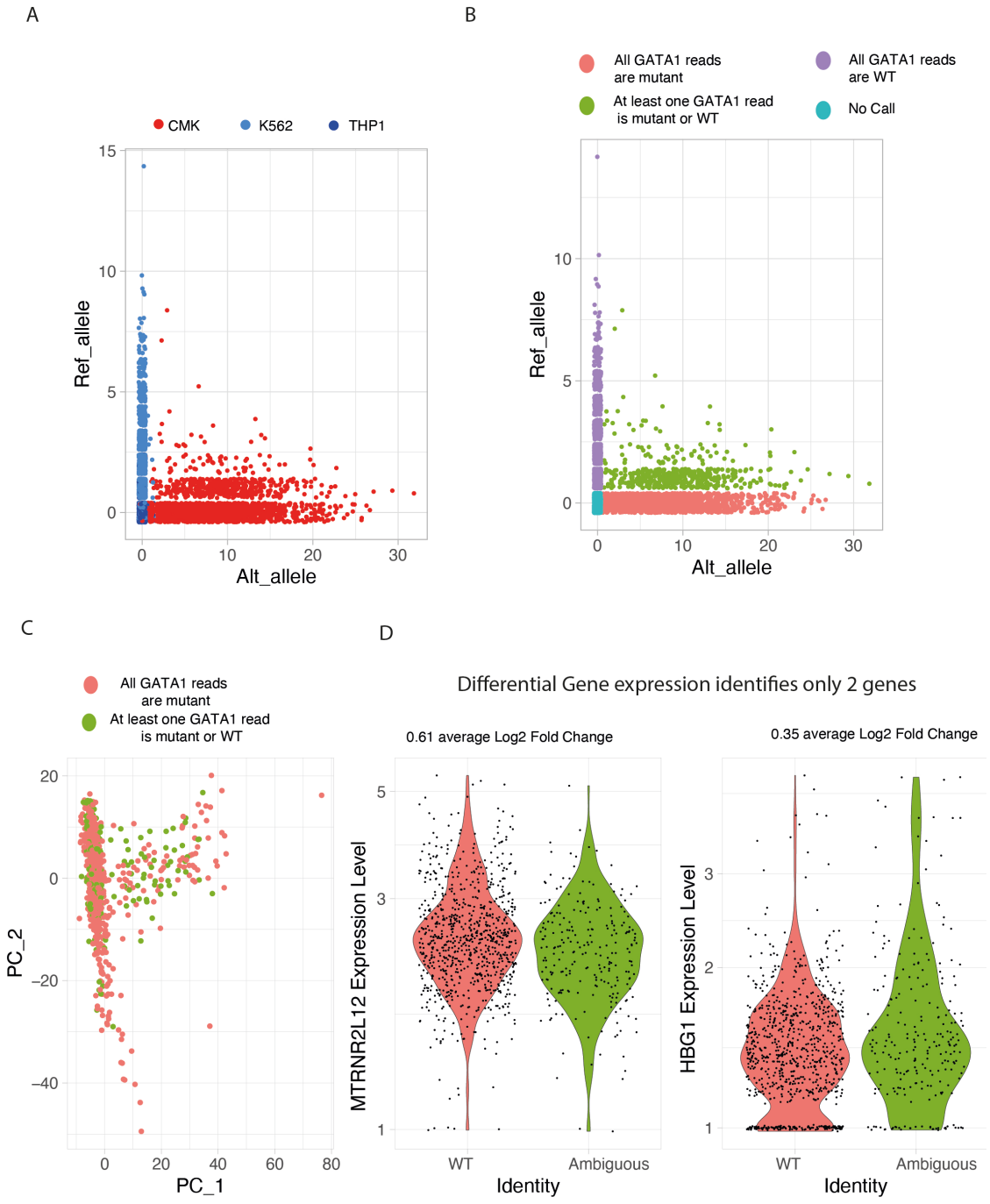


Figure 4.19 VarTrix assigns ambiguous genotypes to a fraction of mutant CMKs

A) Scatter plot of single cells with wild type UMI counts (y-axix : Ref_allele) and mutant UMI counts (x-axis: Alt_allele) colour coded by cell type as in figure 4.18A. B) Scatter plot of single cells with wild type UMI counts (y-axix : Ref_allele) and mutant UMI counts (x-axis: Alt_allele) colour coded by genotype as in figure 4.18C. C) Principal component analysis (PCA) fails to between mutant and the ambiguous genotype. D) Differential gene expression identified only two differentially expressed genes, but they are not statistically significant (adjusted p-value = 1)

4.5 Variant Calling of *GATA1s* mutations with VarTrix in TAM cells

Having determined that VarTrix can identify mutant *GATA1s* reads in CMK cells and that this is inherently depended on the number of *GATA1* UMI counts a given cell expresses, I applied VarTrix to identify *GATA1s* mutant reads in single TAM cells. VarTrix identified 1,138 out of 1,750 cells as having *all mutant reads,* representing 65% of all cells of the sample (Figure 4.20 C, D). Importantly, this mutant cell frequency is highly correlated to the percentage of mutant cells (57.7%) identified by intracellular flow of the same sample (Figure 4.15 C). VarTrix also identified 85 out of 1,750 cell or 4.8% of the cells as wild type. Similarly, iFC detection identified 3.82 % of cells as GATA1 wild type in the same TAM sample. 98.5 % of all mutant TAM cells mapped to either erythroid progenitor cells (456 cells), megakaryocyte progenitor cells (248 cells), EMMPPs (150 cells), EMMPPs-cycling (244 cells), and eosinophil/basophil/mast cell progenitors (23 cells) (Figure 4.20 D) (Table 10). This analysis demonstrates that *GATA1s* mutant TAM ells are heterogeneous mix of different GATA1 expressing cell types. It also demonstrates haematopoietic stem and progenitor cells with *GATA1s* mutations can give rise to EMMPPs, eosinophil/basophil/mast cell progenitors, erythroid progenitor, and megakaryocyte progenitor cells. Functional consequences of *GATA1s* mutations in TAM are explore in the next chapter.

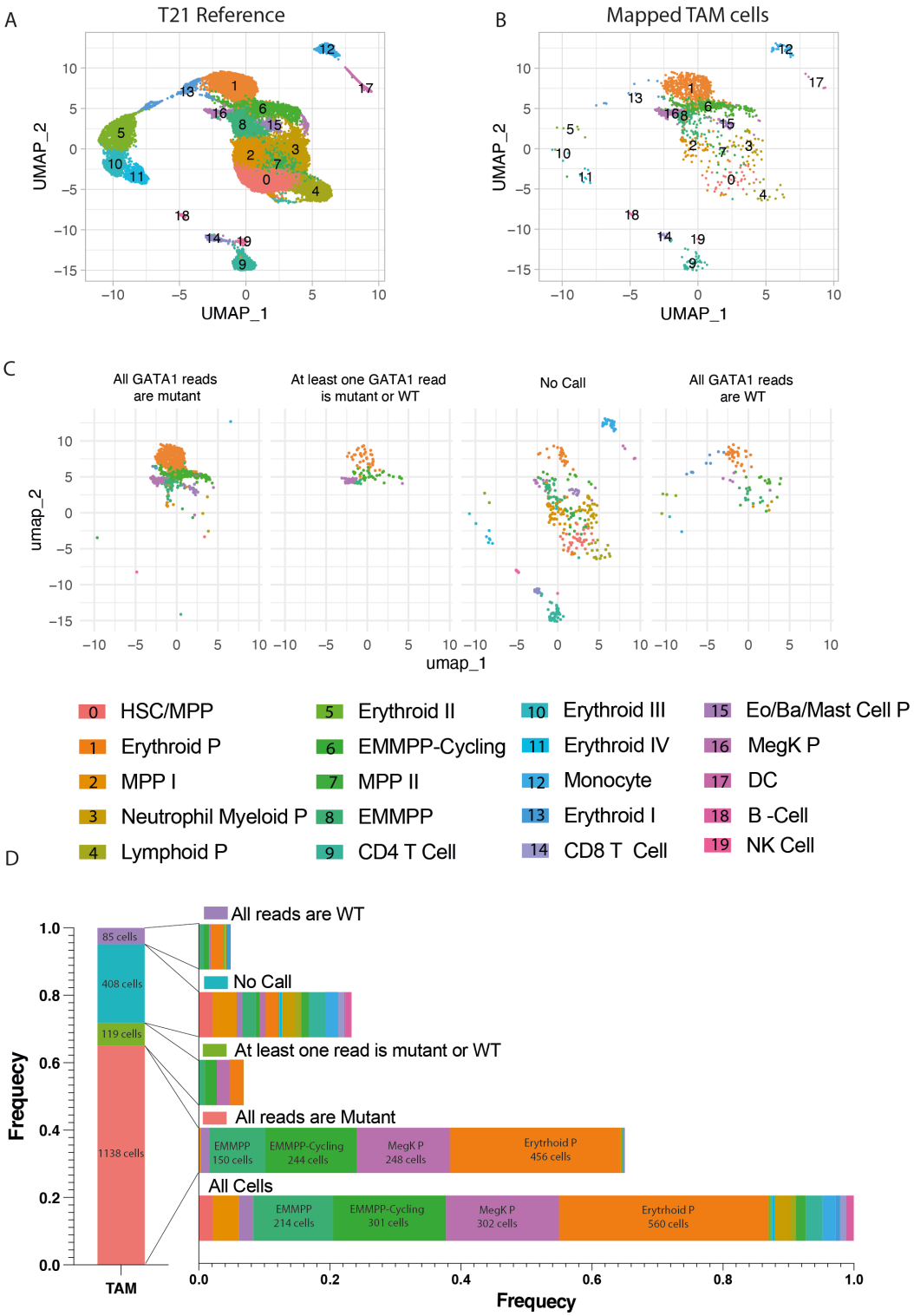


Figure 4.20 VarTrix can identify GATA1s mutation in TAM and find cellular heterogeneity of TAM

A) UMAP visualization of 19 wild type T21 haematopoietic cells clusters. B) UMAP visualization of TAM ells. C) UMAP visualization of TAM cells according to their genotype identify by VarTrix. D) Frequency of TAM cells for each genotype (left), as well as frequency of each cluster per genotype identified and coloured coded as in A.

The genotype of 408 TAM cells (23.3%) was not called as they represented cells with no *GATA1* expression such as HSC/MPP (36 cells), MPP1 (65 cells), MPP II (21 cells), neutrophil-myeloid progenitors (34 cells), lymphoid progenitors (16 cells), as well as other mature lymphoid and myeloid cells (114 cells) (Figure 4.20 C, D, See Table 10 for cell frequencies). No genotype was called in 4.9 % of megakaryocyte progenitor, which express the highest level of *GATA1* counts (Figure 4.21 A). While no genotype was called in 36.8% of eosinophil/basophil/mast cell progenitors, which express the lowest levels of *GATA1* counts (Figure 4.21 A). Thus, the ability of VarTrix to assign a genotype to single cells is inherently dependent on the number of GATA1 reads a given cell express (Figure 4.21 A). As in the cell line mix experiment, 119 out of 1750 TAM cells were called with an ambiguous genotype representing 6.8% of the cells. These cells were denoted by having in most cases only one read mapping to the reference allele (Figure 4.21 B). It is most likely that these cells are also a technical artifact as this TAM sample was taken from a male patient and only has one *GATA1* allele. Only TAM cells with *all reads are mutant* genotype were taken for further studies.

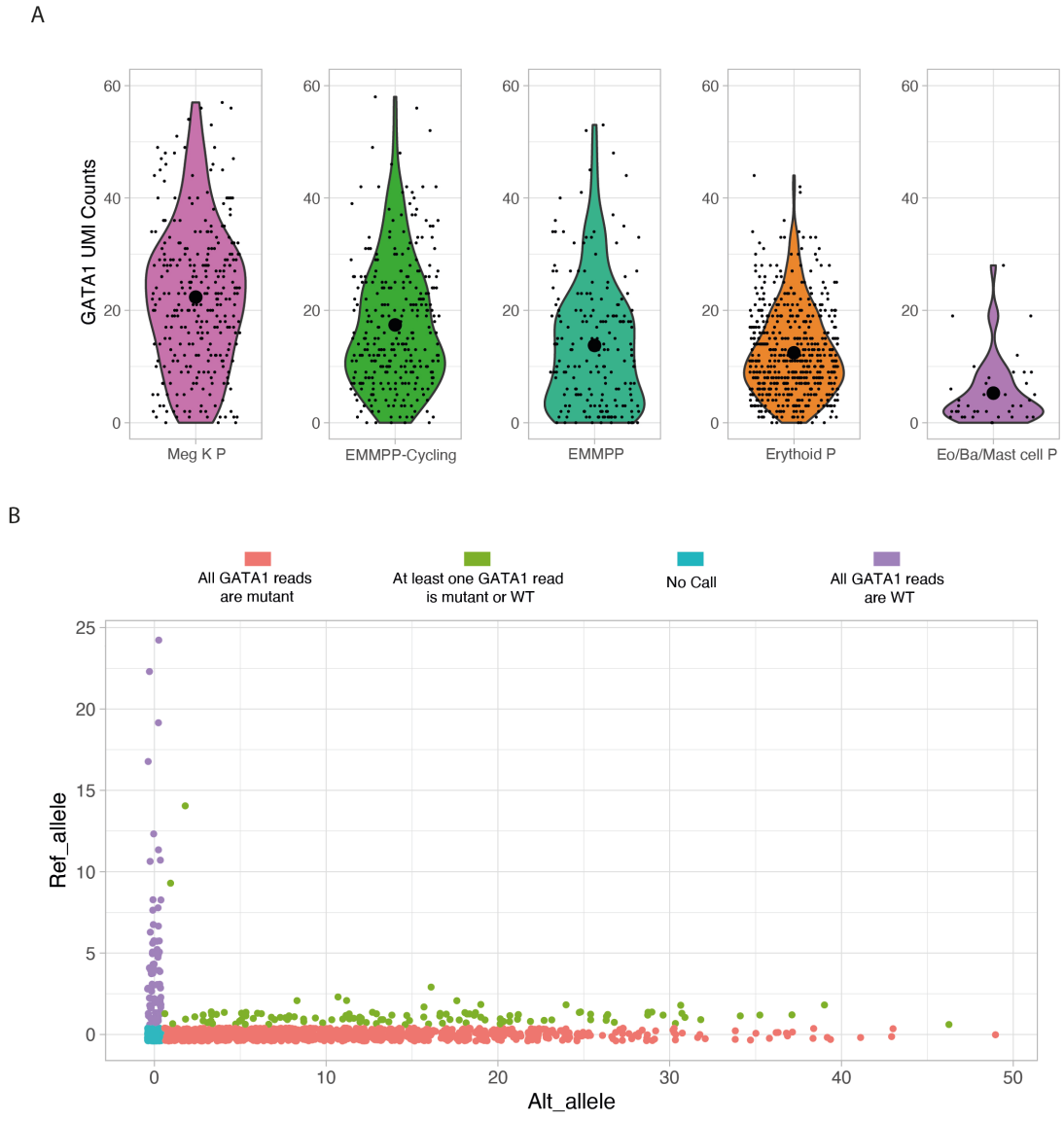


Figure 4.21 Identification of GATA*1s* mutations in single cells depend on the level of *GATA1* expression

A) GATA1 UMI counts (y-axis) per each cell cluster that should express GATA1 (x-axis) with the allelic dropout (the number of cells whit no genotype called that should expresses GATA1) written at the bottom of each cluster. B) Scatter plot of single TAM cells with wild type UMI counts (y-axix : Ref_allele) and mutant UMI counts (x-axis: Alt_allele) colour coded by genotype as in figure 4.18C.

5 Transcriptional and Phenotypic Consequences of GATA1short Expression

To begin to understand the transcriptional consequences of having a *GATA1s* mutation at specific stages of myeloid differentiation, first equal numbers of single cells

from each compartment where *GATA1* is pronouncedly expressed were taken from TAM as well as from T21: EMMPPs (150 cells), EMMPPs-Cycling (244), erythroid progenitor cells (456 cells), megakaryocyte progenitor cells (248 cells), and eosinophil/basophil/mast cell progenitors (23 cells) (Figure 5.1 A). This approach allows comparison of mutant cells to their most transcriptionally similar counterparts.

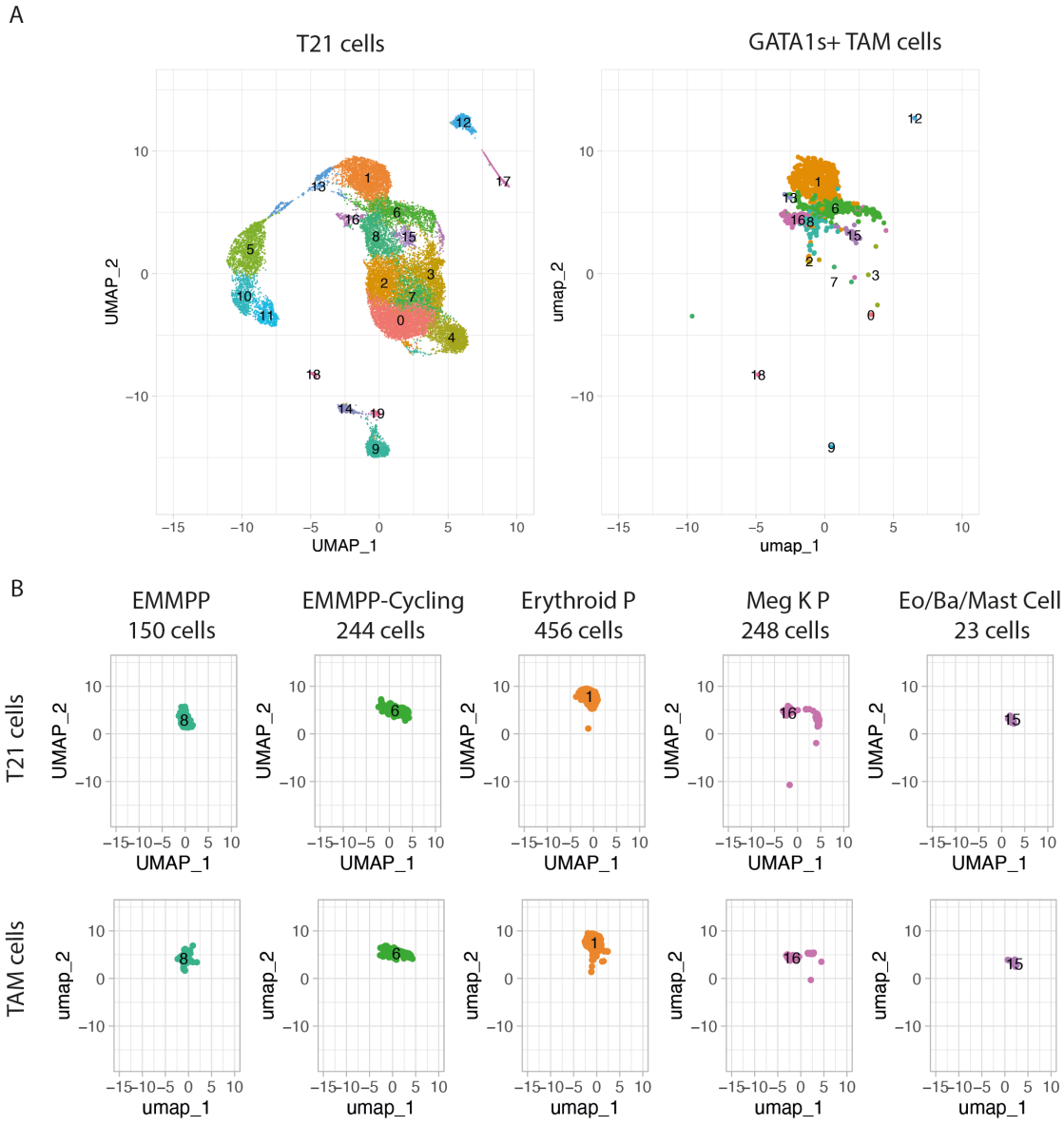


Figure 5.1 Trisomy 21 and GATA1s mutant TAM cells

A) 20 T21 haematopoietic cell clusters (label 0 to 19, left) and GATA1s mutant TAM cells (right) are visualized in UMAP space. B) A similar number of cells was subsetted for each cluster in both TAM and T21 samples. The number of each cell cluster is depicted on top of each compartment.

5.1 Differential Gene Expression Between Trisomy 21 *GATA1* Wild Type Cells and Trisomy 21 *GATA1s* TAM Cells

Differential gene expression analysis between T21 *GATA1* wild type and TAM *GATA1s* mutant cells identified 446 upregulated genes and 118 down regulated genes with a log2 fold change greater than 1 or less than -1, and an adjusted p-value of < 0.05. The cell type with the highest number of upregulated gene were EMMPPs (155 genes), followed by megakaryocyte progenitors (93 genes), erythroid progenitors (83 genes), EMMPPs-cycling (79 genes) and lastly eo/ba/mast cell progenitors (36 genes) (Figure 5.2). Similarly, the cell type with the highest number of down regulated genes was EMMPPs with only 44 genes, followed by megakaryocyte progenitors (40 genes), erythroid progenitors (18 genes), EMMPPs-cycling (12 genes) and lastly eo/ba/mast cell progenitors in TAM had only 4 genes down regulated (Figure 5.2).

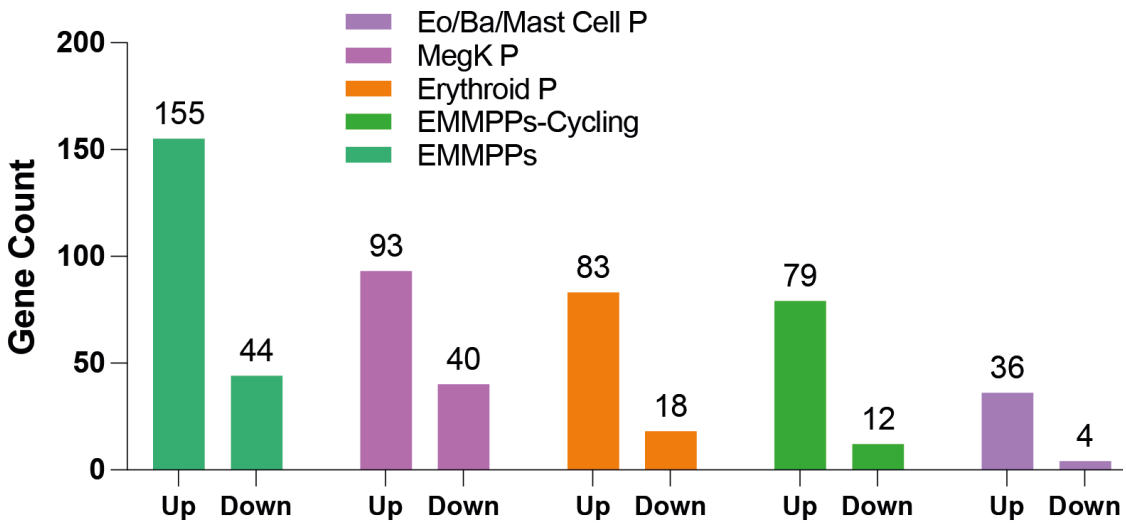


Figure 5.2 Differential gene expression

Differential gene expression testing between GATA1s TAM and T21 cells was performed with MAST. Only genes with a log2 fold change greater than 1 (up) or less than -1 (down) were plotted (y-axis) with an adjusted p-value < 0.05. Colour coded as in figure 4.6.

Many of the upregulated genes in *GATA1s* mutant cells were shared among the different cell compartments. There were 25 shared upregulated genes in most TAM cell compartments except for the eosinophil/basophil/mast cell progenitors, including the p45

subunit of the erythroid/megakaryocytic transcription factor *NFE2*. Also, there were 21 genes that were consistently upregulated in all the cell compartments (Figure 5.3). Some of these genes thought to be specific for the megakaryocytic lineage in normal cells, such as *LAT*, *STXBP5*, *ITGA2B*, were expressed in all mutant *GATA1s* cells. *ITGA2B* was not specific for the megakaryocytic lineage as it was express in erythroid progenitor cells in trisomy 21. Similarly, an erythroid lineage specific *RHAG* gene was also expressed in all mutant TAM cells. There were 4 downregulated genes encoding protein member of the ribosomal 40s subunit (RPS5, RPSA, PRS13, RPS28) and two members of the 60s subunit (*RPL18A* and *RPL30*). This is consistent with recently published evidence suggesting that GATA1 can regulate ribosomal biosynthesis in human erythroid cells[157]. Other genes included *GATA1* itself, *MEIS2*, *FCER1A* and *HPSE2*, among others (Figure 5.3). Enhanced *MEIS2* expression has been observed in induced expression of fusion oncogenes, RUNX1-ETO2[158] and NUP98-KDM5A[159], in human CD34+ cells. Enhanced *MEIS2* expression has also been observed in patients with RUNX1-ETO2 positive leukaemias[160]. It has also been reported that MEIS2 directly binds to the Runt domain of murine RUNX1-ETO2 where it collaborates to induce murine AML[160]. In MN1 positive leukaemias, it has been shown that MEIS2 has essential roles in proliferation, cellular maturation, and self-renewal[161]. *PBX1*, another homeobox member is down regulated due to GATA1s mutation in EMMPPs and megakaryocyte progenitors (Figure 5.5), this is in agreement with previous reports of reduced GATA1s binding at the *PBX1* locus and low *PBX1* expression[162].

The highest overexpressed gene in all TAM cell compartments was *HPSE2*. Unlike HPSE, which can degrade heparan sulphate (HS) chains of cell surface proteoglycan, HPSE2 is a heparinase like protein with no catalytic activity. Mouse models of urofacial syndrome, a congenital disorder with *HPSE2* deletions associated with obstructive urinary tract disease, present with fibrotic tissue in the bladder and increased levels of TGF-B[163]. It has also been speculated that overexpression of HPSE2 due to the lack of

heparinase activity can protect endothelial proteoglycans and at the same time inhibit HPSE activity[164]. Patients with TAM and ML-DS due to *GATA1s* mutation often present liver and bone marrow fibrosis respectively[165]. While tissue fibrosis is mainly due to over production of collagen, accumulation of extracellular matrix components such as the HS-proteoglycan perlecan as well as laminins have been observed[163]. Of note, perlecan can work as bridge between collagen and laminin structures. Moreover, acute megakaryoblastic leukaemia blast cells as well as blast cells in TAM and ML-DS and other myeloproliferative disorders are characterized by increased *TGFB1* expression, the strongest inducer of collagen synthesis as well as an inducer of perlecan. In this study over expression of *TGFB1* (log2FC=1.17) was found, as well as increase *LTBP1* (log2FC=0.99) whose protein product is involved in TGFB processing and activation, in *GATA1s* mutant megakaryocyte progenitors (Figure 5.3 and 5.4). I hypothesize that over expression of HPSE2 as well as TGFB1 and LTBP1 might induce fibrosis in TAM patients by inhibiting the degradation of the extracellular matrix and simultaneously increasing its production. *HPSE2* is not only the highest differentially expressed gene in TAM, but also expressed in CMK cell, an ML-DS derived cell line. Implying that its expression is also a feature of ML-DS (Figure 5.4 C).

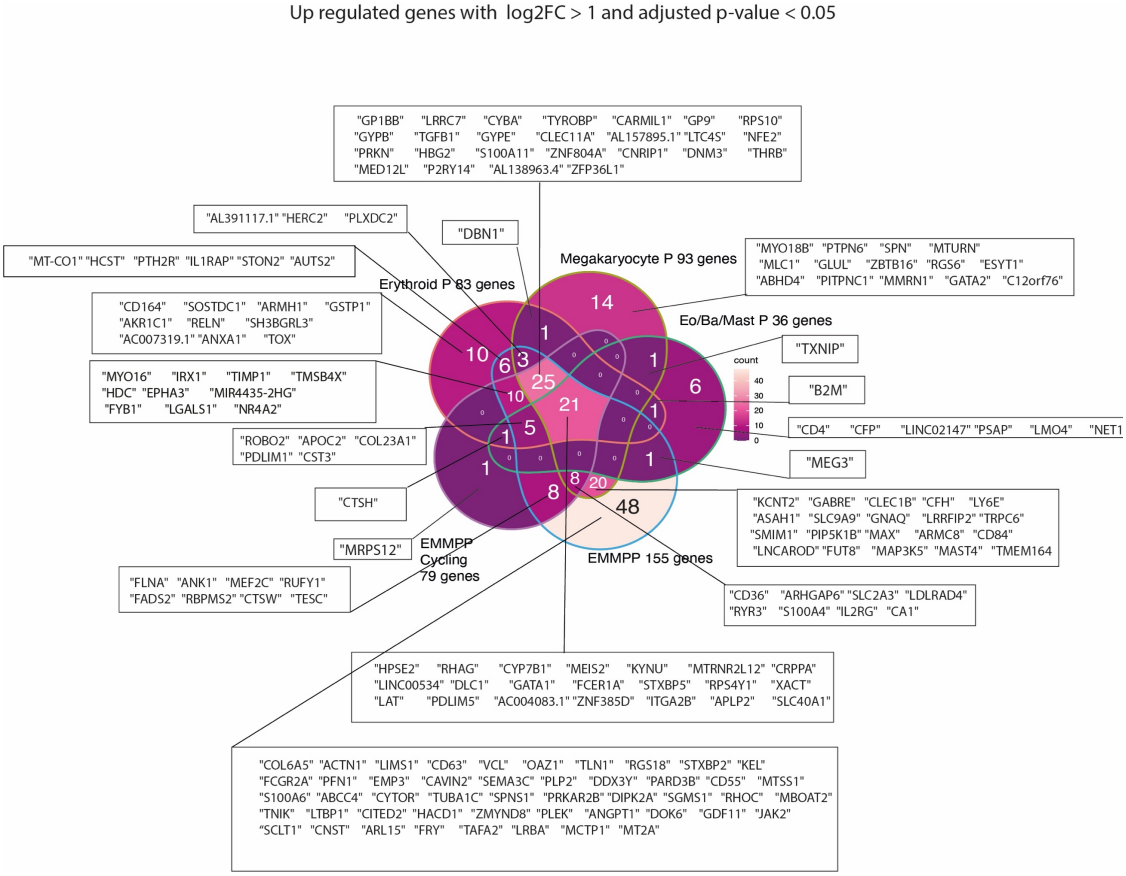


Figure 5.3 Upregulated genes in *GATA1s* mutant TAM cells

Venn diagram of all up-regulated genes with a log2 fold change greater than 1 in TAM GATA1s mutant cells compared to trisomy 21 normal cells. Colour coding correspond to the number of shared genes per compartment.

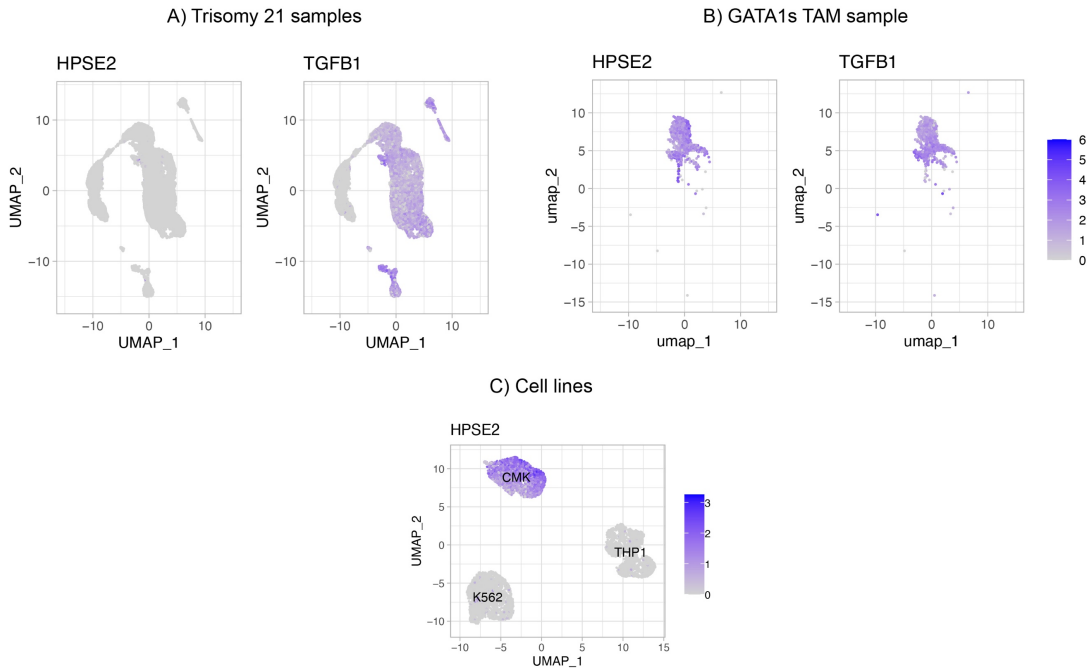


Figure 5.4 *HPSE2* and *TGB1* over expression in Trisomy 21 haematopoiesis and in TAM

HPSE2 and *TGFB1*, two genes that are over expressed in TAM (B) compared to Trisomy 21 (A) are visualize in UMAP space. B) *HPSE2* expression is also uniquely expressed in CMK cells

1.1.1 GATA1s TAM Cells are Marked by a Defective Megakaryocytes

Of the 188 downregulated genes, EMMPPs and the megakaryocyte lineage had the highest (40 genes), follow by the erythroid lineage with 18 genes (Figure 5.5). While the megakaryocytic defect in the Gata1 deficient mouse model is characterized by a decrease in *GP1BA* (CD42b), *GP1BB* (CD42C), the chemokine *PF4* and *MPL*, the TPO receptor[49]. However, in TAM due to *GATA1s* mutations, I observed up regulation of early megakaryocytic markers (*GP1BB* (CD42C), *GP9* (CD42), *ITGA2B* (CD41)). It has been shown that these gene loci are primarily occupied by GATA2, FLI1, LYL1, TAL, ERG, RUNX1 and LMO2, a heptad of transcription factors in megakaryocytic cells[166]. Moreover, TAM cells are also characterized by downregulation of genes involved in pro-platelet formation and release such as *PF4, TPM4, TREML1, MYL9, TUBA4*, and of genes involve in antiviral immunity or inflammation such as *IFITM3* [167], and *CKB*[168], and

PPBP[169] (Figure 5.3 and Figure 5.5). Of note *MYL9* as well as *TREML1* are direct targets of RUNX1[170,171], a chromosome 21 encoded transcription factor and *PF4* and *PPBP* are loci are occupied by the heptad of transcription factors. This differential gene expression analysis points to a maturation and cell dysfunction defect of *GATA1s* megakaryocytic cells as well. This agrees with published data from our group, which showed that due to GATA1s mutations, immature megakaryocytes are delayed in maturation and cells excessively expand[172].

5.1.1 GATA1s TAM Cells are Marked by a Defective Erythropoiesis

A closer inspection of differential gene expression during erythropoiesis shows that *APOC1*, *FAM178B* and *HBB*, are three genes whose expression is induced during the early steps of erythropoiesis (cluster 1 and cluster 13). In the presence of *GATA1s* mutations, the expression of these genes is not seen (Figure 5.6 Erythroid maturation arrest – downregulated gene expression). In terms of cellular erythroid differentiation, it seems that “down regulated” genes are really a depiction of failed transcription rather than suppression of already transcribed genes. Similarly, *CD164, GSTP1, PDLIM1* are three over expressed genes in *GATA1s* cells. In T21, the expression of these three genes is supressed during erythropoiesis (cluster 1 to cluster 13, cluster 5, 10 and11) (Figure 5.7). In the presence of *GATA1s* mutation, it appears that T21 erythroid cells do not complete terminal maturation, instead they are arrested at a cluster 1 like stage, an erythroid progenitor stage. This is consistent with defective erythropoiesis observed in haematopoietic differentiation assays of GATA1s iPCs as well as murine models of GATA1s.

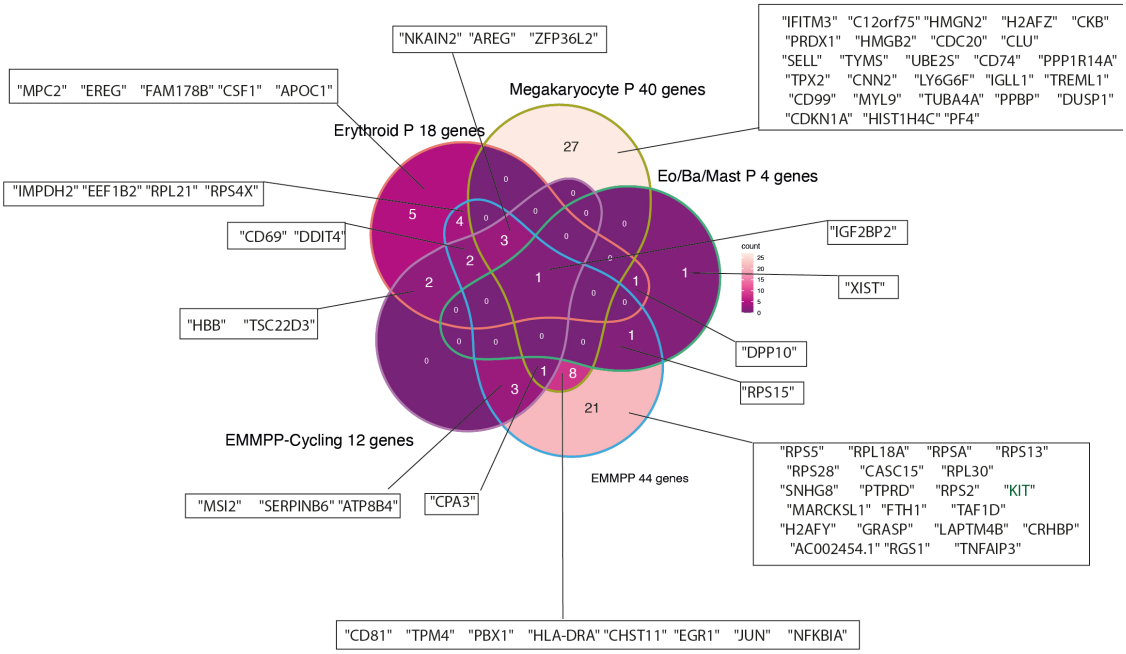


Figure 5.5 Downregulated genes in TAM *GATA1s* mutant cells

Venn diagram of all down-regulated genes with a log2 fold change less than 1 in TAM GATA1s mutant cells compared to trisomy 21 normal cells. Colour coding corresponds to the number of shared genes per compartment.

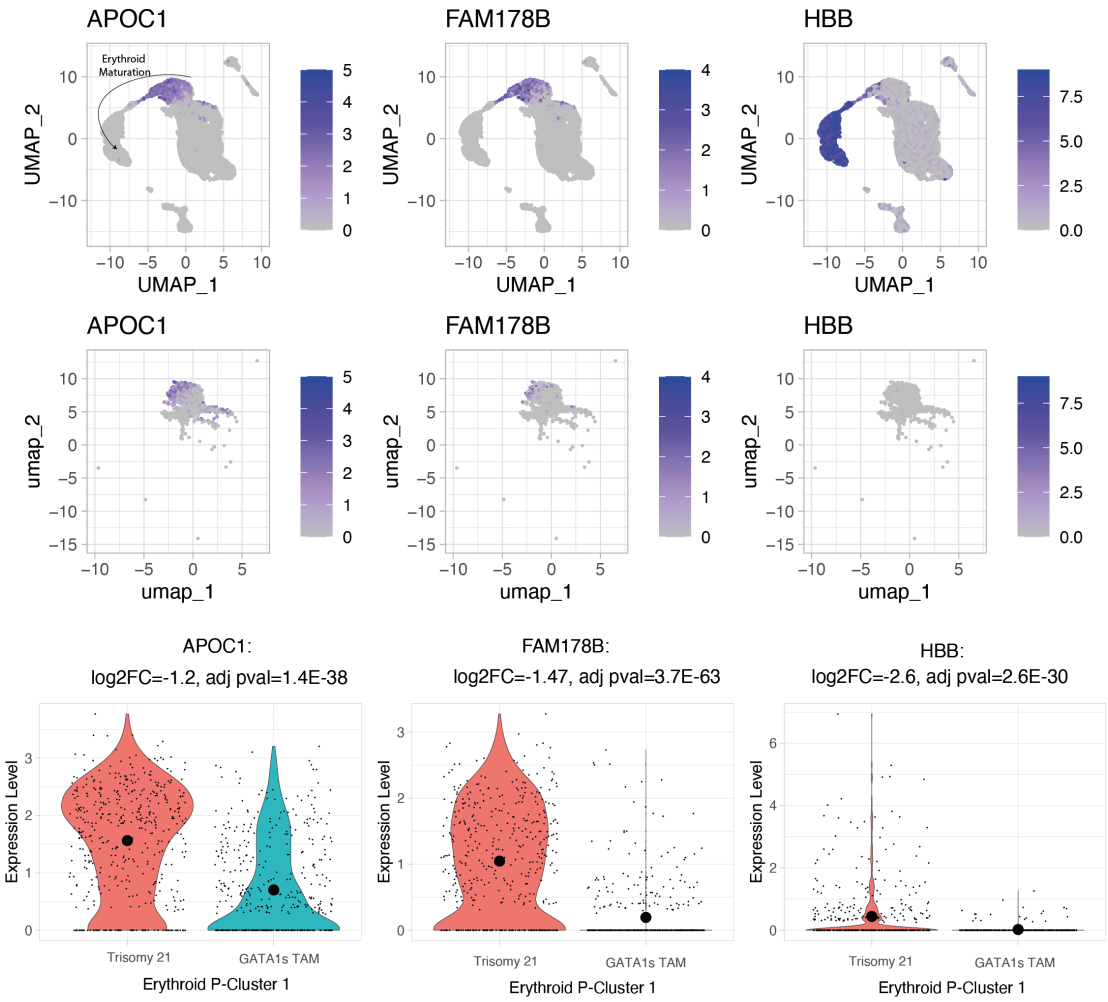


Figure 5.6 Erythroid maturation arrest – downregulated gene expression

APOC1, *FAM1738B* and *HBB* are three genes whose expression is induced during cellular maturation of the erythroid lineage in trisomy 21. Visualisation of single Trisomy 21 cells (top) and TAM cells (bottom) in UMAP space. Bottom, violin plots of *APOC1*, *FAM178B* and *HBB* expression in Trisomy 21 (red) vs GATA1s+ TAM (green) erythroid progenitor cells, cluster 1.

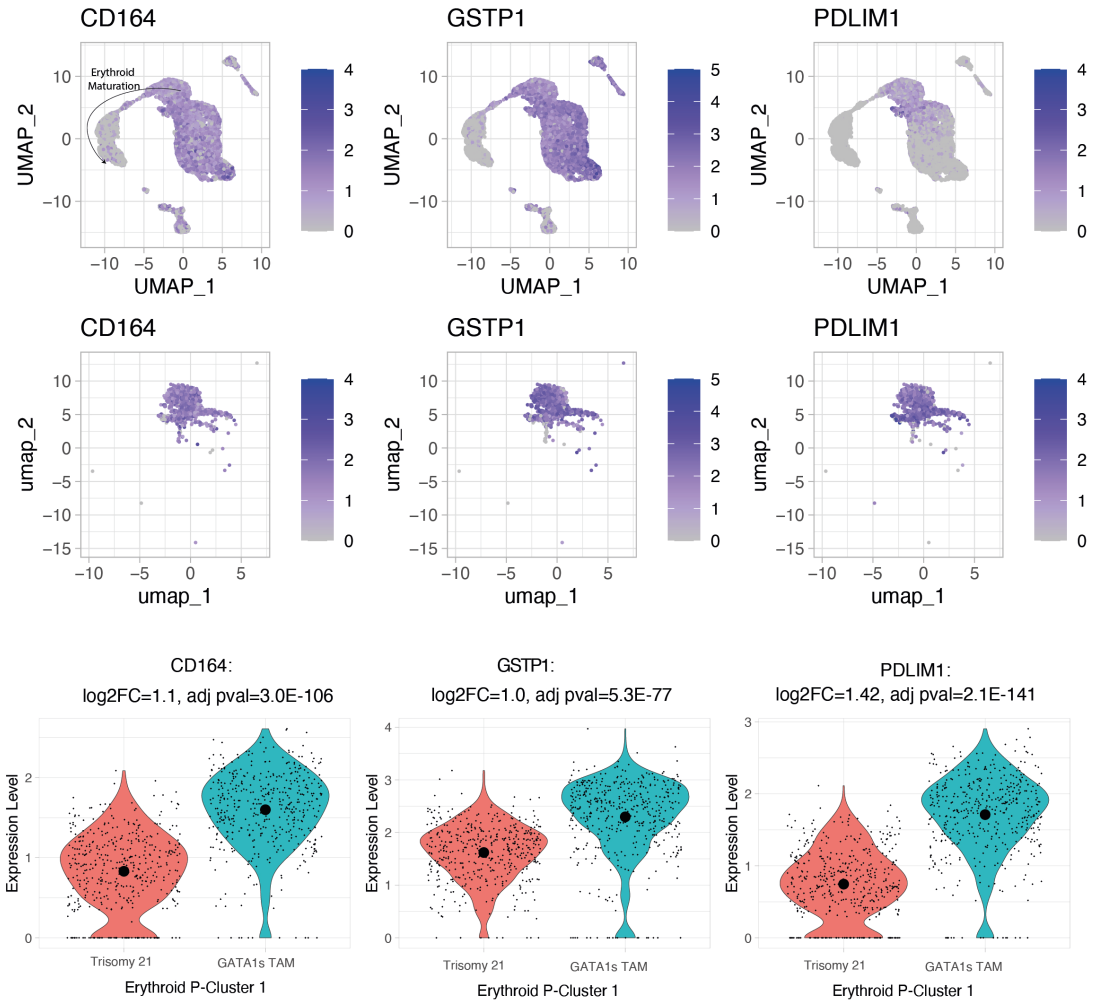


Figure 5.7 Erythroid maturation arrest – upregulated gene expression

CD164, *GSTP1* and *PDLIM1* are three genes whose expression is supressed during cellular maturation of the erythroid lineage in trisomy 21. Visualisation of single Trisomy 21 cells (top) and TAM cells (bottom) in UMAP space. Bottom, violin plots of *CD164*, *GSTP1* and *PDLIM1* expression in Trisomy 21 (red) vs GATA1s+ TAM (green) erythroid progenitor cells, cluster 1.

5.1.2 Functional Consequences of GATA1s Mutations in Terminal Erythropoiesis

What is the molecular mechanism underpinning the bi-phenotypic defects of *GATA1s* mutant cells? Prior data suggest that sustained elevated expression of GATA2 and RUNX1 in GATA1s cells might play a role in explaining these bi-phenotypic defects of the erythroid and megakaryocytic lineages. Experimental evidence from murine cell models of erythropoiesis have identified an anti-correlative gene expression pattern

between GATA1 and GATA2. That is, expression of GATA1 during erythropoiesis coincides with GATA2 suppression at the protein and RNA level. GATA2 and RUNX1 balance haematopoietic self-renewal and differentiation and their expression is downregulated during terminal differentiation of the erythroid but not of the megakaryocyte lineage[104-106]. Chromatin occupancy studies during erythropoiesis have allowed identification of GATA1 and GATA2 direct targets. These studies have defined a replacement of one GATA factor by the another during erythropoiesis and it has been termed GATA factor switching[151,173]. One of the loci that exhibits GATA factor switching is *Gata2* itself. Here, GATA1 replaces GATA2 biding at 5 different regulatory sites of the murine *Gata2* locus[173]. This GATA1 switch at the *Gata2* locus is coupled to transcription repression of *Gata2*. Moreover, murine models of *Gata1s* chromatin occupancy during erythropoiesis have confirmed biding of GATA1s at the GATA2 switch sites to a similar extent as the full-length GATA1 wild type[99]. Nevertheless, exclusive expression of GATA1s in erythroid progenitors leads to GATA2 and RUNX1 overexpression and diminished erythroid maturation[99]. This erythroid defect was rescued by GATA2 reduction[100]. Consequently, the authors hypothesize that erythroid lineage defects in GATA1s murine cells could be due to failed suppression of GATA2 by GATA1s.

Similarly, overexpression of GATA2 in erythroid leukaemia cells lines[105] and primary murine erythroid progenitors[104] impedes terminal erythroid differentiation and leads to phenotypic changes consistent with megakaryopoiesis. Murine models of GATA1s develop a transient megakaryoblastic myeloproliferative disorder that resolves in utero characterized by sustained GATA2, Myc, Myb, and PU.1 expression[97]. Moreover, RUNX1 and GATA2 in association with other transcription factors determine megakaryocytic lineage priming[166]. ML-DS is also characterized by sustained expression of *GATA2*, *MYC* and *KIT*, a direct gene GATA2 target[174]. CMKs cells, an ML-DS cell line is also characterized by high expression of GATA2 (Figure 4.17)

Like disomic erythroid cell differentiation, in trisomy 21, single cell RNA sequencing suggests that there is suppression of *GATA2,* and *RUNX1* expression as erythroid progenitor cells commit to terminal maturation (Figure 5.8), while EMMPPs (cluster 8), and erythroid progenitors (cluster1) co-express *GATA2* and *RUNX1*. In TAM due to *GATA1s* mutation, these two clusters are characterized by over expression of *GATA1* itself (EMMPPs fold change =2.1 and Erythroid P fold change =1.4) and a higher expression of *GATA2* and *RUNX1* compared to their normal counterparts (Figure 5.8). Unlike erythropoiesis, Trisomy 21 megakaryopoiesis is characterized by sustained expression of *GATA2* and *RUNX1* as cells progress from EMMPPs to megakaryocyte progenitors (cluster 16). Like TAM erythroid cells, TAM megakaryocyte cells also over express *GATA1* and *GATA2* and are characterized by elevated *RUNX1* expression (Figure 5.9).

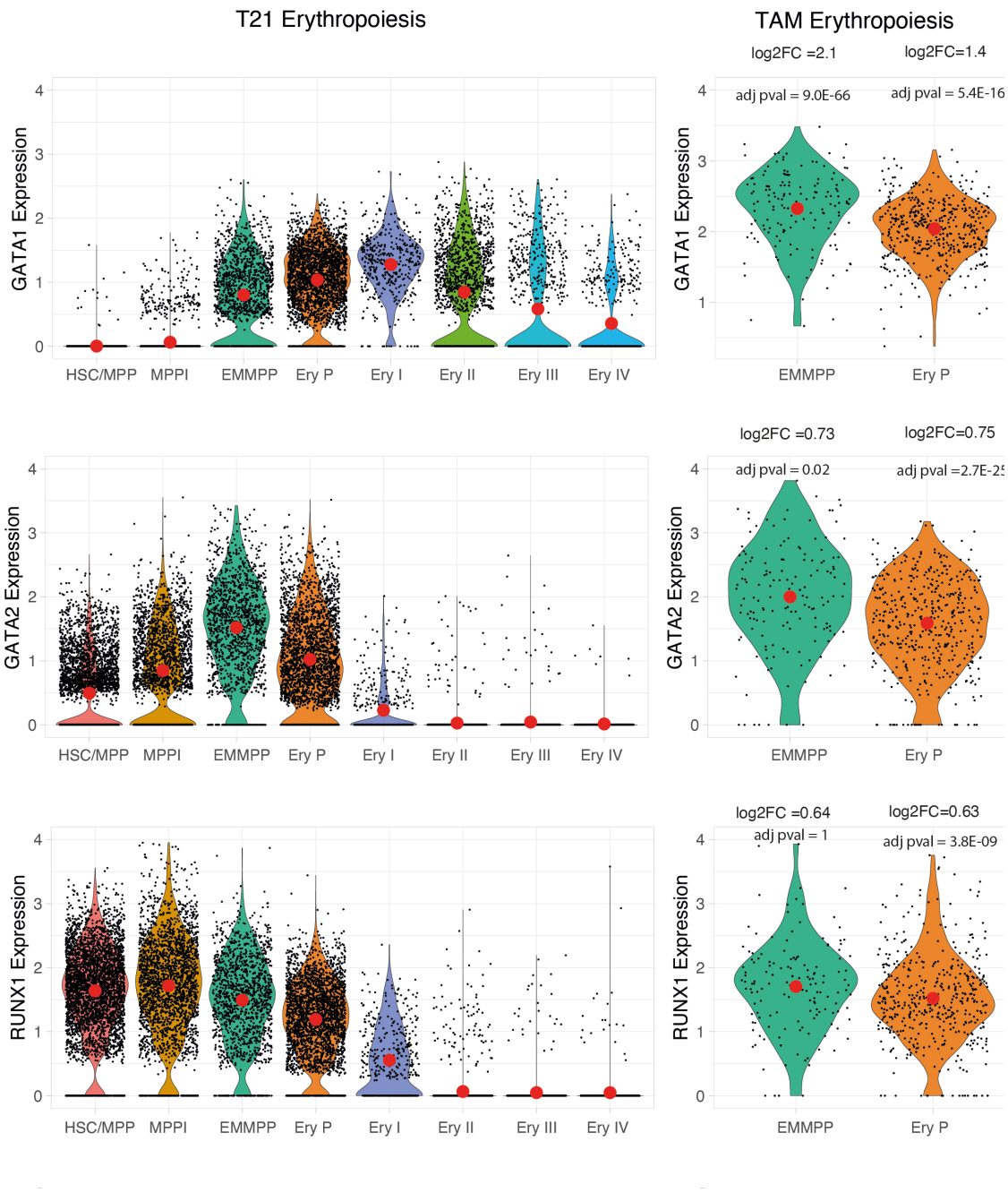


Figure 5.8 *GATA1, GATA2* and *RUNX1* expression in T21 and TAM erythropoiesis

Normalized, log transformed expression of *GATA1*, *GATA2* and *RUNX1* (y-axis) in single T21 (left) and TAM (right) cells during Erythropoiesis. Each dot represents a single cell. The log2 fold change expression (log2FC) and its corresponding adjusted p value in TAM relative to T21 for the EMMPP and Erythroid P (Ery P) are depicted on top of each cluster. Red dots denote the mean expression for each gene per cluster.

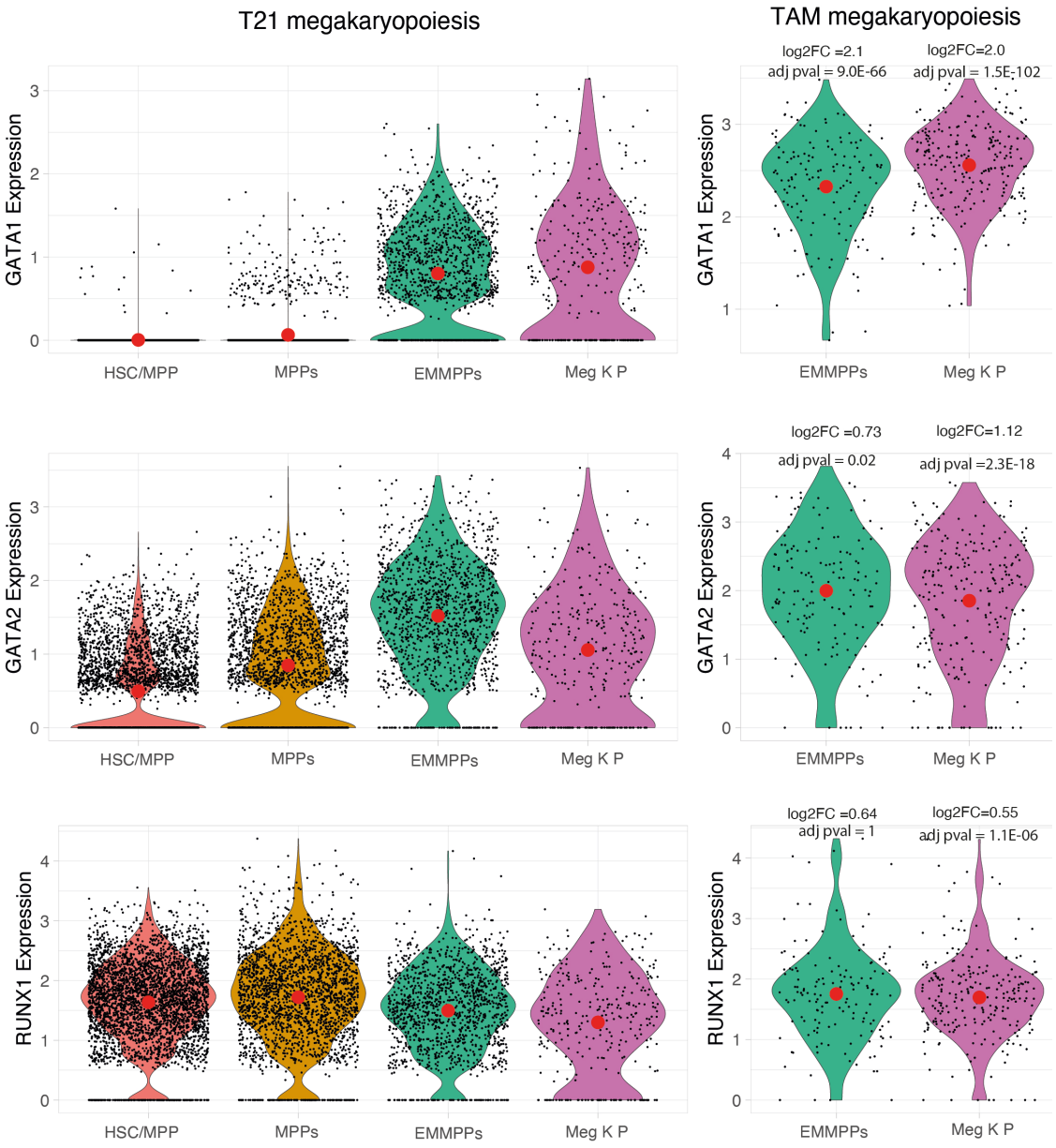


Figure 5.9 *GATA1, GATA2* and *RUNX1* expression in T21 and TAM megakaryopoiesis

Normalized, log transformed expression of *GATA1*, *GATA2* and *RUNX1* (y-axis) in single T21 (left) and TAM (right) cells during Megakaryopoiesis. Each dot represents a single cell. The log2 fold change expression (log2FC) and its corresponding adjusted p value in TAM relative to T21 for the EMMPP and Erythroid P (Ery P) are depicted on top of each cluster. Red dots denote mean expression for each gene per cluster.

To functionally validate the erythroid defect due to *GATA1s* mutations and explored GATA2 sustained expression, I employed CRISPR Cas9 with sgRNAs directed to exon 2 of GATA1 with the goal of recreating GATA1s mutations in primary haematopoietic

stem and progenitor cells (CD34+ cells). I employed the ribonucleoprotein (RNP) version of CRISPR Cas9 with purified spCas9 complexed to chemically modified sgRNAs. First, I designed multiple sgRNAs that target exon 2 of *GATA1*. I chose pairs of these sgRNAs that when delivered simultaneously can entirely delete exon 2 of GATA1 (Supplementary Figure 24). I also employed two control sgRNAs that target the HBB and AAVS1 locus and noticed that the indel rate plateaus after 72 hours post electroporation. Moreover, after 72 hours, one of the pair of sgRNAs (sgRNA A1-A10) was highly efficient at recreating GATA1s^{only} cells in K562 cells (supplementary figure 4, appendix I), with an expected exon deletion of 205 nucleotides. I employed this pair of sgRNAs in sorted cord blood derived CD34+ enriched cells with greater than 95% purity. After 6 days in erythroid promoting cytokines (SCF, EPO, IL3, IL6), GATA1 N and C' terminus, GATA2 and glycophorin A (GlyA), a marker of terminal erythroid differentiation, were assessed using flow cytometry. I observed a 30 % decrease in GATA1 wild type cells and an appearance of 28% of cells with exclusive GATA1s expression (GATA1s^{only}) compared to cells that were electroporated with a control RNP (Figure 5.10 A). There were two GATA1 wild type populations of cells with low (1181 MFI +/- 12.7) and high (7601 MFI +/- 218) GATA1 mean fluoresce intensity (MFI) (Figure 5.10 B). By gating out the GATA1 wild type cells with either low or high MFI, and measuring the MFI of GATA2-FITC signal, I observed that cells with high GATA1 wild type expression had a lower GATA2 MFI (2179 MFI +/- 141) than cells with a low GATA1 expression (GATA2 MFI= 3471 +/- 418). These two populations of cells most likely correspond to erythroid committed cells (GlyA^{+} GATA1^{high}, GATA2$^{low/negative}$) and erythroid progenitor cells (GlyA^{-}, GATA1^{low}, GATA2^{+}). GATA1s^{only} cells on the other hand, are marked by a high GATA2 expression at the same level of wild type cells with low GATA1 expression (Figure 5.10 B and D).

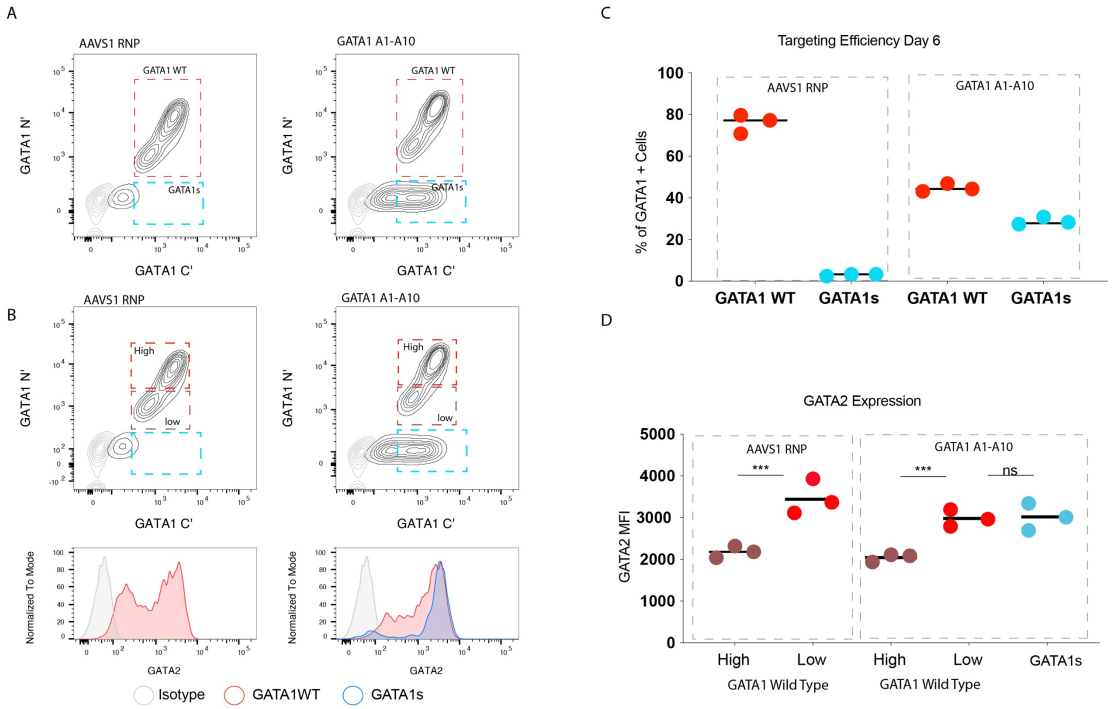


Figure 5.10 Gene engineering of GATA1s mutations

CD34+ cells were FACS sorted with a purity of >95%. After 48 of cell expansion, a AAVS1 RNP control (ribonucleoprotein) as well as a pair of RNPs (A1-A10) targeting exon 2 of GATA1 were electroporated. After overnight recovery, edited cells were forced to differentiate into erythroid cells. At day 6, AAVS1 (control) and A1-A10 edited cells were prepared for iFC to detect GATA1s^{only} and GATA1^{WT} cells. A) Red gate represents GATA1 wild type cells and blue gate represents GATA1s^{only} cells gated from live, singlets. C) Editing efficiency as measure by percentage of GATA1s^{only} cells generated. B) GATA1 wild type cells can be divided into high GATA1 expression (red gate) and low GATA1 expression (brown gate) or GATA1s^{only} cells (blue gate) (top). Overlayed histogram of normalized GATA2 signal of either isotype control (grey), GATA1^{WT}(red), or GATA1s^{only} (blue) (bottom). D) The mean fluorescence intensity (MFI) of GATA2 FITC signal is measure for each gate. Three technical replicates were run and the two tail T-test with p value style APA: 0.12 (ns), 0.033 (*), 0.002 (**), <0.01 (***)

The efficiency of the CRISPR Cas9 system allowed me to compare GATA1s^{only} and GATA1^{WT} cells within the sample. Erythroid differentiation of the GATA1 wild type cells shows that as cells gain GATA1 expression they also gain glycophorin A expression (Figure 5.11 A). Similarly, as cells gain glycophorin A expression they dramatically supressed GATA2 expression (Figure 5.11). Unlike GATA1^{WT} cells, GATA1s^{only} do not express substantial glycophorin A and retain a high GATA2 expression (Figure 5.11 B). Thus, in trisomy 21 GATA1^{WT} cells gain of glycophorin A expression coincides with

suppression of GATA2 expression and defines terminal differentiation of the erythroid lineage. This process is aberrant in GATA1s^{only} cells (Figure 5.11 B).

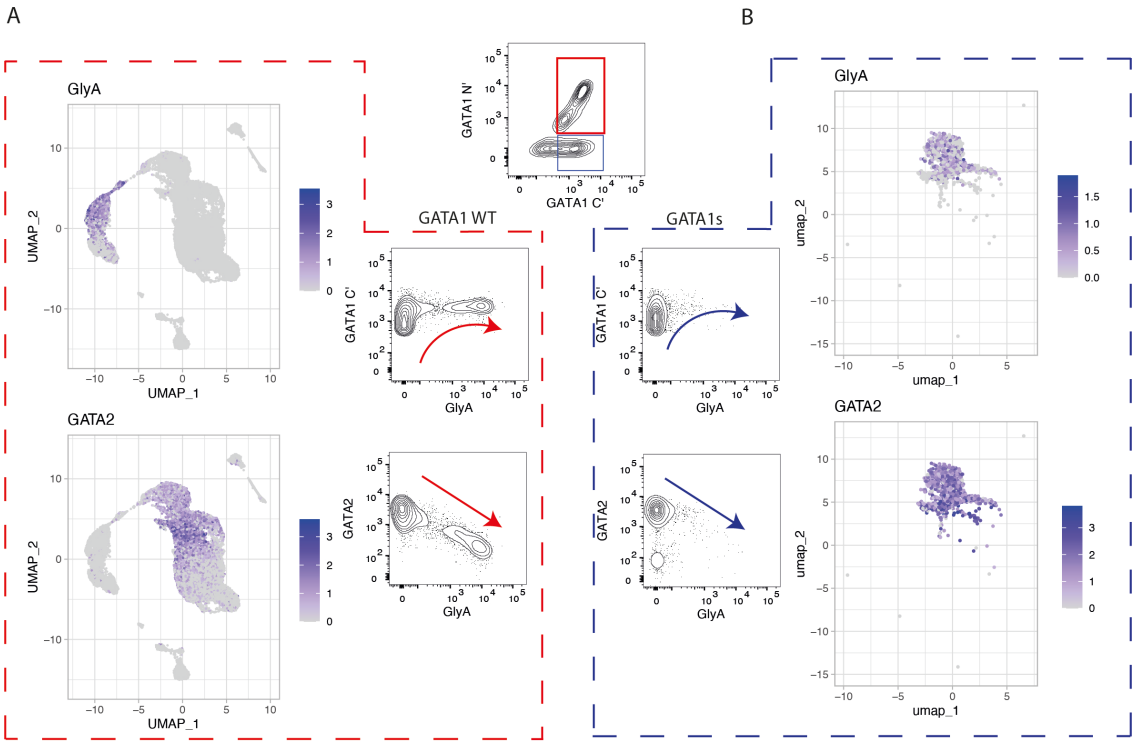


Figure 5.11 Erythroid maturation arrest and sustained GATA2 expression in GATA1^{WT} trisomy 21 and GATA1s TAM cells

A and B) Gene edited CD34+ cells were exposed to erythroid promoting cytokines for 6 days, then glycophorin A, GATA1 and GATA2 expression were assessed by flow cytometry. GATA1 wild type cells (red square) and GATA1s^{only} (blue square) where plotted as either GATA1 C' (y-axis) and glycophorin A/GlyA (x-axis) or GATA2 (y-axis) and GlyA (x-axis). Red arrows refer to erythroid maturation. Normalized log transformed *GATA2* and *GlyA* expression are also shown in UMAP space for trisomy 21 (red dotted square) or TAM (blue dotted square).

To verify the erythroid defect in disomic CD34+ cells after the engineering of *GATA1s* mutations, I sorted 500 edited CD34+ cells (purity >95%) into methocult media to promote colony formation of erythroid and myeloid lineages. After 14 days, the colonies were scored as either erythroid or myeloid colonies, based on morphology. Their phenotype was corroborated by measuring glycophorin A or CD14/CD15 expression by flow cytometry respectively. The percentage of myeloid colonies from cells

electroporated with *GATA1* exon 2 targeting RNP (mean 50.8 percent) was higher than the *AAVS1* targeting RNP control (mean 49.1 %). Also, the percent of erythroid colonies was lower in *GATA1* exon 2 targeting (mean 49.1%) compared to the control (mean 57.2%) (Figure 5.12 A). Surprisingly, for every 100 colonies I found about 4 colonies with an excessive size possibly of hyper proliferation and a different immunophenotype than normal erythroid or myeloid colonies (Figure 5.12 C). They were characterized by lack of myeloid (CD14, CD15), megakaryocytic (CD41) and eosinophil/basophil/mast cell (CD125) but expressed KIT, CD71 and low levels of Glycophorin A. They resemble immature blast cells even after 14 days of the colony forming assay (Figure 5.12 B). In summary, I propose that the erythroid defect seen in GATA1s cells is most likely due to GATA1s failure to supress GATA2. To test this, chromatin profiling studies of the GATA2 locus are required. In the next and final chapter, I explore how new technologies will allow to profile genome-wide chromatin occupancy of transcription factors to be performed in rare primary haematopoietic cell types.

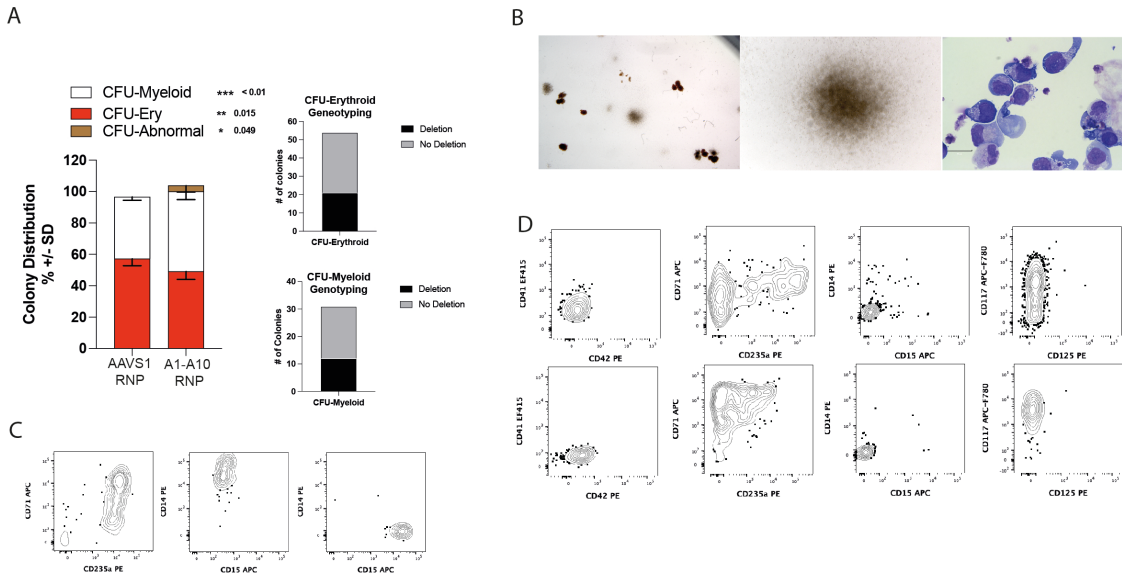


Figure 5.12 Hyperproliferative blast colonies due to GATA1s mutations in disomic cells

A) Colony distribution of a colony-forming assay of CD34+ cells (n= 3 disomic cord bloods) showing a significant increase in abnormal, and myeloid colonies and a significant decrease in erythroid colonies in *GATA1* exon 2 targeting compared to AAVS1 targeting control. B) Erythroid colonies are defined by morphology and red colour due to haemoglobin expression (C) and corroborated by CD71 and GlyA/CD235A expression by flow cytometry while myeloid colonies are not red and their immunophenotype is defined by CD15+ or CD14+ expression by flow cytometry. B) A representative picture of an abnormal colony surrounded by erythroid colonies (left), zoomed in of a picture of an abnormal colony (middle) and H&E of an abnormal colonies (right). D) Representative immunophenotype of an abnormal colony.

6 Genome-Wide Identification of GATA1 Biding Sites by CUT&TAG

Chromatin immunoprecipitation followed by next generation sequencing (Chip-seq) has allowed the genome-wide assessment of transcription factor binding as well as multiple chromatin modifications. Unfortunately, Chip-seq suffers from a low signal and high background and requires formaldehyde fixation that can mask the epitope being assessed. Because of these issues, Chip-seq requires millions of cells (~10^7), which prohibits its use on primary rare cell types such as megakaryocyte or erythroid progenitor cells. Alternative methods of chromatin profiling that can work with a low number of cells have been developed such as ChIPmentation[175]. Other methods that do not require fixation have also been developed. These methods are based on enzymatic activity

tethered directly to the protein being profiled or to antibodies directed to them. These methods are DamID, ChEC-seq, CUT&RUN[176]. Recently, a new method that can overcome the limitations of Chip-seq and CUT&RUN was published[111]. This method is referred as CUT&TAG, in which hyperactive Tn5 transposase is covalently bound to Protein A/G. Then, the fusion pAG-Tn5 is loaded with sequencing ready adaptors. In situ antibody-mediated tethering of pAG-Tn5 to chromatin associated proteins results in tagmentation of surrounding DNA. DNA fragments are ready for PCR amplification and sequencing. CUT&TAG has been used to profile a variety of chromatin modifications in low cell numbers[111] and even single cells[177,178]. CUT&TAG has even been used to simultaneously profile multiple chromatin modifications in the same cells[179]. Yet, the nature of the CUT&TAG protocol has limited its used to profile transcription factor DNA binding.

6.1 Optimization of CUT&TAG For GATA1 N' and C' Terminus In Low Cell Numbers

The protocol can be divided into 8 steps, two of which require a high salt buffer (300mM NaCl) which can hinder transcription factor DNA biding. Nonetheless, CUT&TAG has been used to profile at least three transcription factors so far, NPAT[176], RAD21 and OLIG2[177], of which RAD21 and OLIG2 have been done in single cells.

Table summarizing the protocol steps in CUT&TAG

1. Nuclei isolation
2. Incubation of nuclei with primary rabbit antibody
3. Incubation of nuclei with secondary anti rabbit antibody
4. Incubation of nuclei with pAG-Tn5 fusion protein in high salt buffer (300mM NaCl)
5. In situ tagmentation in high salt buffer
6. pAG-Tn5 release and DNA isolation
7. PCR amplification
8. DNA quantification and sequencing

In order to develop an optimized CUT&TAG protocol to profile GATA1 binding in as low as 50,000 K562 cells I utilized the same anti GATA1 C' and anti GATA1 N' antibodies (Cell Signalling clones: D52H6 and D24E4) as used in the iFC protocol

because they are highly specific (Figure 6.1 A). Reasoning that the Omni-ATAC[180] nuclei isolation protocol optimized for a variety of primary human tissues, would be suitable for CUT&TAG to help remove mitochondrial DNA from the transposition reaction and can help with nuclei clumping. The Omni-ATAC nuclei isolation buffer contains 0.1% NP40, 0.1% Tween-20, and 0.01% digitonin. Digitonin was resuspended in DMSO instead of the recommended water to limit digitonin precipitation and ensures a consistent concentration. Next, a DAPI and acridine orange mixture was used to measure cell permeabilization. While DAPI binds DNA, acridine orange binds RNA and the use of both can identify intact nuclei and nuclei coming from dead cells (Figure 6.1 B). 1% bovine serum albumin was added as it has been shown to help with reduced clumping of nuclei without altering the tagmentation efficiency[177].

The current CUT&TAG protocols recommend primary antibody incubation at 4 degrees Celsius overnight. This incubation time might be suitable for epigenetic modification which are stable but not for transcription factors whose DNA residence is measure in minutes. To optimize the primary antibody incubation time for GATA1, I carried out an experiment in which nuclei isolated from a mixture of thee cell types mixed at equal ratios (K562s, THP1s, CMKs) were exposed to anti GATA1 N' and anti GATA1 C' termini antibodies conjugated two different fluorochromes. K562 cells express both GATA1 isoforms while CMK cells express only the short isoform, and THP1 cells no not express GATA1. After incubation of mixed nuclei with primary antibodies for a series of times at either 4 or 22 degrees Celsius, flow cytometry was used to distinguish the three types of nuclei (Figure 6.1 C). I reasoned that the time and temperature required to distinguish the three nuclei by flow cytometry would be ideal for CUT&TAG. From this I found out that incubation of nuclei with primary anti GATA1 antibodies required at least 3 hours at 4 C for the signal to be stabilized and the three cell lines to be identified. With these optimized parameters I performed CUT&TAG to profile H3K27me3 epigenetic mark as a positive control in 50,000 K562 cells as well GATA1 binding in 50,000 K562,

CMK and THP1 cells (Figure 6.1 D); rabbit IgG isotype was also used as a negative control. After pA/G-Tn5 release and PCR amplification for H3K27me3 CUT&TAG, I obtained 45 ng DNA yield for the IgG and 623 ng in average when an antibody against H3K27me3 was used (Figure 6.1 D). Moreover, capillary electrophoresis of the amplified DNA was characterized by a nucleosomal pattern for H3K27me3 but not for IgG. Similarly, when D52H6 antibody clone was used which recognizes the GATA1 N' terminus only present in K562 cells, I observed an average of 103 ng DNA yield for K562 cells, and an average of 60 and 53 ng for CMK and THP1 respectively. When the isotype was used in CUT&TAG, I obtained an average of average of 43.6 ng, 43.3 ng and 61ng for K562, CMK, and THP1 cells, respectively (Figure 6.1 D). This agrees with K562 cells expressing the full length GATA1 isoform. When the D24E4 antibody clone was used, which recognized the C' terminus of GATA1 and consequently both GATA1 isoforms, I obtained an average of 139 ng DNA yield for K562 cells, and an average of 158 ng in CMKs and only 24 ng for THP1s. Consistent with the fact that both CMKs and K562 cells express the GATA1s isoform. Moreover, capillary electrophoresis failed to detect a substantial signal when the IgG control was used (Figure 6.1 D) for all cell lines and for THP1s when both anti GATA1 antibodies were used.

6.2 C&T Detect: A qPCR Based Method to Quantify Known Transcription Factor Binding Sites

To further validate CUT&TAG, I designed a method referred as C&T detect. Like Chip-qPCR, it is based on the relative quantification by qPCR of a known GATA1 biding site. After tagmentation, a primer specific for the Tn5 adaptor and a reverse primer specific to the *GATA1* promoter were used for qPCR amplification based on cyber green. The cycle threshold or CT was used to quantify the relative yield of a known GATA1 biding sit at the *GATA1* promoter, where a higher CT means a low DNA yield. There was a 150-fold increment relative to the IgG control in K562 cells when the anti GATA1 C' or N' termini antibodies were used. While for CMKs there was a 15-fold increment when

the anti GATA1 C' antibody was used but no fold change when the anti GATA1 N' was used. No relative change was observed when for either antibody in THP1 cells (Figure 6.1 D). Therefore, this is consistent with the fact that K562 cells express both GATA1 isoforms and that GATA1 binding can be detected at the *GATA1* promoter, while CMK cells only express the short GATA1 isoform which is known to retain the two zinc fingers DNA binding domains and can also bind at the *GATA1* promoter.

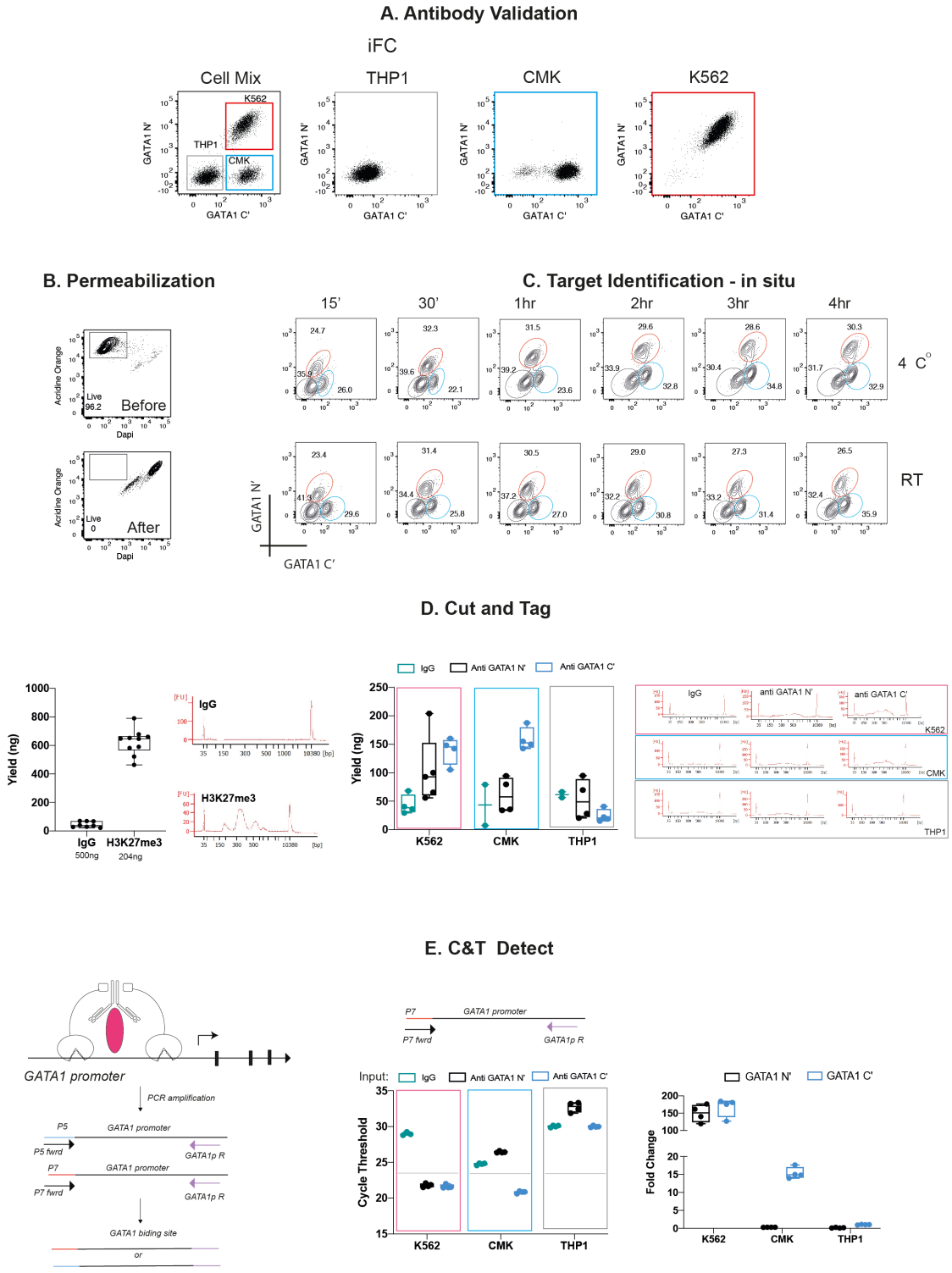


Figure 6.1 CUT&TAG optimization

A) iFC of THP1s, CMKs, K562, with anti GATA1 N' (clone: D52H6, y-axis) and C' (clone:D24E4, x-axis). B) Flow cytometry of isolated nuclei stained for DNA (DAPI, y-axis) and RNA (Acridine orange, y-axis) before permeabilization (top) and after (bottom). C)Intracellular flow cytometry of THP1s, CMKs and K562 cell mixed at equal ratio, with no fixation stained with anti GATA1 N' (y-axis) and C' (x-axis) terminus for 15, 30 minutes or 1, 2,3 and 4 hours at room temperature (top) or 4 C (bottom). D) CUT&TAG was performed in K562 nuclei with anti

H3K27me4 (clone: C36B11) or in CMKs, K562s, and THP1s with anti GATA1 N' and C' termini, the DNA yield as well as the fragment size was measured. E) Schematic representation of C&T detect (left), cycle threshold of qPCR analysis of CUT&TAG DNA yield (middle) and fold change(right) for each cell line and each GATA1 N' or C' terminus as well as IgG control.

CUT&TAG is characterized by a low yield of DNA after tagmentation specially for transcription factors. This low DNA yield must be PCR amplified to prepare libraries as well as to have enough DNA for sequencing. One must PCR amplify the tagmented DNA with an optimal PCR cycle number to avoid reducing the complexity of the DNA library and a high number of PCR duplicates. To obtain enough DNA for sequencing but at the same time limit the number of PCR cycles, I optimized a qPCR-based approach to determine the appropriate number of PCR cycles required.

To test this, I ran a series of CTU&TAG experiments with 2000 K562 cells with GATA1 anti C' and N' terminus antibodies. After tagmentation, the DNA yield was initially amplified for 15 PCR cycles, then 5 ul of each 50 ul reactions was taken and run for 40 qPCR cycles with cyber green to determine the number of cycles required for fluorescent signal to plateau. Then the left over 45 ul of the initial reaction was run for 4, 6, 8 or 10 additional PCR cycles. As seen from Figure 6.2, for either GATA1 C' or GATA1 N' terminus, at 15 PCR cycles no DNA signal is observed by capillary electrophoresis. Only after 19 cycles of PCR amplification from 2000 K562 cells (15+4 cycles) we can see a signal in capillary electrophoresis. This signal increases by additional PCR cycles but at the same time, the average size of the fragment is also reduced as PCR preferentially amplifies small fragments. This approach allows to determine the minimum number of PCR cycles required before qPCR exponential amplification for example 19 cycles for 2000 cells.

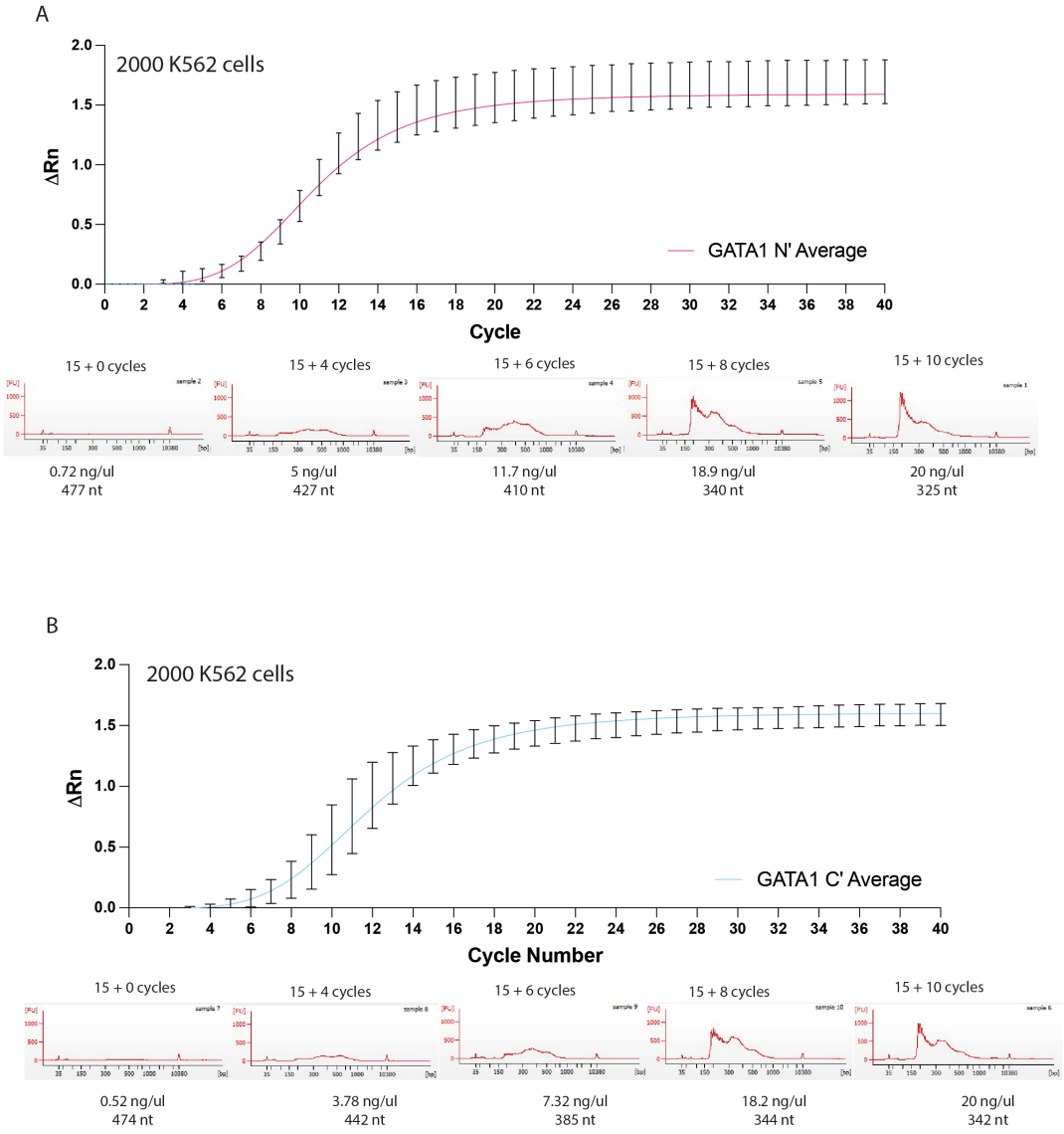


Figure 6.2 PCR cycle optimization of CUT&TAG DNA yield by qPCR

CUT&TAG was performed against GATA1 N' (A) and GATA1 C' (B) in 2000 K562 cells and the tagmented DNA was PCR amplified for 15 cycles, then 5 ul were taken and further amplified for 4, 6, 8 or 10 cycles with the LUNA qPCR master mix to identify the appropriate cycle number before exponential amplification. The final product was run on a high sensitivity bioanalyzer to measure the size and yield. The bioanalyzer traces are shown under the qPCR curve.

6.3 CUT&TAG Bioinformatic Analysis and Quality Control

Having optimized the nuclei permeabilization, antibody incubation and PCR cycle number, I next sequenced CUT&TAG from 1,000, 10,000 and 50,000 K562 cells to profile H3K27me3 epigenetic marks. I also sequenced CUT&TAG fragments from 50,000 K562, CMK and THP1 cells to profile N' and C' GATA1 binding sites aiming to sequence 5 million reads per sample, which is much fewer sequencing requirements than Chip-Seq. No PCR duplicates were removed as it is known that Tn5 has a stereotypical cutting preference, which means that some PCR fragments with similar ends come from different cells. To test the replicate reproducibility, after first mapping the DNA fragments, the Pearson correlation of the log2 transformed of number of reads mapping to windows of 500 bases long for each replicate were determined. Hierarchical clustering was performed on the Pearson correlation value. This analysis separated samples according to the antibody used, IgG vs H3K27me3 independent of cell number (Figure 6.3 A). For H3K27me3, there was a high concordance (Person correlation >0.88) for all samples. Similarly, when anti-GATA1 antibodies were used, samples clustered according to the antibody used and cell type. More importantly, full-length GATA1 binding site in CMKs were clustered together with IgG control (Figure 6.4 A). This is expected as CMKs do not express theGATA1 full-length isoform.

H3K27me3 is an epigenetic modification of heterochromatin regions, thus it is considered a repressive mark. Consequently, the fragment size distribution of H3K27me3 should show a nucleosome pattern. Fragment distribution of mapped reads from H3K27me3 CUT&TAG shows at least three peaks, a nucleosome free region (>140 nucleotides), the first nucleosome with a higher density of fragments and a third nucleosome peak (Figure 6.3 B). CUT&TAG for GATA1 should not be characterized by nucleosome pattern as H3K27me3 does, instead it should be characterized by an enrichment of a nucleosome free region. As expected, the fragment distribution of

GATA1 C' and GATA1 N' terminus was enriched for a nucleosome free region and substantially differed from an H3K27me3 pattern or an ATAC-seq pattern (Figure 6.4 B).

The % of unique reads is high (> 80%) for either in 1,000, 10,000 or 50,000 cells for H3K27me3 epigenetic marks (Figure 6.3 C). These marks are abundant, therefore PCR amplification of tagmented DNA is less likely to lead to a high number of PCR duplicates. While GATA1 biding sites are much rarer in K562 or CMKs, consequently they are more likely to have a high number of PCR duplicates. I observed about 30% unique reads for GATA1 N' or C' terminus in K562 or CMKs (Figure 6.4 C).

Chip-Seq data analysis involves peak calling algorithm which identify genomic regions where the signal is higher than non-targeting antibody isotype controls. In Chip-seq the sequencing requires are high, consequently they lead to high background signal. Thus, peak calling algorithms are primarily concern on statistical methods that allow identification of signal enrichment over background. Both CUT&TAG and CUT&RUN are characterized by genomic regions with no coverage, consequently new peak calling algorithms have been designed for this purpose. Sparse Enrichment Analysis for CUT&RUN or SEACR is a package design to call peaks from data with very low background such as CUT&TAG[116]. We applied SEACR to call peaks for H3K27me3 and GATA1 compared to isotype controls. First, H3K27me3 was characterized by a high percent of reads mapping to peaks (FRIP) (>50%) (Figure 6.3 C). While for GATA1, there was at most 10% of reads mapping to peaks (Figure 6.4 C). Although the number of peaks varied from sample to sample and with different cell numbers used, there were 47,049 shared peaks between all samples for H3K27me3 if stringent criteria were used and 98,733 peaks if relax criteria were used (Figure 6.3 E). Similarly, for GATA1 the number of peaks varied across cell types and antibodies used. I identified 4,916 shared peaks between GATA1 C' and GATA1 N' in two replicates of K562 cells for each antibody, and 3,206 shared peaks between two replicates of GATA1 C' terminus with stringent criteria (Figure 6.4 E, F). Importantly, no shared peaks were called when the

anti GATA1 N' antibody was used in CMKs and no shared peaks were called for either anti GATA1 antibodies in THP1s cells (Figure 6.4 E, F).

To quantitatively compare the signal intensity and the background level for all peak within a 3000-nucleotide window for each sample, I generated heatmap around genomic regions called for either H3K27me3 or GATA1 C' and N' terminus. We noted that the H3K27me3 signal is narrowly cantered with a high signal to noise ratio, and it minimally decreased with low cell input (Figure 6.3 D). The background signal had at most a score of 2 arbitrary units generated by SEACR. Similarly, for GATA1, I noted a high signal to noise ratio and that the peak signal was much narrower for the N' terminus than the C' terminus, with a background signal less than 1 arbitrary units generated by SEACR (Figure 6.4 D). Nonetheless motif enrichment analysis identified the (A/T) GATA (A/G) motif for each antibody used (Figure 6.5).

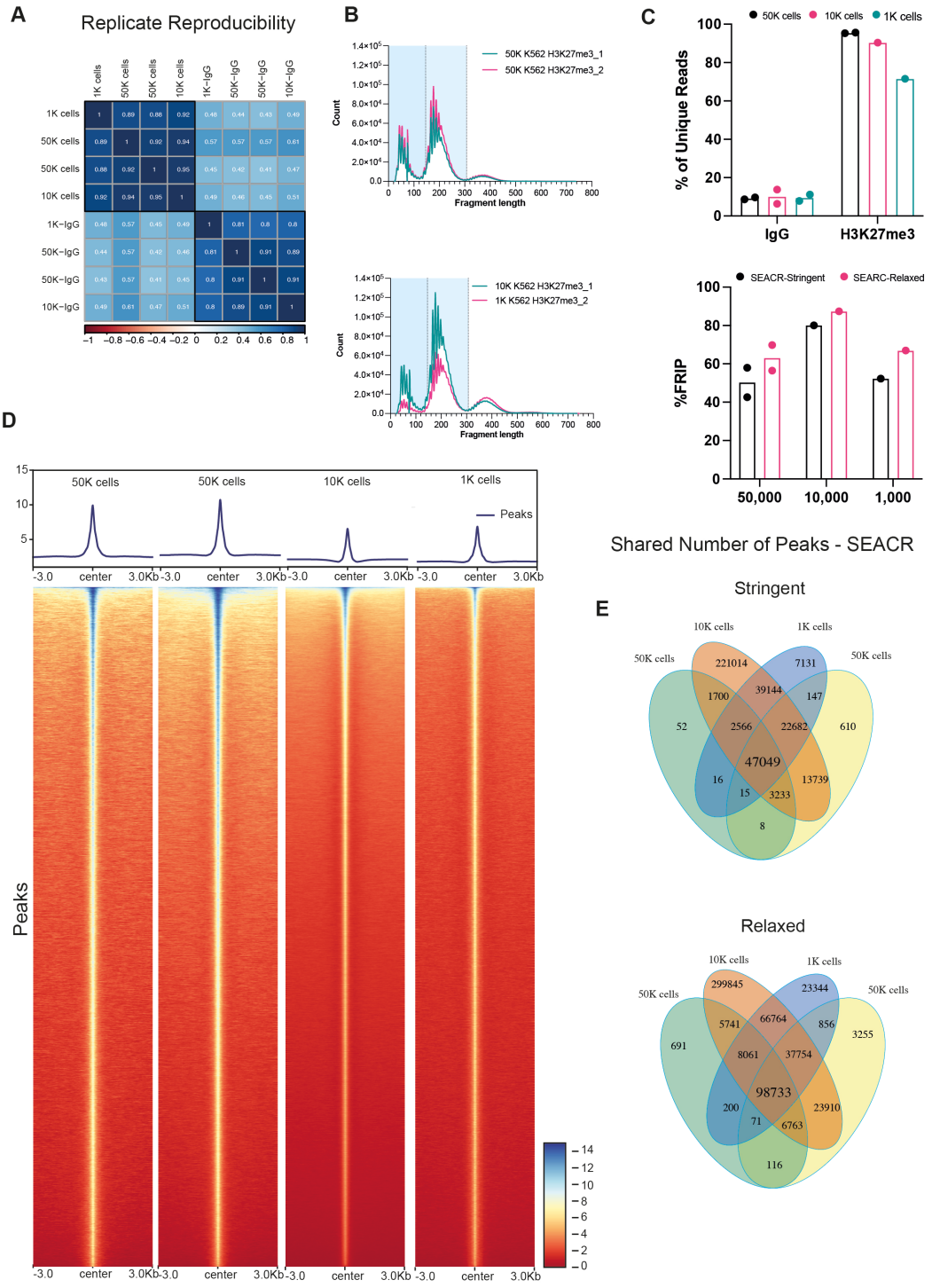


Figure 6.3 CUT&TAG profiles genome wide H3K27me3 epigenetic marks in low cell numbers

A) Pearson correlation and hierarchical clustering of the log2 transformed read count for 500 bp bins for each H3K27me3 replica. B) Fragment size distribution from 50000 (top), 10000 and 1000 cells (bottom). C) Percent of unique fragment reads and percent of fragment reads in peaks (FRIP) after fragment mapping for each sample. D) Heatmap of CUT&TAG signal on all peak called per sample. E) Shared number of peaks called with SEACR with stringent (top) or relaxed (bottom) criteria.

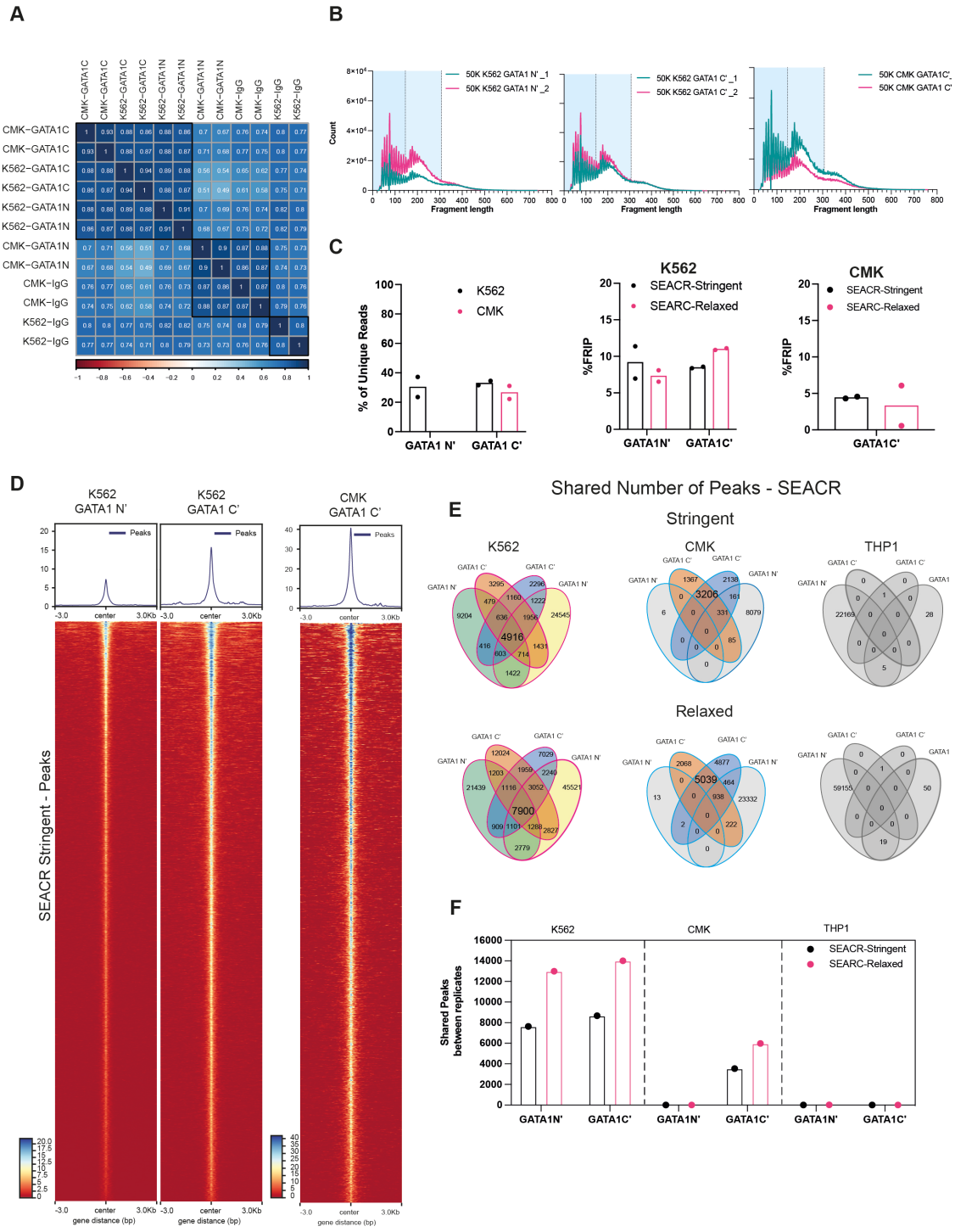
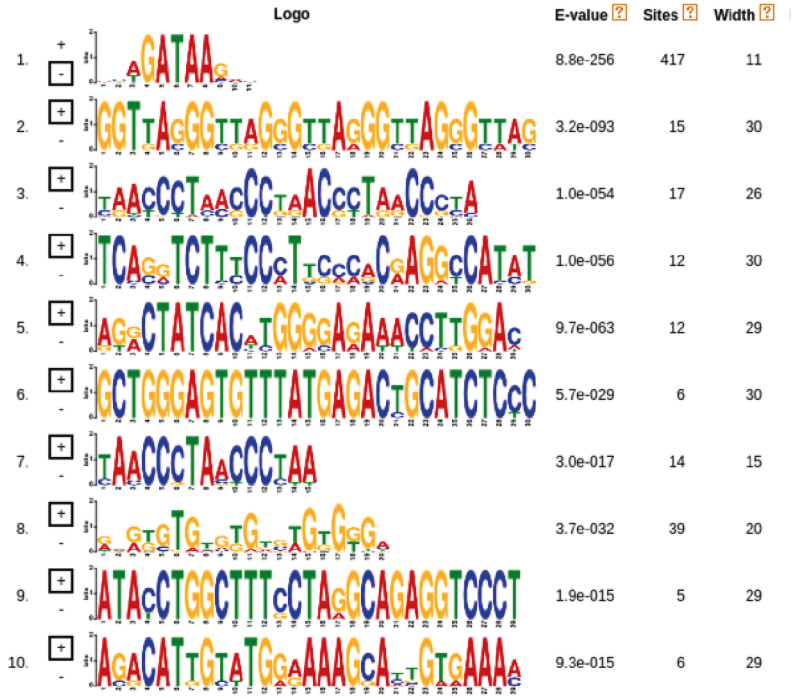


Figure 6.4 CUT&TAG profiles genome wide GATA1 binding sites in low cell numbers.

A) Pearson correlation and hierarchical clustering of the log2 transformed read count in 500 bp bins for each of the GATA1 N' and C' terminus. B) Fragment size distribution from 50,000 cells with GATA1N' (left), or GATA1 C' (middle) in K562s and GATA1 C' in CMK cells (right). C) Percent of unique reads and percent of reads in peaks (FRIP) after fragment mapping for each cell line. D) Heatmap of CUT&TAG signal on all peak called per sample. E) Shard number of peaks called with SEACR with stringent (top) or relaxed (bottom) criteria. F) Bar plot of the number of shared peaks for each cell line and each GATA1 antibody.



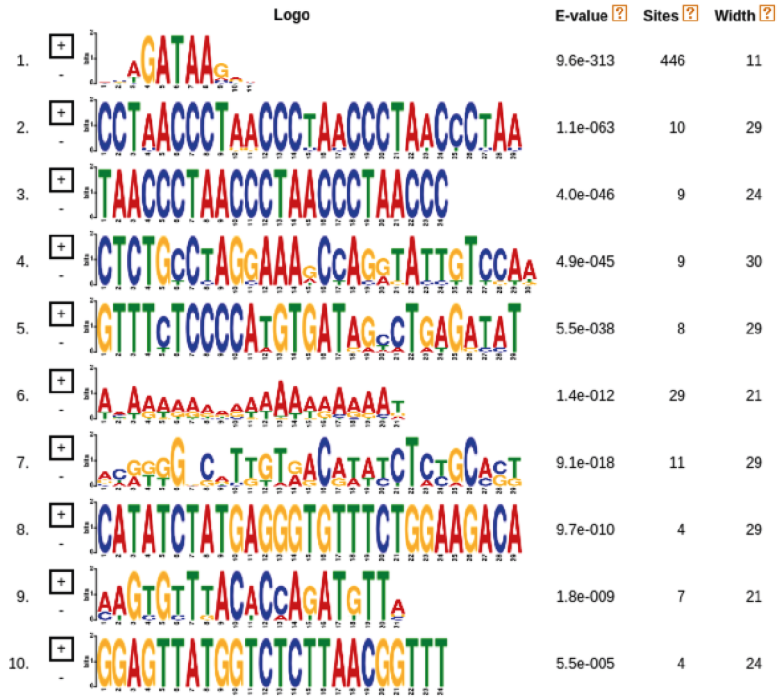


Figure 6.5 Top 10 motifs identified by MEME in GATA1 CUT&TAG

MEME-chip suit was used to identify de novo motifs centrally located in CUT&TAG peaks

6.4 GATA1s Binds at the -77 GATA2 Enhancer in an ML-DS Cell line

Visualization of CUT&TAG read density for GATA1 N' and C' terminus at the *GATA1* locus in K562 and CMK cells identified the known regulatory GATA1 binding sites[181-183]: a -3.6 kb peak upstream of the promoter corresponding to the HS1 site, and a -0.6kb peak corresponding to the IE promoter. Despite the observed overexpression of *GATA1* in *GATA1s* mutant single TAM cells (Figure 5.8, Figure 5.9) as well as in single CMK cells relative to K562 and THP1 cells (Figure 4.17), no differential binding of GATA1 was observed at the *GATA1* locus in CMK cells compared to K562 (Figure 6.6). On the contrary, CUT&TAG read density for GATA1 N' and C' termini in K562 and CMK cells at the *GATA2* locus identified differential binding of GATA1 at the -77kb GATA2 enhancer (Figure 6.7). My data suggest that sustained elevated GATA2 expression might explain in part the bi-phenotypic defects of GATA1s TAM cells but the mechanism behind this is not yet clear. A recent study using targeted deletion of the conserved GATA2 enhancers, has shown that 3 of 5 GATA2 cis regulatory elements, the -1.8, -2.8 and -3.9kb upstream of the GATA2 promoter had little or no consequences for *Gata2* expression[184], while deletion of a conserved GATA motif +9.5kb cis element, was previously shown to abrogated HSC emergence[185]. The -77kb cis element located upstream of the *RPN1* gene has been shown to be a distal enhancer of *GATA2* rather than *RNPN1*[186]. Similar to the +9.5kb element, deletion of the -77 is embryonic lethal but HSC emergence is not affected[187]. While the +9.5kb element regulates HSC emergence, it is the -77 enhancer that confers GATA2 expression in progenitor cells. Point mutations in the GATA2 -77 enhancer have not been reported, yet a translocation of the -77 enhancer to the *MECOM* locus encoding the EVI gene is a disease driver in 3a21; q26 AML[188,189]. It has been proposed that both an elevated EVI expression and reduced GATA2 expression are responsible for leukaemia development[190]. Taken together, these data suggest that that GATA1s biding at the GATA2 77kb enhancer in human cells may be responsible for sustaining GATA2 expression in GATA1s expressing cells.

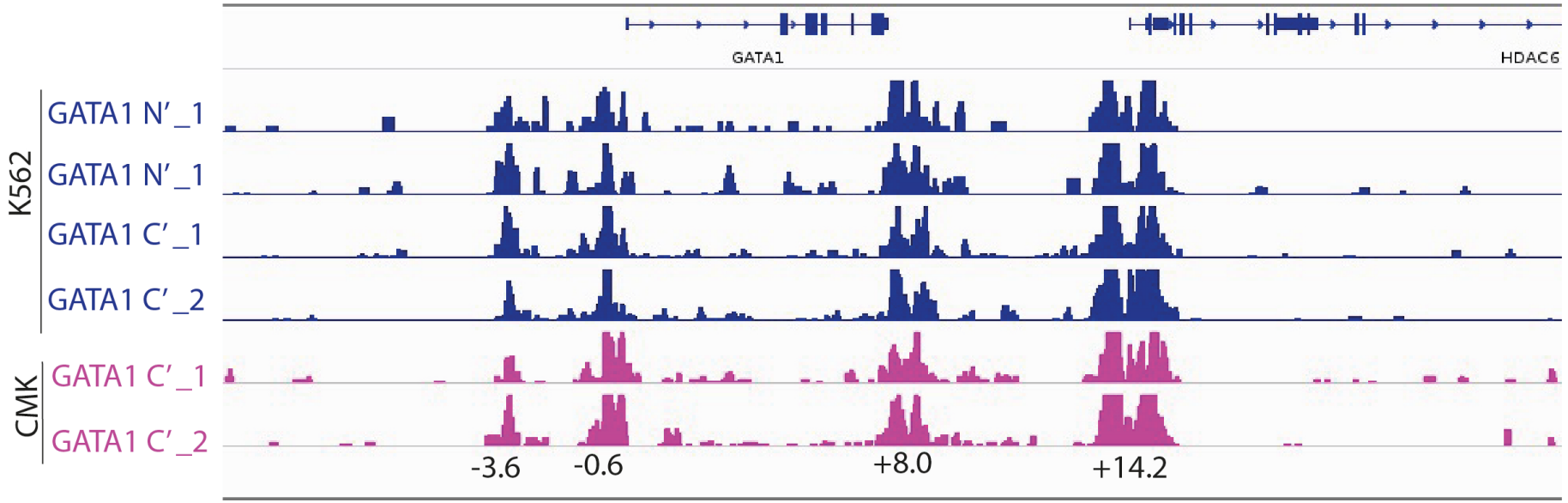


Figure 6.6. GATA1 binding at the GATA1 locus identified by CUT&TAG

CUT&TAG read density at the *GATA1* locus in K562 (blue) and CMKs (pink) cells identifies known GATA1 biding sites.

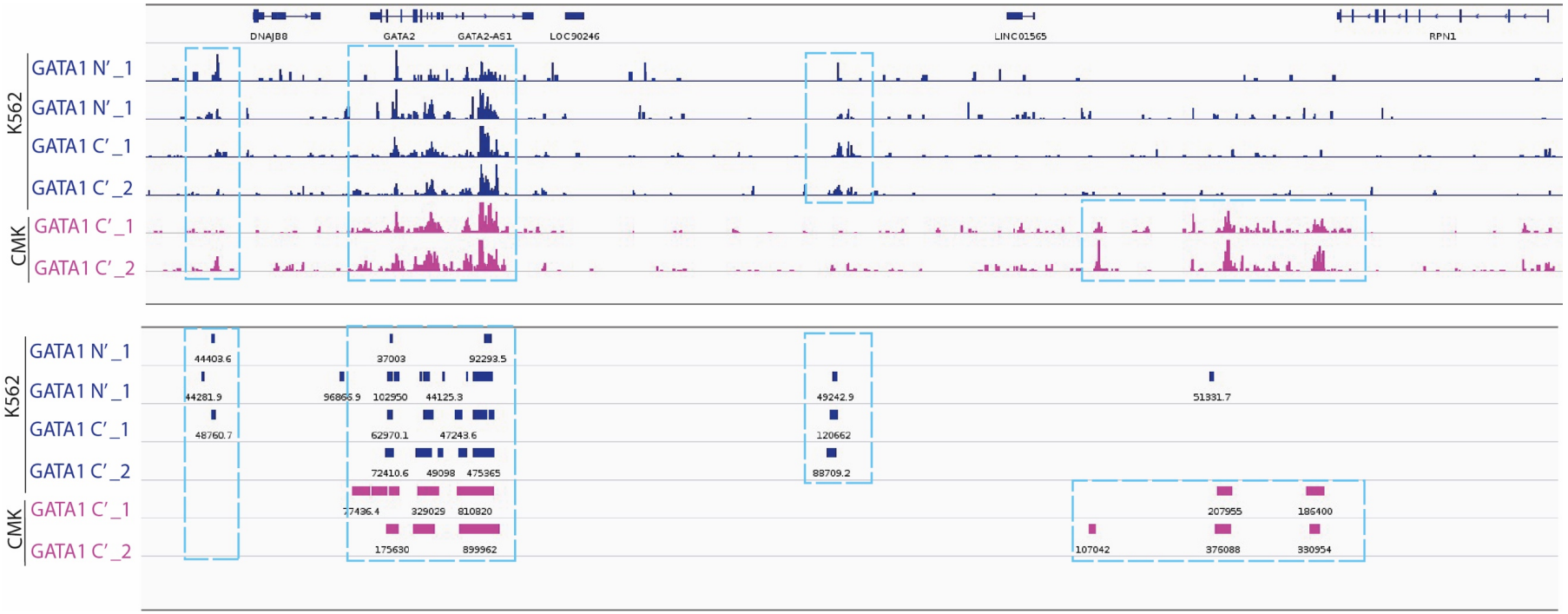


Figure 6.7 GATA1s biding at a distal GATA2 enhacer site in CMKs

CUT&TAG read density (top) and peak calling (bottom) at the GATA1 locus in K562 (blue) and CMKs (pink) cells identifies known GATA1 biding sites at the GATA2 locus (dotted squares)

7 Discussion

In the field of blood cancer and cancer in general, there are two important questions that need to be addressed: first, why a set of genetic changes can transform a specific cell type but not another? and second, how two or more genetic changes synergize to induce transformation? TAM and ML-DS are genetically simple disease models of blood cancer in which these two questions can be addressed. Like other forms of blood cancers, TAM and ML-DS involve a step wise acquisition of genetic changes that result in advantageous phenotypes such as excessive proliferation and a block in cellular maturation. Unlike adult AML, which is genetically complex, ML-DS only requires three genetic changes for leukaemia transformation: an extra copy of chromosome 21, an N' terminally truncating *GATA1* mutation, and in 50% of the cases, a single allele mutation in any member of the cohesin complex. Because *GATA1s* mutations are not leukaemogenic in the absence of trisomy 21, and germline variants in genes encoding members of cohesin complex do not lead to an increased risk of leukaemia development or other forms of cancer, it is most likely that these three genetic changes are synergistic. That is, a single genetic change leads to modest haematopoietic alterations, but when they co-occur in a precise cell type, their combinatorial effect results in leukaemia transformation. Although, the molecular mechanism behind the synergistic effect between trisomy 21, GATA1s and a dysfunctional cohesin complex in a precise cell type is still yet to be clarified; in this thesis I have prioritised the cell types where the synergy between trisomy 21 and GATA1s have the greatest impact on haematopoietic cell maturation. This is an important first step in answering questions about the mechanism of this cooperation.

A constitutional extra copy of chromosome 21 in the haematopoietic system has a relatively modest effect on its own. In fetal life, it leads to the abnormal formation of virtually all blood lineages, with the megakaryocytic and the B-cell lineages being reduced, while the erythroid lineage cells much more abundant. In molecular terms, an

extra copy of chromosome 21 in Down syndrome causes genome-wide perturbation of gene expression and no single gene can explain the phenotypes observed. Outside of trisomy 21, N' terminally truncating mutation in the *GATA1* gene do occur, but here despite being germline, they mainly affect the erythroid lineage leading to anaemia, macrocytosis, and reticulocytopenia. Germline variants in genes encoding the cohesin complex cause Cornelia de Lange syndrome (CLS) and Roberts syndrome, two genetic disorders characterized by intellectual disability and developmental abnormalities. In CLS, the only haematopoietic abnormality reported is mild thrombocytopenia and like trisomy 21, disruption of the cohesin complex leads to genome-wide perturbation of gene expression and no single gene can explain the phenotypes observed.

In what precise cell type then, trisomy 21, GATA1s and a dysfunctional cohesin complex synergize to cause ML-DS? High throughput single cell RNA sequencing can deconvolute highly heterogenous populations of haematopoietic cells, allowing to create transcriptional landscapes of inferred cellular differentiation. With this approach I have identified multiple populations with robust *GATA1* expression: EMMPPs, cycling-EMMPPs, megakaryocyte progenitors, Ba/Eo/Mast cell progenitors and a series of erythroid cells ranging from progenitors to terminally differentiated cells. When *GATA1s* mutation occurs in Trisomy 21, aberrant gene expression was observed in EMMPPs, cycling-EMMPPs, megakaryocyte and erythroid progenitors. Although erythroid and megakaryocytic cells formed in *GATA1s* mutant trisomy 21 cells, they failed to mature properly. This maturation defect of the erythroid and megakaryocytic lineage may be due to dysregulated gene expression of transcription factors such as *E2F*, *RUNX1*, *GATA2*, and *GATA1* itself. By narrowing the cell types where trisomy 21 and GATA1s have a maximal synergy, more specific questions about the role of T21 could be addressed. What are the transcriptional differences between disomic and trisomy 21 cells for each of these compartments where *GATA1s* mutation have the most drastic effects? This

question can be address once single disomic cells are captured and their transcriptome is sequenced and compared to the reference single cell T21 transcriptional landscape.

Methods to test for differential gene expression in single cells are currently being developed and optimized. Consequently, obtaining a differentially expressed gene set between disomic cells and trisomy 21 cells or between trisomy21 and GATA1s TAM cells at specific stages of haematopoietic cell maturation could readily be achieved. But this raises the question as to what genes among those that are differentially expressed in each compartment are direct targets of GATA1? A direct gene target of GATA1 could be defined as one whose transcription increases or decrease upon binding of GATA1 at one or more of its regulatory DNA elements. In this thesis I have optimized CUT&TAG to profile genome wide biding of GATA1 in low cell numbers. CUT&TAG, a method published by the Henikoff lab, has recently drawn attention because it is highly modular. It not only allows for genome wide interrogation of single protein-DNA interactions in single cells, but it also allows for interrogation of multiple proteins-DNA interactions in the same cell. Moreover, multiple laboratories such as the Fazzio lab, the Satija lab, and the Castelo-Branco group have used commercially available 10X Genomics kits to encapsulate and sequence single cells upon CUT&TAG. Although I did not have enough time to apply this technique in TAM or ML-DS cells, I envision that my preliminary optimization of CUT&TAG specifically for human GATA1 coupled to single cell encapsulation with the 10X Genomics kits will allow for genome-wide profiling of GATA1-DNA interaction in EMMPPs, cycling-EMMPPs, megakaryocyte progenitors, Ba/Eo/Mast cell progenitors and erythroid progenitors. This experimental approach should allow identification of the direct targets of GATA1 at crucial states of cellular maturation in trisomy 21 and disomic cells. This approach will also prioritize differentially expressed GATA1 direct gene targets between trisomy 21 and TAM cells to offer mechanistic insights.

It has been proposed that first mutations that lead to leukaemia must be acquired in long-lived haematopoietic stem cells. Although, preliminary data suggest that in fact *GATA1s* mutations are acquired in HSCs, here they might not have any noticeable effects as HSC do not express GATA1. What HSCs have is an intrinsic ability to pass this and other genetic perturbation to daughter cells where they can accumulate. Once a haematopoietic stem cell gives rise to the first cell where GATA1 is expressed, GATA1 will be able to remodel the epigenome by i repressive and activating complexes to precise genomic loci. In trisomy 21, I found that the first cell with robust GATA1 expression is the EMMPP or cluster 8. Several key questions follow from this: once expressed in the EMMPP cluster, what genes will GATA1 bind first? And how does GATA1 biding will change as EMMPPs progress to cycling-EMMPPs and differentiate to erythroid progenitor and megakaryocyte progenitor cells? Thus, if a *GATA1s* mutation does in fact occur in the HSC, then mutant HSC will propagate the mutation into EMMPPs, and EMMPPs into erythroid and megakaryocyte progenitors. Consequently, in TAM the epigenome of EMMPPs and thus of erythroid and megakaryocyte progenitors, will be initially remodelled by a mutant form of GATA1, GATA1s. If this were to be the case, this phenomenon raises a further question: Are the leukaemogenic consequences of a *GATA1s* mutation different if it occurs in HSCs or if they occur in a cell type where GATA1 wild type has already remodelled the epigenome?

In this thesis I presented and optimized CRISPR Cas9 system to recreate *GATA1s* mutations in primary haematopoietic stem and progenitor cells. This showed that the erythroid maturation defects identified by single cell RNA sequencing of TAM samples can be recapitulated *in vitro*. I also generated a detailed single cell transcriptional landscape of trisomy 21 haematopoiesis that I have used to map single TAM cells and could be used to map single ML-DS cells to identify at which state of cellular maturation ML-DS cells are blocked. It can also be used to mapped single cells taken at different time points of an *in vitro* differentiation assay starting from gene edited HSCs. Lastly, I

showed that robust GATA1 expression at the protein level coincides with gain of CD38 expression in CD34+ HSPCs and GATA1 expression can be shown in CMPs, and MEPs but not in lympho/myeloid CD45RA+ cells. I envision that the use of these tools, intracellular detection of GATA1 and gene editing of HSPCs to recreate *GATA1s* mutations coupled to the analysis chosen for single cell RNA sequencing mapping to known single cell transcriptional landscapes of haematopoiesis will prove useful in answering why a set of genetic changes can transform a specific cell type but not another? and, how two or more genetic changes synergize to induce transformation?

Finally, in the field of AML, the cell population of leukaemic transformation is regarded as the cell type in which complete maturation arrest has occurred. Like in adult AML in which leukaemia stem cells are arrested at a similar stage to the LMPP[35], it is plausible that in ML-DS, leukaemia stem cells could be arrested at an EMMPP like stage or cluster 8. Limited evidenced for this was recently presented in which 6 ML-DS patients harboured driver mutations that changed alanine 455 to aspartate (A455D) in the *CSF2RB* gene[10]. The *CSF2RB* gene encodes the beta common subunit for the IL3, IL5 and the GM-CSF cytokines receptors, it is also known as CD131, which is over expressed in ML-DS[174]. Importantly, onset of CD131 expression marks a purifiable GATA1-expressing progenitor cell population within the CMP compartment with megakaryocyte, erythroid, mast cell, and basophil potential[36]. It is highly likely that CD131+ CMPs or EMPPs are the population of cells referred here as EMMPPs or cluster 8 or CD131 expression in CMPs enriches for EMMPPs. Reference mapping of cohesin mutant single ML-DS cells with Symphony as well as variant calling of *GATA1s* mutations with VarTrix will determine if ML-DS cells are arrested at the EMMPPs like stage and will allow the identification of a population of cells where there is maximal synergy between trisomy 21, GATA1s and a dysfunctional cohesin complex. This in turn will allow the mechanism of this interaction to be dissected and, potentially, specifically targeted.

8 Conclusion

In this thesis I have presented an optimized intracellular flow cytometric (iFC) protocol for the detection of TAM and ML-DS cells with exclusive GATA1s expression as well as for the identification of wild type cells expressing both GATA1 isoforms. When target next generation is not available, iFC may help diagnose TAM and ML-DS, and it is much more specific than conventional cell surface flow cytometry. iFC can be used to monitor shrinkage or expansion of the TAM/ML-DS clone during the early months after birth. I hope that one day the intracellular detection of GATA1 isoforms can server as a clinically useful tool to monitor measurable residual disease to predict ML-DS infants at risk of relapse after chemotherapy and so improve their long-term outlook.

9 References

1. Lange BJ, Kobrinsky N, Barnard DR, et al. Distinctive demography, biology, and outcome of acute myeloid leukemia and myelodysplastic syndrome in children with Down syndrome: Children's Cancer Group Studies 2861 and 2891. *Blood*. 1998;91(2):608-615.
2. Hasle H, Friedman JM, Olsen JH, Rasmussen SA. Low risk of solid tumors in persons with Down syndrome. *Genet Med*. 2016;18(11):1151-1157.
3. Wechsler J, Greene M, McDevitt MA, et al. Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. *Nat Genet*. 2002;32(1):148-152.
4. Ahmed M, Sternberg A, Hall G, et al. Natural history of GATA1 mutations in Down syndrome. *Blood*. 2004;103(7):2480-2489.
5. Mundschau G, Gurbuxani S, Gamis AS, Greene ME, Arceci RJ, Crispino JD. Mutagenesis of GATA1 is an initiating event in Down syndrome leukemogenesis. *Blood*. 2003;101(11):4298-4300.
6. Rainis L, Bercovich D, Strehl S, et al. Mutations in exon 2 of GATA1 are early events in megakaryocytic malignancies associated with trisomy 21. *Blood*. 2003;102(3):981-986.
7. Roberts I, Alford K, Hall G, et al. GATA1-mutant clones are frequent and often unsuspected in babies with Down syndrome: identification of a population at risk of leukemia. *Blood*. 2013;122(24):3908-3917.
8. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. *Science*. 2019;366(6465).
9. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med*. 2016;374(23):2209-2221.
10. Labuhn M, Perkins K, Matzk S, et al. Mechanisms of Progression of Myeloid Preleukemia to Transformed Myeloid Leukemia in Children with Down Syndrome. *Cancer Cell*. 2019;36(2):123-138.e110.
11. Tenen DG. Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer*. 2003;3(2):89-101.
12. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip Rev Syst Biol Med*. 2010;2(6):640-653.
13. Ogawa M. Differentiation and proliferation of hematopoietic stem cells. *Blood*. 1993;81(11):2844-2853.
14. Rosenbauer F, Tenen DG. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat Rev Immunol*. 2007;7(2):105-117.
15. Takei H, Kobayashi SS. Targeting transcription factors in acute myeloid leukemia. *Int J Hematol*. 2019;109(1):28-34.
16. Orkin SH. Diversification of haematopoietic stem cells to specific lineages. *Nature Reviews Genetics*. 2000;1(1):57-64.
17. Porcher C, Swat W, Rockwell K, Fujiwara Y, Alt FW, Orkin SH. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell*. 1996;86(1):47-57.
18. Shivdasani RA, Mayer EL, Orkin SH. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*. 1995;373(6513):432-434.
19. Okuda T, van Deursen J, Hiebert SW, Grosveld G, Downing JR. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell*. 1996;84(2):321-330.
20. Tsai F-Y, Keller G, Kuo FC, et al. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature*. 1994;371(6494):221-226.
21. Palii CG, Cheng Q, Gillespie MA, et al. Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate. *Cell Stem Cell*. 2019;24(5):812-820.e815.
22. Hock H, Hamblen MJ, Rooke HM, et al. Intrinsic requirement for zinc finger transcription factor Gfi-1 in neutrophil differentiation. *Immunity*. 2003;18(1):109-120.

23. Frelin C, Herrington R, Janmohamed S, et al. GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat Immunol*. 2013;14(10):1037-1044.
24. Tindemans I, Serafini N, Di Santo JP, Hendriks RW. GATA-3 function in innate and adaptive immunity. *Immunity*. 2014;41(2):191-206.
25. Notta F, Doulatov S, Laurenti E, Poeppl A, Jurisica I, Dick JE. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science*. 2011;333(6039):218-221.
26. Majeti R, Park CY, Weissman IL. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell*. 2007;1(6):635-645.
27. Notta F, Zandi S, Takayama N, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*. 2016;351(6269):aab2116.
28. Romeo PH, Prandini MH, Joulin V, et al. Megakaryocytic and erythrocytic lineages share specific transcription factors. *Nature*. 1990;344(6265):447-449.
29. Martin DI, Zon LI, Mutter G, Orkin SH. Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. *Nature*. 1990;344(6265):444-447.
30. Zon LI, Yamaguchi Y, Yee K, et al. Expression of mRNA for the GATA-binding proteins in human eosinophils and basophils: potential role in gene transcription. *Blood*. 1993;81(12):3234-3241.
31. Krause DS, Fackler MJ, Civin CI, May WS. CD34: structure, biology, and clinical utility. *Blood*. 1996;87(1):1-13.
32. Sonoda Y. Human CD34-negative hematopoietic stem cells: The current understanding of their biological nature. *Exp Hematol*. 2021;96:13-26.
33. Manz MG, Miyamoto T, Akashi K, Weissman IL. Prospective isolation of human clonogenic common myeloid progenitors. *Proc Natl Acad Sci U S A*. 2002;99(18):11872-11877.
34. Xavier-Ferrucio J, Krause DS. Concise Review: Bipotent Megakaryocytic-Erythroid Progenitors: Concepts and Controversies. *Stem Cells*. 2018;36(8):1138-1145.
35. Goardon N, Marchi E, Atzberger A, et al. Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell*. 2011;19(1):138-152.
36. Drissen R, Thongjuea S, Theilgaard-Mönch K, Nerlov C. Identification of two distinct pathways of human myelopoiesis. *Sci Immunol*. 2019;4(35).
37. Psaila B, Barkas N, Iskander D, et al. Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol*. 2016;17:83.
38. Karamitros D, Stoilova B, Aboukhalil Z, et al. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat Immunol*. 2018;19(1):85-97.
39. Mori Y, Iwasaki H, Kohno K, et al. Identification of the human eosinophil lineage-committed progenitor: revision of phenotypic definition of the human common myeloid progenitor. *J Exp Med*. 2009;206(1):183-193.
40. Miyawaki K, Iwasaki H, Jiromaru T, et al. Identification of unipotent megakaryocyte progenitors in human hematopoiesis. *Blood*. 2017;129(25):3332-3343.
41. Sanada C, Xavier-Ferrucio J, Lu YC, et al. Adult human megakaryocyte-erythroid progenitors are in the CD34+CD38mid fraction. *Blood*. 2016;128(7):923-933.
42. Popescu DM, Botting RA, Stephenson E, et al. Decoding human fetal liver haematopoiesis. *Nature*. 2019;574(7778):365-371.
43. Tusi BK, Wolock SL, Weinreb C, et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*. 2018;555(7694):54-60.
44. Ranzoni AM, Tangherloni A, Berest I, et al. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell*. 2021;28(3):472-487.e477.
45. Qin P, Pang Y, Hou W, et al. Integrated decoding hematopoiesis and leukemogenesis using single-cell sequencing and its medical implication. *Cell Discov*. 2021;7(1):2.

46. Kulessa H, Frampton J, Graf T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev*. 1995;9(10):1250-1262.
47. Pevny L, Simon MC, Robertson E, et al. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*. 1991;349(6306):257-260.
48. Fujiwara Y, Browne CP, Cunniff K, Goff SC, Orkin SH. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc Natl Acad Sci U S A*. 1996;93(22):12355-12358.
49. Vyas P, Ault K, Jackson CW, Orkin SH, Shivdasani RA. Consequences of GATA-1 deficiency in megakaryocytes and platelets. *Blood*. 1999;93(9):2867-2875.
50. Majewski IJ, Metcalf D, Mielke LA, et al. A mutation in the translation initiation codon of Gata-1 disrupts megakaryocyte maturation and causes thrombocytopenia. *Proc Natl Acad Sci U S A*. 2006;103(38):14146-14151.
51. Shivdasani RA, Fujiwara Y, McDevitt MA, Orkin SH. A lineage-selective knockout establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet development. *Embo j*. 1997;16(13):3965-3973.
52. Migliaccio AR, Rana RA, Sanchez M, et al. GATA-1 as a regulator of mast cell differentiation revealed by the phenotype of the GATA-1low mouse mutant. *J Exp Med*. 2003;197(3):281-296.
53. Ohneda K, Moriguchi T, Ohmori S, et al. Transcription factor GATA1 is dispensable for mast cell differentiation in adult mice. *Mol Cell Biol*. 2014;34(10):1812-1826.
54. Yu C, Cantor AB, Yang H, et al. Targeted deletion of a high-affinity GATA-binding site in the GATA-1 promoter leads to selective loss of the eosinophil lineage in vivo. *J Exp Med*. 2002;195(11):1387-1395.
55. Nei Y, Obata-Ninomiya K, Tsutsui H, et al. GATA-1 regulates the generation and function of basophils. *Proc Natl Acad Sci U S A*. 2013;110(46):18620-18625.
56. Yu M, Riva L, Xie H, et al. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell*. 2009;36(4):682-695.
57. Evans T, Reitman M, Felsenfeld G. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc Natl Acad Sci U S A*. 1988;85(16):5976-5980.
58. Tsai SF, Martin DI, Zon LI, D'Andrea AD, Wong GG, Orkin SH. Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature*. 1989;339(6224):446-451.
59. Ito E, Toki T, Ishihara H, et al. Erythroid transcription factor GATA-1 is abundantly transcribed in mouse testis. *Nature*. 1993;362(6419):466-468.
60. Halsey C, Tunstall O, Gibson B, Roberts I, Graham G. Role of GATA-1s in early hematopoiesis and differences between alternative splicing in human and murine GATA-1. *Blood*. 2010;115(16):3415-3416.
61. Calligaris R, Bottardi S, Cogoi S, Apezteguia I, Santoro C. Alternative translation initiation site usage results in two functionally distinct forms of the GATA-1 transcription factor. *Proc Natl Acad Sci U S A*. 1995;92(25):11598-11602.
62. López AJ. Developmental role of transcription factor isoforms generated by alternative splicing. *Dev Biol*. 1995;172(2):396-411.
63. Martin DI, Orkin SH. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev*. 1990;4(11):1886-1898.
64. Kaneko H, Kobayashi E, Yamamoto M, Shimizu R. N- and C-terminal transactivation domains of GATA1 protein coordinate hematopoietic program. *J Biol Chem*. 2012;287(25):21439-21449.
65. Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*. 2016;44(W1):W344-350.

66. Whyatt DJ, deBoer E, Grosveld F. The two zinc finger-like domains of GATA-1 have different DNA binding specificities. *Embo j*. 1993;12(13):4993-5005.
67. Omichinski JG, Clore GM, Schaad O, et al. NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science*. 1993;261(5120):438-446.
68. Omichinski JG, Trainor C, Evans T, Gronenborn AM, Clore GM, Felsenfeld G. A small single-"finger" peptide from the erythroid transcription factor GATA-1 binds specifically to DNA as a zinc or iron complex. *Proc Natl Acad Sci U S A*. 1993;90(5):1676-1680.
69. Trainor CD, Ghirlando R, Simpson MA. GATA zinc finger interactions modulate DNA binding and transactivation. *J Biol Chem*. 2000;275(36):28157-28166.
70. Tsang AP, Visvader JE, Turner CA, et al. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell*. 1997;90(1):109-119.
71. Ling T, Crispino JD. GATA1 mutations in red cell disorders. *IUBMB Life*. 2020;72(1):106-118.
72. Crispino JD, Horwitz MS. GATA factor mutations in hematologic disease. *Blood*. 2017;129(15):2103-2110.
73. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*. 1989;245(4916):371-378.
74. Sigler PB. Transcriptional activation. Acid blobs and negative noodles. *Nature*. 1988;333(6170):210-212.
75. Cress WD, Triezenberg SJ. Critical structural elements of the VP16 transcriptional activation domain. *Science*. 1991;251(4989):87-90.
76. Staller MV, Holehouse AS, Swain-Lenz D, Das RK, Pappu RV, Cohen BA. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst*. 2018;6(4):444-455.e446.
77. Staller MV, Ramirez E, Holehouse AS, Pappu RV, Cohen BA. Design principles of acidic transcriptional activation domains. *bioRxiv*. 2021:2020.2010.2028.359026.
78. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res*. 2018;46(W1):W329-w337.
79. Kadri Z, Shimizu R, Ohneda O, et al. Direct binding of pRb/E2F-2 to GATA-1 regulates maturation and terminal cell division during erythropoiesis. *PLoS Biol*. 2009;7(6):e1000123.
80. Marlow EC, Ducore J, Kwan ML, et al. Leukemia Risk in a Cohort of 3.9 Million Children with and without Down Syndrome. *J Pediatr*. 2021;234:172-180.e173.
81. Langebrake C, Creutzig U, Reinhardt D. Immunophenotype of Down syndrome acute myeloid leukemia and transient myeloproliferative disease differs significantly from other diseases with morphologically identical or similar blasts. *Klinische Padiatrie*. 2005;217(3):126-134.
82. Girodon F, Favre B, Couillaud G, Carli PM, Parmeland C, Maynadié M. Immunophenotype of a transient myeloproliferative disorder in a newborn with trisomy 21. *Cytometry*. 2000;42(2):118-122.
83. Karandikar NJ, Aquino DB, McKenna RW, Kroft SH. Transient myeloproliferative disorder and acute myeloid leukemia in Down syndrome. An immunophenotypic analysis. *American journal of clinical pathology*. 2001;116(2):204-210.
84. Antonarakis SE, Skotko BG, Rafii MS, et al. Down syndrome. *Nature Reviews Disease Primers*. 2020;6(1):9.
85. Korbel JO, Tirosh-Wagner T, Urban AE, et al. The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proc Natl Acad Sci U S A*. 2009;106(29):12031-12036.
86. Webb D, Roberts I, Vyas P. Haematology of Down syndrome. *Archives of Disease in Childhood - Fetal and Neonatal Edition*. 2007;92(6):F503-F507.

87. Henry E, Walker D, Wiedmeier SE, Christensen RD. Hematological abnormalities during the first week of life among neonates with Down syndrome: data from a multihospital healthcare system. *Am J Med Genet A*. 2007;143a(1):42-50.
88. Jardine L, Webb S, Goh I, et al. Intrinsic and extrinsic regulation of human fetal bone marrow haematopoiesis and perturbations in Down syndrome. *bioRxiv*. 2021:2021.2006.2025.449771.
89. Chou ST, Byrska-Bishop M, Tober JM, et al. Trisomy 21-associated defects in human primitive hematopoiesis revealed through induced pluripotent stem cells. *Proc Natl Acad Sci U S A*. 2012;109(43):17573-17578.
90. Maclean GA, Menne TF, Guo G, et al. Altered hematopoiesis in trisomy 21 as revealed through in vitro differentiation of isogenic human pluripotent cells. *Proc Natl Acad Sci U S A*. 2012;109(43):17567-17572.
91. Roy A, Cowan G, Mead AJ, et al. Perturbation of fetal liver hematopoietic stem and progenitor cell development by trisomy 21. *Proc Natl Acad Sci U S A*. 2012;109(43):17579-17584.
92. Hasaart KAL, Manders F, van der Hoorn ML, et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. *Sci Rep*. 2020;10(1):12991.
93. Sankaran VG, Ghazvinian R, Do R, et al. Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J Clin Invest*. 2012;122(7):2439-2443.
94. Ludwig LS, Gazda HT, Eng JC, et al. Altered translation of GATA1 in Diamond-Blackfan anemia. *Nat Med*. 2014;20(7):748-753.
95. Parrella S, Aspesi A, Quarello P, et al. Loss of GATA-1 full length as a cause of Diamond-Blackfan anemia phenotype. *Pediatr Blood Cancer*. 2014;61(7):1319-1321.
96. Klar J, Khalfallah A, Arzoo PS, Gazda HT, Dahl N. Recurrent GATA1 mutations in Diamond-Blackfan anaemia. *Br J Haematol*. 2014;166(6):949-951.
97. Li Z, Godinho FJ, Klusmann JH, Garriga-Canut M, Yu C, Orkin SH. Developmental stage-selective effect of somatically mutated leukemogenic transcription factor GATA1. *Nat Genet*. 2005;37(6):613-619.
98. Birger Y, Goldberg L, Chlon TM, et al. Perturbation of fetal hematopoiesis in a mouse model of Down syndrome's transient myeloproliferative disorder. *Blood*. 2013;122(6):988-998.
99. Ling T, Birger Y, Stankiewicz MJ, et al. Chromatin occupancy and epigenetic analysis reveal new insights into the function of the GATA1 N terminus in erythropoiesis. *Blood*. 2019;134(19):1619-1631.
100. Byrska-Bishop M, VanDorn D, Campbell AE, et al. Pluripotent stem cells reveal erythroid-specific activities of the GATA1 N-terminus. *J Clin Invest*. 2015;125(3):993-1005.
101. DeVilbiss AW, Sanalkumar R, Hall BD, Katsumura KR, de Andrade IF, Bresnick EH. Epigenetic Determinants of Erythropoiesis: Role of the Histone Methyltransferase SetD8 in Promoting Erythroid Cell Maturation and Survival. *Mol Cell Biol*. 2015;35(12):2073-2087.
102. Malik J, Getman M, Steiner LA. Histone methyltransferase Setd8 represses Gata2 expression and regulates erythroid maturation. *Mol Cell Biol*. 2015;35(12):2059-2072.
103. Vicente C, Vazquez I, Conchillo A, et al. Overexpression of GATA2 predicts an adverse prognosis for patients with acute myeloid leukemia and it is associated with distinct molecular abnormalities. *Leukemia*. 2012;26(3):550-554.
104. Briegel K, Lim KC, Plank C, Beug H, Engel JD, Zenke M. Ectopic expression of a conditional GATA-2/estrogen receptor chimera arrests erythroid differentiation in a hormone-dependent manner. *Genes Dev*. 1993;7(6):1097-1109.
105. Ikonomi P, Rivera CE, Riordan M, Washington G, Schechter AN, Noguchi CT. Overexpression of GATA-2 inhibits erythroid and promotes megakaryocyte differentiation. *Exp Hematol*. 2000;28(12):1423-1431.

106. Kuvardina ON, Herglotz J, Kolodziej S, et al. RUNX1 represses the erythroid gene expression program during megakaryocytic differentiation. *Blood*. 2015;125(23):3570-3579.
107. Nikolaev SI, Santoni F, Vannier A, et al. Exome sequencing identifies putative drivers of progression of transient myeloproliferative disorder to AMKL in infants with Down syndrome. *Blood*. 2013;122(4):554-561.
108. Yoshida K, Toki T, Okuno Y, et al. The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nat Genet*. 2013;45(11):1293-1299.
109. Wagenblast E, Araújo J, Gan OI, et al. Mapping the cellular origin and early evolution of leukemia in Down syndrome. *Science*. 2021;373(6551).
110. Grimm J, Heckl D, Klusmann JH. Molecular Mechanisms of the Genetic Predisposition to Acute Megakaryoblastic Leukemia in Infants With Down Syndrome. *Front Oncol*. 2021;11:636633.
111. Kaya-Okur HS, Wu SJ, Codomo CA, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*. 2019;10(1):1930.
112. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048.
113. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359.
114. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
115. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842.
116. Meers MP, Tenenbaum D, Henikoff S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin*. 2019;12(1):42.
117. Yu F, Sankaran VG, Yuan GC. CUT&RUNTools 2.0: A pipeline for single-cell and bulk-level CUT&RUN and CUT&Tag data analysis. *Bioinformatics*. 2021.
118. Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
119. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e3529.
120. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst*. 2019;8(4):329-337.e324.
121. Hippen AA, Falco MM, Weber LM, et al. miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. *bioRxiv*. 2021:2021.2003.2003.433798.
122. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289-1296.
123. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2018.
124. Korsunsky I, Nathan A, Millard N, Raychaudhuri S. Presto scales Wilcoxon and auROC analyses to millions of observations. *bioRxiv*. 2019:653253.
125. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. *bioRxiv*. 2021:060012.
126. Roy A, Wang G, Iskander D, et al. Developmental stage- and site-specific transitions in lineage specification and gene regulatory networks in human hematopoietic stem and progenitor cells. *bioRxiv*. 2021:2021.2004.2020.440420.
127. Andreatta M, Carmona SJ. UCell: Robust and scalable single-cell gene signature scoring. *Comput Struct Biotechnol J*. 2021;19:3796-3798.
128. Kang JB, Nathan A, Zhang F, et al. Efficient and precise single-cell reference atlas mapping with Symphony. *bioRxiv*. 2021:2020.2011.2018.389189.
129. Bull MJ. Down Syndrome. *N Engl J Med*. 2020;382(24):2344-2352.
130. Gamis AS, Alonzo TA, Gerbing RB, et al. Natural history of transient myeloproliferative disorder clinically diagnosed in Down syndrome neonates: a report

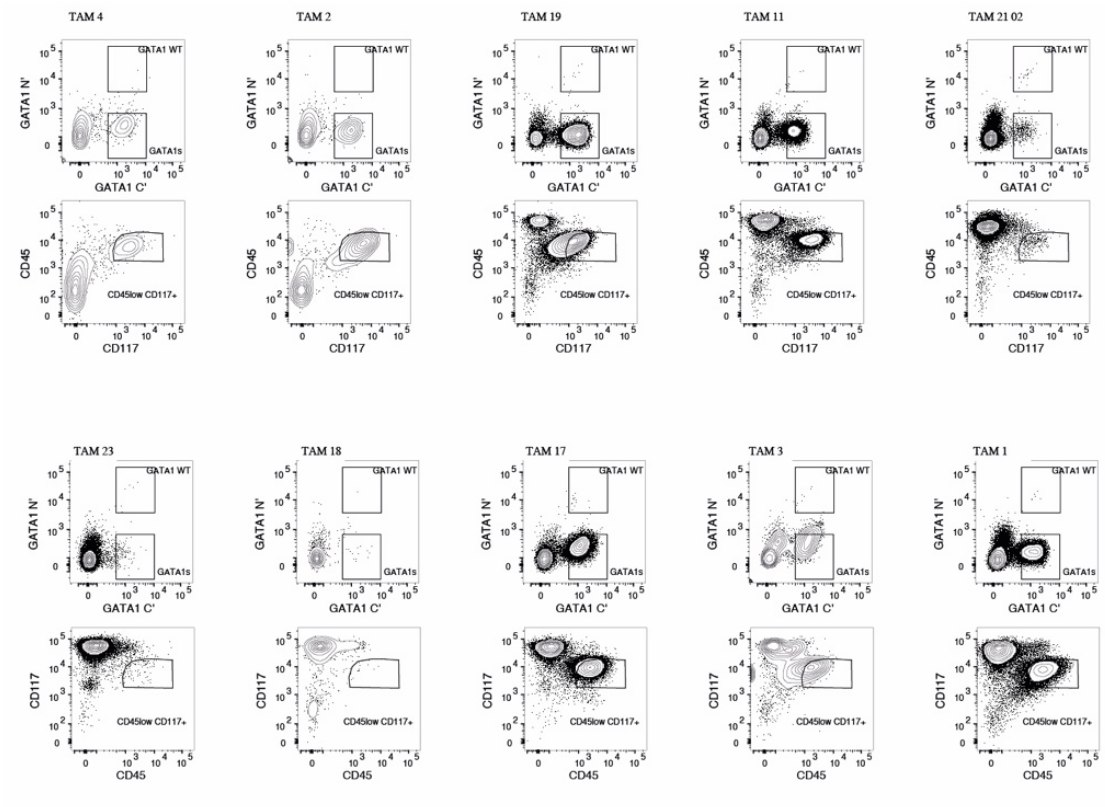
from the Children's Oncology Group Study A2971. *Blood*. 2011;118(26):6752-6759; quiz 6996.
131. Boztug H, Schumich A, Pötschger U, et al. Blast cell deficiency of CD11a as a marker of acute megakaryoblastic leukemia and transient myeloproliferative disease in children with and without Down syndrome. *Cytometry Part B, Clinical cytometry*. 2013;84(6):370-378.
132. Roberts I, Alford K, Hall G, et al. GATA1-mutant clones are frequent and often unsuspected in babies with Down syndrome: identification of a population at risk of leukemia. *Blood*. 2013;122(24):3908-3917.
133. Goemans BF, Noort S, Blink M, et al. Sensitive GATA1 mutation screening reliably identifies neonates with Down syndrome at risk for myeloid leukemia. *Leukemia*. 2021;35(8):2403-2406.
134. Bacher U, Shumilov E, Flach J, et al. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. *Blood cancer journal*. 2018;8(11):113-113.
135. Litzenburger UM, Buenrostro JD, Wu B, et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol*. 2017;18(1):15.
136. Lee WY, Weinberg OK, Evans AG, Pinkus GS. Loss of Full-Length GATA1 Expression in Megakaryocytes Is a Sensitive and Specific Immunohistochemical Marker for the Diagnosis of Myeloid Proliferative Disorder Related to Down Syndrome. *American journal of clinical pathology*. 2018;149(4):300-309.
137. Zola H. High-sensitivity immunofluorescence/flow cytometry: detection of cytokine receptors and other low-abundance membrane molecules. *Curr Protoc Cytom*. 2004;Chapter 6:Unit 6.3.
138. Gillespie MA, Palii CG, Sanchez-Taltavull D, et al. Absolute Quantification of Transcription Factors Reveals Principles of Gene Regulation in Erythropoiesis. *Mol Cell*. 2020;78(5):960-974.e911.
139. Jaso JM, Wang SA, Jorgensen JL, Lin P. Multi-color flow cytometric immunophenotyping for detection of minimal residual disease in AML: past, present and future. *Bone Marrow Transplant*. 2014;49(9):1129-1138.
140. Sandes AF, Kerbauy DM, Matarraz S, et al. Combined flow cytometric assessment of CD45, HLA-DR, CD34, and CD117 expression is a useful approach for reliable quantification of blast cells in myelodysplastic syndromes. *Cytometry B Clin Cytom*. 2013;84(3):157-166.
141. Creutzig U, Reinhardt D, Diekamp S, Dworzak M, Stary J, Zimmermann M. AML patients with Down syndrome have a high cure rate with AML-BFM therapy with reduced dose intensity. *Leukemia*. 2005;19(8):1355-1360.
142. Taub JW, Matherly LH, Stout ML, Buck SA, Gurney JG, Ravindranath Y. Enhanced metabolism of 1-beta-D-arabinofuranosylcytosine in Down syndrome cells: a contributing factor to the superior event free survival of Down syndrome children with acute myeloid leukemia. *Blood*. 1996;87(8):3395-3403.
143. Uffmann M, Rasche M, Zimmermann M, et al. Therapy reduction in patients with Down syndrome and myeloid leukemia: the international ML-DS 2006 trial. *Blood*. 2017;129(25):3314-3321.
144. Taub JW, Berman JN, Hitzler JK, et al. Improved outcomes for myeloid leukemia of Down syndrome: a report from the Children's Oncology Group AAML0431 trial. *Blood*. 2017;129(25):3304-3313.
145. Schuurhuis GJ, Heuser M, Freeman S, et al. Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. *Blood*. 2018;131(12):1275-1291.
146. Alford KA, Reinhardt K, Garnett C, et al. Analysis of GATA1 mutations in Down syndrome transient myeloproliferative disorder and myeloid leukemia. *Blood*. 2011;118(8):2222-2238.

147. Pine SR, Guo Q, Yin C, et al. GATA1 as a new target to detect minimal residual disease in both transient leukemia and megakaryoblastic leukemia of Down syndrome. *Leukemia research*. 2005;29(11):1353-1356.
148. Flasinski M, Scheibke K, Zimmermann M, et al. Low-dose cytarabine to prevent myeloid leukemia in children with Down syndrome: TMD Prevention 2007 study. *Blood Adv*. 2018;2(13):1532-1540.
149. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315-1316.
150. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
151. Bresnick EH, Lee HY, Fujiwara T, Johnson KD, Keles S. GATA switches as developmental drivers. *J Biol Chem*. 2010;285(41):31087-31093.
152. Alquicira-Hernandez J, Powell JE. Nebulosa recovers single cell gene expression signals by kernel density estimation. *Bioinformatics*. 2021.
153. Park CS, Lewis A, Chen T, Lacorazza D. Concise Review: Regulation of Self-Renewal in Normal and Malignant Hematopoietic Stem Cells by Krüppel-Like Factor 4. *Stem Cells Transl Med*. 2019;8(6):568-574.
154. Kikushige Y, Yoshimoto G, Miyamoto T, et al. Human Flt3 is expressed at the hematopoietic stem cell and the granulocyte/macrophage progenitor stages to maintain cell survival. *J Immunol*. 2008;180(11):7358-7367.
155. Lehnertz B, Chagraoui J, MacRae T, et al. HLF Expression Defines the Human Hematopoietic Stem Cell State. *Blood*. 2021.
156. Fares I, Chagraoui J, Lehnertz B, et al. EPCR expression marks UM171-expanded CD34(+) cord blood stem cells. *Blood*. 2017;129(25):3344-3351.
157. Amanatiadou EP, Papadopoulos GL, Strouboulis J, Vizirianakis IS. GATA1 and PU.1 Bind to Ribosomal Protein Genes in Erythroid Cells: Implications for Ribosomopathies. *PLoS One*. 2015;10(10):e0140077.
158. Krejci O, Wunderlich M, Geiger H, et al. p53 signaling in response to increased DNA damage sensitizes AML1-ETO cells to stress-induced death. *Blood*. 2008;111(4):2190-2199.
159. Cardin S, Bilodeau M, Roussy M, et al. Human models of NUP98-KDM5A megakaryocytic leukemia in mice contribute to uncovering new biomarkers and therapeutic vulnerabilities. *Blood Adv*. 2019;3(21):3307-3321.
160. Vegi NM, Klappacher J, Oswald F, et al. MEIS2 Is an Oncogenic Partner in AML1-ETO-Positive AML. *Cell Rep*. 2016;16(2):498-507.
161. Lai CK, Norddahl GL, Maetzig T, et al. Meis2 as a critical player in MN1-induced leukemia. *Blood Cancer J*. 2017;7(9):e613.
162. Chlon TM, McNulty M, Goldenson B, Rosinski A, Crispino JD. Global transcriptome and chromatin occupancy analysis reveal the short isoform of GATA1 is deficient for erythroid specification and gene expression. *Haematologica*. 2015;100(5):575-584.
163. Mak KM, Mei R. Basement Membrane Type IV Collagen and Laminin: An Overview of Their Biology and Value as Fibrosis Biomarkers of Liver Disease. *Anat Rec (Hoboken)*. 2017;300(8):1371-1390.
164. Kiyan Y, Tkachuk S, Kurselis K, et al. Heparanase-2 protects from LPS-mediated endothelial injury by inhibiting TLR4 signalling. *Sci Rep*. 2019;9(1):13591.
165. Bhatnagar N, Nizery L, Tunstall O, Vyas P, Roberts I. Transient Abnormal Myelopoiesis and AML in Down Syndrome: an Update. *Curr Hematol Malig Rep*. 2016;11(5):333-341.
166. Pimkin M, Kossenkov AV, Mishra T, et al. Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res*. 2014;24(12):1932-1944.
167. Campbell RA, Schwertz H, Hottz ED, et al. Human megakaryocytes possess intrinsic antiviral immunity through regulated induction of IFITM3. *Blood*. 2019;133(19):2013-2026.

168. Gangaraju R, Song J, Kim SJ, et al. Thrombotic, inflammatory, and HIF-regulated genes and thrombosis risk in polycythemia vera and essential thrombocythemia. *Blood Adv*. 2020;4(6):1115-1130.
169. Wang D, Zhang Z, Cui S, et al. ELF4 facilitates innate host defenses against Plasmodium by activating transcription of Pf4 and Ppbp. *J Biol Chem*. 2019;294(19):7787-7796.
170. Jalagadugula G, Mao G, Kaur G, Goldfinger LE, Dhanasekaran DN, Rao AK. Regulation of platelet myosin light chain (MYL9) by RUNX1: implications for thrombocytopenia and platelet dysfunction in RUNX1 haplodeficiency. *Blood*. 2010;116(26):6037-6045.
171. Glembotsky AC, Sliwa D, Bluteau D, et al. Downregulation of TREM-like transcript-1 and collagen receptor α2 subunit, two novel RUNX1-targets, contributes to platelet dysfunction in familial platelet disorder with predisposition to acute myelogenous leukemia. *Haematologica*. 2019;104(6):1244-1255.
172. Juban G, Sakakini N, Chagraoui H, et al. Oncogenic Gata1 causes stage-specific megakaryocyte differentiation delay. *Haematologica*. 2021;106(4):1106-1119.
173. Grass JA, Jing H, Kim SI, et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol*. 2006;26(19):7056-7067.
174. Bourquin JP, Subramanian A, Langebrake C, et al. Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proc Natl Acad Sci U S A*. 2006;103(9):3339-3344.
175. Gustafsson C, De Paepe A, Schmidl C, Månsson R. High-throughput ChIPmentation: freely scalable, single day ChIPseq data generation from very low cell-numbers. *BMC Genomics*. 2019;20(1):59.
176. Policastro RA, Zentner GE. Enzymatic methods for genome-wide profiling of protein binding sites. *Brief Funct Genomics*. 2018;17(2):138-145.
177. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol*. 2021;39(7):825-835.
178. Zhang B, Srivastava A, Mimitou E, et al. Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *bioRxiv*. 2021:2021.2009.2013.460120.
179. Gopalan S, Wang Y, Harper NW, Garber M, Fazzio TG. Simultaneous profiling of multiple chromatin proteins in the same cells. *Mol Cell*. 2021.
180. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods*. 2017;14(10):959-962.
181. Gutiérrez L, Caballero N, Fernández-Calleja L, Karkoulia E, Strouboulis J. Regulation of GATA1 levels in erythropoiesis. *IUBMB Life*. 2020;72(1):89-105.
182. Valverde-Garduno V, Guyot B, Anguita E, Hamlett I, Porcher C, Vyas P. Differences in the chromatin structure and cis-element organization of the human and mouse GATA1 loci: implications for cis-element identification. *Blood*. 2004;104(10):3106-3116.
183. Fujiwara T, O'Geen H, Keles S, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell*. 2009;36(4):667-681.
184. Bresnick EH, Johnson KD. Blood disease-causing and -suppressing transcriptional enhancers: general principles and GATA2 mechanisms. *Blood Adv*. 2019;3(13):2045-2056.
185. Johnson KD, Hsu AP, Ryu MJ, et al. Cis-element mutated in GATA2-dependent immunodeficiency governs hematopoiesis and vascular integrity. *J Clin Invest*. 2012;122(10):3692-3704.
186. Grass JA, Boyer ME, Pal S, Wu J, Weiss MJ, Bresnick EH. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc Natl Acad Sci U S A*. 2003;100(15):8811-8816.

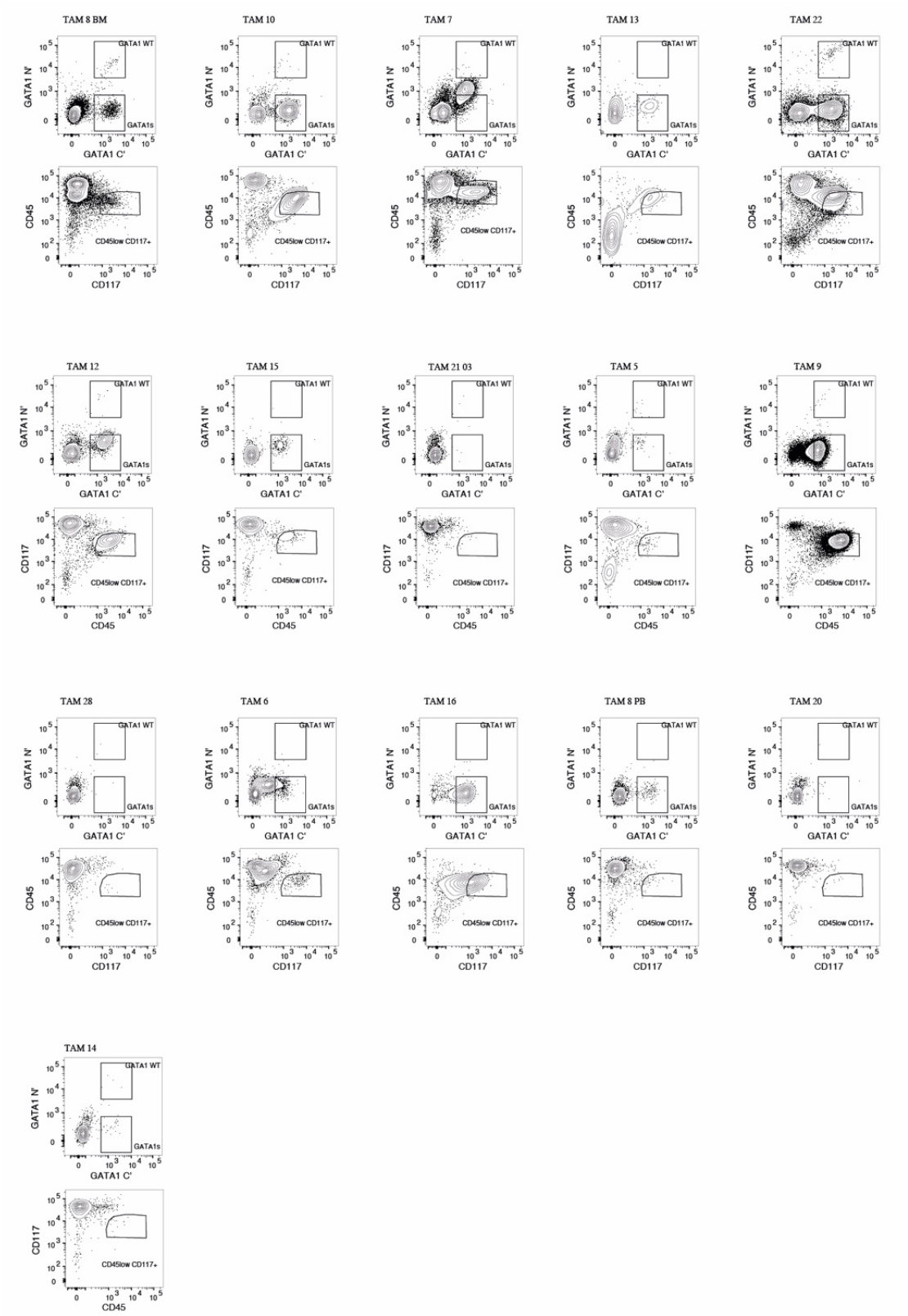
187. Johnson KD, Kong G, Gao X, et al. Cis-regulatory mechanisms governing stem and progenitor cell transitions. *Sci Adv*. 2015;1(8):e1500503.
188. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.
189. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.
190. Katayama S, Suzuki M, Yamaoka A, et al. GATA2 haploinsufficiency accelerates EVI1-driven leukemogenesis. *Blood*. 2017;130(7):908-919.

10 Appendix I. Supplementary Figures

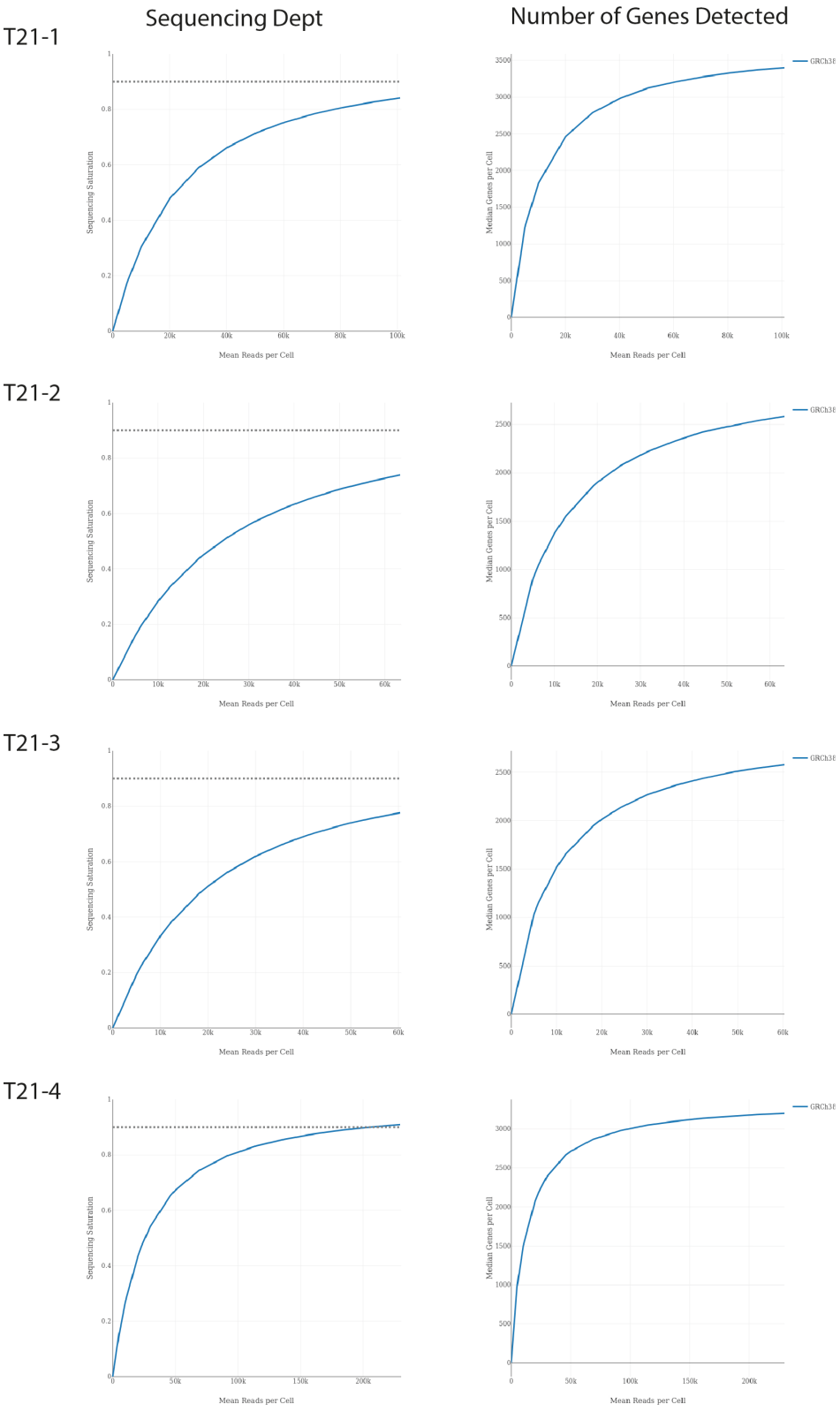


Supplementary Figure 1. iFC in TAM samples

Representative flow cytometry plots of 26 TAM peripheral blood samples where live cells are plotted either on the GATA1 N' (y-axis) vs GATA1 C' (x-axis) (top) or the CD117 (y-axis) CD45(x-axis) bottom. Event frequency for each gate can be found on the Table 6.

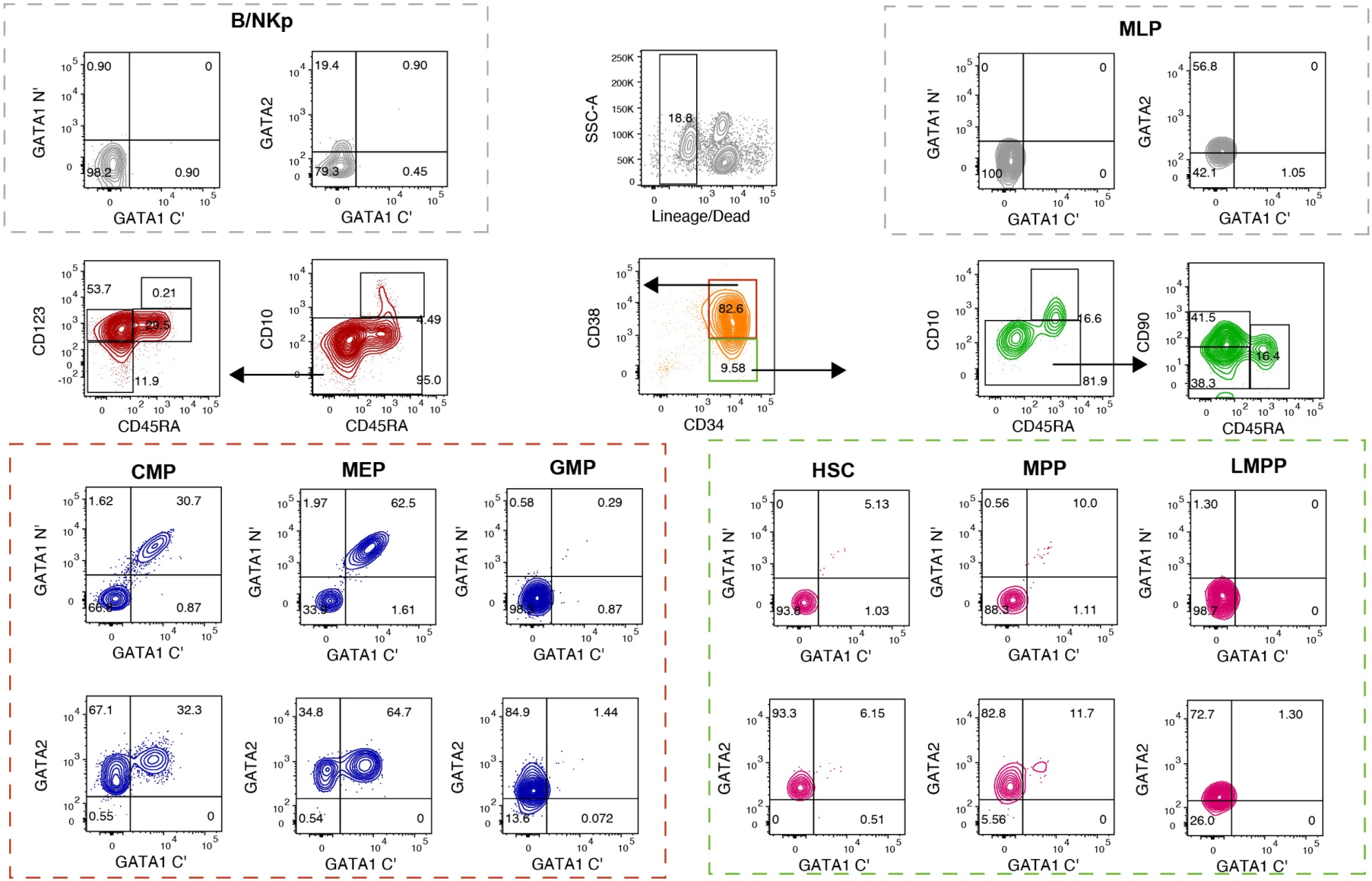


Supplementary Figure 1. iFC in TAM samples, continuation



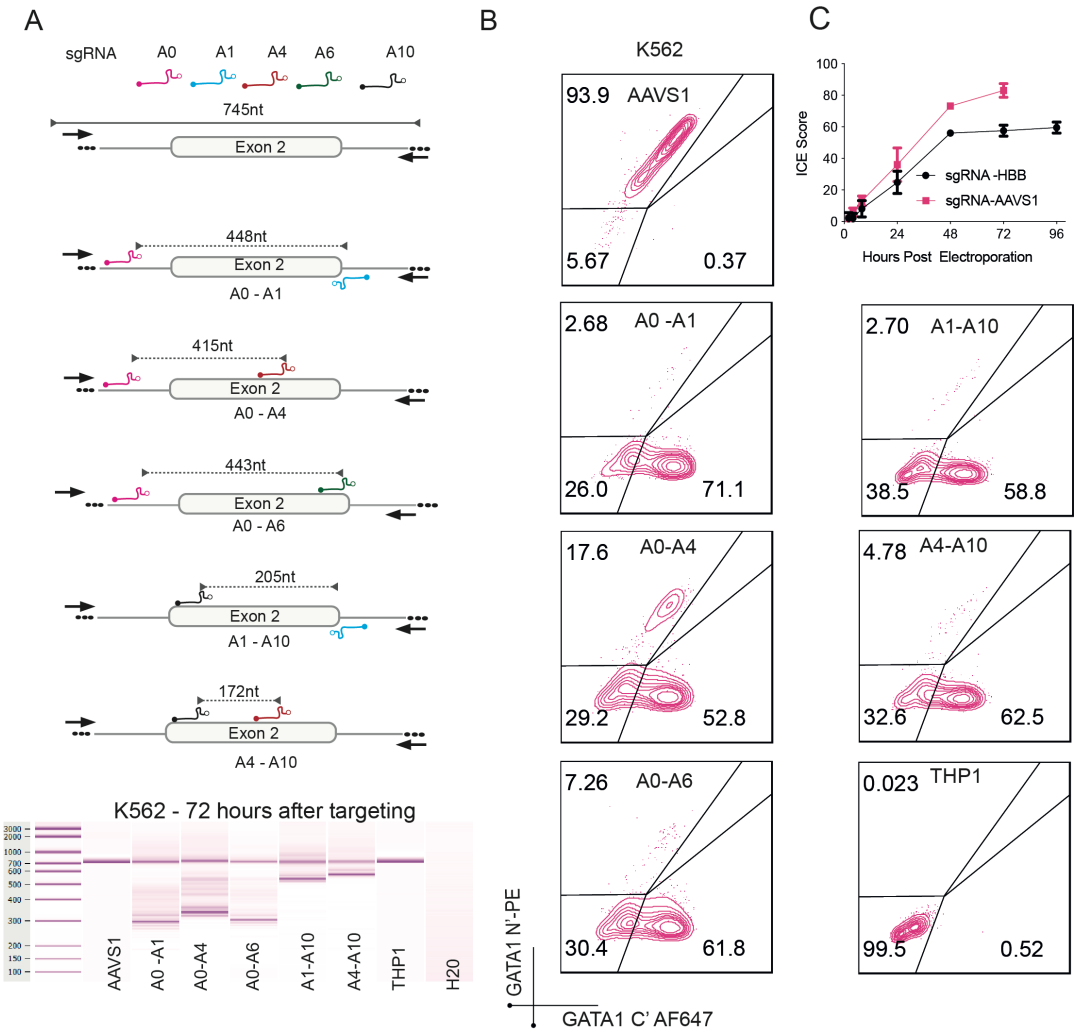
Supplementary Figure 2. Single cell RNA sequencing metrics in 4 T21 samples

Sequencing dept (right) and number of genes detected (left) of 4 trisomy 21 samples labelled T21-1, T21-2, T21-3, T21-4. Dotted lines represent sequencing saturation.


B/NKp
GATA1 N'
GATA1 C'
0.90
0
98.2
0.90
GATA2
19.4
0.90
79.3
0.45
SSC-A
Lineage/Dead
18.8
MLP
0
0
100
0
56.8
0
42.1
1.05
CD123
CD45RA
53.7
0.21
11.9
CD10
4.49
95.0
CD38
CD34
82.6
9.58
CD10
CD45RA
16.6
81.9
CD90
41.5
16.4
38.3
CMP
MEP
GMP
1.62
30.7
0.87
1.97
62.5
1.61
0.58
0.29
0.87
67.1
32.3
0.55
0
34.8
64.7
0.54
0
84.9
1.44
13.6
0.072
HSC
MPP
LMPP
0
5.13
93.8
1.03
0.56
10.0
88.3
1.11
1.30
0
98.7
0
93.3
6.15
0
0.51
82.8
11.7
5.56
0
72.7
1.30
26.0
0

Supplementary Figure 3. Intracellular detection of GATA1 N' and C' termini as well as GATA2 in haematopoietic stem and progenitor cells.

First a set of antibodies conjugated to APC-Cy7 against markers of mature cells types (CD2, CD3, CD4, CD8a, CD11b, CD14, CD19, CD20, CD15) were used to exclude lineage positive cells. Dead cells were also excluded (Fixable near IR +). Then, progenitor cells were defined as CD38+ (brown gate) and primitive cells were defined as CD38- (green gate) of the CD34+ compartment. CD10+ CD45RA+ mark natural killer and B cell immature (B/NKp) cells in the progenitor compartment, and MLP (myloid-lymphod progenitor) in the primitive compartment. CMP, GMP and MEP (blue contour plot) compartments were defined based on the expression of CD123 and CD45RA within the CD10- subset of progenitor cells. HSC, MPP and LMPP (pink contour plot) were defined based on the expression of CD90 and CD45RA within the CD10- subset of primitive cells. Finally GATA1 N' and C' termini as well as GATA2 expression were analysed in each compartment.



Supplementary Figure 4. CRISPR Cas9 with sgRNAs directed to exon 2 of GATA1 to recreate GATA1s cells.

A) 4 sgRNA were designed directed to exon 2 of human GATA1 and 1 sgRNA targets upstream of the exon 2. The size of between the expected cutting size of a pair of sgRNAs is depicted with dotted lines. Bioanalyzer traces of PCR amplicons after 72 hours of targeting with each pair of sgRNAs, THP1 genomic DNA is used a control. Representative intracellular flow plots of targeted K562 cells with anti GATA1 antibodies directed against GATA1 N' and C' termini. Time course of two control sgRNAs against *HBB* or *AAVS1* loci, with indel rate (ICE score) on y-axis and time after electroporation on the x-axis.

11 Appendix II. Supplementary Tables

Table 1. Flow cytometry and Intracellular Flow Cytometry Antibodies for TAM and ML-DS

Marker	Color	Clone	CAT NO:	VENDOR	Staining, ul	
CD36	BUV395	CB38	565422	BD	1/100	
CD7	BUV805	M-T701	742002	BD	2/100	
CD45	BV421	HI30	304032	Biolegend	3/100	
CD34	BV650	582	743533	BD	2/100	
CD41	FITC	HIP8	303704	Biolegend	1/100	
GlyA	Pe-Cy5	HIR2	306606	Biolegend	1/100	
CD117	PE-Cy7	104D2	313212	Biolegend	1/100	
N' GATA1	PE	D52H6	13353S	Cell signaling	24ng/100	
C' GATA1	AF647	D24E4	4589S	Cell signaling	30ng/100	
rabbit IgG	PE	DA1E	5742	Cell signaling	24ng/100	
rabbit IgG	AF6467	DA1E	3452	Cell signaling	30ng/100	
Dead cells	Near IR	NA	L34975	ThermoFisher	1/500	

Table 2. NGS GATA1 primer sets for genotyping TAM and ML-DS samples

Table 2. List of primer spanning GATA1 exon 2 and exon 3				
	Gene	Primer Name	Forward Primer 5' -3'	Reverse Primer 5' -3'
1	GATA1	GATA1_ex2_A	**ACACTGACGACATGGTTCTACA**AGGAAGAGGAGCAGGTGAAAG	**TACGGTAGCAGAGACTTGGTCT**AGAGGGGAAGAAAACCCCTGATT
2	GATA1	GATA1_ex2_B	**ACACTGACGACATGGTTCTACA**TTCTGTGTCTGAGGACCCCTT	**TACGGTAGCAGAGACTTGGTCT**GGGCAGTGGAGGAAGCTG
3	GATA1	GATA1_ex2_C	**ACACTGACGACATGGTTCTACA**GGACCTCAGAGCCCCTCC	**TACGGTAGCAGAGACTTGGTCT**CGTCCCTGTAGTAGGCCAGT
4	GATA1	GATA1_ex2_D	**ACACTGACGACATGGTTCTACA**GAATCAGGGGTTTTCTTCCCCTC	**TACGGTAGCAGAGACTTGGTCT**GATCTCCATGGCAACCCCAAC
5	GATA1	GATA1_ex3	**ACACTGACGACATGGTTCTACA**TGACGTGCGCTGACCCTA	**TACGGTAGCAGAGACTTGGTCT**GGTGGGACACACAGTTGAGG
CS1 and CS2 are marked in Bold				

Table 3. CRISPR Cas9 sgRNAs for GATA1 exon 2 deletion and PCR amplification

Name	sgRNA sequence 5'- 3'	Forward PCR Primer	Reverse PCR Primer
GATA1_A0	tggaacggggagatgcagga	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA
GATA1_A1	CAGGTAACTCCATTGAGTGG	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA
GATA1_A4	tgCGGCACTGGCCTACTACA	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA
GATA1_A6	GAGGCCtACAGACACtCCCC	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA
GATA1_A10	CCAUGGAGUUCCCUGGCCUG	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA
AAVS1/PPP1R2C	GGGGCCACtAGGGACAGGAt	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA
HBB	CTTGCCCCACAGGGCAGTAA	gacaggaggaaaagggaagg	AAGCTTCCAGCCATTTCTGA

Table 4. C&T Detect and CUT&TAG primer sequences

NAME	SEQUENCE 5' -3'	Order Desalted standard	100uM resuspension	Comment
P5F C&Tdetect	AATGATACGGCGACCACCGAGATCTACAC	YES	yes	CandT-qPCR
P7F C&Tdetect	CAAGCAGAAGACGGCATACGAGAT	YES	yes	candT-qPCR
GATA1pR	GAGCCTGTGGGATACCTTGA	YES	yes	candT-qPCR
GATA1pF	CAGTTTCCCGCCTCCAAG	YES	yes	candT-qPCR
Universal i5 primer	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGATGTGTAT	YES	yes	Amplify tagged DNA
i7_1	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_2	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_3	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_4	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_5	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_6	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_7	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_8	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_9	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA
i7_10	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGTG	YES	yes	Amplify tagged DNA

Table 5 Haematopoietic specific lineage gene signatures from Roy A. et al[12]

G2M		S_phase		Erythroid	Myeloid	Lymphoid	Megakaryocyte	HSPC_MPP		Eosinophil_Basophil_Mast		Endothelial_cells		Quiescence				
HMGB2	HJURP	MCM5	RFC2	EPOR	ELANE	VPREB1	MPIG6B	AVP	NRIP1	CLC	NLK	TCIM	IGFBP4	ADAM28	CRHBP	GLIPR1	MIR22HG	STAT4
CDK1	CDCA3	PCNA	RPA2	KLF1	AZU1	VPREB3	PF4	HOPX	GUCY1A1	EPX	CSF2RB	OIT3	S100A16	ADAM8	CTSO	GPR126	MLLT3	SVIL
NUSAP1	HN1	TYMS	NASP	TFR2	PRTN3	LTB	GP9	MLLT3	BEX2	IL5RA	CXCL3	GPR182	WWTR1	AIF1	CXCL1	GUCY1A3	MPZL2	TCEAL2
UBE2C	CDC20	FEN1	RAD51AP1	CSF2RB	CFD	DNTT	VWF	RBPMS	MDK	GRAP2	RNASE2	LYVE1		AKR1C3	CXCL10	H1F0	MYO5C	TCF4
BIRC5	TTK	MCM2	GMNN	APOE	MPO	EBF1	SELP	BEX1	IL18	HDC	RNASE3	PCAT19		ALDH1A1	CXCL11	HCP5	NEAT1	TCF7L2
TPX2	CDC25C	MCM4	WDR76	APOC1	CST7	CD79A		CAVIN1	TNFSF10	IL1RL1		FCN3		ALOX5	CXCL13	HIST1H2AC	NFAT5	TFPI
TOP2A	KIF2C	RRM1	SLBP	CNRIP1	CTSG	CD79B		HMGA2	GNAI1	PRG2		KDR		ANKRD12	CXCL2	HIST1H2BK	NR4A2	TGFB1I1
NDC80	RANGAP1	UNG	CCNE2		CLEC11A	MME		NPR3	ADGRG1	BACE2		MMRN2		ANXA5	CXCL3	HIST2H2BE	NRIP1	TM4SF1
CKS2	NCAPD2	GINS2	UBR7		RNASE2	IRF8		MAFF	BST2	ALOX5		STAB2		APP	CXCL5	HLA-DMA	OPTN	TNFSF10
NUF2	DLGAP5	MCM6	POLD3		RNASE3	LCK		CEBPB	RARRES3	SLC18A2		NTS		AREG	CXCL6	HLA-DMB	PCDH9	TNFSF4
CKS1B	CDCA2	CDCA7	MSH2		KCNE5	SYK		CD52	ALDH1A1	ALOX5AP		APLNR		ARMCX2	CYTIP	HLA-DPA1	PMCH	TPBG
MKI67	CDCA8	DTL	ATAD2			ADA		MSRB3	RGS10	TRDC		COL4A2		BAALC	DDX17	HLA-DPB1	PPFIBP1	TRA2A
TMPO	ECT2	PRIM1	RAD51			HLX		ADAM28	PCBD1	MS4A3		CLEC4M		BCL11A	DMD	HLA-DQA1	PROM1	TRA2B
CENPF	KIF23	UHRF1	RRM2			LST1		HOXB2	ID1	EXD3		RBP1		BIRC3	DST	HLA-DQB1	PTGER4	TRIM22
TACC3	HMMR	MLF1IP	CDC45			ITGAL		JAML	ICA1	LMO4		CD14		C3orf14	DUSP6	HLA-DRB4	PTGS2	TSPAN31
FAM64A	AURKA	HELLS	CDC6							SLC45A3		CDH5		CCL19	EBI3	HLF	PTPRC	TSPAN6
SMC4	PSRC1		EXO1							HPGD		FGF23		CCL20	EMP1	HLX	RABGAP1	VNN1
CCNB2	ANLN		TIPIN							CPA3		COL4A1		CCL8	ENPP4	IDS	RBPMS	ZBTB10
CKAP2L	LBR		DSCC1							RHEX		TM4SF18		CCNG2	ERG	IFI16	RGS1	
CKAP2	CKAP5		BLM							ANXA1		NR2F2		CD200	EVI2A	IL18R1	RPL31	
AURKB	CENPE		CASP8AP2							CCL3		ECSCR		CD40	FAM129A	IL1B	SCHIP1	
BUB1	CTCF		USP1							AL157895.1		ANGPTL4		CD44	FHL1	IL8	SELL	
KIF11	NEK2		CLSPN							PLIN2		FEZ1		CD52	FLJ10038	INHBA	SH3BP5	
ANP32E	G2E3		POLA1							MEIS2		CLDN5		CD59	FNBP1	JUN	SOD2	
TUBB4B	GAS2L3		CHAF1B							HPGDS		DNASE1L3		CH25H	FTH1	KIAA012KYN	SORL1	
GTSE1	CBX5		BRIP1							ANGPT2		ECM1		CLEC2B	FXR1	LIMCH1	SPINK2	
KIF20B	CENPA		E2F8							CXCL8		PLPP3		CREM	GBP2	MAFF	SPTBN1	

6

Table 6. Patient samples, genotyping and flow cytometric features

Table 1. Patient Sample Information													
Sample No	Days Old	Sex	Source	Blast %	Adjusted VAF % by sex	VAF % 1)	VAF % 2)	VAF % 3)	Mutation	CD45low CD117+ % of live	CD45low CD117+ CD7+ % of live	GATA1 C' % of live	GATA1 N' % of live
TAM 1	20	M	PB		15.32	15.32	21.32	1.83	1) c.36_37del:p.S12fs 2) c.45dupC:p.L15fs 3) c.220+1G>T (exon 2 splice donor)	20.4	17.4	23.8	0.01
TAM 2	3	M	PB	73% Morphology	92.11	92.11			c.184_185insCCCTGTAGGCCT:p.Y62delinsSLX: nonsense mutation	44.2	21.7	32.5	0.08
TAM 3	0	M	PB	58% Morphology	Sanger	Sanger			c.1-20 g>c: SA	24.9	22.1	30.9	0.37
TAM 4	1	M	PB	80% Morphology	77.8	77.8	4.12		1) c.A1G:p.M1V 2) c.36dupA:p.S12fs	15.5	7.59	21	0.2
TAM 5	31	F	PB		0.52	0.26			c.184_217delTACTACAGGGACGCTGAGGCCTACAGACACTCCC:p.Y62fs	3.06	0.96	2.63	0.43
TAM 6	9	F	PB	5 % Morphology	25.4	12.7			c.C89A:p.S30X	3.25	1.97	13.4	0.03
TAM 7	7	F	PB		22.64	11.32			c.151_161dupAGCACAGCCAC:p.T54fs	16.3	3.69	36.2	0.13
TAM 8 -PB	42	F	PB		7.4	3.7			exon2:c.G3T:p.M1I	1.01	0.74	2.94	0.01
TAM 8 - BM	42	F	BM		4.18	2.09			exon2:c.G3T:p.M1I	3.26	1.57	4.01	0.24
TAM 9	2	M	PB	61% by FC	81.5	81.5	1.05		1) c.90_91del:p.S30fs 2) c.1-22c>g (intronic, 1bp upstream of SA)	93.2	64	71.6	0.03
TAM 10	6	F	PB	25% byt Morphology	43.18	21.59			c.1-33_1-11del (deletion of splice acceptor and 5'UTR)	50.7	28	59.3	0.29
TAM 11	11	F	PB	7% by Morphology	40.44	20.22			c.146_180del:p.A49fs	17.9	11.3	16.7	0.02
TAM 12	2	M	PB		60.02	60.02			c.T2C:p.M1T	30	18.9	31	0.19
TAM 13	3	F	PB	20% by Morphology	13.72	6.86	6.57		1) c.G3A:p.M1I 2) c.172_220+42del	9.7	6.44	17	0.04
TAM 14	3	M	PB		18.5	18.5			c.A1G:p.M1V	0.93	0.55	1.19	0.13
TAM resolving/ML-DS evolving 15	146	M	PB		0.86	0.86			c.61dupC:p.D20fs	8.55	5.72	12.2	0.06
TAM 16	3	M	PB	40% FC	68.22	68.22			3611	60.1	15.9	79.9	0.02
TAM 17	2	M	PB		60.02	60.02			c.1-22:C>G	35.4	33.7	39.7	0.04
TAM 18	29	F	PB		2.62	1.31			c.1-46_1-25delGTCTGAGGACCCCTTCTGTCCT	0.62	0.37	0.78	0.11
TAM 19	11	F	PB	72% FC	87.4	43.7			c.1-38_1-19del: loss of Exon 2 splice acceptor (intronic deletion of 20 bp that takes out Exon2 splice acceptor site)	46.7	29.4	77.7	0.03
TAM 20	34	F	PB		0.32	0.16			c.151_161dupAGCACAGCCAC p.54Tfs	0.55	0.93	0.34	0.05
TAM 21-02	34	F	PB	No Blast	1.56	0.78			c.C207G:p.Y69X	0.59	0.47	0.79	0.04
TAM 21-03	62	F	PB		0.2	0.1			c.C207G:p.Y69X	0.16	0.14	0.15	0.07
TAM 22	10	M	PB	10 % by Morphology	51.5	51.5			c.1_10delATGGAGTTCC p.1Mfs (Deletion of 10 bases which takes out the initiating methionine)	41	27.1	49.3	0.57
TAM 23	103	M	PB		0.42	0.42			c.152_153insGTGAG:p.S51_T52delinsRX	0.29	0.21	0.34	0.02
TAM 28	63	M	PB		1.04	1.04			c.220_221insGT:p.V74fs	0.27	0.074	0.21	0.06
ML-DS 23	463	M	BM		8.78	8.78			c.152_153insGTGAG:p.S51_T52delinsRX	7.66	5.89	4.74	0.01
ML-DS 24Pre	563	F	BM		51.1	25.55			c.90_91delAG:p.S30fs	24.6	19.8	23	0.18
ML-DS 24 Post	598	F	BM		0.06	0.03			c.90_91delAG:p.S30fs	8.22	1.04	0.13	1.95
ML-DS 24 Post	710	F	BM		N.D	N.D				9.89	1.12	0.042	1.78
ML-DS 25 Pre	453	M	BM	7% FC, 20 % Morphology	8.95	8.95			c.196_203delGCTGAGGC:p.A66fs	18.4	10.1	13.2	0.28
ML-DS 25 Post	538	M	BM		N.D	N.D				18.7	0.93	0.035	1.29
ML-DS 26	313	F	BM		42	21			c.120_124dupCTTGG and c.147_164delCCCGAGCACAGCCACCGC : p.D42fs	35.1	30.7	24.1	0.18
ML-DS 27 Pre	564	M	BM		6.26	6.26			c.C186G:p.Y62X	12.3	6.77	9.47	0.39
ML-DS 27 Post	613	M	BM		ND	ND				31.8	4.95	0.071	0.67
DS1	11	Uknown	PB	1.2% Morphology		N.D				4.22	1.26	0.0071	0.94
DS2	12	Female	PB			N.D				10.9	1.87	0	1.28
DS3	20	Male	PB			N.D				0.28	0.18	0	0.02
DS4	10	Male	PB			N.D				1.08	0.14	0	0.16
DS11	11	F	PB	1 % Morphology		N.D				1.24	0.54	0.0039	0.06
DS12	11	F	PB	1 % Morphology		N.D				2.03	0.64	0	0.4
DS13	3	M	PB	1 % Morphology		N.D				1.4	0.69	0	0

N.D = Not Detected
PB= Peripheral Blood
BM= Bone Marrow
VAF: Variant Allele Frequency
FC = blast detected by flow cytometry
Morphology = blast detected by morphology

Table 7. Receiver Operating Characteristic (ROC) analysis

Cutt-off	Sensitivity%	95% CI	Specificity%	95% CI	Youden's Index
> 0.1070	96.55	82.82% to 99.82%	0	0.000% to 22.81%	95.55
> 0.1600	96.55	82.82% to 99.82%	15.38	2.734% to 42.23%	110.93
> 0.1950	96.55	82.82% to 99.82%	23.08	8.180% to 50.26%	118.63
> 0.2900	93.1	78.04% to 98.77%	23.08	8.180% to 50.26%	115.18
> 0.4200	89.66	73.61% to 96.42%	23.08	8.180% to 50.26%	111.74
> 0.5050	86.21	69.44% to 94.50%	23.08	8.180% to 50.26%	108.29
> 0.5450	86.21	69.44% to 94.50%	30.77	12.68% to 57.63%	115.98
> 0.5950	82.76	65.45% to 92.40%	30.77	12.68% to 57.63%	112.53
> 0.6650	82.76	65.45% to 92.40%	38.46	17.71% to 64.48%	120.22
> 0.7150	82.76	65.45% to 92.40%	46.15	23.21% to 70.86%	127.91
> 0.8350	79.31	61.61% to 90.15%	46.15	23.21% to 70.86%	124.46
> 0.9450	79.31	61.61% to 90.15%	61.54	35.52% to 82.29%	139.85
> 1.000	75.86	57.89% to 87.78%	61.54	35.52% to 82.29%	136.4
> 1.080	75.86	57.89% to 87.78%	69.23	42.37% to 87.32%	144.09
> 1.190	75.86	57.89% to 87.78%	76.92	49.74% to 91.82%	151.78
> 1.415	75.86	57.89% to 87.78%	84.62	57.77% to 97.27%	159.48
> 1.720	72.41	54.28% to 85.30%	84.62	57.77% to 97.27%	156.03
> 1.920	72.41	54.28% to 85.30%	92.31	66.69% to 99.61%	**163.72**
> 2.830	68.97	50.77% to 82.72%	92.31	66.69% to 99.61%	160.28
> 4.320	65.52	47.35% to 80.06%	92.31	66.69% to 99.61%	156.83
> 5.335	65.52	47.35% to 80.06%	100	77.19% to 100.0%	164.52
> 5.805	62.07	44.00% to 77.31%	100	77.19% to 100.0%	161.07
> 6.165	58.62	40.74% to 74.49%	100	77.19% to 100.0%	157.62
> 6.605	55.17	37.55% to 71.59%	100	77.19% to 100.0%	154.17
> 7.180	51.72	34.43% to 68.61%	100	77.19% to 100.0%	150.72
> 8.845	48.28	31.39% to 65.57%	100	77.19% to 100.0%	147.28
> 10.70	44.83	28.41% to 62.45%	100	77.19% to 100.0%	143.83
> 13.60	41.38	25.51% to 59.26%	100	77.19% to 100.0%	140.38
> 16.65	37.93	22.69% to 56.00%	100	77.19% to 100.0%	136.93
> 18.15	34.48	19.94% to 52.65%	100	77.19% to 100.0%	133.48
> 19.35	31.03	17.28% to 49.23%	100	77.19% to 100.0%	130.03
> 20.75	27.59	14.70% to 45.72%	100	77.19% to 100.0%	126.59
> 21.90	24.14	12.22% to 42.11%	100	77.19% to 100.0%	123.14
> 24.60	20.69	9.846% to 38.39%	100	77.19% to 100.0%	119.69
> 27.55	17.24	7.598% to 34.55%	100	77.19% to 100.0%	116.24
> 28.70	13.79	5.497% to 30.56%	100	77.19% to 100.0%	112.79
> 30.05	10.34	3.581% to 26.39%	100	77.19% to 100.0%	109.34
> 32.20	6.897	1.225% to 21.96%	100	77.19% to 100.0%	105.897
> 48.85	3.448	0.1769% to 17.18%	100	77.19% to 100.0%	102.448

Table 8. FACS Sorting Panel for Single Cell RNA sequencing

Marker	Color	Catalogue number	Clone	Staining ul per 100ul
CD38	BV421	HIT2	BioLegend	2.5
CD41	BV605	HIP8	BioLegend	2
CD45	BV650	HI30	BioLegend	2
CD123	Pe-Cy7	6H6	BioLegend	2.5
CD45RA	BB515	HI100	BD bioscience	1.25
CD131	PE	1C1	BioLegend	2.5
CD71	PE-Cy5		BD bioscience	0.5
CD117	BV711	104D2	BioLegend	1
CD34	APC	581	BioLegend	0.4
Viability	Fixable Near IR		Thermo	1
Anti-human CD2	APC cy7	RPA-2.10	BioLegend	0.25
Anti-human CD3	APC cy7	HIT3a	BioLegend	1.5
Anti-human CD4	APC cy7	RPA-T4	BioLegend	0.25
Anti-human CD8a	APC cy7	RPA-T8	BioLegend	0.5
Anti-human CD11b	APC cy7	ICRF44	BioLegend	1
Anti-human CD14	APC cy7	63D3	bioLegend	1
anti Human CD15	APC Cy7	W6D3	BioLegend	2
Anti-human CD19	APC cy7	HIB19	BioLegend	1
Anti-human CD20	APC cy7	2H7	BioLegend	1

Table 9. FACS cell sorting for single cell RNA sequencing

Cell Type	Immunophenotype	T21_1	T21_2	T21_3	T21_4
EMPPs	CMP CD131	1158	474	1022	1903
MegP	CMP CD41	2343	1139	19	2129
MK-MEP	MEP CD41 CD71	2516	4080	364	3280
Primitive	CD34+ CD38-	1317	483	1256	2189
Progenitors	CD34+ CD38+			4829	
HSPCs	CD34+	5000	25463	11127	13001
Precursors	CD34- CD117+	5983	17138	8817	7045
lineage negative	lin -	30882			
	Total sorted cells	49199	48777	27434	29547

Table 10. Genotyping frequenting per cell cluster TAM

Cell cluster number	Cell Cluster	All reads are mutant	At least one read is mutant or wild type	No Call	All reads are wild type	Total Cells	Allelic drop out per cluster
0	HSC/MPP	1	0	36	0	37	97.3
1	Erythroid P	456	37	34	33	560	6.1
2	MPP I	4	0	65	1	70	92.9
3	Myeloid P	3	0	34	3	40	85.0
4	Lymphoid P	0	0	16	0	16	100.0
5	Erythroid II	0	0	2	4	6	33.3
6	Cycling	244	32	11	14	301	3.7
7	MPP II	3	0	21	2	26	80.8
8	EMPP	150	16	37	11	214	17.3
9	CD4 T cells	1	0	44	0	45	97.8
10	Erythroid III	0	0	2	1	3	66.7
11	Erythroid IV	0	0	7	1	8	87.5
12	Monocyte	1	0	34	0	35	97.1
13	Erythroid I	3	0	0	9	12	0.0
14	CD8 T Cell	0	0	16	0	16	100.0
15	Eo/Ba/Mast Cell	23	0	14	1	38	36.8
16	Megakaryocyte P	248	34	15	5	302	5.0
17	Dendritic Cell	0	0	4	0	4	100.0
18	B-Cells	1	0	15	0	16	93.8
19	Naturall Killer	0	0	1	0	1	100.0
	Total sum	1138	119	408	85	1750	
	Frequency	65.03	6.80	23.31	4.86	100.00	
	Blue font corresponds to cell cluster with known GATA1 expresison						

12 Appendix III. Publication Obtained During My DPhil Project

Gaëtan Juban, Nathalie Sakakini, Hedia Chagraoui, **David Cruz Hernandez**, Qian Cheng, Kelly Soady, Bilyana Stoilova, Catherine Garnett, Dominic Waithe, Georg Otto, Jessica Doondeea, Batchimeg Usukhbayar, Elena Karkoulia, Maria Alexiou, John Strouboulis, Edward Morrissey, Irene Roberts, Catherine Porcher, Paresh Vyas. Oncogenic Gata1 causes stage-specific megakaryocyte differentiation delay. **Haematologica** 2020;106(4):1106-1119.

David Cruz Hernandez, Marlen Metzner, Anne Pieta de Groot, Batchimeg Usukhbayar, Natalina Elliott, Irene Roberts, Paresh Vyas; Sensitive, rapid diagnostic test for transient abnormal myelopoiesis and myeloid leukemia of Down syndrome. **Blood** 2020; 136 (12): 1460–1465.

Garnett, C, **Cruz Hernandez, D**, Vyas, P. GATA1 and cooperating mutations in myeloid leukaemia of Down syndrome. **IUBMB Life**. 2020; 72: 119– 130.

Cruz Hernandez, D. Vyas, P. Oncogenic Drivers and Development. **Cancer Discovery**.2019; CD-19-1082.

Labuhn, M. et al. Mechanisms of progression of myeloid preleukemia to transformed myeloid leukemia in children with Down syndrome. ***Cancer Cell***; **2019**. 36, 123–138.e10