# Exploring the expression and function of RNA-binding proteins and splicing factors in haematopoiesis

**MRes Thesis**

By Nicole Forrester 100349138

Supervised by Dr Iain Macaulay and Dr Wilfried Haerty



**September 2023**

# TABLE OF CONTENTS

# ACKNOWLEDGEMENTS

I'd like to thank my supervisors Dr Iain Macaulay and Dr Wilfried Haerty for their support and discussions and particularly for supporting me in my decisions. Thanks to Iain for his help with FACS and to Wilfried for his help with RNA-seq data analysis and thanks to both for the constant supply of chocolate and sweet treats for everyone!

Thank you to the Macaulay group for their help, feedback and their patience whilst I single-handedly kept the 37°C incubator stocked full with numerous culture plates and as I ran around the lab each day, trying to prepare samples in time for the 2:30pm Sanger sequencing collection. In particular, thank you to Dr Laura Mincarelli for the use of her data in my analysis and to Dr Dave Wright for his crystal clear explanations and for his help with the eCLIP data analysis. Also, thank you to Dr Angela Man from the Haerty group for the CRISPR chats and help with the gel electrophoresis and to Dr Conrad Nieduszynski for the use of the equipment for gel electrophoresis and for often checking in on me.

Thank you to Genomics Pipelines for generating and sequencing the libraries in record time – my thesis wouldn't be the same without your rapid turnaround.

To all the students at EI, a massive thank you for providing regular social reliefs from lab work and a lot of laughs! Thank you to Anita Scoones, Matt Madgwick, Peter Osbourne, Sam Witham and Becky Shaw for giving me (a lot of) your time and kind words of support and encouragement - I really needed them.

As always, thank you to my family for their love and understanding and their support for my decisions.

A very special thank you to Dr Guy Pearson from the Carlton Lab at the Francis Crick Institute for his extensive expertise in all things CRISPR and FACS. Your ideas and suggestions, your proof-reading and our countless conversations about my project have been invaluable and I couldn't have made it through these past two years without you. Thank you for helping me to improvise, adapt and overcome the hard times. Thank you for always believing in me no matter what and for showing me a new world of opportunities and possibilities.

# ABSTRACT

Alternative splicing (AS) contributes to the generation of functional diversity in higher eukaryotic proteomes by producing multiple isoforms of the same protein from the same gene locus. AS is influential in cell lineage commitment in multiple systems and differentiation processes, including haematopoiesis. RNA-binding motif protein 15 (RBM15) is a known regulator of AS in haematopoietic stem cells, influencing homeostasis in this compartment and particularly regulating megakaryocyte differentiation. This project investigates the regulatory functions of RBM15, and furthermore examines the expression of other splicing factor genes, identifying any lineage-specific patterns within the haematopoietic system. RNA-seq was performed on control and CRISPR-Cas9 edited HAP1 cells to explore differential expression. Editing the *RBM15* locus impacted the transcriptome dramatically with 8,710 genes upregulated and 7,941 downregulated, including key haematopoiesis genes such as *HCLS1* and *PDGFRA*. Additionally, an inverse relationship between *RBM15* and *RBM15B* expression was discovered, highlighting that these genes appear closely related. Next, a single-cell short-read dataset from haematopoietic stem cells was explored for patterns of RNA-binding protein and splicing factor (RBP/SF) expression. 83 out of 428 RBP/SF genes were found to be expressed and three displayed lineage-specificity: *Carhsp1* and *Larp1b* were upregulated in erythroid progenitors and *Khdrbs3* was upregulated in megakaryocyte progenitors. To investigate the transcriptomic influence of these genes, they were targeted for knockout in HAP1 cells using CRISPR-Cas9 and guides were confirmed to successfully edit two of the three genes. Overall, this research highlights the broad influence of RBM15 on transcriptional regulation and has identified RBP/SFs preferentially acting in specific haematopoietic lineages.

# CHAPTER 1: INTRODUCTION

Proteins are long, folded chains of amino acids which perform a vast array of crucial functions in the body. From catalysis of chemical reactions to providing cellular structure, cell signalling to nutrient transport, pathogen defence to cell division, proteins are essential macromolecules in all tissues of an organism. Some proteins are ubiquitous across cell type while others are tissue-specific and enable specialised functions to be carried out only in certain cell types. Proteins are encoded by genes, sequences of deoxyribonucleic acids (DNA), and are synthesised following two processes: transcription and translation. Transcription generates an intermediate molecule known as messenger ribonucleic acid (mRNA), a single-stranded copy of the gene with some chemical modifications. Translation describes the reading of the RNA in groups of three nucleotides called codons and the consequent production of an amino acid polypeptide chain, which later folds to form a protein. The regulation of these two processes can control how much protein is formed and even which proteins are formed.

Post-transcriptional regulatory mechanisms occur prior to translation, either in the modification of nascent mRNA into mature mRNA, processing the molecule to be ready for translation, or in the export of the mature mRNA molecule from its site of synthesis in the nucleus to its site of translation in the cytoplasm (Corbett, 2018). These mechanisms include: mRNA splicing, the removal of introns from the mRNA and the reassembly of the exons; mRNA capping, the modification of the 5' end of the mRNA to protect it from exonucleases and aid in ribosomal binding at the initiation of translation; polyadenylation, the addition of a poly(A) tail to the 3' end of the mRNA, which also protects the molecule and aids in translation initiation, but can be varied in length to control mRNA half-life; and nuclear export, the exit of the mature mRNA via pores in the nuclear membrane. The focus of this thesis is on splicing, and alternative splicing in particular.

## *1.1 What is alternative splicing?*

Genes are segmented into protein-coding exons with introns interspersed between them. Human genes contain an average of 8 exons, ranging in size from 2 to 12,000 nucleotides (nt; median of 150 nt; Lander *et al.*, 2001). Introns can vary from 10 to over 100,000 nt in length and, taken together, amount to a quarter of the human genome (Singh, 2018). Prior to translation of the mRNA into a protein, introns must be removed

from precursor mRNA (pre-mRNA) sequences as a key step in the formation of mature mRNA molecules.

Splicing is the removal of introns from pre-mRNA and the joining of the remaining exons. Different combinations of exons from a gene can be spliced to produce multiple different mature mRNA sequences, typically with one prevalent form and a number of lesser abundance transcripts. This process is called alternative splicing (AS) and the collection of transcripts generated are known as isoforms. Isoforms originating from the same gene often encode proteins with different and even opposing functions, sometimes causing downstream consequences for the cell. One such example of antagonistic isoforms is from the AS of exon 2 of Bcl-2-like protein 1 (*BCL2L1*) which can produce either the long BCL-XL or the short BCL-XS transcripts, encoding anti-apoptotic and pro-apoptotic proteins, respectively (Boise *et al.*, 1993).

The first example of AS was discovered in 1977 in studies of type 2 adenovirus (Berget, Moore and Sharp, 1977; Chow *et al.*, 1977). AS was initially thought to occur in only 5-40% of eukaryotic genes (Sharp, 1994; Mironov, Fickett and Gelfand, 1999; Brett *et al.*, 2000), but now it is known that 95% of human multi-exonic genes are alternatively spliced (Pan *et al.*, 2008).

## *1.2 Patterns of AS*

There are 5 common patterns of AS: 1) exon skipping; 2) alternative 5' (donor) splice site; 3) alternative 3' (acceptor) splice site; 4) intron retention; 5) mutually exclusive exons (Fig.1).

Exon skipping occurs when one or more exons are excluded from the final mature mRNA transcript. In mammals, this is the most prevalent AS event (40%) (Mehmood *et al.*, 2020). Alternative 5' donor and 3' acceptor sites can be found within some pre-mRNA sequences and are used to produce alternative isoforms. Usage of these sites results in longer or shorter exons being included as compared to constitutive splicing. In about 5% of higher eukaryotic AS events, an intron or portion of intronic sequence is retained in the mature mRNA (Kim, Goren and Ast, 2008). Sometimes these intron-retaining mRNAs are degraded by nonsense-mediated decay, however, others remain and can encode new protein isoforms. Intron retention often produces dysfunctional proteins which may contribute to disease, such as a truncated Tau protein resulting from retention of intron 11 in Alzheimer's disease; a frameshift causes a premature stop codon to be

encountered (Adusumalli *et al.*, 2019; Ngian *et al.*, 2022). Mutually exclusive exons do not appear in the same mature mRNA together as only one of the exons can be retained in the transcript after splicing; the splicing events for each mutually exclusive exon are dependent on each other, with one event being blocked if the other takes place. Within the exon 6 cluster of the Dscam gene in Drosophila, there are 48 alternative exons and the single exon to be included is chosen via docking site-selector sequence interactions (Graveley, 2005).

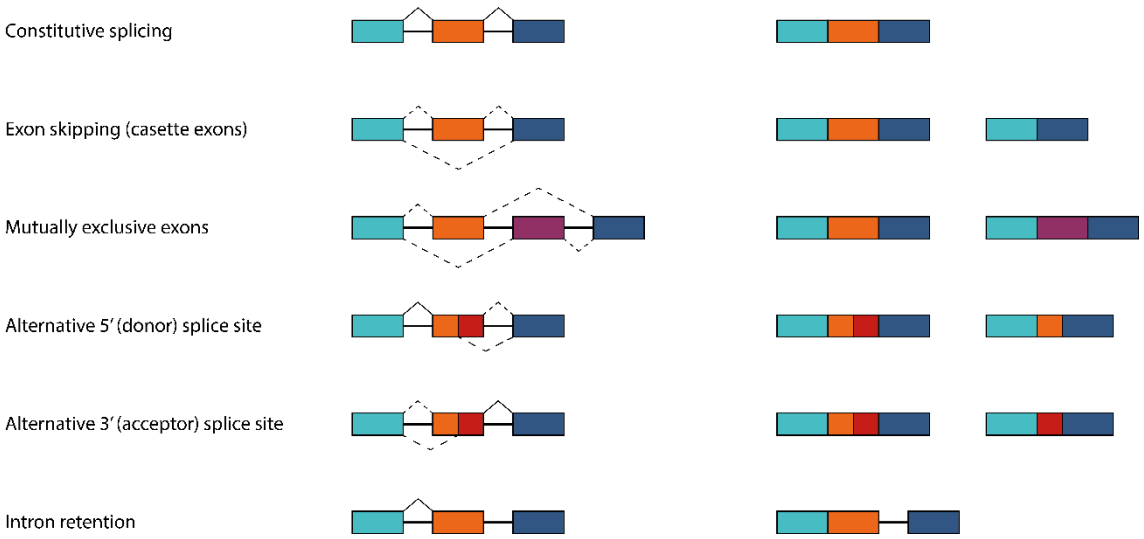


***Figure 1 – Types of alternative splicing events**. The relevant splicing paths for each AS event (left) along with the resulting isoform products (right). The top path in each case corresponds to the first isoform; the bottom path responds to the second (if applicable). The solid path lines show instances of constitutive splicing whilst the dashed lines show possible splicing paths for the purpose of simply illustrating the events and their possible outcomes.*

## *1.3 Splicing mechanism and its regulation*

Splicing is a highly regulated process with hundreds of proteins and RNA molecules contributing to the mechanism and its regulation. As discussed in detail below, disruptions to splicing due to dysfunction of splicing machinery components or regulators have been proven to be directly associated to disorders and diseases (Dvinge *et al.*, 2016; Urbanski, Leclair and Anczuków, 2018). Therefore, it is vital for the health of the organism that each splicing component is functioning correctly.

### 1.3.1 Mechanism of splicing

RNA splicing is performed in the nucleus by the spliceosome, a ribonucleoprotein (RNP) complex composed of 5 small nuclear RNPs (snRNPs): U1, U2, U4, U5 and U6. Each of

the small nuclear RNAs (snRNAs) are less than 200 nt in length and are complexed with numerous protein subunits to form each snRNP, including 7 core or Sm proteins. There are four sites within pre-mRNA molecules that are recognised by the spliceosome: the 5' splice site (5’ SS), the 3' splice site (3' SS), the branch point sequence (BPS) and the polypyrimidine tract. The first two sites demarcate the start and end of an intron respectively. The 5’ SS is denoted by a GU dinucleotide and the 3’ SS is indicated by AG. The BPS marks a significant adenine base within the intron, nearer to the 3' SS than to the 5' SS, which is important for the splicing mechanism. The polypyrimidine tract is found between the BPS and 3’ SS.

Many proteins and RNPs are involved in the mechanism of splicing and therefore there are several steps within the mechanism, as each component contributes to the process. First, splicing factor 1 (SF1; AKA branch-point binding protein, BBP) binds to the BPS, U2 auxiliary factor 2 (U2AF2) binds to the polypyrimidine tract, U2AF1 binds to the 3' acceptor SS and U1 binds with the 5' donor SS by base-pairing, thus forming Complex E or the commitment complex. With the help of U2AF (which subsequently leaves the complex), U2 (including SF3A and SF3B) displaces SF1, forming base pairs with the BPS: this forms Complex A or the pre-spliceosome complex. U4/U6-U5 joins the complex as a triple snRNP, with U4 and U6 being bound together by base pair interactions and U5 being loosely bound by protein interaction, causing rearrangements of Complex A to form Complex B or the precatalytic spliceosome. More conformational changes see the base-pairing between U4 and U6 become broken, U6 displacing U1 and binding the 5' SS as well as U4 dissociating from the complex, allowing U6 and U2 to bind via base-pairing: Complex B* or the activated/catalytic spliceosome is formed, presenting the active site of the spliceosome and positioning the pre-mRNA for the first transesterification reaction. The 2'-hydroxyl group at the BPS attacks the phosphate at the 5' SS, resulting in cleavage between the 5' end of the intron and the upstream exon and the subsequent formation of a phosphodiester bond between the free intron end and the BPS, thus forming the lariat. Once the reaction has occurred, the spliceosome is referred to as Complex C. Further rearrangements take place, promoting the second transesterification reaction and the subsequent formation of the post-spliceosomal complex. The 3'-hydroxyl group of the upstream exon attacks the phosphate at the 3' SS, completing the removal of the intron and ligating the two exons together by formation of a phosphodiester bond. Following completion of splicing, the intron RNA sequence is degraded and the snRNPs are recycled.

### 1.3.2 Splicing regulatory factors

The complex process of AS is primarily regulated by the interactions between cis-acting elements within the pre-mRNA sequences and the trans-acting proteins which bind to them (Barash *et al.*, 2010). The two important classes of cis-acting sequences are enhancers and silencers (Wang *et al.*, 2015). Enhancers can be found in exons (exonic splicing enhancer, ESE) or in introns (intronic splicing enhancer, ISE) (Fairbrother *et al.*, 2002) and help to promote splicing of proximal splice sites when bound to by splicing activator proteins. A major class of splicing activator proteins are the family of serine/arginine-rich nuclear phosphoproteins (SR proteins). Silencers are also exonic or intronic (ESS/ISS) and are bound to by splicing repressor proteins to block splicing or cause exon skipping (Wang *et al.*, 2004). Heterogeneous nuclear ribonucleoproteins (hnRNPs) are the main splicing repressors. Importantly, this delineation of SR proteins as activators and hnRNPs as repressors is not exclusive; some of these trans-acting regulatory proteins have the ability to act as both activators and inhibitors of splicing. For example, hnRNPA1 sterically prevents exon inclusion by binding to an ESS/ISS or to both ends of an exon, forming a loop and causing the exon to be skipped (Blanchette and Chabot, 1999). Yet for the proapoptotic Fas receptor mRNA, hnRNPA1 interrupts donor site selection of exon 5 thereby enabling exon 6 inclusion in the mature mRNA (Oh *et al.*, 2013). hnRNPL is another dual-acting regulatory factor as it can both activate and repress inclusion of exon 5 of the *CD45* gene (Motta-Mena, Heyd and Lynch, 2010). It has been demonstrated that the function of these dual-acting splicing factors depends on their position of action on the pre-mRNA sequences; a splicing factor may repress splicing by binding to an exon within one pre-mRNA molecule but may enable splicing in a different pre-mRNA by binding to the intron of that molecule.

Given that AS is dynamic, it is not surprising that it can be influenced by means other than the direct action of regulatory factors. One such additional influence is the RNA structure itself, where the secondary structure of pre-mRNA can aid splicing by bringing splice sites into proximity with one another or keep them apart to hinder splicing (Jacobs, Mills and Janitz, 2012). Another influence is the rate of transcription. Because splicing occurs as the pre-mRNA is being transcribed by RNA Pol II, slower transcription rates enable the assembly of the spliceosome on weak splice acceptor sites in addition to the strong splice sites which are favoured during higher transcription rates (Kornblihtt, 2007). A third and related additional influence on AS is the structure of chromatin from which the pre-mRNA is transcribed (Luco *et al.*, 2011; Naftelberg *et al.*, 2015). The nucleosome structure itself can affect transcription rate, with nucleosomes acting as barriers to modulate transcription speed by causing RNA Pol II to stall (Hodges *et al.*, 2009).

Epigenetic modifications on chromatin further affect AS, with histone modifications such as H3K36me3 less prominently enriched in alternatively spliced exons than in constitutive exons (Kolasinska-Zwierz *et al.*, 2009) and emerging evidence showing that chromatin can act as direct adapters for AS machinery (Luco *et al.*, 2010).

Ultimately, the diversity of isoforms generated is due to the complex interplay between all of the aforementioned regulatory methods. The competition of SFs and spliceosome components on the pre-mRNA, in addition to the epigenetic structure of the gene, its transcription rate, and the RNA structure generated, all operate to yield the large diversity of mRNAs in AS.

### 1.3.3 Consequences of splicing dysregulation

AS enables expanded gene functionality through increased transcriptomic and proteomic diversity: from the 19,988 human protein-coding genes, 87,814 protein-coding transcripts are produced (*GENCODE - Human Release 40*, 2022; GRCh38.p13). This vital process drives fundamental biological processes including differentiation and proliferation mechanisms in organism development and aging, sex determination, tissue differentiation and disease development. Over 30% of tissue-dependent transcript variations are due to local splicing variations (Vaquero-Garcia *et al.*, 2016).

However, AS can become dysfunctional. This arises either through mutations in cis-elements such as mutations of a splice site, or through changes in the activity or abundance of trans-acting SFs. As mutations in trans-acting splicing machinery affect splicing on a larger scale compared to mutations in a single mutated cis sequence, these mutations are rarer and more lethal (Tazi, Bakkour and Stamm, 2009). Ultimately, the dysfunction of AS results in altered protein isoform ratios, or the production of toxic proteins through dysfunctional intron removal or incorrect exon inclusion.

Unsurprisingly, AS dysfunction is intimately associated with disease: it is linked with approximately 15% of hereditary diseases and is commonly described as a hallmark of cancers in humans (Cui, Cai and Stanley, 2017; Urbanski, Leclair and Anczuków, 2018). Examples of non-cancerous diseases associated with AS dysfunction include in non-alcoholic fatty liver disease (Yufeng Li *et al.*, 2021), ulcerative colitis (Li and Tan, 2021), and autism spectrum disorder (Leung *et al.*, 2023). However, the main diseases associated with dysfunctional splicing are cancers. These cancers seem unrestricted to tissue type, with dysfunctional AS seen in breast, lung, colon, bladder, and blood cells, and form because the isoforms generated through dysfunctional AS favour cancer

development (Malcovati *et al.*, 2011; Anczuków *et al.*, 2012). Examples of this include: facilitation of cell proliferation, as with the generation of the Mnk2-b isoform that promotes growth by phosphorylation of the translation initiation factor eIF4E (Maimon *et al.*, 2014); failure of appropriate tissue differentiation, as with mutations of the U2 snRNP component SF3B1 associated with myelodysplastic syndromes (Malcovati *et al.*, 2011); with cancerous mutations generating toxic products such as a novel isoform of the erythroid lineage transcription factor TAL1 causing dysfunctional erythroid differentiation (Wahl and Lührmann, 2015; Jin *et al.*, 2017); prevention of cell death, as with the generation of the BCL-xL isoform in lymphomas that prevents apoptosis (Boise *et al.*, 1993); promotion of angiogenesis, as with AS of blood vessel forming growth factor VEGFA (Houck *et al.*, 1991); and the enabling of cell invasion and metastatic dissemination, as with the generation of actin nucleation and polymerisation isoforms MENA-INV and MENAΔv6 which have been associated with breast tumours (Tanaka *et al.*, 2014). Despite these associations, it is worth noting that whilst dysfunctional AS is a core and extensive disease mechanism, many non-cancer related diseases have been linked to AS only by association studies, and molecular mechanistic understanding for many examples is lacking.

## *1.4 Current approaches and challenges to studying AS*

RNA-seq, typically the conversion of RNA from cells and tissues into complementary DNA (cDNA) and its subsequent sequencing, is used to investigate differential gene expression and has been widely used in the study of AS. These methods include short-read and long-read sequencing, and the sequencing of bulk and single-cell populations. The advantages and disadvantages of each are detailed below.

### 1.4.1 Bulk short-read RNA-seq

Bulk RNA-seq replaced microarray methods in the late 2000s as a method of quantifying transcriptomes from a population of cells or tissue (Kuksin *et al.*, 2021). In this method, bulk short-read RNA-seq libraries are prepared by the lysis of cells or tissues and RNA extraction. mRNA is then enriched by poly(A) isolation and ribosomal RNA depletion, with the purified mRNA then reverse-transcribed into double-stranded cDNA which is then fragmented, and sequencing adaptors are then ligated onto each fragment. The processed cDNA is then amplified by PCR and sequenced using next generation sequencing.

Bulk RNA-seq has several advantages over single-cell RNA-seq. One advantage is in the simplicity of the collection and processing of RNA. Bulk RNA-seq can be performed on samples obtained from surgical biopsies preserved in formalin or freshly frozen, meaning that all samples can be collected and processed simultaneously without time constraints. Conversely, single-cell RNA-seq requires viable cells to allow robust cell isolation meaning samples cannot be fixed or frozen, and the different sensitivities of different cell types to isolation may affect their representation in final datasets (Kuksin *et al.*, 2021). Single-cell RNA sequencing requires additional steps to allow the identification of which transcripts belong to which cell, whereas bulk RNA-seq is substantially simpler in that the only identifiers necessary are those to differentiate between samples. Another advantage of bulk RNA-seq over single-cell RNA seq is there is less impact of the "dropout effect", wherein transcripts are falsely undetected ("zero inflation"). Due to the technical limitation of sampling minute levels of RNA content in single cells compared to a population of cells as in bulk RNA-seq, only 10% of total genes are effectively measured in single-cell RNA-seq experiments (Kolodziejczyk *et al.*, 2015). This means that for total transcriptome coverage, bulk RNA-seq is far more powerful than single-cell methods. On a more practical side, bulk RNA-seq can be simpler to analyse and is sufficient when looking for overarching effects. In addition, bulk RNA-seq is a cheaper method by a factor of six (Kuksin *et al.*, 2021).

In the context of measuring AS, one major advantage that bulk RNA-seq has over single-cell methods is in its comparatively low number of dropouts and in the greater sequencing depth (Kuksin *et al.*, 2021). This means that bulk methods have greater ability to detect rare isoforms generated through AS, albeit with the major limitation of struggling to deconvolve the subpopulation of cells in which alternative isoforms are present and being ultimately contingent on a sufficient length of transcript being sequenced to allow for accurate distinguishing between isoforms (discussed in the short-read *vs* long-read section below). This makes bulk methods useful in *in vitro* experimental systems when attempting to look at the consequence on AS by different treatments, but less beneficial when looking to delineate the complexities of heterogeneous populations, such as in haematopoiesis characterisation.

### 1.4.2 Single-cell short-read RNA-seq

Single-cell RNA-seq (scRNA-seq) enables the high-throughput profiling of gene expression in large numbers of individual cells, allowing insight into cellular heterogeneity. This is particularly important for the identification and study of cellular

subtypes within phenotypically homogeneous tissues. Single-cell resolution data enables the identification of subtype-specific cellular activities and the subsequent mapping of developmental and disease trajectories.

There have been many different approaches to single-cell library preparation. Two popular methods are Drop-Seq, commercialised in the 10x Genomics Chromium platform (Zheng *et al.*, 2017) and the Smart-Seq2 library preparation protocol (Picelli *et al.*, 2013).

10x Genomics provides solutions for single-cell sequencing of hundreds to tens of thousands of cells per sample which utilise a droplet system for single-cell isolation. Single cells are individually encapsulated in microscopic aqueous droplets in oil emulsion containing enzymes and a Gel Bead coated in barcoded oligonucleotides (oligos). The oligos for the standard 3' RNA-seq protocol each contain a poly(dT) sequence complementary to the poly(A) tail of mRNA molecules, a barcode unique to each Gel Bead, a barcode unique to each oligo and a primer for sequencing. Following encapsulation, the cells lyse and the RNA is released. The oligos of the Gel Bead capture the RNAs by their 3' ends, tagging them with the cell- and molecule-specific barcodes (Zheng *et al.*, 2017).

Smart-Seq2 relies on single-cell isolation via fluorescence-activated cell sorting (FACS), often into 96-well plates before library preparation. Unlike the 10x approach, Smart-Seq2 generates full-length cDNA (up to 10 kb) and therefore improves the read-coverage of both ends of the transcripts, increases the length, and yield of cDNA and ameliorates the 3' capture bias which often occurs in data from bulk poly(A) isolation RNA-seq and limits AS detection.

The major benefit of single cell RNA-seq is that it allows the deconvolution of transcriptomic results by cell, theoretically making it a particularly powerful tool in the field of haematopoiesis given the complexities of cell commitment (Fig.2). However, the major limitation of the method is that the datasets are less complete and are more noisy than bulk methods due to the high prevalence of dropouts and poor sequencing depth. 10x Genomics Chromium library preparation does not allow the sequencing of full-length transcripts, introducing the can be processed through the Chromium Controller simultaneously (Stegle, Teichmann and Marioni, 2015; Hicks *et al.*, 2018). As for Smart-Seq, libraries are often sequenced using short-read sequencers, again introducing the limitations of short-read sequencing. Furthermore, both methods only capture polyadenylated RNAs, ignoring other functional RNAs including long non-coding RNAs.

These factors are particularly problematic in the field of AS where it is necessary to accurately measure the presence of rare events.

### 1.4.3 Long-read RNA-seq

Despite the different advantages of bulk and single-cell RNA-seq short-read sequencing, they are both limited in the same way in the context of AS. Specifically, due to the short sequencing length, using short-read sequencing can make it impossible to determine the diversity of isoforms present in a sample. Attempts to identify AS in short-read datasets focus on identification of splice junctions, as seen in software such as SplAdder and Bisbee (Kahles *et al.*, 2016; Halperin *et al.*, 2021), but these fail to identify complex events that involve more than one type of alteration and cannot distinguish between differentially spliced isoforms that use the same splice junction in samples where multiple isoforms are simultaneously present. To this end, long-read transcript sequencing offers some solutions to these problems.

Full-length cDNAs are produced during RNA-seq library preparation, albeit sometimes as an intermediate. Long-read sequencing (also known as third-generation sequencing) enables direct sequencing of the full-length cDNA without fragmentation and PCR. Pacific Biosciences (PacBio) is one of the two main long-read sequencing technology companies. It uses single-molecule real-time (SMRT) sequencing to generate up to 4 million high quality reads typically of 10kb and above in length using its Sequel sequencers. SMRT sequencing is performed in SMRT cells which contain tens of thousands of zero-mode waveguides (ZMWs). These well-like structures facilitate real-time sequencing as a DNA-template-polymerase complex is immobilised at the bottom of each well and phospholinked fluorophore-labelled nucleotides are introduced into the ZMW. As sequential light pulses are produced, each nucleotide in the template is identified by the attached fluorophore. The phosphate chain is then cleaved and the fluorophore is released, allowing another phospholinked nucleotide to be incorporated and identified. The other widely used long-read sequencing platform was developed by Oxford Nanopore Technologies (ONT) and uses thousands of synthetic polymer membranes studded with nanopores, through which single DNA molecules can pass. The DNA is directed towards the nanopores by an electrical potential until one molecule is captured within a pore. Then a motor protein pulls the DNA through the pore at a consistent speed, which disrupts the electrical current by different amounts depending on which nucleotide base is passing through the nanopore at that moment. For each nanopore, the current perturbations at each base are recorded and a recurrent neural

network (RNN) converts them into a sequence of bases to make up each read. The ONT MinION can generate reads of up to 4 Mb in length in higher quantities in comparison to PacBio systems, but the MinION has a higher sequencing error rate.

However, there are limitations with current long-read sequencing technologies. Whilst they facilitate the easy assembly of transcripts and clear isoform identification, one problem is that there is substantially lower throughput. This results in fewer reads per cell, and a lower power in detecting low to moderately expressed transcripts. Improvements have been made to both the PacBio and ONT technologies to target these limitations by circular sequencing methods, whereby circularised DNA is sequenced in repeated passes, improving sequencing accuracy. Further advancement in long read sequencing will likely decrease cost, increase throughput, and increase accuracy.

### 1.4.4 Combined approaches

The disadvantages of short-read sequencing are complemented by the advantages of long-read sequencing and vice versa. Using combined approaches of these two technologies allows for more detailed and accurate characterisations of AS and other biological processes. Novel splice sites discovered using long-read sequencing can be validated using the high accuracy and coverage of short-read sequencing. Creating accurate full-length isoform datasets in this way fills gaps in gene annotations in public databases and enables the analysis of all AS events including exon-skipping, intron retention, alternative 3' SSs and 5' SSs. Complex isoforms generated by multiple splicing events can also be analysed with a combined approach as they may not be detected in short-read RNA-seq but are when sequencing full-length transcripts. Long-read sequencing is good at identifying novel isoforms but generally lacks sufficient accurate coverage to characterise lowly expressed genes; this is also a problem for some short-read scRNA-seq methods, including 10x. In such cases, the desired isoforms can be enriched for during library preparation before sequencing.

RNA-seq is excellent at providing transcript-level insights into expression of genes and potential function within a cell but omits information within the context of the resulting proteins. AS of pre-mRNA transcripts, post-translational modifications to proteins and interactions between proteins yields a vast complexity in the proteome of a cell and thus in its function. Transcriptomics fails to capture much of this complexity. Approaches combining transcriptomic and proteomic data are beginning to appear and will become more prevalent in AS research in the future to validate the existing knowledge and

facilitate new discoveries. Such approaches include iCLIP (König *et al.*, 2010) and proteogenomics (Miller *et al.*, 2022).

## *1.5 Haematopoiesis*

The closed circulatory system of mammals connects different tissues via a mutual route of metabolite and waste exchange. The circulation fluid, blood, contains dissolved nutrients such as glucose and amino acids, and cells that perform specialised functions. Such functions include the transport of oxygen and carbon dioxide between tissues and the lungs, clotting when blood vessels are damaged to avoid infection and loss of blood, and innate and adaptive immune responses to destroy pathogens, debris, dead cells, or tumours to keep the organism healthy. These blood cells are formed constantly throughout the life of an organism in the bone marrow in a process termed haematopoiesis.

### 1.5.1 Haematopoietic lineages

The classical description of haematopoiesis is as a hierarchical process of differentiation, whereby multipotent self-renewing haematopoietic stem cells (HSCs) differentiate into four main blood lineages. These are: the erythroid lineage, which generates erythrocytes; the megakaryocyte lineage, which generates platelets; the myeloid lineage, which functions mainly in innate immunity by generation of granulocytes and monocytes; and the lymphoid lineage, which functions mainly in adaptive immunity by generation of B and T lymphocytes (Pucella, Upadhaya and Reizis, 2020; Fig.2).

Differentiation from bone-marrow located multipotent HSCs to terminally differentiated circulating specialised blood cells happens via a series of progenitor cell intermediates which restrict fate and commit into mature lineages (Rankin and Sakamoto, 2018). Cells from the same lineage display similarities in their transcriptomic profiles and the presentation of certain cell-surface proteins. By measuring the transcriptome of each cell by single-cell RNA-sequencing (described above) or the abundance of cell surface proteins by antibody labelling and flow cytometry, it is possible to quantify the relative amounts of different blood lineages, how treatments affect haematopoietic differentiation and the generation of cell fate maps to trace the haematological history of a terminally differentiated cell.

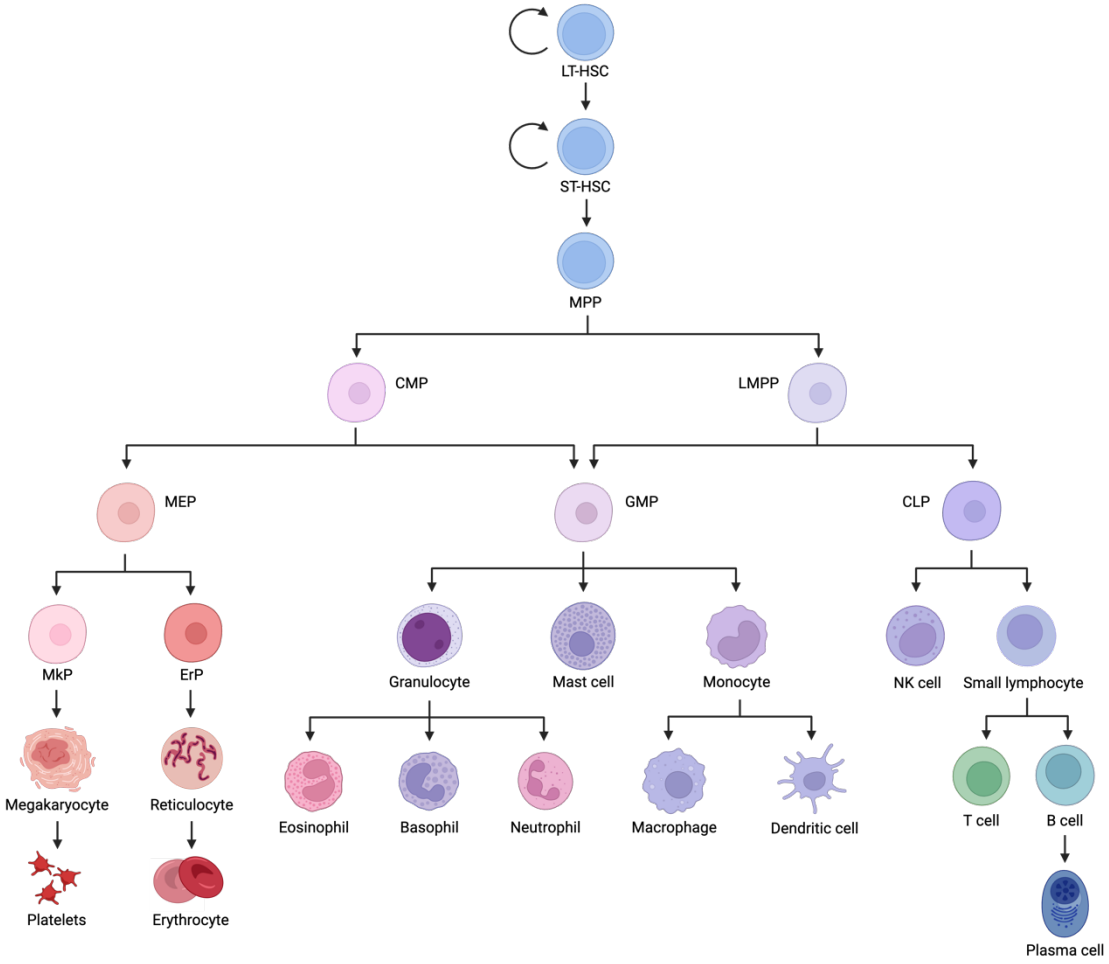


***Figure 2 - Hierarchical organisation of haematopoietic cell lineage commitment.*** *Tree diagram showing the possible cell types a haematopoietic stem cell has the ability to differentiate into and the intermediate progenitor cells for each lineage. Created in BioRender. LT-HSC = long-term HSC, ST-HSC = short-term HSC, MPP = multipotent progenitor, CMP = common myeloid progenitor, LMPP = lymphoid-primed MPP, MEP = megakaryocyte-erythroid progenitor, GMP = granulocyte-macrophage progenitor, CLP = common lymphoid progenitor, MkP = megakaryocyte progenitor, ErP = erythroid progenitor.*

At the top of the haematopoietic hierarchy are the multipotent HSCs which are the least common blood cell type, numbering only 1 per 1,000,000 cells in bone marrow (Doulatov *et al.*, 2012). These are subdivided into the long-term repopulating HSCs (LT-HSCs) which are able to reconstitute the haematopoietic system of irradiated mice via their ability to differentiate and self-renew, and short-term HSCs, which are formed from LT-HSCs and give rise to multipotent progenitor cells (MPPs). MPPs generate both the myeloid lineages and the lymphoid lineages via the generation of oligopotent cells, common myeloid progenitors (CMPs) or LMPPs (lymphoid-primed MPPs). LMPPs have lympho-myeloid potential as they give rise to common lymphoid progenitors (CLPs) as well as granulocyte-monocyte progenitors (GMPs). Progenitors can be defined from blood fractions by the cell surface proteins they display, with CMPs being negative in Lin and Sca but positive for Kit ($Lin^-Sca1^-Kit^+$), and CLPs being negative in Lin, low in Sca and Kit, and positive for IL7Rα ($Lin^-Sca1^{low}Kit^{low}IL7R\alpha^+$).

CMPs differentiate into two lineages: the megakaryocyte-erythroid progenitors (MEPs) lineage which is defined $FcγRII/III^{low}CD34^-$ and gives rise to the lineage-restricted progenitors of megakaryocytes and reticulocytes; and GMPs which are defined $FcγRII/III^{high}CD34^+$ and give rise to lineage-restricted granulocytes and monocytes (Fiedler and Brunner, 2012). In turn, megakaryocytes terminally differentiate into platelets that function in blood clotting, reticulocytes into erythrocytes that function in oxygen transport, granulocytes into eosinophils, basophils and neutrophils, mast cells that function in the innate immune response via release of toxic granules and inflammation, and monocytes into the professional antigen-presenting cells, macrophages and dendritic cells (Table 1).

The other major set of lineages in haematopoiesis derive from CLPs which differentiate into either cytotoxic Natural Killer (NK) cells, or small lymphocyte progenitors which differentiate into the cytotoxic, helper, or regulatory T cells, or B cells and plasma cells which produce antibodies. Further functions of each of these terminally differentiated blood cells are detailed in Table 1.

***Table 1 - A table to show the function and immediate precursors of terminally differentiated lineages.***

| Terminal Differentiated Lineage | Precursor | Function |
|---|---|---|
| Platelets | Megakaryocytes | • Blood clotting |
| Erythrocytes | Reticulocytes | • Oxygen and Carbon Dioxide Transport |
| Eosinophils | Granulocytes | • Release of toxic granule proteins in immune defence<br>• Multicellular pathogen defence |
| Basophils | Granulocytes | • Inflammation reactions<br>• Immune response coordination |
| Neutrophils | Granulocytes | • Cytokine release at sites of infection<br>• Phagocytosis of microbes<br>• Toxic granule release<br>• Extracellular microbial trap formation (NETs) |
| Mast cells | Granulocyte-macrophage progenitor | • Inflammation reactions |
| Macrophages | Monocytes | • Phagocytosis of dying or dead cells and cellular debris<br>• Phagocytosis of pathogens in innate immunity<br>• Proinflammatory cytokine secretion<br>• Antigen presentation for T cells |
| Dendritic cells | Monocytes/Lymphoid primed multipotent progenitors | • Phagocytosis of pathogens for antigen presentation to T cells |
| Natural killer (NK) cells | Lymphoid primed multipotent progenitors | • Cytotoxic granule release for killing of virally infected cells and intracellular pathogens<br>• Tumour surveillance<br>• Clearance of senescent cells |
| T cells | Small lymphocytes | *CD4+ helper T cells:*<br>• Maturation of B-cells into plasma cells and memory cells<br>• Activation of cytotoxic T cells<br>*CD8+ cytotoxic T cells:*<br>• Destruction of virally infected cells and tumour cells<br>*Memory T cells:*<br>• Immune memory<br>*Regulatory CD4+ T cells:*<br>• Immune suppression |
| Plasma cells | B cells | • Antibody production |

## 1.5.2 Mechanisms of lineage commitment

The commitment of a progenitor cell to differentiate into a particular cell type necessitates the activation of lineage-specific pathways and the silencing of other differentiation options. This commitment occurs through an interplay between extrinsic and intrinsic factors to lead to a remodelling of the transcriptome (Fiedler and Brunner, 2012). Extrinsic factors, such as cytokines or nutrient concentrations, are extracellular ligands that stimulate responses by binding to receptors displayed on the cell surface of haematopoietic cells (Rankin and Sakamoto, 2018). Binding triggers the receptor to either dimerise (e.g. CSF-receptor; Li and Stanley, 1991), oligomerise with a signalling

subunit (e.g. IL-6 receptor; Uciechowski and Dempke, 2020), or change in conformation (e.g. EPO receptor; Lu, Gross and Lodish, 2006), triggering a kinase cascade that affects transcription. Because kinase cascades amplify the effect of cytokine-receptor activation, and because the binding affinities between cytokines and their receptors are in the picomolar range, differentiation cytokines are kept at low levels in blood serum and require only a small change in concentration to drive large-scale directed differentiation of certain lineages (Fiedler and Brunner, 2012). Some cytokines act broadly on a range of lineages, with one example being interleukin-7, which drives the differentiation of HSCs into CLPs, and CLPs into B-cell progenitors and T-cell and NK progenitors (Akashi *et al.*, 1997; Appasamy, 1999; Cool *et al.*, 2020; Chen *et al.*, 2021). Other cytokines specifically drive the differentiation of only one lineage, with classic examples being thrombopoietin (TPO) that drives platelet formation and erythropoietin (EPO) that drives erythroid differentiation, the latter being famous for its common misuse in professional cycling to boost oxygen carrying capacity by increasing red blood cell numbers.

Intrinsic differentiation factors conversely lie within each lineage and are often the downstream targets of extrinsic factors (Fiedler and Brunner, 2012). These typically exist as gene regulatory networks driven by master transcription factors which are specific for many genes, many of which may be transcription factors themselves. The expression of these master transcription factors thus leads to a cascade of transcriptional change, ultimately driving large-scale alterations in the proteome of differentiating blood cells to drive commitment. There are numerous examples of these master transcriptional regulators. One is the transcription factor CCAAT-enhancer binding protein α (C/EBPα), the expression of which drives the transition of HSCs to MPPs and oligopotent progenitors. C/EBPα acts as an antimitotic factor to block self-renewal (Timchenko *et al.*, 1996) and upregulates transcription factor PU.1 that triggers CMP and CLP generation (Ohlsson *et al.*, 2016). Another example is GATA-1 whose expression drives commitment into erythroid-megakaryocyte lineages (Iwasaki *et al.*, 2003). Germ-line ablation of GATA1 results in embryonic lethality due to a block in erythroid differentiation resulting in fatal anaemia (Fujiwara *et al.*, 1996), and deletion of an enhancer in the *GATA1* locus led to a block in megakaryocyte differentiation, causing mice to suffer from thrombocytopenia (Shivdasani *et al.*, 1997).

Transcription factors must be able to bind to DNA in order to initiate or suppress transcription. The condensing of DNA between forms accessible and inaccessible to RNA polymerase II is controlled by epigenetic DNA modifications, and it is therefore unsurprising that these also have a crucial role in determining lineage commitment (Rankin and Sakamoto, 2018). There are three main types of modification: DNA

methylation, histone modification, and chromatin remodelling. Firstly, DNA methylation at CpG dinucleotides typically blocks transcription by either directly obstructing the binding of transcription factors to DNA, or by recruitment of epigenetic modifiers to a particular gene locus (Greenberg and Bourc'his, 2019). DNA methylation is crucial in HSCs to maintain their ability to self-renew (Celik, Kramer and Challen, 2016), with evidence being that the removal of DNA methyltransferase Dnmt1 that functions to copy DNA methylation status during replication leads to a loss of self-renewal in HSCs (Trowbridge *et al.*, 2009). Secondly, histone modifications are post-translational modifications to histones such as acetylation and phosphorylation which change the affinity of nucleosomes and DNA to either have activating and repressing properties (Vu, Luciani and Nimer, 2013). One example of their relevance to haematopoiesis is in the promotion of lymphoid commitment via histone acetyltransferases (HATs) and histone deacetylases (HDACs). The HAT GCN5 regulates B-cell differentiation by modulating the transcription factors Btk and Syk (Okkenhaug *et al.*, 2002; Kikuchi *et al.*, 2011), and the HAT p300 and HDAC SIRT1 regulate Treg differentiation by controlling the expression of *Foxp3* (van Loosdregt *et al.*, 2010). Finally, chromatin remodelling complexes are also important in lineage commitment. These multiprotein complexes function to change nucleosome location or conformation. These have been implicated in the transcriptional activation of genes, including during LMPP to CLP transition and in B-cell differentiation, with one example being the recruitment of the SWI/SNF complex to the upstream enhancer of B-cell receptor CD19 to facilitate its transcription and trigger B-cell commitment (Walter, Bonifer and Tagoh, 2008).

The stability of lineage-defining transcripts also appears to play a crucial role in haematopoietic commitment. There has been increasing evidence that microRNAs (miRNA), small single-stranded non-coding RNAs of 21-23 nucleotides, are used extensively during haematopoiesis to control differentiation (Fiedler and Brunner, 2012). These molecules work to silence genes via the destruction of mRNA by cleavage of mRNA or the destabilisation of mRNA by shortening its polyA tail, thus acting as post-transcriptional differentiation control mechanisms in the context of haematopoiesis (Fabian, Sonenberg and Filipowicz, 2010). The earliest evidence of the relevance of miRNAs was the lineage-specific ablation of miRNA biogenesis enzyme Dicer in lymphoid progenitors in mice leading to severe defects in B and T cell development (Cobb *et al.*, 2005; Koralov *et al.*, 2008), and that miRNA transcription varies during haematopoiesis (Chen *et al.*, 2014). The specific functions of these miRNAs has been gradually revealed, and examples include the expression of miR-126 which degrades HOAX9 transcripts to allow HSC to MPP transition (Felli *et al.*, 2005), the decreased

expression of miR-24 which allows expression of activin type I receptor to promote erythropoiesis (Wang *et al.*, 2008), and expression of miR-150 in MEPs which drive megakaryocyte differentiation by targeting transcription factor c-Myb mRNA (Lu *et al.*, 2008).

An emerging additional mechanism of haematopoiesis is alternative splicing (AS). As described above, this allows the formation of multiple protein isoforms from the same gene via the inclusion or exclusion of different exons in the final mRNA. These isoforms can have different and even opposing functions via the inclusion or exclusion of different protein domains, and in the context of haematopoiesis typically link to the regulation of differentiation-associated transcription factors or proteins that control the stability of differentiation-associated transcripts (Gao, Vasic and Halene, 2018; Chen and Abdel-Wahab, 2021). The role of AS in driving blood lineage commitment is explored in detail in the introduction of Chapter 3.

### *1.5.3 MPL* and *RBM15 in haematopoiesis*

Utilising a mouse model, the Macaulay lab demonstrated that key haematopoietic genes are alternatively spliced (Mincarelli *et al.*, 2023). This includes *Mpl* which encodes the receptor for thrombopoietin, the hormone which stimulates megakaryocyte and platelet production (thrombopoietin response). *Mpl* is an example of a gene which is alternatively spliced into isoforms with antagonistic functions. Lacking exons 9 and 10, *Mpl-204* produces a truncated protein missing a transmembrane domain, which inhibits the normal Mpl signalling by downregulating expression of the primary isoform, *Mpl-201* (*Mpl-FL*), which encodes the full-length transmembrane Mpl receptor (Coers, Ranft and Skoda, 2004). The AS of *Mpl* pre-mRNA to produce the truncated *Mpl-204* (*Mpl-TR*) isoform depends on a protein called RBM15, which negatively regulates the thrombopoietin response through the mediation of N6-methyladenosine (m6A) methylation and potentially histone acetylation and methylation at the *Mpl* locus (Xiao *et al.*, 2015).

RBM15 is an RNA-binding protein (RBP) that has a diverse set of functions in mRNA processing including not only N6-adenosine methylation (Patil *et al.*, 2016), but also mRNA export (Lindtner *et al.*, 2006; Zolotukhin *et al.*, 2009), histone association (Lee and Skalnik, 2012), and recruitment of specific SFs to introns (Zhang *et al.*, 2015), with each of these functions explored in detail in the introduction to Chapter 2. In addition to the evidence linking RBM15 to haematopoiesis via its involvement with Mpl, a link between RBM15 and haematopoiesis has been clearly established from associative

studies of blood diseases. For example, *RBM15* was found to be mutated or vary in copy number in 25 cancer types, with poor prognosis associated with its overexpression, and is substantially upregulated in multiple myeloma cells, the second most common haematopoietic malignancy (Bai, Xu and Chen, 2021; Zhao *et al.*, 2022). In addition, *RBM15* is the fusion partner of the myocardin gene *MKL1* in the t(1;22) translocation found in acute megakaryoblastic leukaemia which comprises 10% of childhood acute myeloid leukaemia (AML) leading to a median survival of 8 months (Baruchel *et al.*, 1991; Carroll *et al.*, 1991; Lion *et al.*, 1992; Ma *et al.*, 2001). Experimental evidence has validated the associative data between RBM15 and haematopoiesis. Homozygous knockout of *Rbm15* in mice led to embryonic lethality at E9.5, and its conditional knockout caused a loss of peripheral B cells due to a block in pro/pre-B differentiation, a myeloid and megakaryocytic expansion in the spleen and bone marrow, an increase in the haematopoietic stem cell compartment, and a shift in progenitor fate towards granulocyte differentiation (Raffel *et al.*, 2007). Furthermore, RBM15 was found to decrease in abundance when haematopoietic cell lines EML and 32DWT18 cells differentiate into promyelocytes and more mature neutrophils, and artificial RBM15 depletion by shRNA in 32DWT18 cells replicated this phenotype with their decreased proliferation and differentiation into metamyelocytes and maturing neutrophils (Ma *et al.*, 2007). Ma *et al.* (2007) showed this association was driven by the effect of RBM15 on Notch signalling, interesting in the context of haematopoiesis given that Notch receptors and ligands are widely expressed in the haematopoietic system and Notch signalling facilitates bone marrow expansion and inhibits myeloid differentiation (Milner *et al.*, 1996; Bigas, Martin and Milner, 1998; Duncan *et al.*, 2005). However, despite the clear importance of RBM15 in haematopoiesis, the association of RBM15 and AS as shown in the case of Mpl, and the importance of AS in haematopoiesis, a full understanding of the function of RBM15 in the intersection between AS and haematopoiesis remains to be elucidated.

# PROJECT AIMS

This MRes project builds upon preliminary work by the Macaulay lab to address questions of the function and mechanism of AS in haematopoiesis. This project is separated into two main aims.

The first project aim was to characterise the transcriptomic consequences resulting from depletion of the protein RBM15 in cell culture models. RBM15 is an RBP that has functions in haematopoiesis by maintaining long-term HSCs and mediating the differentiation of megakaryocytes and B-cells, with mutations in *RBM15* resulting in blood cancers such as chronic and acute myeloid leukaemias. RBM15 is thought to function in AS through affecting the efficiency of RNA splicing and processing as part of the WMM complex that affects N6-mRNA methylation. Despite these insights, one clear gap in our understanding is a full description of which transcripts are affected by RBM15. This is addressed here by the generation of *RBM15* knockout cells and bulk RNA-seq measurement of the transcriptomic consequences, with data analysis paying particular attention to differences in the expression of genes with known relation to haematopoiesis.

The second project aim was to explore a single-cell short-read haematopoietic stem cell dataset to identify cell-type specific patterns of RBP and SF expression. This dataset from the Macaulay lab (Mincarelli *et al.*, 2023) isolated lineage negative, cKit/Cd117 (LK) cell fractions which contain haematopoietic stem and progenitor cells from young and adult mice by FACS and performed parallel short- and long-read sequencing. However, this dataset had previously not been analysed specifically in the context of searching for changes in factors relevant for AS. Doing so would provide further understanding of how changes in AS impact upon haematopoietic commitment. If time allowed, the key identified genes would then be depleted individually in cell culture models, and the transcriptomic consequences measured by RNA-seq to understand their individual importance in haematopoiesis.

# CHAPTER 2: *RBM15* KNOCKOUT

## *2.1 Introduction*

*RBM15* (RNA-binding protein 15), also known as OTT1 (One-twenty-two protein 1) is a gene located on chromosome 1 that encodes the 977 amino acid protein RBM15. The protein belongs to the Spen family of proteins which also includes SMRT/HDAC1 associated repressor protein (SHARP), and RBM15B, a paralogue of RBM15. This protein family is characterised by the presence of a series of N-terminal RNA recognition motifs (RRMs), of which RBM15 has three, and a C-terminal Spen orthologue and paralogue C-terminal (SPOC) domain of 15-20 kDa that allows binding to phosphoserines associated with transcription machinery and co- and post-transcriptional regulators (Appel *et al.*, 2023).

RBM15 has been reported to interact with a variety of proteins giving it a diverse set of mRNA processing functions in cells. Firstly, it is a component of the MACOM and MACOM-like WTAP subcomplexes alongside WTAP, ZC3H13, CBLL1, and VIRMA, which combines with the MAC subcomplex of METTL3 and METTL14 to make the N6-methyltransferase complex (MTC; Patil *et al.*, 2016; Knuckles *et al.*, 2018). N6-methyladenonsine (m6A) accounts for the most abundant mRNA internal modifications in higher eukaryotes (Dominissini *et al.*, 2012; Zaccara, Ries and Jaffrey, 2019), with dedicated complexes for writing and erasing m6A modifications, and acts to regulate RNA fate by altering their stability, splicing and translation (Huang *et al.*, 2021). In particular, RBM15 has been shown to function in the MTC by facilitating binding to specific mRNA regions with U-rich consensus motifs (Patil *et al.*, 2016). Whilst the full impact of RBM15 regulating MTC recruitment to specific mRNAs is not understood, it is known to be necessary in the correct m6A formation on the X chromosome transcriptional silencer XIST which, if disrupted by RBM15 depletion, leads to impaired XIST-mediated gene silencing (Patil *et al.*, 2016). Secondly, investigations into the molecular pathology underlying the association of acute megakaryoblastic leukaemia (AMKL) with a RBM15-MKL1 fusion protein led to the discovery of RBM15 as able to interact with the LPDSD motif of Setd1b histone H3-Lys4 methyltransferase (KMT2G) via its SPOC domain (Lee and Skalnik, 2012), although the wider transcriptional consequences of this interaction are not understood. Thirdly, RBM15 is strongly implicated in mRNA export via its interaction with the crucial mRNA exportins NXF1 and DBP5 (Lindtner *et al.*, 2006; Zolotukhin *et al.*, 2009). In a model proposed by Zolotukhin *et al.* (2009), RBM15 acts as a core component of mRNA export, as silencing of RBM15

led to a cytoplasmic depletion and nuclear accumulation of general mRNAs and is thought to operate at nuclear pore complexes by bridging the interaction between the export receptor NXF1 that guides messenger ribonucleoproteins to the nuclear pore complexes and DBP5 which allows NXF1 dissociation during pore translocation. The association of RBM15 and mRNA export seems particularly convincing, especially given that RBM15 has been shown to interact with the viral protein EB2 of Epstein-Barr virus to allow the efficient nuclear export of viral mRNAs (Hiriart *et al.*, 2005). Finally, and of relevance to AS, RBM15 has been shown to interact with splicing factor SF3B1, which is a core component of the RNA splicing A complex that binds to BPS (Zhang *et al.*, 2015). Investigations of acute megakaryocytic leukaemia cell lines by Zhang *et al.* (2015) revealed that the aberrant overexpression of PRMT1 led to downregulation of RBM15 by trigging its ubiquitylation and proteasomal degradation, and in turn, this led to decreased presence of RBM15 at pre-messenger RNA intronic regions of genes crucial for megakaryopoiesis such as *GATA1*. Furthermore, it was shown that RBM15 recruits SF3B1 to these regions, thereby suggesting that RBM15 may function more broadly as a regulator of AS by mediating the recruitment of SFs.

As detailed in the main introduction, the connection between RBM15 and haematopoiesis is clear. *RBM15* was found to be mutated or vary in copy number in 25 cancer types and was found to be substantially upregulated in myeloid leukaemias (Bai, Xu and Chen, 2021). Furthermore, studies in conditional *RBM15* deletion mouse models have shown that its removal results in dramatic effects on haematopoietic commitment. Finally, work from the Macaulay lab showed that the AS of key haematopoiesis gene *Mpl,* which controls megakaryocyte production, is alternatively spliced by RBM15. Importantly however, the mechanistic link between RBM15 function, AS activity (either through regulation of SF expression or through direct association with SFs), and haematopoiesis remains largely unexplored.

Based on previous work in the Macaulay lab and the described functions of RBM15 in mRNA processing and AS, it was hypothesised that a deletion of *RBM15* would result in widespread impacts on gene and isoform expression. Identifying these genes would, in turn, provide candidates for future experiments concerning RBM15 and its function in blood stem and progenitor cells. This project begins this line of enquiry. RBM15 was depleted using CRISPR-Cas9 gene knockout in the myelogenous leukaemia cell lines HAP1 and K562, and the transcriptomic changes measured by bulk short-read RNA-sequencing. Particular attention during data analysis was given to highlighting changes in transcriptome abundance of genes related to AS.

### 2.1.1 CRISPR-Cas9

This project used CRISPR-Cas9 to deplete *RBM15*. CRISPR-Cas9 is a gene editing method often used to explore the consequences of the deletion of genes, known as knockout experiments (Fig.3). Cas9 is an endonuclease which precisely cuts DNA to form a double-stranded break when it is guided to specific locations by a guide RNA (gRNA); the locations are determined based on the sequence of the gRNA. Once the gRNA (with Cas9 bound) has bound to the DNA by complementary base pairing, the Cas9 detects a neighbouring protospacer adjacent motif (PAM) sequence and cuts three bases upstream of the PAM. The Cas9 used in knockout experiments was derived from the bacteria *Streptococcus pyogenes* and its PAM sequence is denoted by 5'-NGG-3'. This double-stranded break promotes low-fidelity repair by non-homologous end joining (NHEJ) which fixes the damage but also results in insertions or deletions (indels) at the cut-site. Indels of a number of bases indivisible by three disrupt the reading frame as well as the sequence of DNA. This generates transcripts that have premature termination codons resulting in transcribed mRNA to be degraded by nonsense mediated decay, resulting in no mRNA being present for protein production.

There are multiple methods of delivering the CRISPR-Cas9 system into cells. One method is to nucleofect cells with ribonucleoproteins (RNPs) in which the single guide RNA (sgRNA) is already bound to the Cas9. This method typically results in high editing efficiency and is particularly useful in cells which are difficult to transfect by chemical methods, such as primary cells, but is more complex than other methods to prepare. Alternatively, the CRISPR-Cas9 components can be transfected into cells encoded on plasmids, with the sgRNA expressed from a U6 promoter either from the same plasmids as the Cas9, or from its own vector. This method is the simplest to perform for simple perturbations such as single gene knockouts as it typically requires the simple cloning of a sgRNA into a plasmid already encoding Cas9, the U6 promoter and the gRNA scaffold. However, this method suffers from reduced efficiency because not all cells will be transfected, the expression of the Cas9 is transient, and the editing response is delayed compared to RNP nucleofection as it relies on plasmid DNA being transcribed and translated before editing can occur. A third method, and the approach used in this project, was to use cells already stably expressing Cas9 and transiently transfected pre-synthesised gRNAs. In this method, Cas9 is knocked into the genome of the cell line in a 'safe harbour' locus such as the AAVS1 locus, meaning it is constitutively expressed, but unable to function without the adding of gRNAs. These gRNAs are synthesised already complete with their scaffold sequence and can be transfected directly into cells

using reagents such as DharmaFECT, Lipofectamine 3000 or RNAiMAX. The major benefits of this method are that it is highly effective as all cells express the Cas9 and the small RNA sequence transfection is highly efficient; it is simple to perform given that most cell lines can be purchased already with Cas9 knocked in and the gRNAs can be synthesised and easily transfected; and there is minimal lag time between gRNA delivery and gene editing because all components are already present.

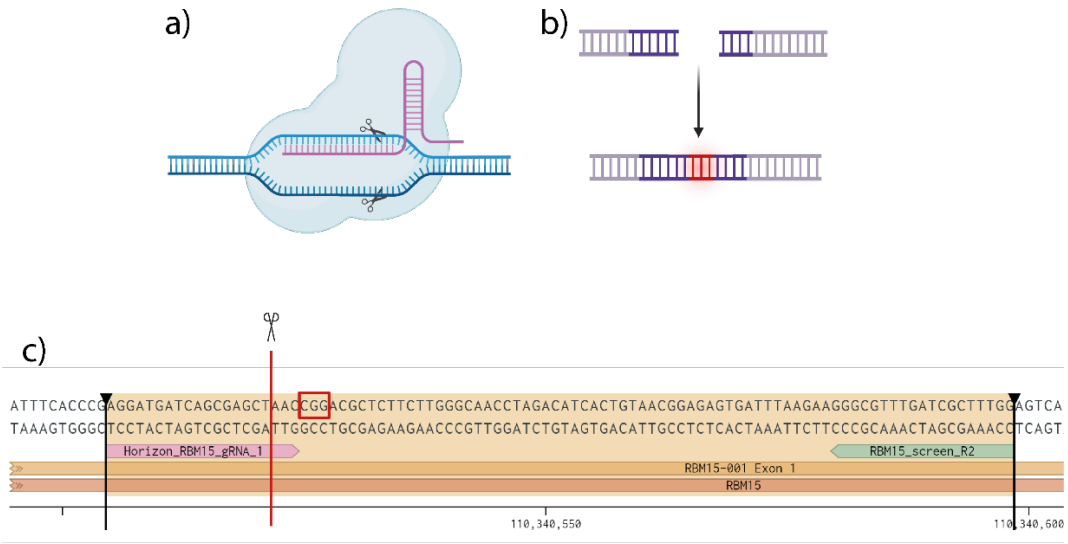


***Figure 3 - The process of gene editing by CRISPR-Cas9.*** *a) Image from BioRender. The sgRNA (pink) is bound to by Cas9 nuclease (grey) and guides Cas9 to the target sequence by complementary base-pairing. Cas9 recognises the PAM sequence NGG and makes a double-stranded break three bases upstream. This type of break promotes low-fidelity repair by non-homologous end-joining (among other mechanisms) which commonly fixes the break but leaves insertions or deletions (indels) of a few bases at the cut-site. The indels alter the DNA sequence and often the reading frame, if the indel is not that of a multiple of 3 bases; a shifted reading frame can lead to a premature stop codon downstream which can knock out the function of the affected gene. b) Image from BioRender. Depiction of a double-stranded break which has been repaired but with the insertion of a few bases. c) Screenshot from Benchling illustrating the location of the PAM (red box, CGG), gRNA target sequence (pink) and the cut-site (red line) for the amplicon generated to assess the editing of RBM15 by transfection with gRNA_1. The orange selected region shows the region of interest when analysing the Sanger traces: the sequencing would begin from the primer (green) and continue leftwards. Any evidence of editing in this case would be seen by signal alterations (mismatches, indels) in the Sanger trace immediately to the left of the cut-site.*

It is important to note that CRISPR/Cas9 can have off-target effects if the gRNAs are badly designed. Other genes may share some sequence homology with the target gene. Cas9 may still cut at these genomic locations because it can tolerate up to 3 mismatches between the DNA and the gRNA (Fu *et al.*, 2013; Hsu *et al.*, 2013; Wang, La Russa and Qi, 2016). This means that Cas9 may cut unintended target genes, possibly causing downstream consequences of differential expression which would confound the results

of an RNA-seq experiment. When using tools to design and select gRNAs like CRISPick (Broad Institute, 2022), on-target and off-target scores are calculated (Hsu *et al.*, 2013; Doench *et al.*, 2014), representing the cleavage efficiency of Cas9 with the gRNA and the probability of the gRNA binding off-target, respectively. Both are scored between 0 and 100, and the best gRNAs have high on-target scores and low off-target scores, meaning they are highly specific and will greatly reduce the probability of Cas9 cutting off-target. Caution should be taken when interpreting experimental results following transfections with the gRNAs; it is important to check that the guide is specific, with no off-target effects and is only causing edits in the desired location.

### 2.1.2 HAP1 and K562 Cells

A key consideration of any in vitro experiment is the cell line used. Commonly cultured lines include HeLa, HEK293, U2OS, RPE, COS-7, and CHO cells, with each of these cell lines having advantages and disadvantages. Many experiments will aim to use a cell line derived from a species relevant to their topic area; for human cancer studies, it is more typical to use human cell lines. Another important consideration is ease of editing, with it being substantially less probable to get homozygous edits in cell lines with expanded ploidy such as HeLa cells. A third factor is to use cell lines relevant to the specific area of research, with one classic example being the use of RPE cells for neuronal research as RPE cells and retinal neurons share the same origin and polarised architecture, and adult RPE cells express neuronal markers in vitro.

This project makes use of HAP1 and K562 cell lines. Importantly, these are biologically relevant models of haematopoiesis because they are both human myelogenous leukaemia cell lines (derived from leukaemic cells which develop from mutated myeloid cells in the bone marrow). Specifically, the HAP1 cell line is a human near-haploid line which has a single copy of almost every chromosome. The cells are small, adherent, and fibroblast-like and are easy to culture and transfect. HAP1 cells were derived from the KBM7 cancerous cell line so they are able to divide indefinitely (Kotecki, Reddy and Cochran, 1999). HAP1 cells are a popular choice for knockout experiments because of their near-haploid status, making it easy to determine phenotypic effects after creating mutations as heterozygosity is avoided. This is unlike diploid lines, in which a heterozygous edit can result in the effect of the mutated allele being masked by the unaffected allele. One drawback of HAP1s however is that the original KBM7 cells were isolated from a male patient with chronic myeloid leukaemia (CML) and therefore HAP1 cells too possess the Philadelphia chromosome that is characteristic of this disease: the

reciprocal translocation of chromosomes 9 and 22 which generates BCR-ABL (Fig.4). In addition, HAP1 cells are near-haploid because they possess two alleles for some genes on chromosome 15: all the genes in the original chromosome 15, and 330 genes from chromosome 15 fused to chromosome 19 (Fig.4). A second drawback is that the haploid status may also confer less realistic consequences of experimental mutations when acting as a model for diploid organisms.

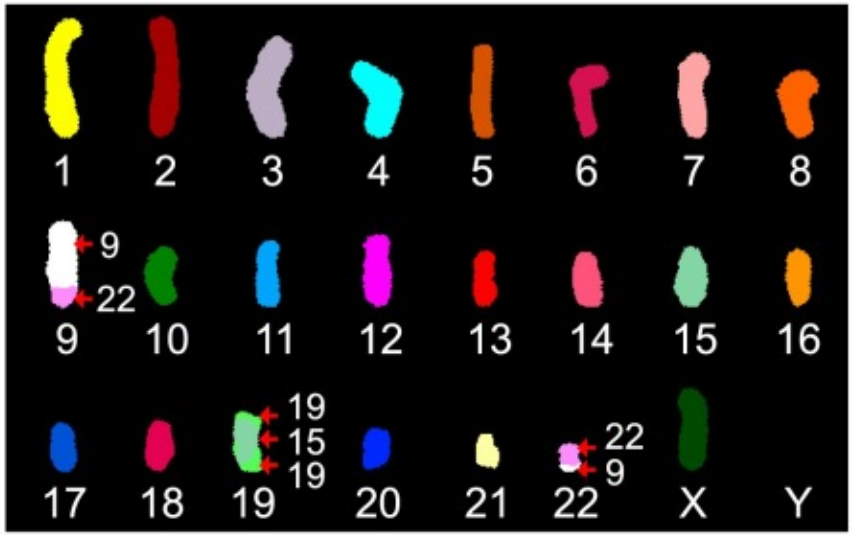


***Figure 4 - HAP1 karyotype.*** *HAP1 karyotype showing the insertion of the Chr15 fragment in Chr19 and the 9;22 translocation (Essletzbichler* et al*., 2014).*

The K562 cell line is a triploid line of human lymphoblasts isolated from the bone marrow of a 53-year-old female CML patient in terminal blast crisis (Lozzio and Lozzio, 1975). These non-adherent rounded cells comprise a highly undifferentiated and immortalised myelogenous leukaemia cell line, particularly erythroleukaemic. They can spontaneously differentiate into erythroid, monocytic and granulocytic progenitors. Like HAP1, K562 cells are easy to culture but also do not represent a completely accurate model for diploid organisms. However, since many primary cancer cells are polyploid, the K562 line can provide useful insights within a disease context. K562 cells also contain the Philadelphia chromosome.

Using both cell lines in combination allowed the study of removing the AS protein RBM15 in biologically relevant haematopoiesis models, with HAP1 representing a population of haematopoietic progenitors, and K562 representing a malignant cell population of erythroid progenitors. Both cell lines used already expressed Cas9 knocked in, in their AAVS1 safe harbour locus. Together, these cell lines were considered useful for studying the downstream effects of knocking out genes encoding SFs acting on erythroid-specific transcripts in particular.

### 2.1.3 Summary

Chapter 2 describes the early development and trial of a CRISPR-KO pipeline, the targeting of the *RBM15* gene with three different guides to establish new processes in the lab, the identification of which guides successfully edit *RBM15*, and the start of exploration into the downstream consequences of knocking out *RBM15* in human cells.

## *2.2 Reagents*

### 2.2.1 Tissue culture

- Gibco Dulbecco's Phosphate-Buffered Saline (DPBS; ThermoFisher)
- Gibco TrypLE Express (ThermoFisher)
- IMDMc (complete) media: Gibco Iscove's Modified Dulbecco Media (IMDM; ThermoFisher) supplemented with 10% heat-inactivated Foetal Bovine Serum (Fisher Scientific) and 1% Penicillin/Streptomycin

### 2.2.2 Transfections

- DharmaFECT1 transfection reagent (Horizon)
- Set of 3 Edit-R predesigned synthetic sgRNAs targeting *RBM15* (Horizon; Table 2)

***Table 2* - RBM15-*KO guide RNA information table.*** *The guide RNAs transfected into HAP1 Cas9+ cells in order to generate an* RBM15-*KO.*

| Name | sgRNA sequence (5' to 3') | Strand of Target | Strand of sgRNA | Orientation | sgRNA Cut Position (1-based) | PAM Sequence | Target Exon Number |
|---|---|---|---|---|---|---|---|
| RBM15_gRNA_1 (SG-010854-01) | AGGATGATCA GCGAGCTAAC | + | + | sense | 110340522 | CGG | 1 |
| RBM15_gRNA_2 (SG-010854-02) | GTGAGCGGAG CAAGAAGTTA | + | + | sense | 110339625 | GGG | 1 |
| RBM15_gRNA_3 (SG-010854-03) | GTCGATCTGG GCCACCAAGT | + | - | antisense | 110340950 | GGG | 1 |

### 2.2.3 Nucleic acid extractions

- DNA extraction kit - ThermoFisher PureLink DNA Mini Kit
- RNA extraction kit - ThermoFisher PureLink RNA Mini Kit

### 2.2.4 Amplicon PCR and primers

- KAPA HiFi 2X PCR MM (Roche)
- Single-sized products were purified from the post-PCR reaction mix using QIAGEN MinElute Purification Kit

- Primers were ordered from IDT as 25 nmole DNA oligos resuspended to 100 µM with TE (Table 3)

***Table 3* - RBM15 *amplicon PCR primers.*** *The primers designed for edit-verification purposes: to generate amplicons surrounding the cut-sites for the gRNAs used in the experiments described in this Chapter. Adapted from the output of Primer-BLAST.*

| Gene | Guide covered | Primer name | Sequence (5' → 3') | Length (bp) | Tm (°C) | GC% | Intended product length (bp) | Number of possible unintended templates | Unintended product lengths (bp) | Distance from primer to guide (incl.) (bp) |
|---|---|---|---|---|---|---|---|---|---|---|
| *RBM15* | gRNA_2 | F1 | CGACCCGCAA CAATGAAGGG | 20 | 61.92 | 60.00% | 1494; 1073 | 0 | NA | 102 |
| *RBM15* | gRNA_3 | R1 | CAGATACTGCT GCTGGTAACGA | 22 | 60.16 | 50.00% | 1494 | 0 | NA | 73 |
| *RBM15* | gRNA_1 | R2 | CCAAAGCGA TCAAACGCCC | 19 | 60.15 | 57.89% | 1073 | 0 | NA | 94 |

### 2.2.5 Gel electrophoresis

1.5% agarose gels were used and run at 120 volts for 45-60 minutes. The NEB 1kb Plus DNA Ladder was used as a sizing reference. Bands were cut from the gels and purified using QIAGEN QIAquick Gel Extraction kit.

### 2.2.6 Cell lines

- HAP1 Cas9+ (Horizon)
- K562 Cas9+ (Horizon)

## *2.3 Methods*

### 2.3.1 CRISPR-KO experiment design and transfections

In order to explore the functions of RBM15, an experimental process was designed to target a gene in a knockout (KO) experiment, validate the KO and subsequently analyse the transcriptomic consequences of removing the gene. These experiments were designed to be performed on two cell lines: HAP1 and K562.

#### *2.3.1.1 Cell culture*

##### *2.3.1.1.1 HAP1 subculturing procedure*

The media was aspirated from the cells and then the cells were washed with DPBS. After aspirating the DPBS, trypsin was added and the cells were incubated at 37°C for 4 minutes then neutralised with IMDMc media. The cells were centrifuged at 300 rcf for 5 minutes then the media aspirated and the cell pellet resuspended with fresh media. The cells were split to need (usually 1:10-1:25 into a new 6-well plate) and topped up with sufficient fresh media for the vessel.

Cells were kept <85% confluency for the first KO experiment and <75% confluency for the second as much as possible, however, greater confluency did occur on occasion.

##### *2.3.1.1.2 K562 subculturing procedure*

The cells were centrifuged at 300 rcf for 5 minutes then the media aspirated and the cell pellet resuspended with fresh media. The cells were split to need (usually 1:5-1:10 into a new T-75 flask) and topped up with sufficient fresh media for the vessel.

#### *2.3.1.2 Experimental design overview*

Three guides per gene were used in the experiment to ensure at least one resulted in successful editing of the target gene. The *RBM15* sgRNAs were obtained from Horizon as a predesigned set of 3. The guides were annotated onto Ensembl gene sequence maps on Benchling (Benchling, 2021) for visualisation and recording-keeping (https://benchling.com/nicoleforresterei/f_/YvRIseur-sf-ko-experiment/).

Both the HAP1 and K562 cell lines were transfected separately with each of the three guides. The transfected cells were sorted using FACS to obtain single-cells by gating for cells, singlets, and living cells by use of a propidium iodide (PI) viability stain that allows exclusion of PI-positive dead cells. A single cell was sorted into each well of a 96-well plate for each of the KO cell lines. Whilst these cells grew up, the remainder of the cells were screened to check which of the three guides used to target the gene had successfully enabled desired cutting by the Cas9 protein.

The remaining cells from the single-cell sorts were each plated into a well of a 6-well plate; in the time the single-cell clones were growing up into purified putative KO cell lines, these "bulk" transfected cells were used to screen the guides for successful editing.

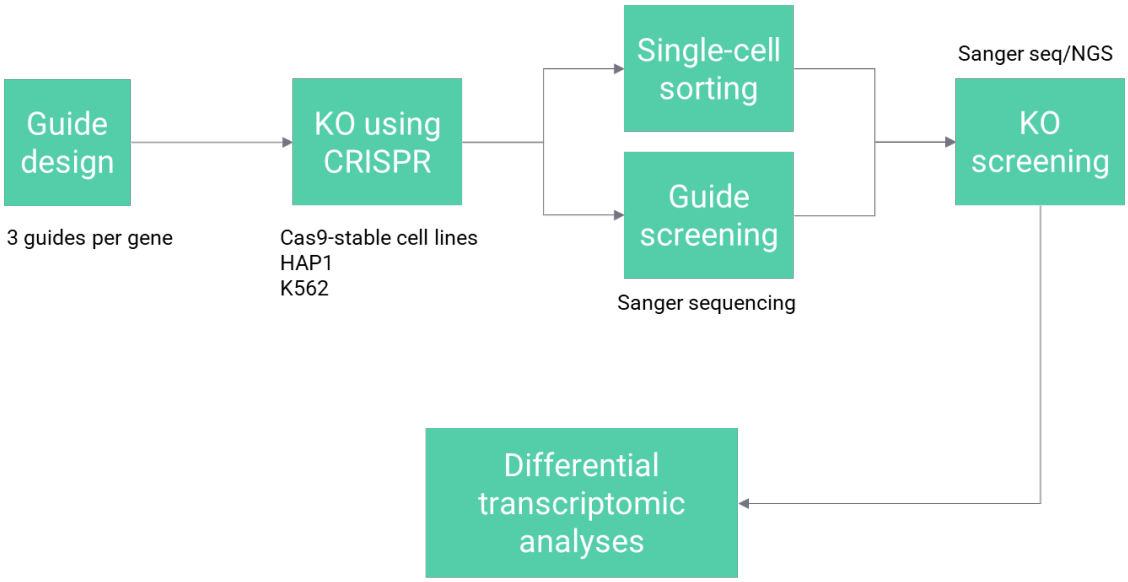


***Figure 5 - Experimental design overview for the knockout experiments.***

### *2.3.1.3 Transfections*

Transfections were performed based on the guidelines recommended by Horizon: the DharmaFECT Transfection Reagents - siRNA Transfection Protocol. Steps described below for transfection of cells seeded in a 6-well plate (volumes are per 6-well sample).

Cells were counted and plated into a 6-well plate 24-48 hours prior to transfection then incubated at 37°C with 5% $CO_2$. Stock sgRNAs were diluted with 200 µl 10 mM Tris to make up 10 µM sgRNA. Tube 1 was prepared as follows: 10 µM sgRNA (5 µl) was diluted with Opti-MEM (195 µl) to make up 0.25 µM sgRNA; this was carefully pipette mixed then incubated for 5 minutes at room temperature. Tube 2 was prepared as follows: 1

mg/ml stock DharmaFECT1 (DF1) transfection reagent (8 µl) was diluted with Opti-MEM (192 µl) to make up 40 µg/ml DF1; this was carefully pipette mixed then incubated for 5 minutes at room temperature. The contents of Tube 1 were added to Tube 2, carefully pipette mixed and incubated for 20 minutes at room temperature. 1600 µl antibiotic-free medium (IMDM 10% FBS, no pen/strep) was added to the combined tubes to bring the transfection mix up to 2 ml total volume. The culture medium on the cells was then replaced with this transfection mix. The cells were incubated at 37°C with 5% $CO_2$ for 24 hours, then the transfection mix on the cells was replaced with antibiotic-free medium and allowed to continue to incubate. 24-48 hours post-transfection the cells were harvested, sorted and split. Cells were harvested for making freezer vials and for genomic DNA extraction. Cells were sorted (FACS) into a 96-well plate and left to grow for 2 weeks (until colonies have formed). The remaining cells were split to maintain a live culture.

Cells were seeded at a density of 0.5 M cells per 6-well for the first KO experiment and 0.4 M cells per 6-well for the second KO experiment.

#### *2.3.1.4 FACS for sub-cloning*

Sorting was performed using the BD FACSMelody. Cells were stained using Propidium Iodide (ThermoFisher, PI Ready Flow Reagent) and individual viable cells were sorted into round-bottomed 96W plates prefilled with IMDMc media. See Chapter 2.3.3 for details of subculturing of single-cell clones following FACS.

##### *2.3.1.4.1 Checking ploidy by FACS*

To investigate their ploidy, cells were stained with violet blue cell cycle analysis DNA dye (ThermoFisher, Vybrant DyeCycle Violet Ready Flow Reagent) and analysed on the BD FACSMelody.

### 2.3.2 Screening the guides for successful editing at the intended cut-site

Once the sgRNA had been delivered to the cells, a method was required to determine whether the gRNA had guided the Cas9 in the cells to cut at the intended loci. If this cut had occurred, insertions or deletions (indels) of a few base pairs should have occurred directly following the protospacer adjacent motif (PAM) sequence in the target sequence (Fig.3). Extracting gDNA from the transfected cells, generating amplicons containing the

gRNA cut-site and Sanger sequencing them enables analysis for the presence of the indels and thus verification that the targeted gene had been successfully edited.

The bulk transfected cells were washed in PBS, trypsinised and neutralised with media then centrifuged and the supernatant aspirated. The gDNA was extracted from the cells using the ThermoFisher PureLink DNA Mini extraction kit. Amplicons were generated using KAPA HiFi 2X Ready Mix (Roche) and custom primers designed to surround the gRNA cut-sites (Fig.6). The PCR products were examined for unintended PCR products using gel electrophoresis (1.5% agarose). Samples with a clean single band were purified using QIAGEN MinElute PCR Purification kit. The correct band was excised from the gel for multiple-band samples and purified using QIAGEN QIAquick Gel Extraction kit. The purified amplicons were sent externally for Sanger sequencing (Azenta/GENEWIZ), sequencing from the screen PCR primer closest to the gRNA cut-site in each case. The resulting .ab1 files were uploaded to Benchling using the Alignment tool and aligned to the reference genomic sequence using default parameters.

A bulk sample of transfected cells including cells that had been correctly edited would produce a Sanger trace with multiple signal peaks at each base immediately following the gRNA cut-site. The sample would contain: 1. some untransfected cells; 2. some transfected but unedited cells; 3. some edited cells. The Sanger trace for pure populations of the first two would each match the wild-type sequence. The Sanger trace for the edited cells would match the wild-type sequence up until the cut-site, when there would be indels, normally of a few bases, before resuming with the wild-type sequence. In a sample with all three of these populations present, the alignment tool (in this case, Benchling) would not be able to separate the different signals into a wild-type and an edited trace but instead show the two in parallel. The result would be a wild-type sequence until the cut-site, then multiple signal peaks for the bases following as the indels present in some of the cells will have shifted the sequence out of alignment of the wild-type sequence (see Sanger trace for sample A in Fig.7 or that for A2 in Fig.12).

#### *2.3.2.1 Screening primer design*

Primers used in the process of amplicon generation and Sanger sequencing analysis of the areas surrounding the guide cut-sites were designed using Primer-BLAST (NCBI, 2022). When possible, a single amplicon was designed to cover as many of the guides for the gene as possible. See Fig.6 below for a visual illustration of the primer and guide sequence locations within *RBM15*.

To isolate the cut-site for assessing the success of each guide, the sequence surrounding the cut-site was copied from the Ensembl gene sequence in Benchling into Primer-BLAST, with around 150-350 bases either side of the cut-site. The forward primer was searched for within the first X number of bases that allowed for at least 50 bp before the closest guide sequence. This was to ensure good quality Sanger sequencing at the cut-site, since the first 50 bp of a Sanger trace exhibits poor quality as the reaction begins. Similarly, the reverse primer was searched for at a sensible distance from the other end of the closest guide cut-site. The other parameters for the tool were kept at the default settings except the minimum PCR product size was set to around 250 bp and the database used to check the primer specificity against was set to “Genomes for selected organisms (primary reference assembly only)”.

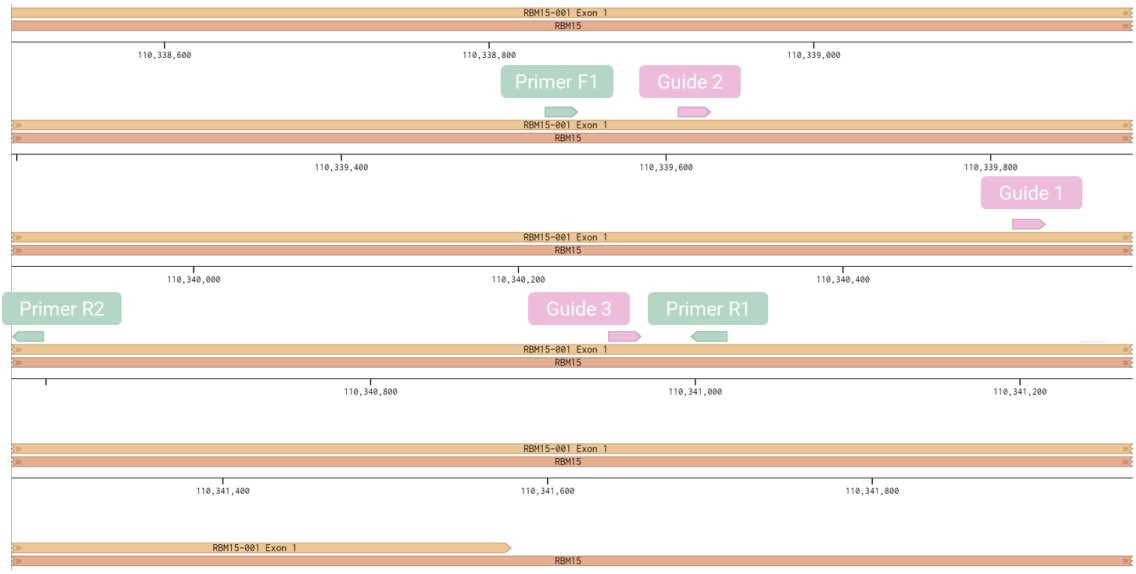


***Figure 6* – RBM15 *amplicon PCR primers.*** *Screenshot from Benchling sequence map depicting the primer (green) locations for screening the* RBM15 *guides (pink) to determine which of them was most successful in editing the target gene. The darker orange line symbolises a 4.1 kb section of the* RBM15 *gene, from its 5’ start, with the lighter orange line above it marking the entirety of Exon 1. The sequences of the predesigned sgRNAs from Horizon all targeted Exon 1. Three primers (one forward, two reverse) were designed to generate two amplicons which could be Sanger sequenced from one end to assess whether each of the guides had cut or not. An amplicon produced from primers F1 and R2 would enable assessment of Guide 1 (sequencing from R2) or Guide 2 (sequencing from F1 whilst an amplicon produced from F1 and R1 would enable assessment of Guide 3 (sequencing from R1).*

### 2.3.3 Identification and analysis of edited single-cell clone (putative knockout)

Following confirmation that the CRISPR-Cas9 had cut correctly in cells within a sample, a purified edited clone could be obtained for further analysis. The edited cells that had been single-cell sorted using the BD FACSMelody were allowed to grow for two weeks.

Colonies formed from the single-cells were transferred into individual wells of 24-well plates when large enough. Upon the next subculturing, half the clone cells underwent gDNA extraction; the edited single-cell clone gDNA was screened by the same process described above for the bulk cells (PCR amplification, purification, Sanger sequencing).

### 2.3.4 RNA-seq and differential transcriptomic analysis of KO cell line

Once an edited clone (A2F7) and a wild-type clone (A2E4) had each been identified, the transcriptomic differences between the two clones could be assessed to explore the consequences of knocking out the targeted RBP/SF gene. Cells were split into 4 equal batches and allowed to grow until the next split. RNA was then extracted from each clone batch and submitted to Genomics Pipelines, Earlham Institute for RNA-seq library preparation and sequencing. The library preparation protocol used was NEBNext Ultra II Directional RNA-seq (poly-A isolation). Sequencing was performed on an Illumina NovaSeq 6000, SP v1.5 flow cell (both lanes), 150 bp paired-end reads, estimated to yield ~100 M reads per sample.

The sequencing reads underwent quality control and adapter trimming using Trim Galore (Babraham Bioinformatics, 2022) before being mapped to the human reference genome GRCh38 using STAR (Dobin *et al.*, 2013). Next, the reads were merged from the two SP flow cell lanes for each sample using samtools merge (Danecek *et al.*, 2021). After sorting and converting the BAM files to SAM files, the reads mapping to genes were counted using htseq count, producing a raw counts matrix.

The raw counts matrix was read into R and differential gene expression analysis was performed using DESeq2, comparing the edited clone to the wild-type clone. Genes with no counts in any samples were removed from the data beforehand. DESeq2 reduces the high false positive rate associated with multiple testing (testing multiple genes for differential expression at $p<0.05$) by adjusting the p-values: the genes are ranked by p-value then each p-value is multiplied by the total number of tests/rank (Benjamini and Hochberg, 1995). Differentially expressed (DE) genes were defined by the adjusted p-value for the difference between the edited and wild-type conditions being less than 0.05; a table of DE genes was generated, along with plots showing the normalised counts for significant genes across the samples, grouped by condition.

The results from the differential expression analysis were further assessed by integrating eCLIP data from ENCODE (Stanford University, 2022; ENCODE Project Consortium, 2012; accession ENCFF006CBO). Enrichment of RBM15-linked genes (those containing

binding sites used by RBM15) was checked for in the DE genes from A2F7 edited clone against the list of total expressed genes analysed as part of the DE analysis. A hypergeometric test was used to statistically assess the difference between the two resulting lists of enriched RBM15-linked genes using the phyper function from the DPQ package in R.

## *2.4 Results*

### 2.4.1 HAP1 *RBM15*-KO experiment

Two iterations of the RBM15-KO experiment were performed in HAP1 Cas9+ cells. The first iteration tested transfection conditions and determined that gRNA_1 cuts *RBM15* effectively. However, the verified edited single-cell clones from this experiment were found to be diploid (described and discussed in detail later) and must have been so just prior to or upon transfection. In the second iteration, cells were grown more sparsely to avoid them becoming diploid and were transfected with the optimum transfection conditions and verified sgRNA (gRNA_1) established from the first attempt to try to generate haploid *RBM15* KOs.

#### *2.4.1.1 Testing the effectiveness of different RBM15 gRNAs*

It was important to establish which sgRNAs were able to effectively target Cas9 to *RBM15*. To do this, three different guides – gRNA_1 (Sample A), gRNA_2 (Sample C), gRNA_3 (Sample E) – were transfected into HAP1 Cas9+ cells, genomic DNA extracted, the genomic locus surrounding the three potential Cas9 cut sites amplified by PCR, and the amplicons sequenced by Sanger sequencing. gRNA_1 showed evidence of editing the *RBM15* locus as there was a clear presence of more than one trace following the gRNA_1 cut site, indicating indels in at least some cells (Fig.7– Sample A). This was not visible in the gRNA_1 control of sgRNA addition without DharmaFECT transfection reagent (Fig.7 – Sample B). gRNA_2 may have led to editing of the *RBM15* locus in a small number of cells as a small alternative signal peak was visible in the Sanger sequencing trace after the Cas9 cut site, with this not visible in the control for gRNA_2 (Fig.7 – Samples C, D). The effect of gRNA_3 was unmeasurable because the PCR primers did not successfully produce an amplicon encompassing the cut-site directed by gRNA_3 (Fig.7 – Samples E, F). Because gRNA_1 was already shown to be effective, validation of the effectiveness of gRNA_3 effectiveness was not pursued further. It was concluded that gRNA_1 was the best choice for the most efficient editing of *RBM15* in HAP1 Cas9+ cells.

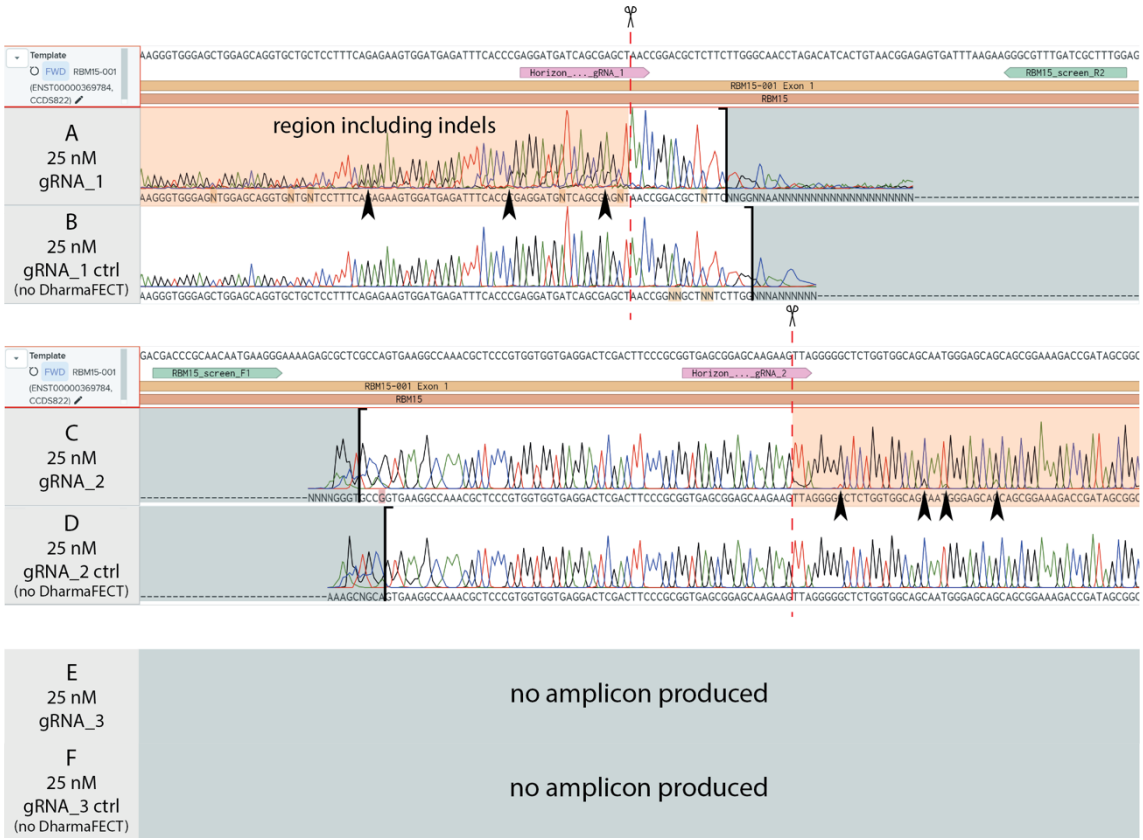


***Figure 7 – HAP1 Cas9+ cells showed evidence of editing in* RBM15 *with 25nM gRNA_1.*** *HAP1 Cas9+ cells were transfected with either 25nM gRNA_1 +/- DharmaFECT (Samples A/B), 25nM gRNA_2 +/- DharmaFECT (Samples C/D), or 25nM gRNA_3 +/- DharmaFECT (Samples E/F), genomic DNA extracted, amplicons generated around the Cas9 cut site by PCR, and amplicons sequenced by Sanger sequencing. Sequencing traces were aligned to the* RBM15 *gene using Benchling and examined for the presence of multiple overlapping sequencing traces. The gRNA annealing site is shown in pink, and the position of the sequencing primer shown in green. Red dashed lines indicate the position of the cut-site. Black arrows indicate example positions where non-wild-type sequencing traces overlap the wild-type trace.*

### *2.4.1.2 Testing gRNA_1 transfection concentrations*

In addition to choice of gRNA, it was important to determine the concentration of gRNA that would give the best editing. This was performed using gRNA_1 as that gave the best editing of the three guides tested. To determine optimal transfection concentrations, cells were transfected with either 25nM gRNA_1 (Sample A) or 50nM gRNA_1 (Sample G), *RBM15* locus amplicons obtained as described above, and the Sanger sequencing traces compared for the prevalence of *RBM15* non-wild-type traces. As seen from Fig.8, alternative signal traces appear after the cut site in both the 25nM gRNA_1 and 50nM gRNA_1 transfection conditions (Fig.8). However, the signal of the alternative trace for the 50nM treatment is generally lower than for the 25nM treatment. In addition, by eye, there was increased cell death following transfection with 50nM gRNA_1 compared to 25nM gRNA_1. It was therefore deemed that 25nM was the optimal final concentration

of sgRNA_1 when transfecting HAP1 Cas9+ cells using 4µg DharmaFECT1 to edit the *RBM15* locus.

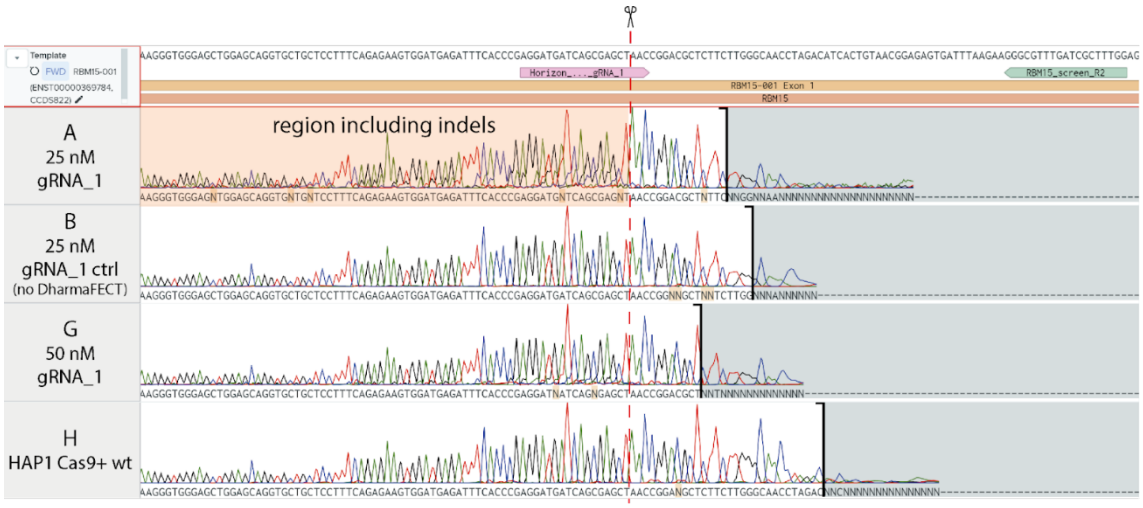


***Figure 8 - Sanger sequencing traces indicating signs of possible editing in samples A and G and no editing in samples B and H.*** *Screenshot from Benchling alignment tool. A: bulk batch of cells that were transfected with 25 nM gRNA_1. B: bulk batch of cells that received a transfection mix with 25 nM gRNA_1 but no transfection reagent (wild-type (wt) control for sample A). G: bulk batch of cells that were transfected with 50 nM gRNA_1, a double-dose version of sample A. H: wt HAP1 Cas9+ cells (untransfected). These traces were produced from sequencing from the RBM15_screen_R2 primer (green). The orange highlighted region shows the region following the cut-site which should contain indels if cells within the sample have been successfully edited. An alternative signal peak for each base can be seen after the gRNA_1 cut-site (red dotted line) in sample A, showing there are edited cells in the sample. Sample B and sample H do not show these alternative signal peaks. Sample G does show them but at a lower signal level.*

### *2.4.1.3 Isolation of an edited RBM15 clone and determining ploidy*

A single clonal population of *RBM15*-KO cells was required for downstream bulk RNA-seq analysis. To generate clonal populations, bulk populations of 25nM gRNA_1-treated HAP1 Cas9+ cells were sorted as single cells into the wells of a 96-well plate. Clones were named by their original name (e.g. Sample A for 25nM gRNA_1 treatment) plus the well coordinate of the 96-well plate that single cell was sorted into. After screening individual clones, one clone was identified (AB12) that showed a non-wild-type *RBM15* sequence by Sanger sequencing (Fig.9). The Sanger trace showed a clear signal that was different from the wild-type sequence. However, only a single trace would be expected for an edited haploid clonal cell population; one that matched the wild-type sequence up until the cut-site followed by an insertion or a deletion before continuing to align with the wild-type sequence. Despite being different from the wild-type trace, the trace for AB12 resembled the bulk sequencing data (Fig.7 – Sample A) of multiple sequencing traces overlayed. This suggested that there was more than one allele present in AB12.

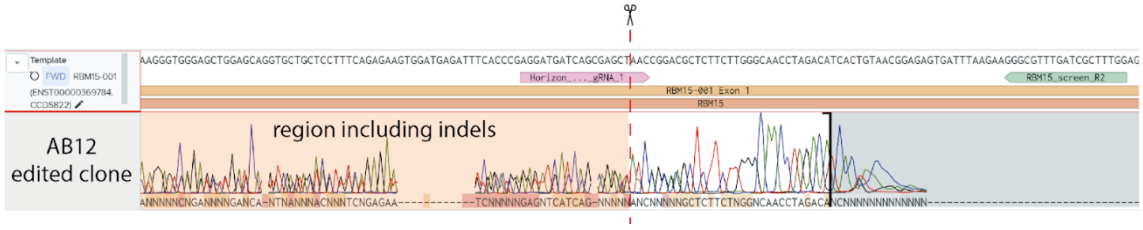


***Figure 9 - Sanger sequencing traces indicating signs of possible editing in sample AB12.*** *Screenshot from Benchling, Sanger trace for AB12, clonal population from the single-cell sorted into well B12 of the 96-well plate for sample A (25 nM gRNA_1). This trace was produced from sequencing from the RBM15_screen_R2 primer (green). An alternative signal trace is present alongside the wild-type sequence trace, showing that these cells cannot be haploid but they are edited on one allele. The trace is also "messier" than expected, with the alternative signal unexpectedly appearing prior to the cut-site (red dotted line) for gRNA_1. Therefore, the results of this transfection cannot be elucidated with certainty.*

It is known that HAP1 cells can revert to being diploid if left to grow to a high confluency (Olbrich, Mayor-Ruiz, Vega-Sendino *et al.*, 2017). Since AB12 unexpectedly showed evidence of multiple overlayed sequencing traces, not possible in a single-cell sorted haploid cell line, it was deemed necessary to determine the ploidy of AB12 to see if they had become diploid. This was done by treating AB12 and haploid wild-type HAP1 Cas9+ cells with a cell-permeable violet DNA-staining reagent (Vybrant CellCycle Violet) and then measuring DNA content per cell by flow cytometry. Haploid cells should have roughly half the DNA content of diploid cells. This analysis revealed that AB12 cells had roughly double the DNA content of control haploid wild-type HAP1 Cas9+ cells as measured by Vybrant CellCycle Violet DNA fluorescence (Fig.10). This implies that AB12 had become diploid. Furthermore, the multiple traces observed of the *RBM15* gRNA_1 target site by Sanger Sequencing (Fig.9) indicate that AB12 was heterozygous diploid with one allele with indels and one wild-type allele.

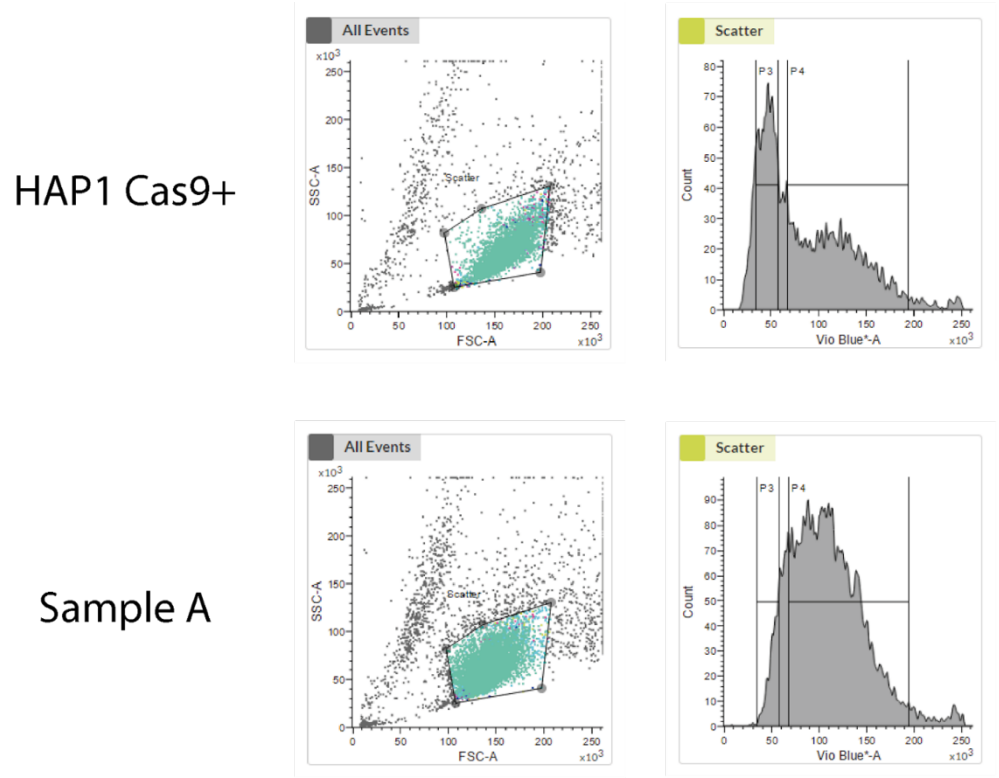


***Figure 10 - Plots showing change in ploidy in sample A compared to HAP1 Cas9+ cells.*** *Haploid HAP1 Cas9+ cells (top) and 25nM gRNA_1 transfected HAP1 Cas9+ cells (Sample A, bottom) were analysed for their DNA content by flow cytometry using Violet Blue DNA staining, giving an indication of ploidy. The left plots above show all events as determined by forward scatter (FSC) and side scatter (SSC) and the gating of live cells by the shown P1 gate (green). From P1 singlets were then gated (P2; not shown). The right plots show the violet blue DNA staining from each cell within P1 and P2 gates, with the likely haploid population as determined by the control HAP1 Cas9+ cells gated P3, and the likely diploid population gated as P4. Images are screenshots from BD FACS Chorus.*

#### *2.4.1.3.1 Second attempt to generate a clonal* RBM15*-KO cell line*

The experiment of transfecting HAP1 Cas9+ cells with *RBM15* gRNA_1 was repeated to attempt to generate a haploid *RBM15*-KO, with cells grown more sparsely to prevent cells becoming diploid upon editing. As before, 25nM of gRNA_1 was transfected into HAP1 Cas9+ cells to generate a bulk population containing both edited and unedited cells (Sample A2). The ploidy of this bulk population (A2) was analysed using violet DNA staining and flow cytometry. This revealed that many, but not all, A2 cells had at least a doubling of violet DNA fluorescence compared to untreated HAP1 Cas9+ cells, and a clearly defined smaller population of A2 cells had the same violet fluorescence as the controls (Fig.11). This indicated that, like the first attempt at generating *RBM15*-KO, many of the gRNA_1 transfected cells had become diploid, but unlike the first attempt, a small population of transfected cells had successfully remained haploid.

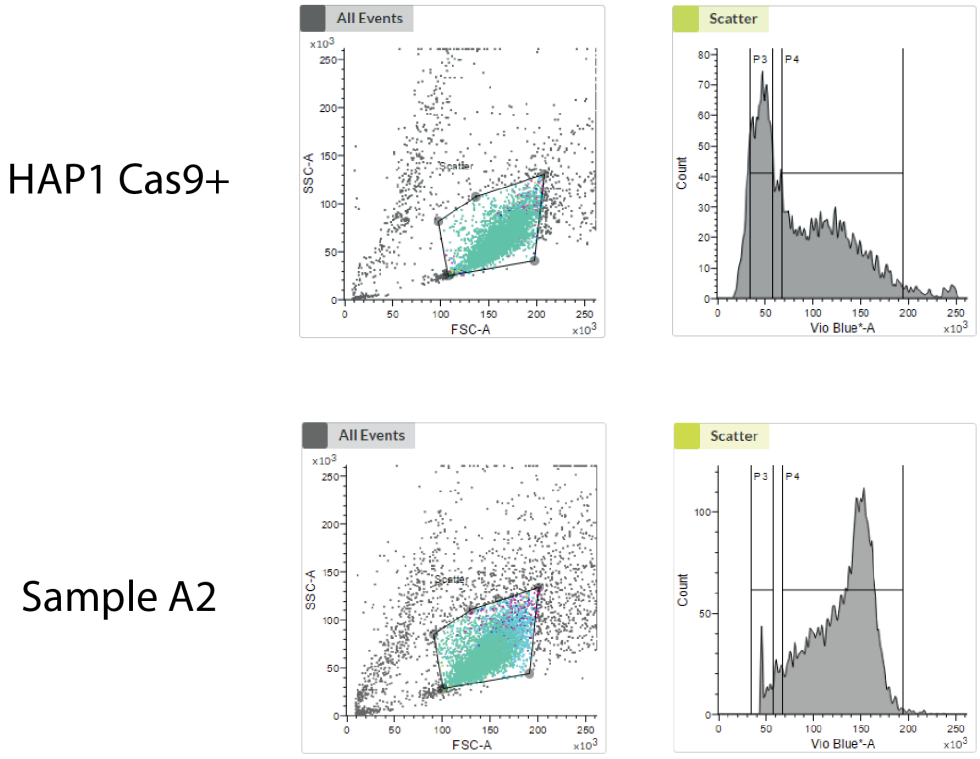


***Figure 11 - Plots showing heterogeneous ploidy in Sample A2 compared to HAP1 Cas9+ cells.*** *Haploid HAP1 Cas9+ cells (top) and 25nM gRNA_1 transfected HAP1 Cas9+ cells (Sample A2, bottom) were analysed for their DNA content by flow cytometry using Violet Blue DNA staining. The left plots above show all events as determined by forward scatter (FSC) and side scatter (SSC) and the gating of live cells by the shown P1 gate (green). From P1 singlets were then gated (P2; not shown). The right plots show the violet blue DNA staining from each cell within P1 and P2 gates, with the likely haploid population as determined by the control HAP1 Cas9+ cells gated P3, and the likely diploid population gated as P4. Images are screenshots from BD FACS Chorus.*

Despite containing both haploid and diploid cells, the A2 population was single-cell sorted into wells of a 96-well plate to isolate clonal cell populations. Expanded clones were screened by Sanger sequencing for whether they contained indels near the gRNA_1 directed Cas9 cut site. For downstream RNA-sequencing analysis, two clones were required: a wild-type clonal population and an *RBM15*-edited clonal population. Analysis by Sanger sequencing revealed two clones of interest: a wild-type clone A2E4, which matched the HAP1 Cas9+ wild-type parental cells, and the edited clone A2F7, which showed a clear difference from both the HAP1 Cas9+ wild-type cells and A2E4 clonal cell line after the gRNA_1 directed Cas9 cut site (Fig.12). Unfortunately, however, the Sanger trace for A2F7 showed the presence of overlapping traces after the cut site, suggesting that it was heterozygous diploid for *RBM15*.

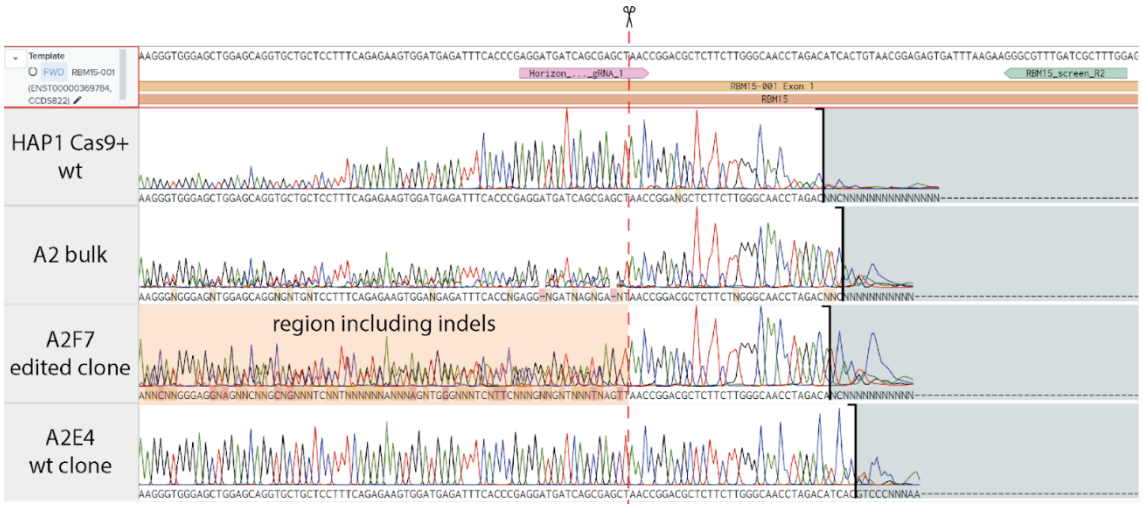


***Figure 12 - Sanger sequencing traces indicating signs of possible editing in samples A2 and A2F7 and no editing in HAP1 Cas9+ and sample A2E4.*** *Screenshot from Benchling alignment tool showing Sanger sequencing traces for the amplicon surrounding the gRNA_1 directed Cas9 cut-site in wild-type HAP1 Cas9+ cells (top), the bulk population of 25nM gRNA_1 transfected HAP1 Cas9+ cells (2nd row), and two single-cell clones A2F7 (third row) and A2E4 (bottom) derived from the A2 bulk population. The Cas9 cut site is labelled by red dotted line, and a region thought to deviate from the wild-type HAP1 Cas9+ cells is indicated for the A2F7 clone.*

### *2.4.1.4 RNA-seq results*

The genotype of the *RBM15*-edited A2F7 cell line and the consequences of the *RBM15* editing required investigation. Bulk RNA-seq was used to gain an initial and overall measure of the transcriptomic differences between *RBM15*-edited A2F7 cells and A2E4, a verified wild-type population of cells subcloned from the same parental sample as A2F7 (Sample A2). Four technical replicates of each sample were sequenced by short-read bulk RNA-seq on a NovaSeq 6000 SP flow cell with an expected 100 M reads per sample. The number of reads obtained for each sample averaged 130 M reads, and all were at least 10% higher than the expected 100 M value (Table 4).

***Table 4 - The number of reads produced from the RNA-seq of A2F7 (edited clone) and A2E4 (wt clone).*** *Sequencing was performed on SP Flowcell of NovaSeq 6000 by Genomics Pipelines, Earlham Institute. The replicates for each clone are numbered 1-4 following the name of the clone they originate from.*

| Sample | Total number of reads |
|---|---|
| A2F71 | 132,102,541 |
| A2F72 | 114,655,117 |
| A2F73 | 140,708,515 |
| A2F74 | 132,183,766 |
| A2E41 | 152,042,756 |
| A2E42 | 135,707,727 |
| A2E43 | 130,905,452 |
| A2E44 | 116,163,461 |

A sample distance heatmap was generated to investigate the clustering of the RNA-seq samples (Fig.13) as a means of quality control. This showed that each of the four technical repeats for A2E4 and A2F7 clustered together. This highlights that the RNA-seq data had good technical reproducibility for A2E4 and A2F7 samples.

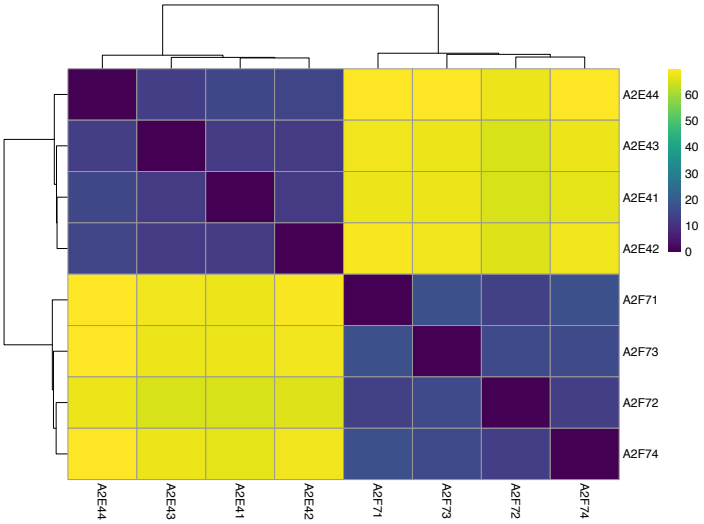


***Figure 13 - Technical replicates clustered by their RNA-seq expression profiles.*** *A sample distance heatmap to visualise the sample relationships in RNA-seq gene expression data. The Euclidean distance with hierarchical clustering between the eight samples was calculated and plotted in R. The colour map indicates sample distance.*

#### *2.4.1.4.1* RBM15 *expression was doubled in the edited clone compared to the wild-type*

An important first point of analysis was to investigate the changes in expression of *RBM15* between the edited A2F7 samples and wild-type A2E4 samples. The normalised read counts for *RBM15* in the edited clone were twice that of the wild-type clone (Fig.14). This result was unexpected since the aim of the experiment was to knock out *RBM15* expression. One explanation is that this doubling reflects a difference in ploidy: A2E4 were haploid, deriving from the small population of haploid cells in the parental A2 population (Fig.11), and A2F7 were diploid, deriving from the major population of diploid cells in the parental A2 population (Fig.11). Transcription from two *RBM15* alleles in the A2F7 clone compared to a single allele in the A2E4 clone could explain the doubling of read number. However, because the counts are normalised by total number of reads per sample, a duplication of gene expression due to diploidy would be cancelled out: a diploid clone producing exactly twice the number of transcripts for each gene as a haploid clone would express each gene in the same proportions. Therefore, simple diploidisation of

A2F7 is not occurring here and it is more likely that expression differences seen between A2F7 and A2E4 is a result of the editing of *RBM15* in A2F7.

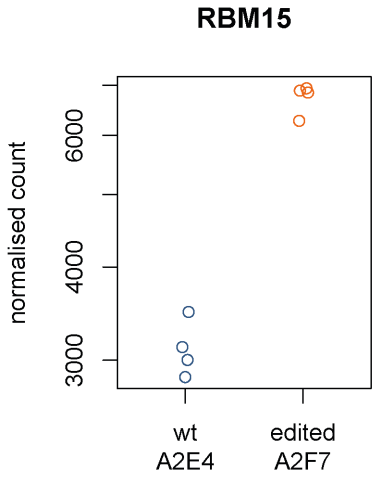


***Figure 14 - RBM15 expression appears doubled in* RBM15*-edited A2F7 cells compared to wild-type A2E4 cells.*** *The normalised read counts for* RBM15 *in wild-type A2E4 and* RBM15*-edited A2F7 samples were plotted.*

#### *2.4.1.4.2 Both* RBM15 *alleles were edited in A2F7*

Although not analysed directly, ploidy could be inferred by analysing the alignment of all transcripts against the *RBM15* locus. This would provide insight into the genomic modifications at the cut-site in RBM15 edited A2F7 cells, and the ploidy of A2F7 and wild-type A2E4. To do this, *RBM15*-edited A2F7 and wild-type A2E4 reads in sorted BAM format and the accompanying .bai files were uploaded to the Integrative Genomics Viewer (IGV; Robinson *et al.*, 2011; Thorvaldsdóttir, Robinson and Mesirov, 2013). For A2F7, particular attention was paid to genomic locations 110340521 and 110340522 as this is where the Cas9 guided by gRNA_1 would have cut. For A2F7, approximately half of the reads contained insertions of a single thymidine (T) directly upstream of the cut-site, and the other half contained a deletion of three bases one base upstream of the cut-site (Fig.15). The insertions and deletions were never found together in the same transcript, showing them to be mutually exclusive. In contrast, A2E4 did not show any indels at the gRNA_1 cut-site (Fig.15).

For A2E4, this analysis showed that *RBM15* was wild-type, although the ploidy could not be estimated using this method. However, for A2F7, the analysis suggested these cells were heterozygous diploid, with both alleles edited, reaffirming the Sanger sequencing data (Fig.12). The three-base deletion maintained the reading frame (as it is a multiple

of three) and resulted in the removal of alanine-372 from the RBM15 protein. Based on AlphaFold's predicted structure of RBM15 (AF-Q96T37-F1), A372 resides in the unstructured beginning of the second RRM domain of RBM15. Given that alanine is typically a non-disruptive amino acid (Park and Cochran, 2009), and that this particular alanine is not within a structurally important part of a core functional domain of RBM15, it is unlikely (but not impossible) that this deletion would affect RBM15 function. However, the single base insertion of thymidine shifts the reading frame as it is not a multiple of three, resulting in the conversion of the codon that should encode asparagine-373 into a stop codon. This would be expected to cause the transcript from this allele to be degraded by nonsense-mediated decay. However, this does not appear to be occurring here as the *RBM15* transcript containing the +1 insertion is as equally represented in the transcriptomic data as the -3 deletion transcripts, and not in the extreme minority as would be expected if it were the target of nonsense-mediated decay (Fig.15). Ultimately, these data suggest therefore that A2F7 is diploid and is producing two different *RBM15* transcripts and two different RBM15 proteins: one transcript from the -3 deletion allele that is generating a protein that is an almost-full length version of RBM15 containing an A372 deletion that would be expected to be fully functional; and one transcript from the +1 insertion allele that is generating a truncated version of *RBM15* up until the novel stop codon at position 373, therefore producing a transcript encoding only the first RBM15 RRM domain. In short, although not the original plan, the RNA-seq analysis here effectively compares the transcriptomic consequences between the production of wild-type RBM15 protein in haploid cells (A2E4) *vs* the production of near-wild-type RBM15 protein and a protein containing the first RBM15 RRM in diploid cells (A2F7).

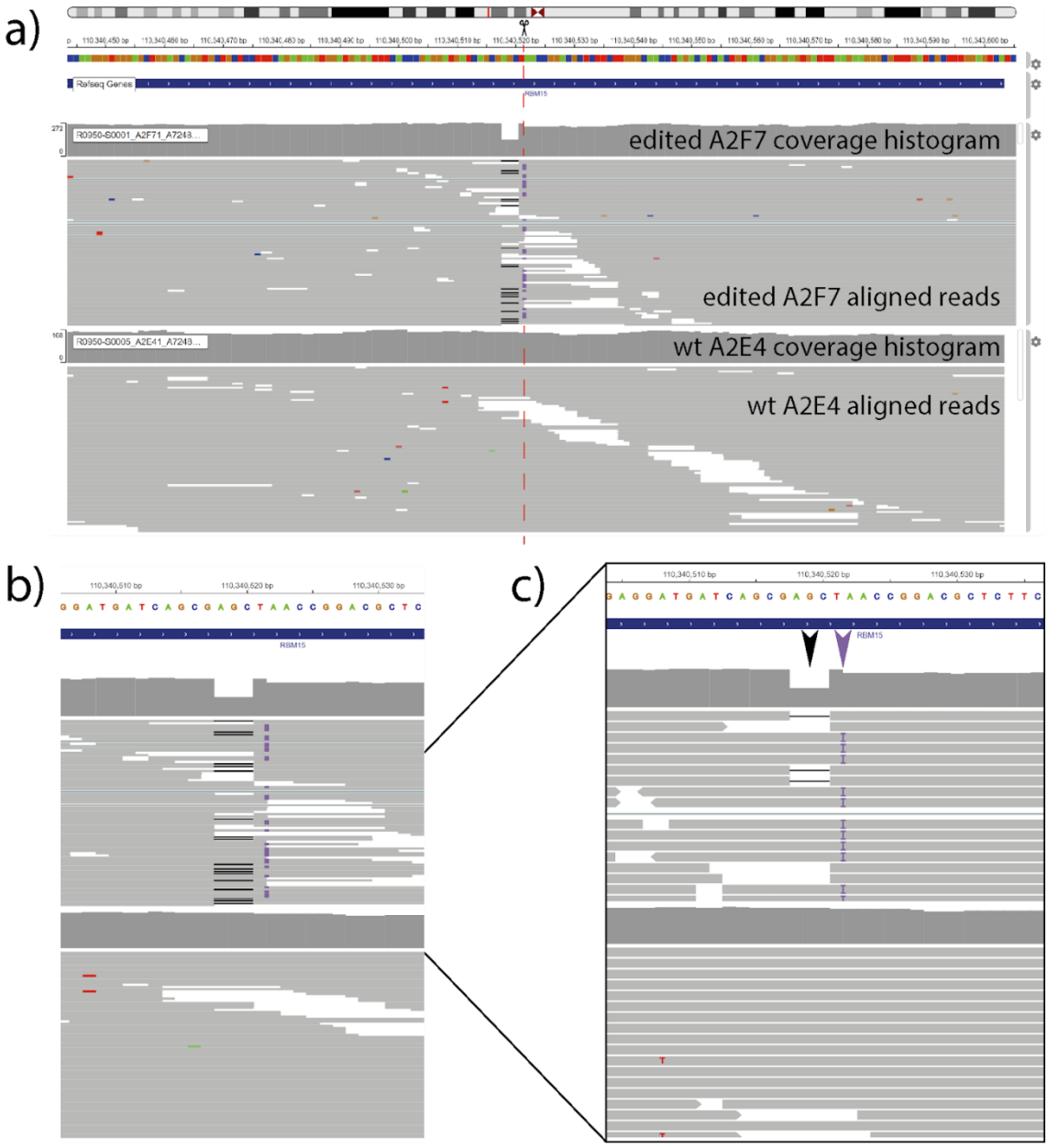


***Figure 15 - A2F7 clones show evidence of being heterozygous diploid for* RBM15 *whilst A2E4 appears* RBM15 *wild-type.*** *Screenshots from IGV showing RNA-seq reads mapped to the* RBM15 *genomic sequence (hg38) at different zoom levels. Genomic information of chromosome position and DNA sequence are displayed at the top of each set of graphics. Aligned reads are visualised as grey blocks with pointed ends indicating the direction of each read. Insertions appear as purple boxes (A, B) or T symbols (C) and deletions are denoted by black lines connecting two sections of a read. The coverage histogram above the aligned reads visualises the number of reads mapping to each nucleotide in the sequence. A2F7 and A2E4 reads are shown in the top and bottom tracks respectively. A) A zoomed-out view, viewing 160 bp in "squished" display mode, where more reads from the sample can be seen without scrolling the track vertically. The top three tracks show the location of Chromosome 1 being viewed (thin red bar on chromosome; top), and the DNA sequence (middle track) with colour blocks indicating base identity. Red dashed line indicates the position of the gRNA_1 directed Cas9 cut-site at position 110,340,522 bp with it not shown over A2F7 to allow viewing of indels. B) A 28bp cropped segment of the 80bp zoomed view using "expanded" display mode. Top tracks showing genomic DNA sequence. C) A 33 bp cropped segment of the 80 bp zoomed view using "expanded" display mode to show a subset of mapped reads. Black and purple arrows indicate the positions of a 3 bp deletion and 1 bp insertion, respectively.*

### *2.4.1.4.3 Differential expression analysis indicates an effect on many genes from the* RBM15 *edit*

The bulk RNA sequencing data was analysed for differentially expressing genes between wild-type A2E4 cells and *RBM15*-edited A2F7 cells using the DESeq2 R package. Out of 33,691 genes expressed across the two conditions, 1,611 genes were significantly differentially expressed (adjusted p-value < 0.05; $Log_2$ Fold Change > 1.5 or < -1.5). Of these, 1,227 genes were upregulated ($Log_2$ Fold Change > 1.5) in the edited clone compared to the wild-type whilst 384 genes were downregulated ($Log_2$ Fold Change < -1.5; Fig.16).

This shows the edit in *RBM15* in A2F7 to be associated with significant and dramatic alteration in gene expression, with a bias towards upregulation. Table 5 shows the top 25 upregulated genes in the *RBM15*-edited clone, and Table 6 shows the top 25 downregulated genes in the wild-type *RBM15* clone. Among these differentially expressed genes, there were several with important functions across the cell, as well as genes that themselves regulate the expression of others. This includes genes with relevance to haematopoiesis including *HCLS1* and *PDGFRA* (discussed in detail in Chapter 2 Discussion).

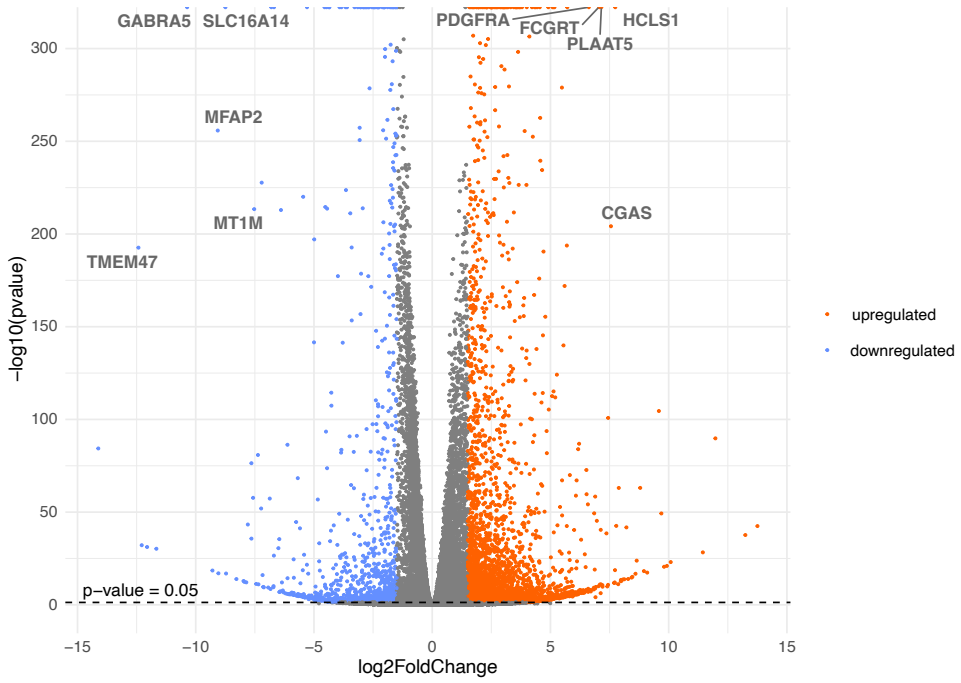


***Figure 16 - Many differentially expressed genes were identified between wild-type A2E4 and* RBM15*-edited A2F7 samples**. A volcano plot plotting the log2FoldChange against -log10(p-value) for each gene. Orange and blue dots indicate statistically significant (p-value <0.05) genes upregulated > log2(1.5) and < log2(-1.5), respectively. Genes for the top 5 statistically significant upregulated and downregulated genes are labelled.*

***Table 5 - Top 25 upregulated DE genes in the edited A2F7 clone replicates vs the wt A2E4 clone replicates.*** *Genes ordered by $Log_2$ Fold Change (shown to 3 decimal places). Gene Symbol and Description are from Ensembl.*

| Ensembl Gene ID | $Log_2$ Fold Change | p-value | Adjusted p-value | Gene Symbol | Gene Description |
|---|---|---|---|---|---|
| ENSG00000180353 | 7.741 | 0 | 0 | *HCLS1* | hematopoietic cell-specific Lyn substrate 1 |
| ENSG00000164430 | 7.559 | 2.24E-132 | 1.38E-129 | *CGAS* | cyclic GMP-AMP synthase |
| ENSG00000168004 | 7.153 | 3.49E-225 | 4.10E-222 | *PLAAT5* | phospholipase A and acyltransferase 5 |
| ENSG00000104870 | 7.071 | 0 | 0 | *FCGRT* | Fc gamma receptor and transporter |
| ENSG00000134853 | 6.623 | 0 | 0 | *PDGFRA* | platelet derived growth factor receptor alpha |
| ENSG00000118246 | 5.712 | 9.23E-266 | 1.40E-262 | *FASTKD2* | FAST kinase domains 2 |
| ENSG00000130508 | 5.485 | 1.30E-148 | 9.85E-146 | *PXDN* | peroxidasin |
| ENSG00000146555 | 5.171 | 2.31E-258 | 3.31E-255 | *SDK1* | sidekick cell adhesion molecule 1 |
| ENSG00000050555 | 5.13 | 0 | 0 | *LAMC3* | laminin subunit gamma 3 |
| ENSG00000148357 | 5.091 | 1.95E-191 | 1.80E-188 | *HMCN2* | hemicentin 2 |
| ENSG00000182985 | 4.912 | 0 | 0 | *CADM1* | cell adhesion molecule 1 |
| ENSG00000144908 | 4.549 | 3.97E-234 | 4.89E-231 | *ALDH1L1* | aldehyde dehydrogenase 1 family member L1 |
| ENSG00000135740 | 4.441 | 1.24E-282 | 2.28E-279 | *SLC9A5* | solute carrier family 9 member A5 |
| ENSG00000185345 | 4.372 | 6.85E-225 | 7.70E-222 | *PRKN* | parkin RBR E3 ubiquitin protein ligase |
| ENSG00000235823 | 4.329 | 0 | 0 | *OLMALINC* | oligodendrocyte maturation-associated lincRNA |
| ENSG00000118946 | 4.219 | 3.77E-141 | 2.63E-138 | *PCDH17* | protocadherin 17 |
| ENSG00000184908 | 3.999 | 2.28E-127 | 1.31E-124 | *CLCNKB* | chloride voltage-gated channel Kb |
| ENSG00000184221 | 3.797 | 4.03E-281 | 6.94E-278 | *OLIG1* | oligodendrocyte transcription factor 1 |
| ENSG00000013364 | 3.712 | 7.03E-133 | 4.43E-130 | *MVP* | major vault protein |
| ENSG00000170689 | 3.706 | 3.15E-206 | 3.25E-203 | *HOXB9* | homeobox B9 |
| ENSG00000188372 | 3.565 | 2.40E-147 | 1.77E-144 | *ZP3* | zona pellucida glycoprotein 3 |
| ENSG00000099365 | 3.443 | 4.65E-235 | 6.01E-232 | *STX1B* | syntaxin 1B |
| ENSG00000278023 | 3.21 | 4.34E-202 | 4.31E-199 | *RDM1* | RAD52 motif containing 1 |
| ENSG00000107186 | 3.047 | 1.93E-133 | 1.25E-130 | *MPDZ* | multiple PDZ domain crumbs cell polarity complex component |
| ENSG00000122378 | 2.955 | 1.09E-130 | 6.38E-128 | *PRXL2A* | peroxiredoxin like 2A |

***Table 6 - Top 25 downregulated DE genes in the edited A2F7 clone replicates vs the wt A2E4 clone replicates.*** *Genes ordered by Log2 Fold Change (shown to 3 decimal places). Gene Symbol and Description are from Ensembl.*

| Ensembl Gene ID | $Log_2$ Fold Change | p-value | Adjusted p-value | Gene Symbol | Gene Description |
|---|---|---|---|---|---|
| ENSG00000147027 | -12.422 | 6.03E-150 | 4.73E-147 | *TMEM47* | transmembrane protein 47 |
| ENSG00000186297 | -10.363 | 0 | 0 | *GABRA5* | gamma-aminobutyric acid type A receptor subunit alpha5 |
| ENSG00000117122 | -9.069 | 2.58E-179 | 2.22E-176 | *MFAP2* | microfibril associated protein 2 |
| ENSG00000163053 | -8.745 | 0 | 0 | *SLC16A14* | solute carrier family 16 member 14 |
| ENSG00000205364 | -7.526 | 5.05E-138 | 3.35E-135 | *MT1M* | metallothionein 1M |
| ENSG00000227128 | -7.21 | 4.91E-144 | 3.53E-141 | *LBX1-AS1* | LBX1 antisense RNA 1 |
| ENSG00000018236 | -6.803 | 0 | 0 | *CNTN1* | contactin 1 |
| ENSG00000101194 | -6.69 | 9.12E-201 | 8.73E-198 | *SLC17A9* | solute carrier family 17 member 9 |
| ENSG00000163032 | -6.395 | 4.99E-126 | 2.74E-123 | *VSNL1* | visinin like 1 |
| ENSG00000163017 | -5.455 | 5.00E-117 | 2.39E-114 | *ACTG2* | actin gamma 2, smooth muscle |
| ENSG00000151689 | -5.294 | 1.79E-182 | 1.60E-179 | *INPP1* | inositol polyphosphate-1-phosphatase |
| ENSG00000101605 | -4.551 | 0 | 0 | *MYOM1* | myomesin 1 |
| ENSG00000144681 | -4.467 | 1.13E-277 | 1.82E-274 | *STAC* | SH3 and cysteine rich domain |
| ENSG00000182050 | -4.364 | 4.47E-219 | 4.82E-216 | *MGAT4C* | MGAT4 family member C |
| ENSG00000124762 | -4.348 | 0 | 0 | *CDKN1A* | cyclin dependent kinase inhibitor 1A |
| ENSG00000040731 | -3.917 | 7.24E-246 | 9.85E-243 | *CDH10* | cadherin 10 |
| ENSG00000168314 | -3.838 | 3.06E-291 | 6.08E-288 | *MOBP* | myelin associated oligodendrocyte basic protein |
| ENSG00000162367 | -3.663 | 0 | 0 | *TAL1* | TAL bHLH transcription factor 1, erythroid differentiation factor |
| ENSG00000146094 | -3.602 | 1.83E-140 | 1.24E-137 | *DOK3* | docking protein 3 |
| ENSG00000115129 | -3.282 | 1.44E-163 | 1.20E-160 | *TP53I3* | tumor protein p53 inducible protein 3 |
| ENSG00000144278 | -3.15 | 1.74E-151 | 1.40E-148 | *GALNT13* | polypeptide N-acetylgalactosaminyltransferase 13 |
| ENSG00000184613 | -3.035 | 4.86E-122 | 2.46E-119 | *NELL2* | neural EGFL like 2 |
| ENSG00000188191 | -3.02 | 2.33E-127 | 1.31E-124 | *PRKAR1B* | protein kinase cAMP-dependent type I regulatory subunit beta |
| ENSG00000080546 | -2.817 | 3.44E-131 | 2.07E-128 | *SESN1* | sestrin 1 |
| ENSG00000101846 | -2.746 | 1.83E-107 | 7.61E-105 | *STS* | steroid sulfatase |

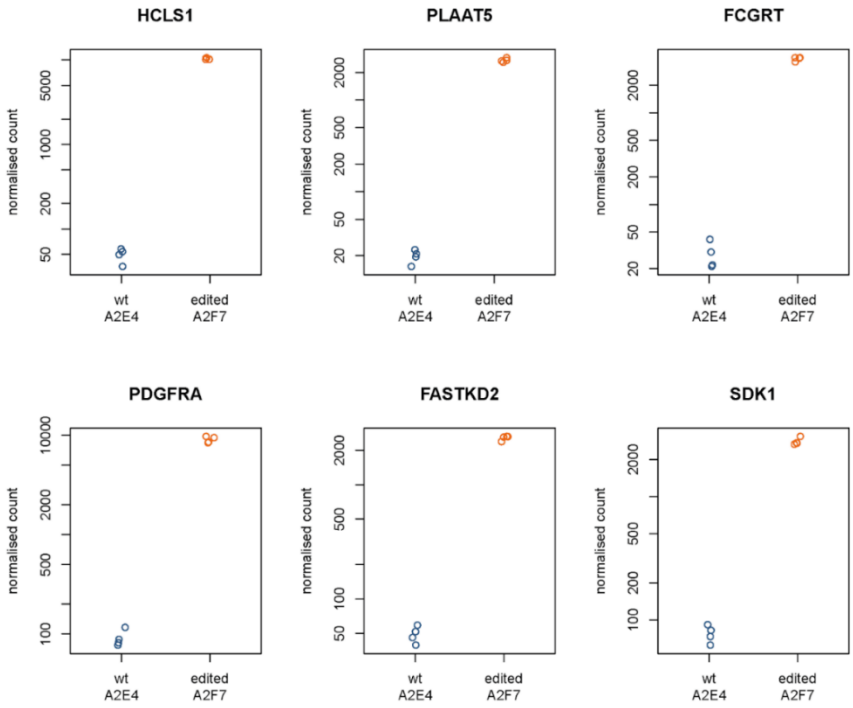


*Figure 17 - Normalised read count plots for a sample of the upregulated DE genes in the edited A2F7 clone replicates compared with the wt A2E4 clone replicates.*

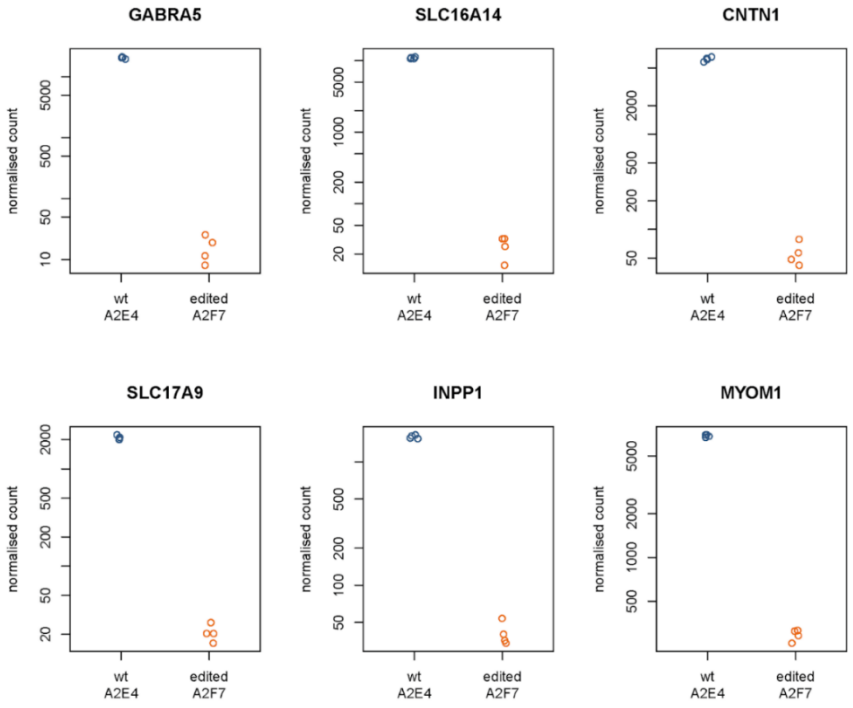


*Figure 18 - Normalised read count plots for a sample of the downregulated DE genes in the edited A2F7 clone replicates compared with the wt A2E4 clone replicates.*

*2.4.1.4.3 DAVID analysis of RNA-seq data to identify ontological themes among differentially expressed genes*

To better interrogate the data, DAVID Analysis was performed to identify whether particular biological features were enriched in the list of differentially expressed genes. To do this, the gene list of the 1,611 genes that were differentially expressed with an adjusted p-value <0.05 and an absolute log2 fold change of greater than 1.5 were uploaded to DAVID and were compared against a background list of all the 33,691 genes identified in RNA-seq dataset. DAVID Analysis then tested whether there was a statistically significant enrichment of certain gene ontological descriptions in the differentially expressed genes compared to these descriptions appearing by chance.

One ontology term selected was the UniProt Keyword “Biological Processes” (BP). This analysis revealed that there was a statistically significant enrichment of cell adhesion, ion transport, differentiation, neurogenesis, and osteogenesis among the differentially expressed genes (Fig.19).

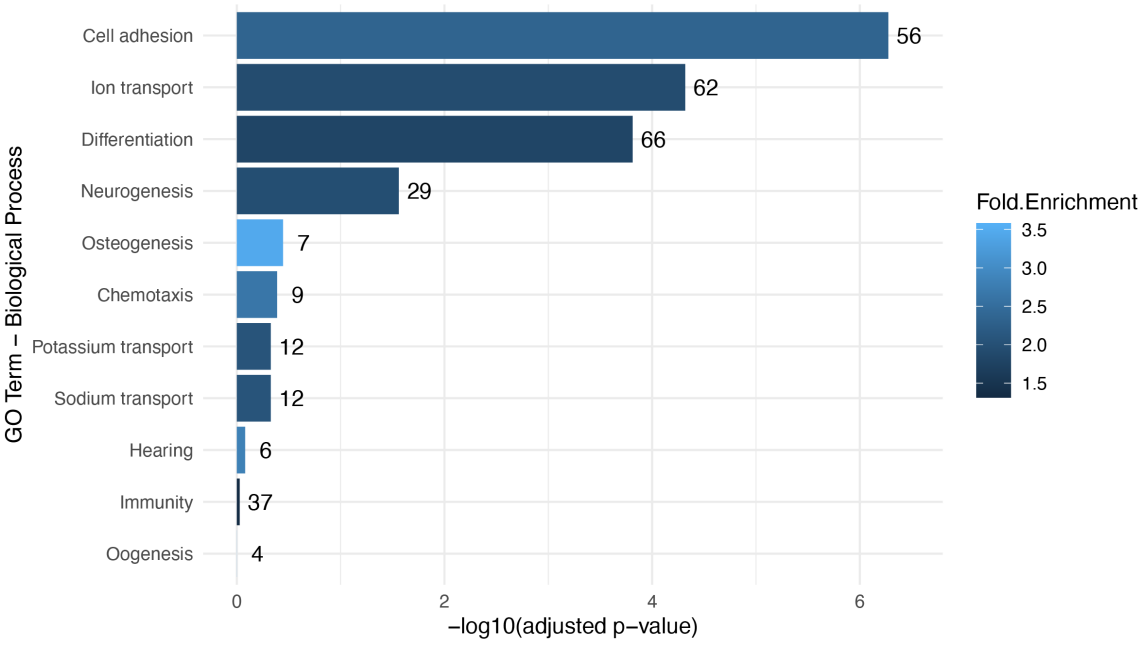


***Figure 19 - Gene ontology terms cell adhesion, ion transport, differentiation, and neurogenesis were significantly enriched in differentially expressed genes.*** *Gene enrichment DAVID Analysis using the UniProt Keyword “Biological Process” was performed to investigate if any biological processes were enriched in differentially expressed genes. The top biological processes enriched were plotted relative to their -log10(adjusted p-value) probability of being enriched compared to being present by chance alone, with the colour scale indicating fold enrichment. Numbers at the end of bars indicate the number of differentially expressed genes found with that ontology.*

Of particular interest was the enrichment of “Differentiation”, given the relationship between RBM15 and haematopoiesis. The gene list for “Differentiation” from DAVID was then used to examine the fold change and p-value of each of the associated differentially expressed genes that were annotated with the “Differentiation” ontology. This revealed that *BMP7*, *NTRK3*, and *SYT4* were the most upregulated “Differentiation” genes in the A2F7 samples compared to the A2E4 samples, and *SOHLH2*, *SEMA3C*, and *LBX1* were the most downregulated (Fig.20).

DAVID Analysis also clusters gene ontology terms that have a related biological process or function and provides an enrichment score for each generated cluster. Interestingly, one of the top clusters contained proteins that had transcription factor TCF3, MYOD, TFAP4, and NHLH1 binding sites (Enrichment Score: 6.87). Each of these terms was strongly statistically significantly enriched in the differentially expressed genes, with adjusted p-values of <0.0001. Interestingly, TCF3 has been shown to interact with TAL1, encoded by one of the strongest downregulated genes in the A2F7 cells, to form a complex that is essential in terminal erythroid expression (Hsu *et al.*, 1994; Wadman *et al.*, 1997), and all four transcription factors are involved in differentiation. Confusingly however, none of these four genes were significantly differentially expressed between the A2E4 and A2F7 populations, implying that the mechanism driving the enrichment of proteins with these transcription factor binding sites in the differentially expressed genes was not through direct changes in the expression of *TCF3*, *MYPD*, *TFAP4* or *NHLH1*.

Other interesting data features highlighted by the DAVID analysis was that over 10% of differentially expressed genes were located on chromosome 19, with many in the region 19q13.33. This enrichment was found to be strongly significant, with an adjusted p-value of <0.0001. Interestingly, the gene *FCGRT* is located in this 19q13.33 region, with this being the fourth highest differentially expressed gene in the transcriptomic dataset. Alone, these data are interesting, but they also highlight that changes in gene expression between the A2E4 and A2F7 clones were unlikely to be solely due to differences in ploidy, otherwise enrichment of any particular chromosomal region should not occur (except Chromosome 15 – Fig.4).

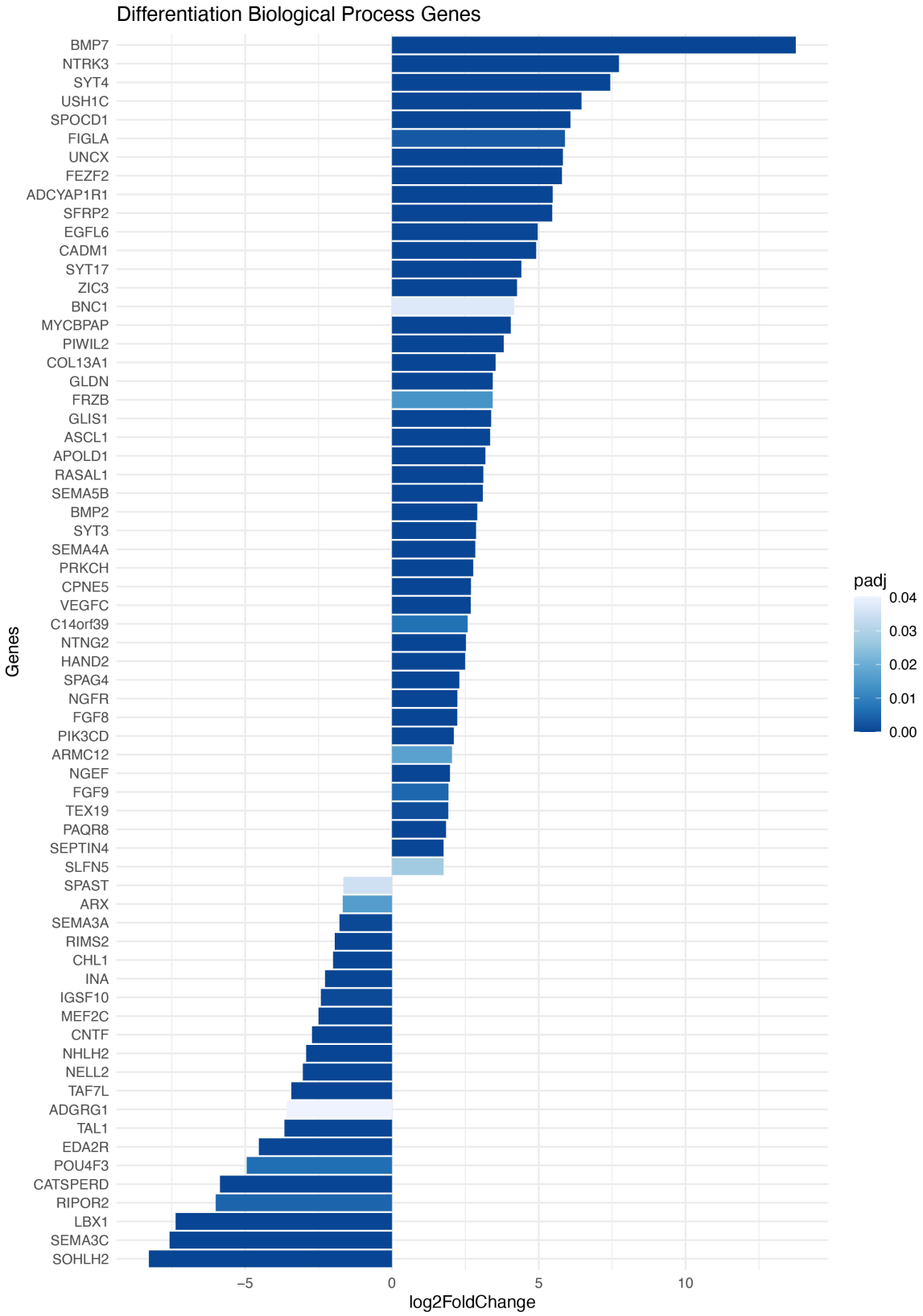


***Figure 20* - BMP7, NTRK3, SYT4 *and* SOHLH2, SEMA3C *and* LBX1 *were the "Differentiation" terms with the greatest up- and downregulation in the acquired RNA-seq dataset.*** *The gene list for the "Differentiation" UnitProt Keyword 'Biological Process' was downloaded and used to filter the differentially expressed genes between A2E4 and A2F7 samples in the RNA-seq dataset. The log2FoldChange for each "Differentiation" annotated differentially expressed gene was plotted on the x-axis, with each gene represented by a different row. Colour scale indicates the adjusted p-value for each differentially expressed gene.*

#### *2.4.1.4.4 Determination of the DE of related paralog* RBM15B *in the edited clone*

*RBM15B* is a paralog of *RBM15*, sharing functional similarities including in XIST-mediated transcriptional silencing (Patil *et al.*, 2016; Uranishi *et al.*, 2009). Therefore, it was investigated whether a change in the expression of *RBM15* was associated with changes in the expression of *RBM15B*. Plotting and comparing the normalised counts for *RBM15B* between the wild-type and edited clones shows that *RBM15B* expression was at least halved in the edited A2F7 clone (Fig.21). This implies that *RBM15* and *RBM15B* expression is inversely correlated.

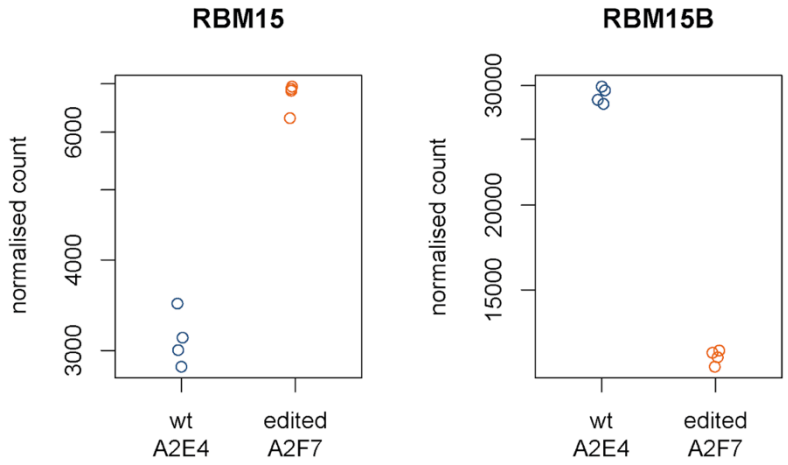


***Figure 21* - RBM15 *and* RBM15B *shows an inverse relationship of gene expression.*** *The normalised counts for* RBM15 *(left graph) and* RBM15B *(right graph) were plotted for wild-type A2E4 (blue circles) and* RBM15*-edited A2F7 samples (orange circles).*

#### *2.4.1.4.5 Comparison with eCLIP data supports* RBM15 *edits as a cause for the differential expression*

Enhanced cross-linking and immunoprecipitation (eCLIP) assays are able to map the binding sites of RBPs, such as RBM15, on their target RNAs. This works by UV-crosslinking of protein and RNA, immunoprecipitation of the stabilised protein-RNA complexes by use of an antibody specific against the protein of interest, and cDNA generation by reverse-transcription of the immunoprecipitated RNA. Therefore, by using an eCLIP dataset, it would be possible to investigate whether genes whose RNA had an RBM15 binding site were significantly altered in expression in the *RBM15*-edited clone A2F7 compared to the wild-type A2E4 population.

To perform this analysis, an eCLIP dataset specific against RBM15 generated in K562 cells was downloaded from ENCODE (ENCODE Project Consortium, 2012; Stanford

University, 2022; accession ENCFF006CBO). This data provided the list of genes encoding RNA transcripts which are bound to by RBM15. Using this list against the transcriptomic dataset generated here, 13,506 genes out the total 33,691 total genes identified produced transcripts that had a RBM15 binding site (40.1%), and 1,122 genes out of the 1,611 differentially expressed genes between *RBM15*-edited A2F7 and wild-type A2E4 populations had transcripts that had an RBM15 binding site (69.6%). Hypergeometric testing was used to assess whether there was an increased probability of a gene whose RNA is associated with RBM15 binding appearing in the list of differentially expressed list of genes compared to chance. This analysis resulted in a p-value of <0.0001, showing that the differentially expressed genes between A2F7 and A2E4 were statistically significantly enriched in genes whose RNA contain RBM15 binding sites. Out of the 13,506 genes whose RNA is bound by RBM15 based on the eCLIP data, 866 genes had transcripts that were significantly increased in abundance, and 256 genes had transcripts that were significantly decreased in abundance (Fig.22). This implies that the differences in *RBM15* expression (likely the overexpression of the first RRM of RBM15) in the A2F7 clone led to a dramatic and significant alteration in the abundance of transcripts with an RBM15 binding site, with a bias towards increased transcript abundance.

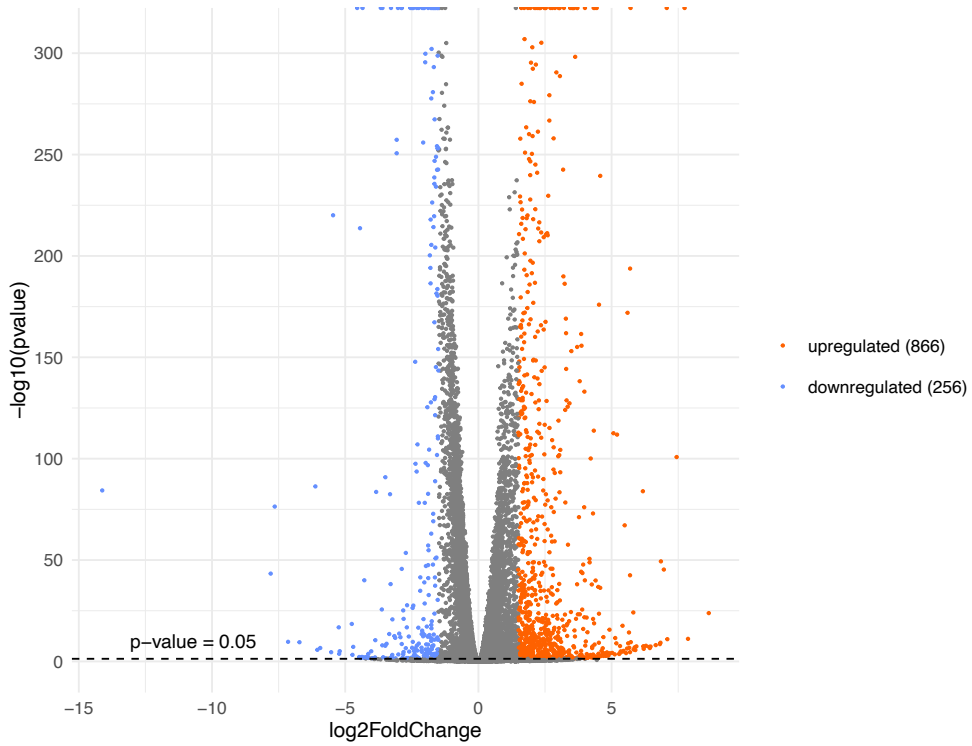


***Figure 22 - Many RBM15-binding site-containing transcripts were detected and show a bias towards increased abundance.*** A volcano plot plotting *log2FoldChange against -log10(adjusted p-value) for each gene annotated to have an RBM15-binding site as identified by an eCLIP dataset (ENCODE accession ENCFF006CBO). Orange and blue dots indicate statistically significant (adjusted p-value <0.05) genes upregulated > log2(1.5) and <log2(-1.5), respectively.*

## 2.4.2 K562 *RBM15*-KO experiment

The editing of *RBM15* was repeated in another cell line. K562 was chosen for these experiments because, as an erythroleukaemic cell line, it represents a malignant haematopoietic progenitor cell population with an erythroid phenotype. Investigating the functions of genes involved in AS in blood cells would be particularly interesting in K562 cells, especially when assessing the functional consequences of knocking out erythroid lineage-related SF genes. Additionally, the effects of changes in dosage of such genes could also be investigated on multiple levels with the hypertriploid K562 line by generating incomplete as well as complete knockouts (knocking out one, two or all three alleles). However, difficulties arose during the subcloning of the transfected K562 cells which meant the experiments in K562 were put on hold whilst the process was first tried in HAP1 cells.

### *2.4.2.1 Testing the effectiveness of different* RBM15 *gRNAs in K562 cells*

It was necessary to test the effectiveness of gRNAs against *RBM15* in K562 Cas9+ cells. The same three gRNAs (gRNA_1, gRNA_2, gRNA_3) as used in the HAP1 editing experiments were transfected into K562 cells, genomic DNA extracted, the locus surrounding the predicted gRNA-targeted Cas9 cut-site amplified by PCR, and the amplicons sequenced by Sanger sequencing. Traces were examined for presence of multiple overlapping peaks after the cut-site.

The sequencing trace for gRNA_1 (Sample K) showed the presence of a weak alternative trace underlying the dominant wild-type sequence, with this starting beyond the cut-site and continuing for the rest of the sequence read length (Fig.23). This alternative trace began further from the cut-site compared to the equivalent gRNA_1 test in HAP1 cells (Fig.23 *vs* Fig.7). The sequencing trace for gRNA_2 (Sample L) did not show the presence of any alternative peaks underlying the dominant wild-type trace, although the low sequence quality around the sgRNA target sequence and cut-site may have masked any low alternative signal peaks present (Fig.23). As in the HAP1 experiment, the efficacy of gRNA_3 was unable to be tested in K562 cells (Sample M) because the PCR primers failed to produce amplicons. Overall, these data suggested that transfection of gRNA_1 in K562 Cas9+ cells was able to produce some *RBM15* edited cells.

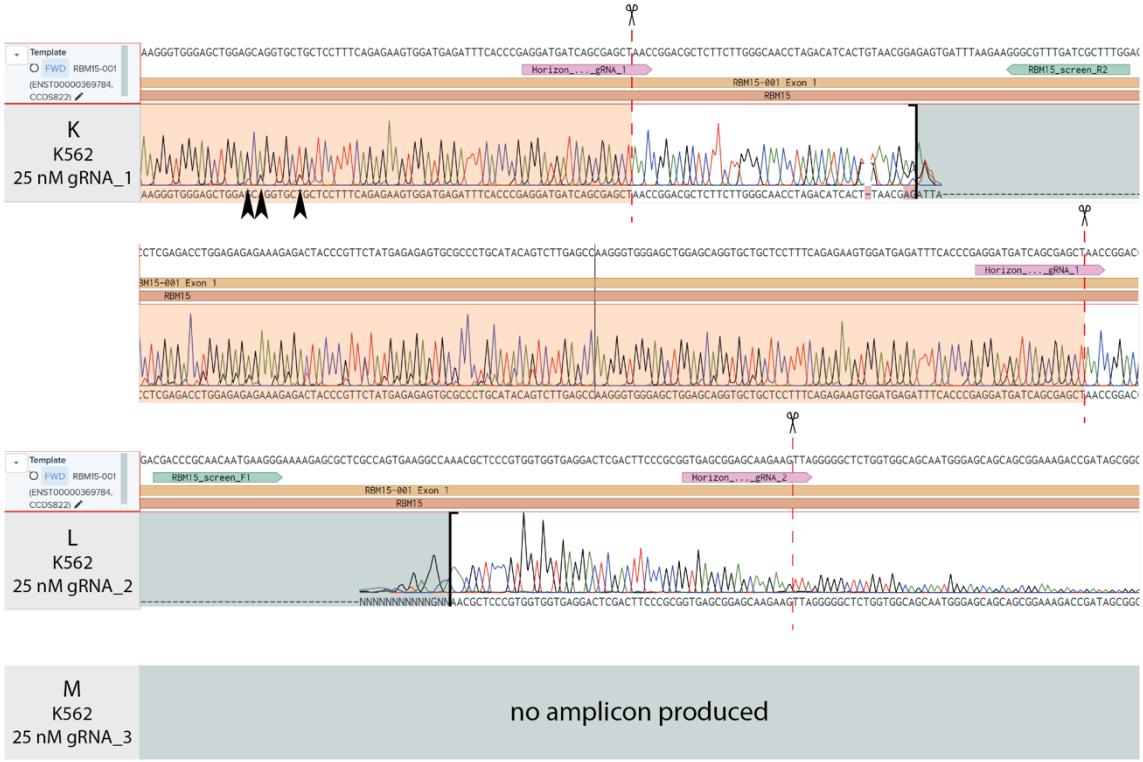


***Figure 23 - Sanger traces indicating signs of possible editing in K562 cells treated with 25nM of gRNA_1, but not 25nM of gRNA_2.*** *K562 cells were transfected with 25nM gRNA_1, gRNA_2, or gRNA_3, genomic DNA extracted, amplicons around the Cas9 cut-site amplified by PCR, and sequenced using Sanger sequencing. Screenshots show sequencing results aligned against the wild-type* RBM15 *sequence. The top trace shows the start of the sequencing read and is continued in the bottom trace. The red line indicates the Cas9 cut-site, and black arrows indicate clear examples of non-wild-type signal trace.*

### *2.4.2.2 Single cell sorting of gRNA_1 transfected K562 cells*

The mixed population of edited and unedited K562 cells treated with 25nM of gRNA_1 was single-cell sorted into 96-well plates in order to generate clonal populations. These populations would be then expanded and screened for editing in *RBM15* by Sanger sequencing and/or next generation sequencing to deconvolve the edits generated in each allele of the hypertriploid K562 cell line. Unfortunately, however, the single-cell clones failed to grow following FACS, despite sorting for live cells (based on propidium iodide staining) and culturing for many weeks. It was decided to continue the editing experiments using only the HAP1 Cas9+ cells and that future experiments would troubleshoot the K562 subcloning.

## *2.5 Discussion*

To investigate the effects on downstream gene expression upon knockout of splicing factor genes in the haematopoietic system, *RBM15* was targeted for editing in a stable Cas9 HAP1 cell line by transfection with single guide RNAs. Editing was confirmed by Sanger sequencing and bulk RNA-seq was performed on a wild-type clonal population and an *RBM15*-edited population.

### 2.5.1 Analysis of the genotype of RBM15 edited HAP1 cells

Whilst *RBM15* was differentially expressed in the edited A2F7 clones compared to wild-type A2E4 clones, the number of transcripts mapping to *RBM15* was unexpectedly increased to double the wild-type expression following the experiment to knock the gene out. A2F7 is suspected to have been diploid since the Sanger sequencing trace showed multiple overlapping traces following the cut-site instead of only a single trace. Upon viewing the RNA-seq reads in IGV, the diploid heterozygous status of A2F7 was confirmed.

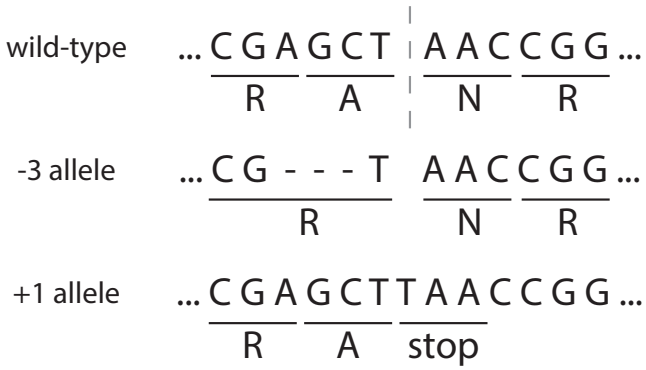


***Figure 24 - RBM15 edits at base and amino acid levels.*** *A 12 bp section of the wild-type RBM15 DNA sequence (top) grouped into codons (underlined) is shown along with the amino acids each codon encodes below. The gRNA_1 cut-site is shown by the grey dotted line. The DNA sequences for the -3 deletion (middle) and +1 insertion (bottom) alleles are shown below for comparison. The 3 deleted bases in the -3 allele, each marked by a hyphen, still result in an arginine (R) residue: the redundancy of the DNA code allows CGT to code for R as well as CGA. The following alanine is deleted, however, the reading frame remains intact and the subsequent sequence unchanged. The inserted T base in the +1 allele creates a stop codon which terminates transcription and creates a truncated amino acid sequence.*

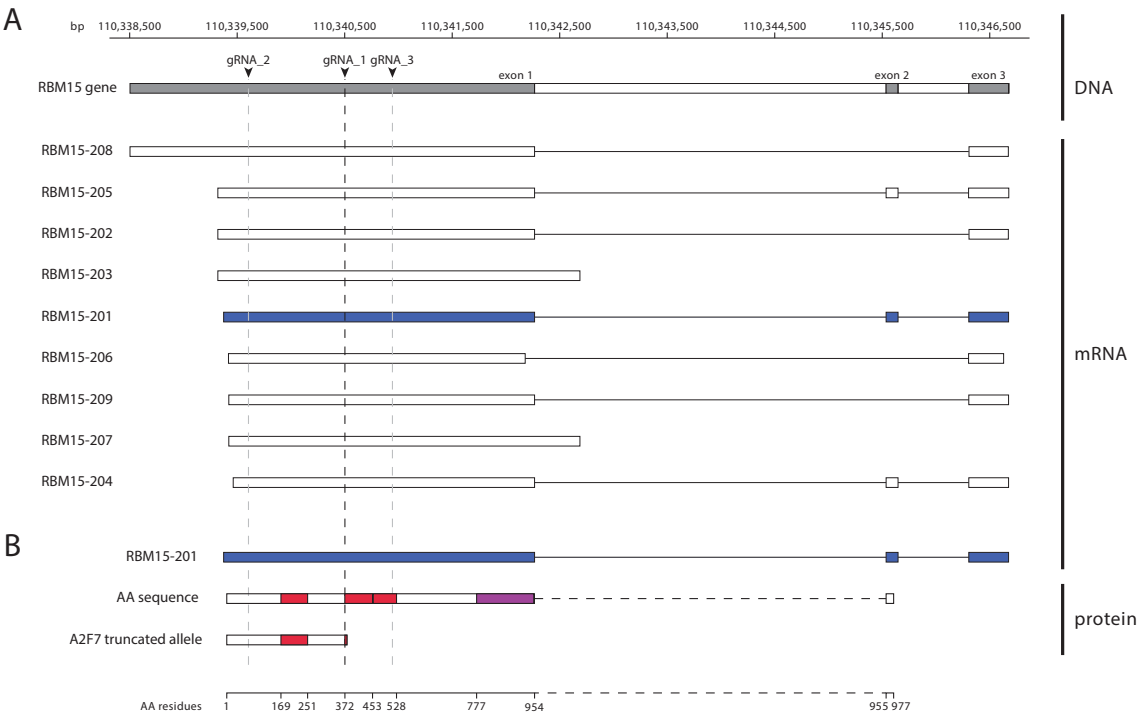


***Figure 25 - Structures of* RBM15 *gene, isoforms, amino acid sequence and A2F7 truncated allele.*** *(A) The structure of the human* RBM15 *gene is annotated with grey sections to mark the three exons. A base-pair (bp) scale is shown above to indicate genomic locus. The three cut-sites for the gRNAs tested in this project are represented by vertical dashed lines: black for gRNA_1 which was selected for further experiments, grey for gRNA_2 and _3. The nine human* RBM15 *transcript isoforms are shown underneath the gene (comprehensive GENCODE transcript set), with boxes representing exons and connecting solid lines representing introns. The primary isoform, RBM15-201, is highlighted in blue. (B) The structure of the RBM15-201 transcript is shown again for comparison against the wild-type and truncated A2F7 amino acid (AA) sequences below it. The boxes in the AA sequence diagrams represent the polypeptide molecule, with a horizontal dashed line to indicate where the intronic sequences are not translated into amino acids. Red sections indicate RRMs and the purple section highlights the SPOC domain. The A2F7 truncated allele ends just after the cut-site, at residue 373, the 7$^{th}$ residue of the second RRM prior to the start of the β pleated sheet at residue 375. An AA residues scale corresponding to the residues in the wild-type sequence is shown at the bottom, with key residues labelled. Labels on the right indicate the molecule represented by each structure schematic: DNA, mRNA or protein.*

A2F7 appears to produce two different *RBM15* transcripts: a -3 deletion transcript and a +1 insertion transcript. The -3 deletion transcript will not have led to any premature stop codons as the deletion is in-frame, and likely produced an almost-full length RBM15 protein, with only Ala-372 removed from the protein (Fig.24). Given the inert nature of alanine and given that Ala-372 is located in an unstructured region of RBM15, it is likely that the protein generated from the "-3 allele" was functionally wild-type. However, contrary to the prediction, A2F7 also generates a +1 transcript that would lead to the generation of a premature stop codon at position 373. However, given that this transcript is equally as abundant as transcripts encoding the -3 deletion, it does not appear to be degraded by nonsense-mediated decay. This means that the +1 allele in A2F7 likely generates a truncated RBM15 protein that encodes only the first 372 amino acids of RBM15: only the first RRM (Fig.25).

It is not possible to confirm the ploidy of A2E4 from the experiments performed. Because it is unedited, viewing *RBM15* RNA-seq reads for A2E4 in IGV only reveals a single set of transcripts generated, but does not distinguish from how many chromosomes these originated. However, the potential difference in ploidy between A2E4 and A2F7 should not impact upon the interpretation of the RNA-seq transcriptomic dataset generated, as the normalisation of transcript abundance by total read count per sample should negate this. Therefore, this suggests that the increase in *RBM15* expression in A2F7 results from the heterozygous edit of *RBM15* to produce one essentially full-length protein (missing a single amino acid), and one truncated protein. Whilst it is unclear what this mechanism may be, is conceivable that if the truncated version of RBM15 was unable to provide the same function as the full-length RBM15, and RBM15 action was dosage-dependent, the edited cells may upregulate *RBM15* to allow the generation of the required two almost full-length versions of RBM15, whilst also inadvertently increasing the production of the RBM15 truncate. Therefore, A2E4 could produce two copies of fully functional RBM15, whilst A2F7 produces four copies comprising of two almost-full length copies (which are required for typical RBM15 function) and two truncated copies. Alternatively, it may be that the expression of the truncated version of RBM15 has a direct effect on the expression of *RBM15* via a positive feedback loop, thereby increasing its expression.

Testing the above hypotheses require several future experiments. Firstly, it is necessary to verify the haploidy of the A2E4 wild-type clone and the diploidy of the A2F7 clone. This could be performed using an established flow cytometry ploidy assay such as that developed by Beigl, Kjosås, Seljeseth and colleagues (2020). Alternatively, genotyping could be performed by short-read next generation sequencing. This would allow the separation of the different allele sequences seen in the Sanger sequencing traces. Ploidy could also be inferred by comparing raw counts from the RNA-seq data between the conditions and a known haploid control. Secondly, it needs to be investigated whether the A2F7 cells produce two different RBM15 proteins: the near-full-length and the truncated proteins. This could be performed by Western blotting lysed A2F7 cells, separating proteins using SDS-PAGE, and identifying RBM15 by use of a polyclonal RBM15 antibody raised against the complete RBM15 protein. If the truncated protein was present, and assuming the polyclonal antibody can still bind the truncated protein, it should present as a smaller molecular weight band compared to the wild-type protein.

### 2.5.2 Editing *RBM15* likely has a large impact on gene expression due to its multiple targets

Although *RBM15* was not knocked out as intended, *RBM15* was successfully edited on both alleles in A2F7 clones with the probable consequence being the generation of a functionally wild-type copy of *RBM15* and a truncated *RBM15* that only produces the first RRM. The RNA-seq analysis is therefore still informative but is likely the comparison between a clone of unknown ploidy that is wild-type for RBM15 and a diploid clone that produces near-wild-type *RBM15* as well as *RBM15* only formed of its first RRM. Differential expression analysis showed 1,611 out of 33,691 genes to be differentially expressed, composed of 1,227 upregulated and 384 downregulated genes. This bias towards upregulated gene expression in the A2F7 clone may be a result of the expression of the truncated RBM15 protein.

Despite these problems, the dataset is worth analysing to see whether haematopoiesis and AS factors were among those differentially expressed in the edited A2F7 clone. In this study, it was hypothesised that by knocking out *RBM15* in HAP1 cells, there would consequently be statistically significant changes in the expression of the genes and RNA transcripts RBM15 influences. In previous studies, knockout of *Rbm15* in mice has had significant effects on haematopoietic processes (Raffel *et al.*, 2007), which would be expected to be reflected in the gene expression profiles of the cells. Similarly, a large number of genes were significantly differentially expressed, including genes encoding known RBM15 targets TAL1, GATA1, RUNX1 and MYC. Notably, MPL was not altered in its expression (Table 7; Zhang *et al.*, 2015; Niu *et al.*, 2009). *RBM15* possesses vital functions including m6A methylation and regulation of AS and megakaryocyte differentiation, among others. It therefore has multiple mechanisms of action (splicing and methylation) upon many genes and its targets are often significant transcription factor genes, which themselves affect the expression of other genes. By editing *RBM15*, there might be a cascade effect as the differential expression of target genes affects the normal expression of its own targets downstream and so on.

***Table 7 – Differential expression of targets of RBM15.*** *A red-blue colour scale has been applied to the log2 fold change column showing high fold change in red and low in blue. P-values and adjusted p-values highlighted in green show statistical significance (<0.05).*

| Gene Symbol | Ensembl Gene ID | $Log_2$ Fold Change | p-value | Adjusted p-value |
|---|---|---|---|---|
| *TAL1* | ENSG00000162367 | -3.66272 | 0 | 0 |
| *MYC* | ENSG00000136997 | 1.144078 | 3.94E-142 | 1.54E-140 |
| *GATA1* | ENSG00000102145 | 5.446037 | 1.05E-03 | 1.87E-02 |
| *RUNX1* | ENSG00000159216 | -0.4073 | 3.81E-05 | 8.45E-05 |
| *MPL* | ENSG00000117400 | 3.132943 | 0.079698 | 0.120291107 |

Some of the genes which are highly up- or down-regulated in the edited A2F7 clone compared to the wild-type A2E4 have important functions, some within the haematopoietic system. One such example is the most highly upregulated gene in A2F7, *HCLS1* (haematopoietic cell-specific lyn substrate 1). Expressed in haematopoietic lineages and important in myelopoiesis, HCLS1 is part of the transcription regulator complex and is also involved in the regulation of signal transduction and phosphorylation (Kitamura *et al.*, 1989; Skokowa *et al.*, 2012). *PDGFRA* is another upregulated gene which encodes the platelet-derived growth factor receptor (alpha). A tyrosine-protein kinase, PDGFRA promotes and inhibits cell proliferation, is involved in differentiation of bone marrow-derived mesenchymal stem cells as well as platelet activation (Selheim *et al.*, 2000; Sévère, Miraoui and Marie, 2011) and the promotion of many signalling cascades, including *AKT1*, *HRAS* and *MAPK* signalling (Heinrich *et al.*, 2003). The upregulation of these genes would have contributed substantially to the large number of differentially expressed genes in the *RBM15*-edited clone; RBM15 is clearly associated with some crucial signalling pathways. It would be interesting to further investigate the association between RBM15 activity and haematopoiesis genes.

RBM15 also may be involved in the regulation of Hox genes (Wiellette *et al.*, 1999), genes responsible for determining the structure of the body during embryonic development. The functions of Hox genes post-development are less well-studied but research has shown that they regulate the differentiation of stem cells (Song *et al.*, 2020). Many of the genes that are differentially expressed when *RBM15* was edited are Hox genes, such as HOXB9 (log2FC = 3.71) and even SOX2 (log2FC = 1.78). As with the genes involved in important intracellular signalling pathways above, altering the expression of the important regulatory Hox genes (as it seems the *RBM15* editing has) could largely explain the differential expression observed.

Gene lists obtained through eCLIP experiments against RBM15 reveal that over half of the differentially expressed genes (adjusted p-value > 0.05) produce RNA bound by RBM15. RBM15, as a member of the N6-methyltransferase complex (MTC), alters the RNA stability of the RNA it binds to. One suggestion therefore is that in the A2F7 clone, the edited RBM15 (likely overexpression of the first RRM) is altering the RNA stability of the transcripts it can bind to, changing their abundance as detected by the RNA-seq. If so, this could reveal mechanisms of RBM15 involvement with RNA stability machineries such as the MTC, and particularly reveal the function of the first RRM of RBM15. Important questions to answer in future analyses are whether there is a global pattern of increased or decreased transcript abundance among the RNAs bound by RBM15 in the A2F7 clone, and whether the subset of RBM15-binding transcripts with increased

abundance have similarities in RNA structure or sequence composition, and how that differs to RBM15-binding transcripts with decreased abundance. This analysis could provide major advances in our understanding of the functions of RBM15 in RNA stability and as part of the MTC.

DAVID analysis revealed that differentially expressed genes in the A2F7 clone were enriched in the ontological term “Differentiation”. Major upregulated “Differentiation” genes in A2F7 were identified to be *BMP7*, *NTRK3*, and *SYT4*, and major downregulated genes included *SOHLH2*, *SEMA3C*, and *LBX1*, some of which have interesting associations to blood cell development. For example, *BMP7* has been described as a key factor inducing the differentiation of immunosurveillance Langerhans cells (Yasmin *et al.*, 2013). *NTRK3*, a tyrosine receptor kinase, has been strongly associated with haematopoiesis, given that its oncofusion with transcription factor ETV6 can lead to acute myeloid leukaemia and acute lymphoblastic leukaemia (Joshi *et al.*, 2019); and transcription factor SOHLH2 has been associated by GWAS with multiple myeloma (Duran-Lozano *et al.*, 2021). Other interesting findings from the analysis was that differentially expressed genes were enriched on chromosome 19, and that there was an overrepresentation of genes with differentiation related *TCF3*, *MYOD*, *TFAP4*, and *NHLH1* binding sites, although the significance of these observations requires further investigation.

### 2.5.3 Inverse relationship in expression of *RBM15* and *RBM15B*

*RBM15* and another gene, *RBM15B*, appear to have redundant function in the regulation of XIST-mediated transcriptional silencing. A double knockdown of the genes inhibited XIST-mediated transcriptional silencing but this did not occur when each gene was individually knocked down (Patil *et al.*, 2016). *RBM15B* shares sequence and domain organisation similarities to *RBM15*, including the presence of three RRMs and a SPOC domain, allowing it to potentially compensate for any loss of RBM15 function, at least in the context of transcriptional silencing. This compensation is explained in part by the apparent ability of RBM15B to play the same function as RBM15 in the MACOM and MACOM-like WTAP subcomplexes in the N6-methyltransferase complex (Knuckles *et al.*, 2018, Patil *et al.*, 2016).

Given this, it was thought prior to the experiment that editing *RBM15* without also editing *RBM15B* may only have mild effects in some pathways. However, this does not appear to be the case, with the occurrence of many differentially expressed genes upon the overexpression of the first RRM of RBM15 in the A2F7 clone. Interestingly, *RBM15B*

expression in the wild-type clone was more than double that in the *RBM15*-edited clone. This suggests that *RBM15B* was accidentally targeted by the gRNA_1 given the large sequence homology between the two genes, or perhaps more tempting, that the overexpression of the first RRM of RBM15 was causing *RBM15B* to be downregulated. It was checked whether gRNA_1 would have accidentally targeted *RBM15B* by alignment of gRNA_1 with *RBM15*. This revealed that there would be 5 mismatches between gRNA_1 and *RBM15B* before the gRNA_1 sequence could anneal to the *RBM15B* locus (Fig.26). Even if this were possible, this binding site is not flanked by a PAM site in the appropriate position, meaning the Cas9 would be unable to cut. This hypothesis can therefore be discounted. If the changed expression was caused by the *RBM15* edit, it would be interesting to know whether a feedback loop involving the two genes could be at play following the overexpression of the first RBM15 RRM. Whilst the mechanism for this is not understood, knowledge of the relationship between these two genes and what other pathways they both function in would benefit from further experimental investigation.

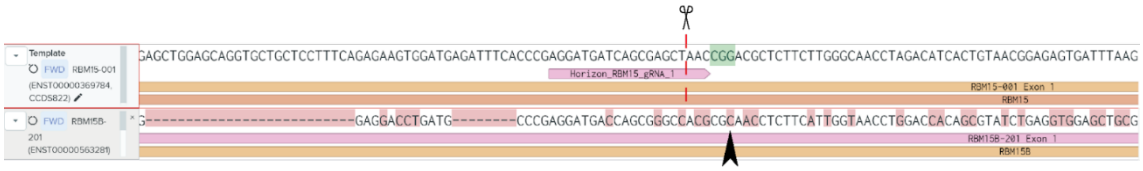


***Figure 26 – gRNA_1 has poor alignment to* RBM15B.** *Screenshot from Benchling showing the alignment of* RBM15B *(bottom track) to* RBM15 *(top) with mismatches and deletions highlighted in red and the RBM15_gRNA_1 PAM sequence highlighted in green. The black arrowhead highlights the lack of PAM sequence in* RBM15B *after the gRNA_1 alignment.*

## 2.5.4 Diploidy of verified edited cells

One priority for future experiments would be to experimentally investigate why the verified edited clones in both *RBM15*-KO experiments became diploid. Two working theories are as follows. First, the cells were simply too confluent upon transfection and had become diploid prior to transfection, as is known to occur (Olbrich, Mayor-Ruiz, Vega-Sendino *et al.*, 2017; Beigl *et al.*, 2020). Alternatively, the cells could not tolerate the knockout and were hence forced into a diploid state to survive.

There is a lower efficiency of editing with increasing ploidy (lower chance of editing both alleles in a cell) and edited polyploid cells may be heterozygous as a result. The first theory could be tested by transfecting at a much lower density in a third experiment. Caution played a factor into transfecting the cells at a higher confluency in the first experiment; it was unknown how toxic the transfection would be for the cells and whether many would survive the process. Given that the high density causing the HAP1 cells to

become more confluent ultimately seemed to be the larger problem, the cells were not allowed to become quite so confluent for the second experiment (Fig.27). The number of passages prior to transfection were also kept as low as possible (the cells had been passaged only twice by transfection). Despite this, the confluency was still too high, if that is what causes the conversion to diploidy.

A stricter restriction on the maximum confluency of the cells at any time during a third experimental repeat and continued minimisation of passage numbers as well as regular ploidy checks by staining and flow cytometry (Beigl *et al.*, 2020) should allow for proper prevention of and monitoring for diploidisation. In addition, the cell-permeable violet DNA staining could be used during the single-cell sorting of cells into the 96-well plate, allowing cells that were haploid at the time of sorting to be selected for. Given that HAP1 cells that have been subcloned are more successful in maintaining their haploid status (Olbrich, Mayor-Ruiz, Vega-Sendino *et al.*, 2017), they would have a high chance of remaining so if haploid at the time of single-cell sorting.

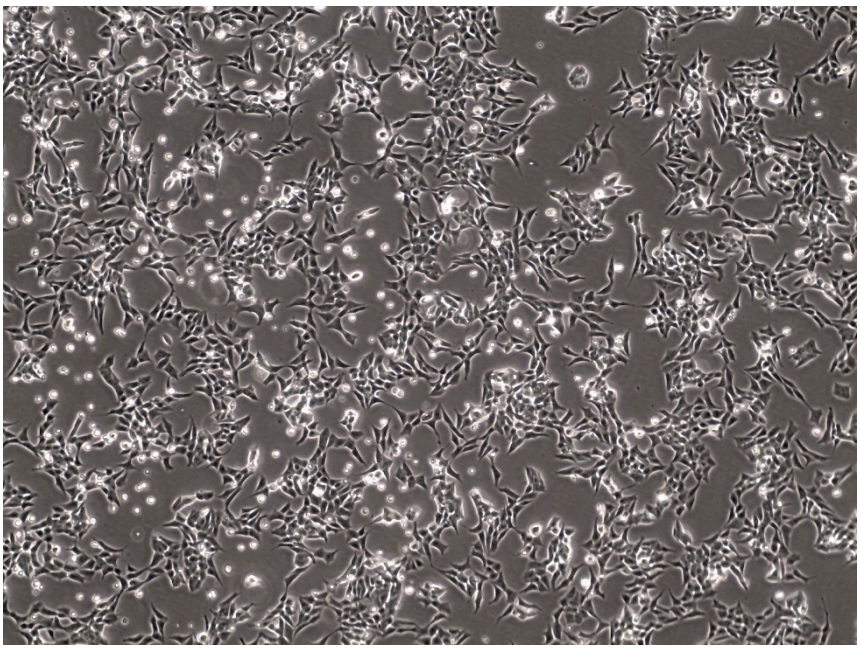

***Figure 27 – Confluency of sample A2 cells just prior to transfection with RBM15_gRNA_1.*** *Image captured using EVOS XL Core Microscope (40x objective).*

HAP1 cells with a gene knocked out will become diploid homozygous knockouts (-/-) if allowed to diploidise (Beigl, Kjosås, Seljeseth *et al.*, 2020). Yet it is uncertain what genotype results if cells cannot tolerate the editing of a gene and are thus forced into a diploid state to survive. Homozygosity might occur but if one allele had already been

edited, a true knockout (-/-) would have been the result. Heterozygosity might occur as the cells revert to diploidy and one of the alleles is correctly repaired by chance but a mechanism for this is so far unclear. The hypothesis would nevertheless be worth testing.

The wild-type clone A2E4 might have remained haploid whereas A2F7 became diploid. It is important to note why this might have happened. HAP1 cells that have been subcloned from single-cells (as in this experiment) have been shown to be more successful in maintaining their haploid status (Olbrich *et al.*, 2017). A2E4 was haploid at the point of single-cell subcloning and therefore had an increased chance of remaining so. This further supports the theory that A2F7 had become diploid prior to subcloning but it is still uncertain as to whether that occurred before or during transfection. Overall, it seems more likely that A2F7 had become diploid and one *RBM15* allele had subsequently been edited although other theories cannot yet be discounted.

### 2.5.5 Future experiments

This experiment has yielded some interesting results that may lead to the elucidation of the importance of RBM15. Further studies must additionally be performed to consolidate the hypotheses that have arisen. As described above, the expression levels of RBM15 (and any other proteins of interest, such as RBM15B) should be measured by a gene expression qPCR assay and also on the protein-level by Western blot. Measuring the final output of RBM15 protein would put the editing of *RBM15* in A2F7 and the subsequent apparent increase in expression into context; clarification of whether the edited allele produces a functional *RBM15* transcript is a priority.

Additionally, the ploidy of the A2E4 wild-type clone must also be confirmed. If the clone is aneuploid, analysis of the differential expression between the clone and wild-type haploid HAP1 cells should take place to check that A2E4 is a good true wild-type to compare the edited clone to.

Optimisation of the transfection and subcloning of K562 would be a lower priority until the HAP1 *RBM15*-KO experiment results have been fully analysed. K562 cells have previously been successful single-cell sorted in the lab but the transfected cells might have required conditioned media for the subcloning here. Given the issues with ploidy encountered with HAP1, successful editing of *RBM15* in the hypertriploid K562 cells in the future may provide another level of investigation.

#### *2.5.4.1 Transfection method*

A challenge encountered during the KO experiment was the lack of a selection method to isolate the cells that had received the sgRNA from those not successfully transfected. Since it was not possible to distinguish between transfected and untransfected cells without subcloning and growing up a clonal population, extracting the gDNA, generated purified amplicons and Sanger sequencing them, it was more difficult and time-consuming to identify an edited clone. The ability to instantly remove untransfected cells from the bulk sample of cells prior to subcloning would simplify the process in future experiments.

A number of alternative transfection methods could be used to enable enrichment or selection for cells that have been successfully transfected. These methods render the Cas9-stable feature of the cell line used in these experiments useless; a standard HAP1 cell line would need to be used instead (although the extra Cas9 expression might not affect the results). Fluorescent Cas9 could be used by endogenously tagging the protein with a GFP, for example, and fluorescence detected in cells under the microscope or by FACS would indicate successful delivery of Cas9. Cells could be transfected with fluorescent sgRNA or transfected with only a single plasmid containing the gRNA cassette, the Cas9 gene, and a fluorescent protein. Cas9 expression-linked fluorescence would identify only those cells that had been transfected, allowing these cells to be isolated by FACS. Another approach would be to use antibiotic resistance to select for transfected cells: the single plasmid could again be transfected into the cells, but instead of containing the GFP sequence, it would include an antibiotic resistance gene. One day following transfection, the cells would be treated with the antibiotic (such as puromycin) and only the transfected cells should survive. Either selection method could be used or both in tandem to ensure minimal contamination of untransfected cells but treating the cells with antibiotics adds an extra step to the process (and possible extra complications or implications) so it would just be simpler to use fluorescence only.

Alternatively, more time could be spent optimising the transfections to have a robust procedure which could be used to edit many other RBPs or SFs. Varying transfection reagent type and concentration, cell confluency and testing more sgRNA concentrations could help to maximise transfection efficiency, thereby increasing the chance of obtaining transfected clones from subcloning. However, this would be time-consuming and is not strictly necessary; fluorescence is the quickest and easiest solution for future experiments.

#### *2.5.4.2 Use of HAP1s as a model cell line*

The HAP1 cell line is often used for CRISPR-Cas9 experiments because of the higher chance of editing alleles following transfections and because it is good for observing potential phenotypes as the effect of any mutations cannot be masked by the presence of another allele that is intact. However, this advantage did not appear to aid these experiments as cells kept becoming diploid before sorting. It is unusual that the diploidisation occurred so early in the experiment (passage 2-3); HAP1s should remain haploid for the first 10 passages according to Beigl, Kjosås, Seljeseth and colleagues (2020). This early diploidisation may point to a possible response to the transfection, as discussed above, but is more likely due to lack of careful ploidy management. Nevertheless, an edited cell line was generated and both alleles were edited; these cells will be useful for further experiments to elucidate the consequences upon editing *RBM15*.

Another caveat of using HAP1s is that *MPL*, an important target of RBM15 in the haematopoietic context, is not expressed in the cell line (0.0 nTPM in the Human Protein Atlas database; RNA-seq data for cell lines (Carette *et al.*, 2011; *The Human Protein Atlas*, 2022) (The Human Protein Atlas, 2022; Carette, Raaben and Wong *et al.*, 2011)). Despite this, MPL did appear as a gene that was expressed in the DESeq2 analysis and was upregulated in the edited clone although this was not statistically significant (Table 7). As a lower priority, experiments to knock out *RBM15* and even *MPL* in a different cell line would be interesting to perform in the future to compare the functional results between the two mutants. This would be best performed in stem cells or using mouse models.

# CHAPTER 3: LINEAGE-SPECIFIC SPLICING FACTORS

## *3.1 Introduction*

The commitment of haematopoietic progenitors into terminally differentiated specialised blood cells requires the coordinated transition of the transcriptome and proteome of a blood cell. As detailed in Chapter 1, commitment occurs through the action of extrinsic factors such as nutrients or cytokines causing changes in the intrinsic differentiation pathways of a progenitor. These pathways include changing the activity or expression of master transcription factors, changing epigenetic modifications on DNA to affect the expression of differentiation-associated genes, and the targeted destruction of differentiation-associated mRNAs by miRNAs. However, compelling evidence suggests that the AS of differentiation-associated pre-mRNAs also plays a crucial role in haematopoietic commitment. This chapter examines the evidence for the function of AS in haematopoiesis and describes an attempt to further investigate how AS gives rise to different lineages.

### 3.1.1 AS and haematopoietic diseases

SFs and AS are important to haematopoiesis by regulating the differentiation of haematopoietic stem cells (HSCs) into the different blood cell lineages (Gao, Vasic and Halene, 2018; Li *et al.*, 2021; Chen and Abdel-Wahab, 2021; Fig.2). This relationship has largely been elucidated through the characterisation of dysfunctional HSC generation, erythropoiesis, granulopoiesis, megakaryopoiesis, and monocyte-to-macrophage differentiation often resulting in cancers (Yapu Li *et al.*, 2021). Mutations in more than 30 SFs have been found in haematological malignancies, with apparent exclusivity of SF mutation per malignancy (Papaemmanuil *et al.*, 2011; Yoshida *et al.*, 2011; Walter *et al.*, 2013; Haferlach *et al.*, 2014). In their analysis of RNA-seq data from The Cancer Genome Atlas, Dvinge and Bradley (2015) found acute myeloid leukaemia (AML) cells to have the highest number of AS mutation events out of 16 tumour types.

The most frequent mutated splicing-related genes in haematological malignancies are those encoding the core spliceosome proteins SF3B1, SRSF2, U2AF1 and ZRSR2 (Dvinge *et al.*, 2016; Visconte, O Nakashima and J Rogers, 2019), with these being the most prevalent mutated SFs in MDS and chronic lymphocytic leukaemia (CLL). Different cancers can acquire SF mutations at different times, with mutations in MDS often acquired during early cancer development (Papaemmanuil *et al.*, 2013; Haferlach *et al.*,

2014), and mutations during CLL typically being more subclonal (Landau *et al.*, 2013). Of all patients with MDS, at least 50% carry a mutation in an RNA SF gene. Beyond a mere association, the importance of the spliceosome proteins SF3B1, SRSF2, U2AF1 and ZRSR2 in haematopoiesis have been shown experimentally through the generation of knockout mice. For example, germline *SF3B1*$^{-/-}$ mice are embryonically lethal, and heterozygotes depleted of SF3B1 by shRNA showed a greater defect in HSC repopulating capacity (Wang *et al.*, 2014). Mice with a conditional *U2AF1* knockout had premature death with defective haematopoiesis and HSC repopulation capacity (Dutta *et al.*, 2021), and conditional *SRSF2* knockout mice developed leukopenia, anaemia, and bone marrow aplasia and were severely compromised in HSC self-renewal in competitive translation (Kim *et al.*, 2015). This work, validating the strong association data, highlights the critical relationship between myeloid malignancies and AS.

### 3.1.2 AS during normal haematopoiesis

Despite a clear relationship, only a small number of studies have attempted to elucidate the function of splicing components during normal haematopoiesis (Chen and Abdel-Wahab, 2021). RNA-seq of normal human HSCs and seven other progenitor populations by Chen *et al.* (2014) showed that each stage of haematopoiesis is defined by cell type-specific differential splicing. One better understood example of this is in erythrocyte differentiation where early erythroid progenitors skip exon 16 in the red blood cell membrane cytoskeleton protein EPB41 but include this in mature erythroblasts (Conboy *et al.*, 1991) as a result of alterations in expression of RNA binding proteins hnRNPA/B, SRSF1, and FOX1/2 (Deguillien *et al.*, 2001; Hou *et al.*, 2002; Yang *et al.*, 2005).

One theme that has emerged is the pattern of intron retention as a driver of haemopoietic stage-specific regulation (Chen and Abdel-Wahab, 2021). In this model, originally derived from studies of erythroid development (Edwards *et al.*, 2016), intron retained transcripts are sequestered away from successful translation, either due to being retained in the nucleus, or by the retained intron containing a premature stop codon and triggering nonsense-mediated decay of the mRNA, resulting overall in a decreased abundance of specific proteins. A study by Ullrich and Guigó (2020) of intron retention during haematopoiesis showed that this was highest in neutrophil and monocyte differentiation, and an increase in intron retention correlated to the differentiation of B-cells. A second theme that has emerged is that stage-specific regulation is controlled by the generation of multiple mRNA isoforms that have opposing functions. One example of this is the AS of *BMP2K* in erythroid differentiation, where long and short isoforms

counterbalance with the long promoting autophagy and erythroid differentiation whilst the short inhibits autophagy and prevents differentiation (Cendrowski *et al.*, 2020). Despite these examples however, a detailed mechanistic description of the relationship of how AS operates across the breadth of haematopoietic cell lineage commitment is currently lacking.

### 3.1.3 Investigation of lineage-specific SFs

One of the aims of this project was to identify important splicing factors in haematopoiesis and understand their function in relevant cell lines. As described in the introduction, there are many proteins which have a part to play in AS and haematopoiesis, from the regulatory SFs to the components of the spliceosome itself. Some of these proteins have been at least identified, with the specific splicing roles of some having already been well-characterised. However, there are still many more SFs functioning in haematopoiesis to be discovered, and the specifics of which SF affects which genes in which lineages is largely unexplored. It would make sense that certain SFs influence the AS of certain genes in certain cell types to drive their differentiation, and advances in single-cell isolation and RNA-sequencing coupled with improvements in RNA sequencing methods and analysis make answering these questions possible. Through computational exploration and analysis of a haematopoietic dataset, it should be possible to identify a list of relevant RNA-binding proteins and splicing factors (RBP/SFs). Then the genes coding for these proteins can be knocked out in subsequent experiments, as with *RBM15* in Chapter 2. For this, a single-cell RNA-seq dataset is required in order to identify RBP/SFs that might act specifically or more prevalently in certain early haematopoietic lineages when splicing decisions and AS is likely to have a large transcriptomic and proteomic impact on the cell. Single-cell RNA-seq data can be analysed in an R package called Seurat (Satija Lab, 2022) that allows quality filtering, differential gene identification between single cells, dimensionality reduction, clustering of cells based on gene expression profiles and assignment of cell types based on results.

Chapter 3 describes the Seurat analysis and interrogation of a single-cell transcriptomic data set from 6 mice sorted for $Lin^{-}cKit^{+}$ haematopoietic stem and progenitor cells from Mincarelli *et al.* 2023 to find haematopoietic lineage-specific RBP/SFs, and the results from an initial experiment to test the efficacy of guide RNAs in knocking them out. This investigation expands our knowledge on lineage-specific RBP/SFs and generates tools for their further investigation.

## *3.2 Reagents*

Reagents for this chapter are the same as in Chapter 2.2 Reagents except for the sgRNA and primer sequences: details below.

### 3.2.1 sgRNA sequences

sgRNAs for the transfection of *CARHSP1*, *KHDRBS3* and *LARP1B* (three guides per gene) were designed using CRISPick and ordered via Horizon's custom Edit-R synthetic sgRNA service (Table 8).

***Table 8 - Lineage-specific RBP/SF genes guide RNA information table.*** *Information on each of the three guides transfected into HAP1 Cas9+ cells in order to generate a* CARHSP1-*KO, a* KHDRBS3-*KO and a* LARP1B-*KO.*

| Name | Alias | sgRNA sequence (5' to 3') | Strand of Target | sgRNA Cut Position (1-based) | PAM Sequence | Target Exon Number | Target Cut Length | Target Total Length |
|---|---|---|---|---|---|---|---|---|
| CARHSP1_ gRNA_rank3 | CARHSP1_ gRNA_1 | AGGACGGTGCGGGCTTCACA | - | 8858458 | GGG | 3 | 173 | 444 |
| CARHSP1_ gRNA_rank10 | CARHSP1_ gRNA_2 | CCCACCCATCAAGCTTCAGT | - | 8859279 | CGG | 2 | 50 | 444 |
| CARHSP1_ gRNA_rank19 | CARHSP1_ gRNA_3 | TCACAGGGCCCCGTCTACAA | - | 8858443 | AGG | 3 | 188 | 444 |
| KHDRBS3_ gRNA_rank1 | KHDRBS3_ gRNA_1 | TTGGAGTTGTAGTACCACGA | + | 135581924 | GGG | 6 | 657 | 1041 |
| KHDRBS3_ gRNA_rank8 | KHDRBS3_ gRNA_2 | GCCCTGCGCCTGGTGAACCA | + | 135457950 | AGG | 1 | 83 | 1041 |
| KHDRBS3_ gRNA_rank9 | KHDRBS3_ gRNA_3 | TCCCGTAAAACAGTTCCCTA | + | 135521351 | AGG | 2 | 202 | 1041 |
| LARP1B_ gRNA_rank3 | LARP1B_ gRNA_1 | TCAAGCTAATAAGCACAAGT | + | 128082177 | GGG | 5 | 229 | 2745 |
| LARP1B_ gRNA_rank4 | LARP1B_ gRNA_2 | TCAAGTAATCAACGTAAGAG | + | 128077958 | AGG | 4 | 212 | 2745 |
| LARP1B_ gRNA_rank11 | LARP1B_ gRNA_3 | GAACGAGACTTCTTTCTTCG | + | 128098231 | GGG | 8 | 713 | 2745 |

### 3.2.2 Primer sequences

Primers for amplicon PCR and Sanger sequencing were ordered from IDT as 25 nmole DNA oligos resuspended to 100 µM with TE (Table 9).

***Table 9 - Lineage-specific RBP/SF amplicon PCR primers.*** *Information on the primers designed for edit-verification purposes: to generate amplicons surrounding the cut-sites for the gRNAs used in the experiments described in this Chapter. Adapted from the output of Primer-BLAST.*

| Gene | Guides covered | Primer name | Sequence (5' → 3') | Length (bp) | $T_m$ (°C) | GC% | Intended product length (bp) | Number of possible unintended templates | Unintended product lengths (bp) | Distance from primer to guide (incl.) (bp) |
|---|---|---|---|---|---|---|---|---|---|---|
| *CARHSP1* | gRNA_1; gRNA_2; gRNA_3 | F1 | AGATGTGCAGGAAGATGTCGG | 21 | 60.13 | 52.38% | 992 | 3 | 2181; 1483; 3738 | 106; 121 |
| *CARHSP1* | gRNA_1; gRNA_2; gRNA_3 | R1 | CTCTTTCCAGGTCAGCCATGT | 21 | 60.00 | 52.38% | 992 | 3 | 2181; 1483; 3738 | 70 |
| *KHDRBS3* | gRNA_2 | F1 | GAGGAGAAGTACCTGCCCGA | 20 | 60.68 | 60.00% | 302 | 20 | 2662; 2013; 1189; 1181; 1746; 2992; 1761; 2120; 1184; 3497; 717; 2595; 666; 3591; 3483; 3475; 1618; 1959; 3656; 3570 | 83 |
| *KHDRBS3* | gRNA_2 | R1 | GCCTTCTCCTTCCAAGTCCC | 20 | 60.03 | 60.00% | 302 | 20 | 2662; 2013; 1189; 1181; 1746; 2992; 1761; 2120; 1184; 3497; 717; 2595; 666; 3591; 3483; 3475; 1618; 1959; 3656; 3570 | |
| *KHDRBS3* | gRNA_3 | F2 | TGCCACTACAGAGCTAACCAGA | 22 | 60.82 | 50.00% | 312 | 0 | NA | |
| *KHDRBS3* | gRNA_3 | R2 | GGGGGATGCTAAGAGAAAACATCT | 24 | 60.39 | 45.83% | 312 | 0 | NA | 134 |
| *KHDRBS3* | gRNA_1 | F3 | GCTTGTGACTTCCTTCTTCCTTTT | 24 | 59.66 | 41.67% | 319 | 0 | NA | |
| *KHDRBS3* | gRNA_1 | R3 | CATAAGTCTCTTGTGTCGGGGG | 22 | 60.42 | 54.55% | 319 | 0 | NA | 159 |
| *LARP1B* | gRNA_2 | F1 | GCAGTTTCAGAGCGTCCTCA | 20 | 60.32 | 55.00% | 314 | 0 | NA | |

## *3.3 Methods*

To identify splicing factors and the haematopoietic cell types they are expressed in, single-cell data from HSCs was required in order to explore the expression patterns of these genes in these specific cells enabling the identification of potential candidate target genes for future knockout experiments.

Therefore, a dataset was obtained from Mincarelli *et al.*, 2023, in which single-cell libraries were generated from lineage negative cKit/Cd117 positive (LK) cells (haematopoietic stem and progenitor cells) from three young and three aged C57/BL6 mice (8 and 72 weeks, respectively). The libraries were sequenced using Illumina short-read and PacBio long-read technologies (minimum 20,000 read pairs per cell). Cells were processed using the 10x Genomics Chromium platform to generate barcoded cDNA pools (3' RNA-seq v2 kit), then libraries were prepared from the cDNA either by completing the 10x library preparation process (short-read libraries) or using the PacBio IsoSeq protocol (long-read libraries from full-length cDNA). The 10x libraries were sequenced on an Illumina NextSeq500 and the raw data were analysed using CellRanger (v3.0.2), outputting a single-cell expression matrix object for each of six samples: biological triplicates of data from young and aged mice (Mincarelli *et al.*, 2023).

### 3.3.1 Short-read data analysis using Seurat

The short-read data were analysed using Seurat v4 (Hao *et al.*, 2021), utilising a standard workflow as described below.

Six separate filtered gene expression matrices were loaded into Seurat. Seurat objects were created from each expression matrix which underwent QC filtering (>1000 features/genes per cell; <5% mitochondrial content), log normalisation (“LogNormalize” method; 10,000 scale factor) and variable feature selection. Next, the six Seurat objects were integrated to mitigate batch effects. The integrated dataset was scaled and Principal Component Analysis (PCA) was performed on the variable features. The top 20 PCs were used to find the shared-nearest-neighbours between cells and this in turn was used to group the cells into clusters with similar patterns of gene expression. The resulting clusters were visualised using a UMAP plot.

Differential expression between clusters was investigated, identifying marker genes as those which are most highly expressed in each cluster. The top five marker genes for

each cluster helped to classify the clusters and assign cell types to each cluster (Table 12). This process was also facilitated with literature searches (Pronk *et al.*, 2007; Woolthuis and Park, 2016; Dahlin *et al.*, 2018), interrogating the Mouse Genome Informatics (MGI) database (The Jackson Laboratory, 2022) and, as a last resort, comparison with the analysis in the paper the data was originally produced as part of (Mincarelli *et al.*, 2023). After cluster identities were assigned within the Seurat object, dimensionality reduction plots were reproduced with the clusters labelled (Fig.29).

### 3.3.2 Generation of candidate gene expression heatmaps

To determine the expression levels of genes in HSCs which might be involved in alternative splicing, a list of relevant genes was required to search the dataset against. Genes encoding RBPs as well as specific splicing factors (SFs) were included in the list to account for RBPs which are not SF themselves but might be involved in the process of splicing as co-factors. This also allows inclusion of RBPs which are SFs but are not yet characterised as such.

A list of 68 splicing factor genes was provided by Dr Wilfried Haerty. This was combined with a list of RBP genes obtained from the RBPDB by searching for all genes encoding for RBP proteins in *Mus musculus* (Berglund *et al.*, 2008; accessed 21$^{st}$ Feb 2022; 415 genes in list). The total number of genes in the combined list after removing duplicates was 428.

The R package pheatmap (Kolde, 2019) was used to create a heatmap which displayed cluster-specific expression of present genes within the combined RBP/SF list, labelled the cell-type clusters and grouped the genes by similar intracluster expression patterns (Fig.32).

Some genes showed relatively high expression values, despite the scaled data from the integrated Seurat object being used for the heatmap. To make the milder intracluster expression patterns appear clearer, another heatmap was generated using the average expression values for each gene within each cluster (Fig.33).

### 3.3.3 Heatmap analysis and candidate gene list reduction

First, the individual cell heatmap was analysed: genes which were particularly upregulated in one or a few clusters were identified visually. The average expression heatmap was then analysed in the same way to add more genes to the shortlist. Literature and database searches were performed to research the general functions of the shortlisted genes and whether they have any relevance in haematopoiesis and disease. It was determined whether each gene is currently known to have any role in splicing and whether the gene is itself alternatively spliced in human, mouse or both.

To reduce the shortlist to a manageable quantity which can be knocked out in cell lines in future experiments, Feature Plots (Seurat::FeaturePlot()) were produced for those genes which appeared to be upregulated in the erythroid clusters and those apparently upregulated in the megakaryocyte cluster (Fig.34). Thus, the cluster- or lineage-specific expression of the shortlisted genes could be confirmed. A lower cut-off of q10 was used for the Feature Plots to discount the lower decile from the plot, removing some of the low-level expression noise and allowing real patterns to appear clearer. Genes which did show cluster-specific expression were confirmed and comprise a “shorterlist” of priority genes for the future KO experiments.

### 3.3.4 CRISPR-KO of lineage-specific genes

The lineage-specific RBP/SF genes were identified as described above in order to be targeted in knockout (KO) experiments as in the *RBM15*-KO experiment described in Chapter 2.3.1. Thus the functional consequences of removing each of those genes can be analysed. Given the difficulties when verifying and subcloning the edited K562 cells, these experiments were only carried out in HAP1 cells.

The conditions previously determined to be optimal were utilised for these transfections: 4 µg DharmaFECT1 transfection reagent and 25 nM sgRNA (final concentrations). Table 10 displays information regarding the sgRNAs used in each sample for this experiment.

***Table 10 - Details of the transfections performed to edit the RBP/SF genes selected from short-read single-cell RNA-seq analysis:* CARHSP1, KHDRBS3, LARP1B.** *The samples were renamed with letters for ease of reference, and the gRNAs were also renamed 1 to 3 for each gene in order of their combined on-target and off-target rank for the same reason.*

| Sample | Gene | Cell line | gRNA | gRNA rank | sgRNA conc | DF1 conc |
|---|---|---|---|---|---|---|
| P | *CARHSP1* | HAP1 | 1 | 3 | 25 nM | 4 µg |
| Q | *CARHSP1* | HAP1 | 2 | 10 | 25 nM | 4 µg |
| R | *CARHSP1* | HAP1 | 3 | 19 | 25 nM | 4 µg |
| S | *KHDRBS3* | HAP1 | 1 | 1 | 25 nM | 4 µg |
| T | *KHDRBS3* | HAP1 | 2 | 8 | 25 nM | 4 µg |
| U | *KHDRBS3* | HAP1 | 3 | 9 | 25 nM | 4 µg |
| V | *LARP1B* | HAP1 | 1 | 3 | 25 nM | 4 µg |
| W | *LARP1B* | HAP1 | 2 | 4 | 25 nM | 4 µg |
| X | *LARP1B* | HAP1 | 3 | 11 | 25 nM | 4 µg |

#### *3.3.4.1 sgRNA sequence design*

The sgRNA sequences were custom designed for this experiment to gain familiarity with the process. Three guides per gene were designed using a web tool, CRISPick (Broad Institute, 2022). The reference genome used was the Human GRCh38 (Ensembl v.105). The other settings for the CRISPick tool included Mechanism (CRISPRko), Enzyme (SpyoCas9; Cas9 from *S. pyogenes* which utilises an NGG PAM sequence), Target (gene name) and Quota. The quota was set to 20, instructing CRISPick to recommend the top 20 candidate gRNA sequences according to raw ranking, cut position and mutual spacing. The output of CRISPick is a text file with the 20 candidate gRNAs listed as rows and with columns including Orientation (of gRNA in relation to target gene), Cut Position, Sequence, PAM Sequence, Exon Number (the exon targeted by the gRNA), On- and Off-Target Ranks and Combined Rank (the overall rank according to the sum of the on- and off-target ranks). See Table 8 in Chapter 3.2.1 for some of this information.

To avoid confusion when ordering the gRNAs and ensuring their correct sequences, only the sense candidate results were considered. The top three sense guides were chosen except when a guide was too similar to a previously chosen guide (one or two bases up- or downstream); the next best guide was selected in its place.

The chosen guides were annotated onto Ensembl gene sequence maps on Benchling (*Benchling*, 2021) for visualisation and recording-keeping

(https://benchling.com/nicoleforresterei/f_/YvRIseur-sf-ko-experiment/).

# *3.4 Results*

## 3.4.1 Seurat analysis

The aim of this part of the project was to analyse the single-cell short-read haematopoietic stem cell dataset from Mincarelli *et al.* (2023) for cell-type specific patterns of RBP and SF expression. This single-cell RNA-seq dataset was from $Lin^-$ $cKit^+$ haematopoietic stem and progenitor cells from 3 young and 3 aged mice but had not been analysed specifically to search for changes in AS factors. Because the dataset is from single-cell RNA-sequencing data, the gene expression of known cell-type-specific biomarker genes could be used to investigate how the expression of each SF varied by haematopoietic cell type. The single-cell RNA-seq data was analysed in single-cell analysis R package, Seurat. This process involved QC filtering, integration of the six separate datasets from each experiment, variable feature (gene) selection, dimensionality reduction (PCA), clustering and cell type assignment. The results of the Seurat analysis are described here.

### *3.4.1.1 Quality control*

The average number of features across the six Seurat objects created for each gene expression matrix was 15,188; the average number of cells following QC filtering was 1,406 (Table 11; Fig.28). There was a variable loss in cells through the QC filters (7.1 - 68.7% cells removed). The samples reduced most from the filtering (m28-31) had much higher proportions of mitochondrial reads, an indicator of stressed and dying cells, likely due to practical problems and delays during the flow-sorting of those particular cells (especially those in Experiment 3). The final total number of cells across the six datasets was 8,436.

***Table 11 - Metadata from pre- and post-QC of scRNA-seq data.*** *Some descriptive metadata, the number of features and cells before cells after QC and the percentage of cells that passed QC for each of the six separate short-read datasets.*

| Experiment | Group | Mouse | No. of features | No. of cells before QC filtering | No. of cells after QC filtering | % cells passed QC |
|---|---|---|---|---|---|---|
| 1 | Young | m26 | 15197 | 1786 | 1659 | 92.9 |
| 1 | Aged | m27 | 15456 | 1871 | 1688 | 90.2 |
| 2 | Young | m28 | 14844 | 3174 | 1533 | 48.3 |
| 2 | Aged | m29 | 16444 | 2631 | 1919 | 72.9 |
| 3 | Young | m30 | 14640 | 1976 | 956 | 48.4 |
| 3 | Aged | m31 | 14128 | 2175 | 681 | 31.3 |

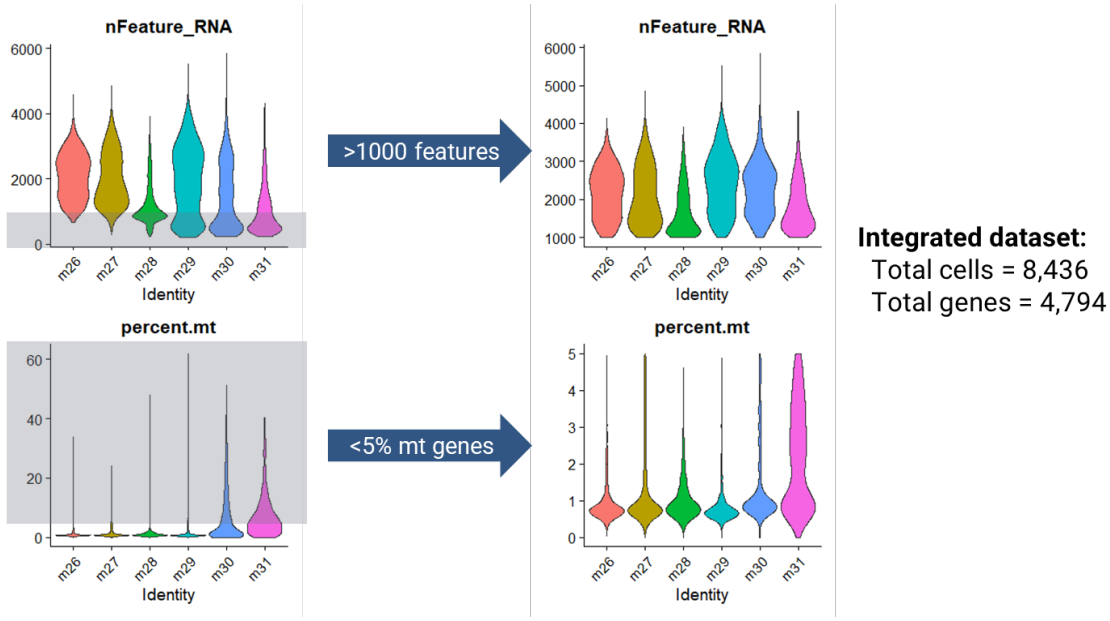


***Figure 28 – Visual representation of consequences of QC filtering on data distributions.*** *Violin plots generated in Seurat showing the number of features (genes) expressed in the cells of each mouse in the experiment before and after QC filtering. The filters removed those cells with fewer than 1000 features and greater than 5% mitochondrial gene content (grey boxes on left images highlight data removed by filters).*

### *3.4.1.2 Integration of datasets*

The six single-cell datasets had to be integrated so that only genes that appeared in all six datasets were taken forward for downstream analysis. In a first attempt, only 26 out of a possible 428 possible RBP/SF genes were found to be expressed in all six datasets, much fewer than expected. However, upon examination of the Seurat analysis R script, it was found that the default settings had been used for integrating the six datasets in Seurat, meaning that only the top 2,000 genes out of the potential 17,757 total genes had been taken forward for integration. Upon discovering this, the Seurat script was altered to integrate all identified genes across the six datasets. This resulted in 4,794 genes in the integrated dataset and 83 RBP/SF genes found to be expressed within it.

### *3.4.1.3 Assignment of cell types to clusters*

The integrated datasets were analysed to investigate the similarities of the gene expression pattern of each cell to that of every other cell in the dataset. This could then be used to cluster the cells by their gene expression profiles, with each cluster representing different haematopoietic cell types. To do this, the integrated dataset was scaled and principal component analysis (PCA) was performed on gene expression values for each gene, and the top 20 principal components used to find the shared-nearest-neighbours for each cell by Uniform Manifold Approximation and Projection

(UMAP) dimensionality reduction analysis. UMAP plots were generated using the first two components identified in the UMAP analysis, with each dot representing a cell (Fig.29). All cells were used in the analysis irrespective of mouse age in order to increase the ability to tightly define cell-type clusters.

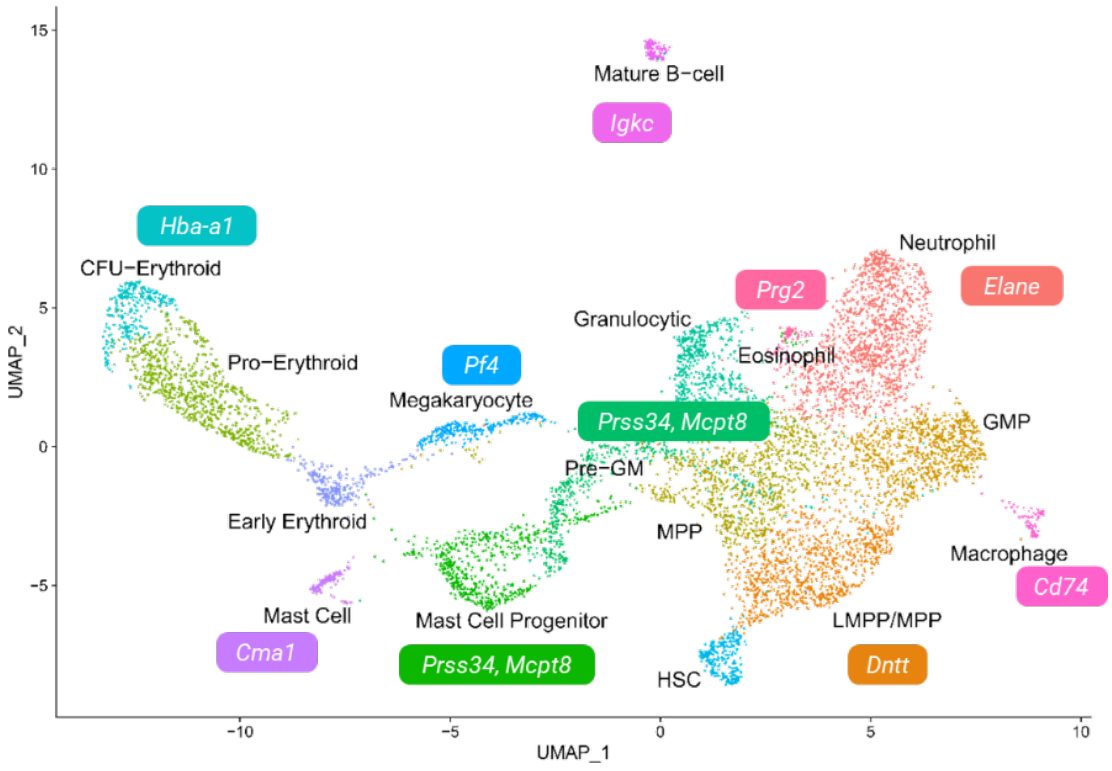


***Figure 29 – Progenitors of major blood cell types are captured within the integrated single-cell haematopoiesis RNA-seq dataset.*** *Gene expression profiles from integrated single-cell data from 3 young and 3 aged mice were processed by UMAP dimensionality reduction to reveal the relationship between each cell and every other cell based on similarities in their gene expression profiles. Plot shows UMAP1 and UMPA2 dimensions, with each point a single cell from the integrated dataset. Cell type for each cluster is shown, and genes whose expression definitively reveal the identity of each cluster are shown in boxes. Note that some clusters are named by their mature cell type for brevity but are in fact the progenitors for those types.*

A total of 16 clusters were identified based on the gene expression patterns from all the cells in the integrated dataset. Clusters were annotated by cell type by looking at the most highly expressed genes within each cluster and researching which cell type they are associated with from the literature and by interrogation of the Mouse Genome Informatics (MGI) database. This annotation was simple for several cell type clusters based on the top few expressed genes, namely neutrophils (*Elane*), the lymphoid-primed multipotent progenitors (*Dntt*), megakaryocyte progenitors (*Pf4*), mature B-cells (*Igkc*, *Igha*, *Jchain*) and eosinophils (*Prg2*, *Prg3*, *Cebpe*). For others, the clear identification of

the clusters from their most highly expressed genes was more difficult, as those genes were not always obvious markers or were shared between clusters. For example, *Prss34* and *Mcpt8* were two of the top five genes for both the pre-granulocyte-monocyte (pre-GM) cluster and the mast cell progenitors. The most highly expressed marker genes are presented in Table 12, and the cell type associated with each cluster displayed on the UMAP plot, with genes that characteristically define cell type clusters also displayed (Fig.29). Many of the potential haematopoietic cell types were represented in this dataset, with this shown graphically in Fig.30. Once the cell identities had been assigned to each cluster, the expression of RBP/SF genes could be analysed across cell types to search for lineage-specific expression patterns.

***Table 12 - The top 5 most highly expressed marker genes per cell type cluster found in the short-read data using Seurat and the names of the identified clusters.***

| **Cell type cluster** | **Most highly expressed marker genes** |
|---|---|
| Neutrophil | *Elane, Ms4a3, Mpo, Gstm1, S100a8* |
| LMPP/MPP | *Wfdc17, Dntt, Cd34, Ctla2a, Myl10* |
| GMP | *Irf8, F13a1, Ms4a6c, Slpi, Ly6c2* |
| MPP | *Stmn1, Tuba1b, Pclaf, H2afy, Hist1h2ap* |
| Pro-Erythroid | *Car1, Car2, Mt1, Blvrb, Mt2* |
| Mast Cell Progenitor | *Ms4a2, Cpa3, Csrp3, Prss34, Mcpt8* |
| Pre-GM | *Ube2c, H2afx, Fcer1a, Prss34, Mcpt8* |
| Granulocytic | *mt-Cytb, mt-Nd1, mt-Nd4, Mpo, Hspa5* |
| CFU-Erythroid | *Car2, C1qtnf12, Blvrb, Hba-a1, Rhd* |
| HSC | *Ltb, Ly6a, Gimap1, Myl10, Mpl* |
| Megakaryocyte | *Pf4, Rap1b, Cd9, Nrgn, Pbx1* |
| Early Erythroid | *Apoe, Vamp5, Fos, Sox4, Gm15915* |
| Mast Cell | *Gzmb, Lmo4, Prr13, Ifitm1, Cma1* |
| Mature B-cell | *Iglc2, Jchain, Igkc, Igha, Ighm* |
| Macrophage | *Cst3, Cd74, H2-Aa, H2-Ab1, H2-Eb1* |
| Eosinophil | *Prg3, Cebpe, Prg2, Cd63, Epx* |

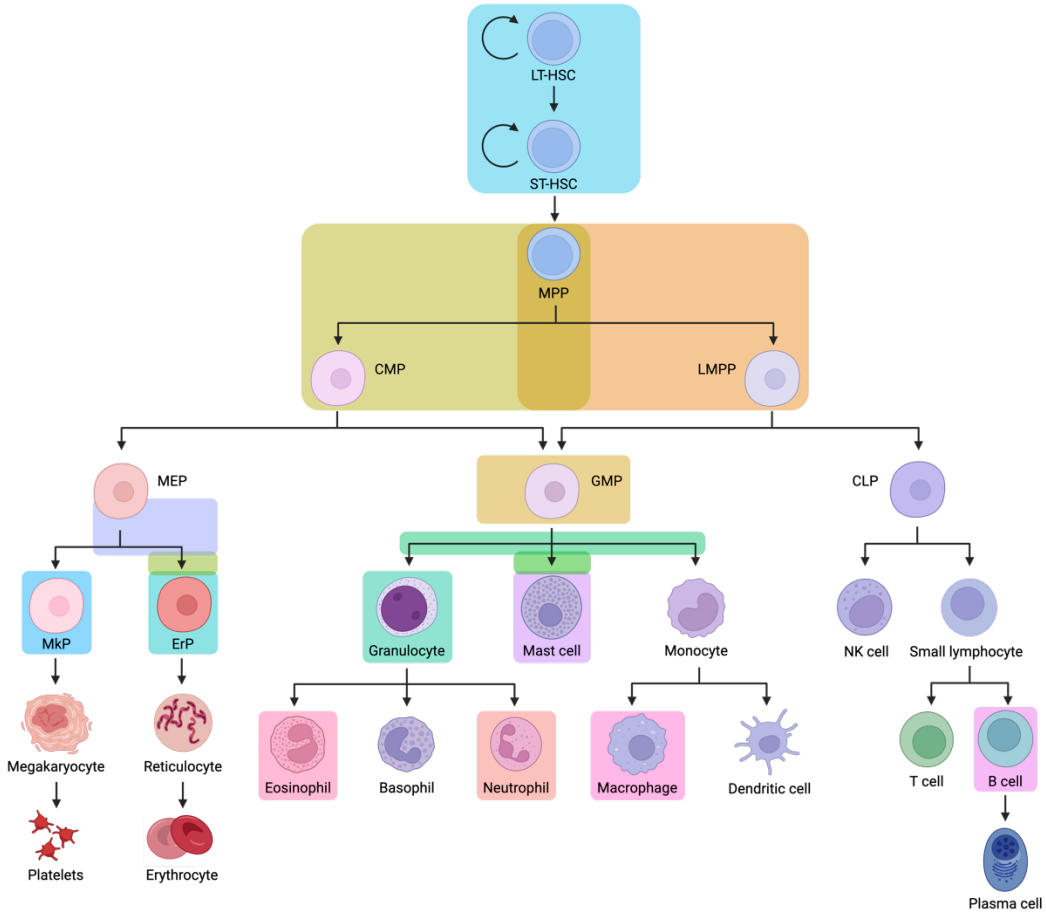


*Figure 30 - Tree diagram of haematopoietic cells with box colours matching UMAP cluster colours.*

## 3.4.2 Splicing factor expression across cell types

### *3.4.2.1 Heatmap analysis*

The expression distribution of SFs was explored in the short-read single-cell data. Gene expression heatmaps were produced that plotted the relative gene expression values for each of the 83 identified splicing factors genes for every cell in the integrated dataset. These graphical depictions of the data were useful in allowing exploratory data analysis, with coloured “hotspots” in the heatmap drawing the eye to potentially interesting results.

Several R packages are available for generating the complex heatmaps that could plot the relative gene expression values for each of the 83 SFs. First, the in-built heatmap plotting function from Seurat was tried: DoHeatmap, with the generated heatmap displayed in Fig.31. Unfortunately, this method displayed genes in the order in which they were input and did not cluster the genes by expression, making visualisation of cell-type specific differences in expression impossible.

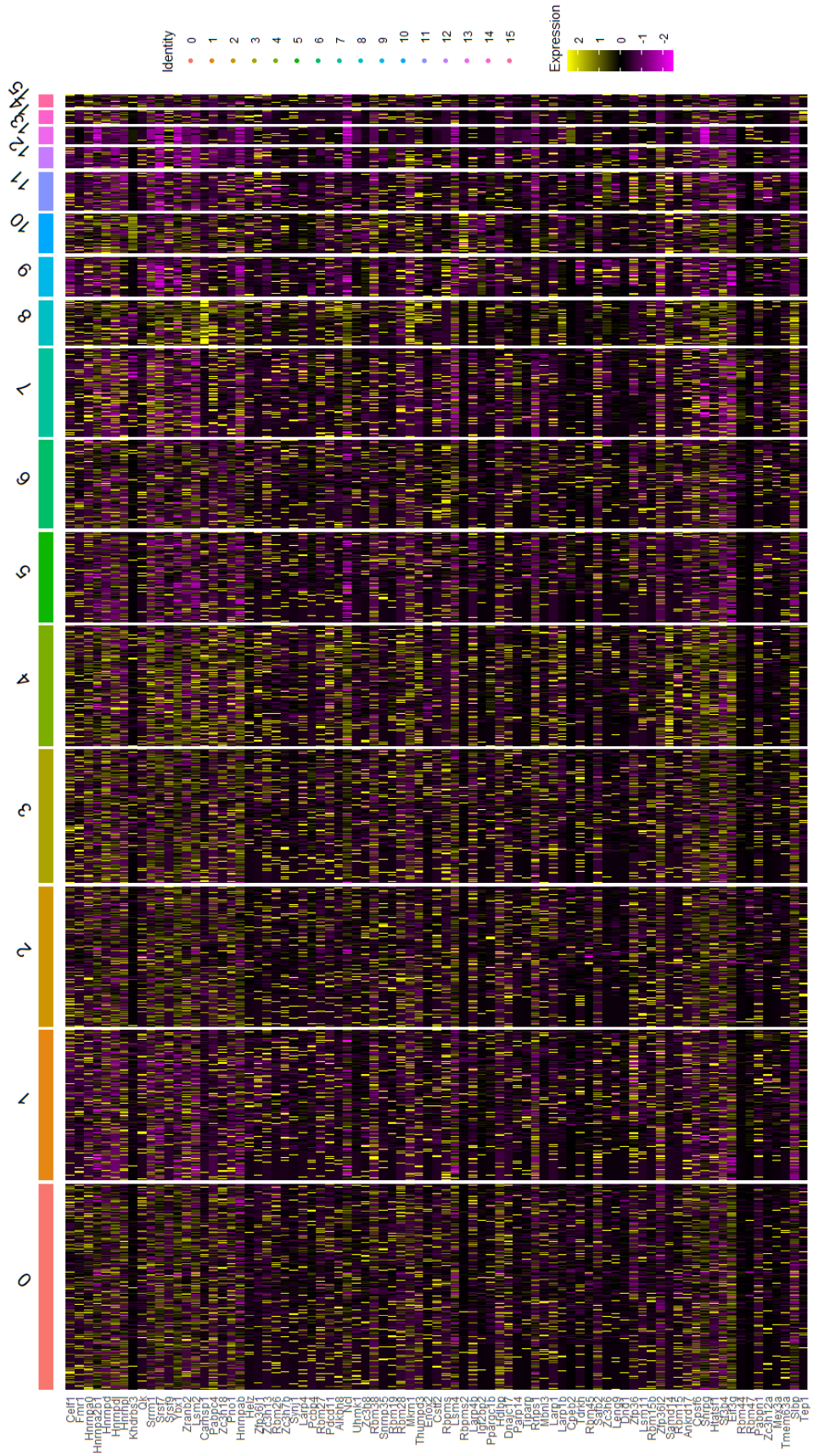


***Figure 31 - Heatmap of gene expression per cluster (unordered)**. The initial heatmap created using Seurat::DoHeatmap, which does not allow for clustering of genes based on similar expression patterns and therefore does not clearly highlight points of interest.*

To overcome this, more R heatmap packages were tried that allowed clustering of cells by their relative gene expression. ComplexHeatmap (Gu, 2021) was tried, but the best display was generated by pheatmap (Kolde, 2019) that allowed genes to be clustered with the cluster_cols argument, and cells to be clustered with the cluster_rows argument (Fig.32). Visually searching the heatmap provided some interesting candidate genes with lineage-specific gene expression patterns, visualised by red areas on the heatmap within a cell-type cluster. There was a marked pattern of upregulation of RBP/SF genes in the top third of the heatmap (outlined by two dashed boxes) in the two erythroid clusters. These included *Carhsp1* that was upregulated in the CFU-erythroid cluster and *Hnrnpab* that was upregulated in the Pro-Erythroid cluster. A less well-defined cluster of megakaryocyte progenitor-specific RBP/SF genes were also located (outlined by small dotted box), with this including *Khdrbs3* and *Rbpms2*. Visual analysis of the heatmap highlighted several genes of interest, which were particularly upregulated in one or a few clusters whilst having relatively low expression levels in the rest: *Carhsp1*, *Rbpms2*, *Khdrbs3*, *Samd14*, *Larp1b*, *Slbp* and *Srrm1*.

To further aid data visualisation, the average intra-cluster gene expression was calculated and plotted using pheatmap (Fig.33). The average expression values per gene per cluster were calculated using medians rather than means so that extreme outlier expression values did not skew the colour scale in the heatmap. This averaging heatmap clearly highlights the group of SFs that were upregulated in the Pro- and CFU-Erythroid lineages, and the strong upregulation of *Rbpms2* and *Khdrbs3* in the megakaryocyte lineage. The use of this averaging heatmap plot highlighted a further group of genes which show more distinct differential expression between cluster, including *Pabpc4*, *Mkrn1*, *Eif3g*, *Srsf7*, *Ncl*, *Hnrnpab*, *Ybx1*, *Snrpg*, *Lsm4*, *Hnrnpa2b1*, *Sf3b4*, *Rnps1*, *Lsm2*, and *Hnrnpdl*.

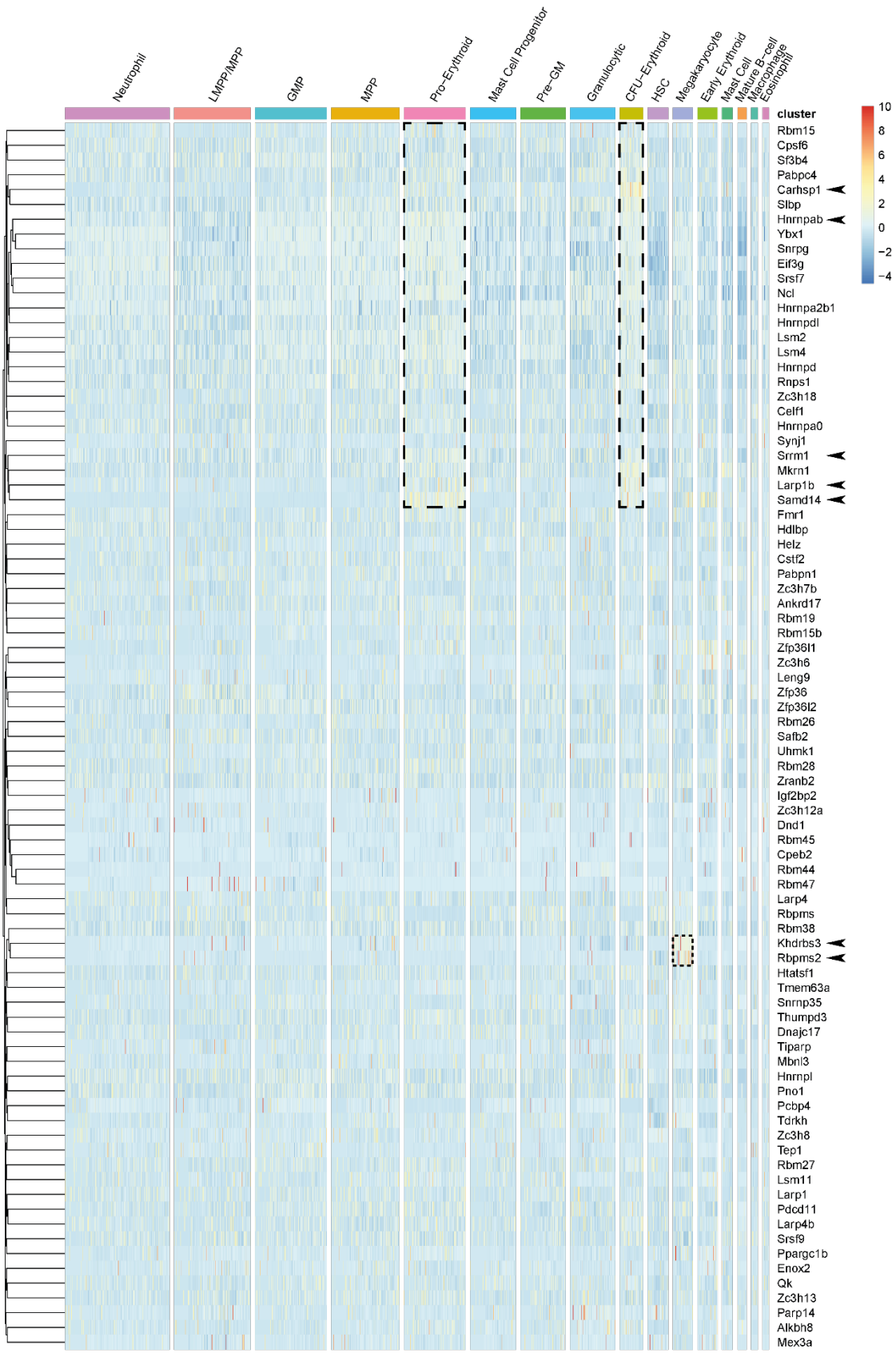


***Figure 32 - Heatmap of gene expression per cluster (ordered).*** *Heatmap created using pheatmap displaying the expression of RBP/SF genes from each single cell within each of the clusters in the short-read data. The genes are clustered to highlight similar expression patterns between genes. Notable clusters of upregulated gene expression specifically in the erythroid and megakaryocytic lineages are indicated by dashed and dotted boxes, respectively, Genes chosen for further investigation are marked by arrowheads.*

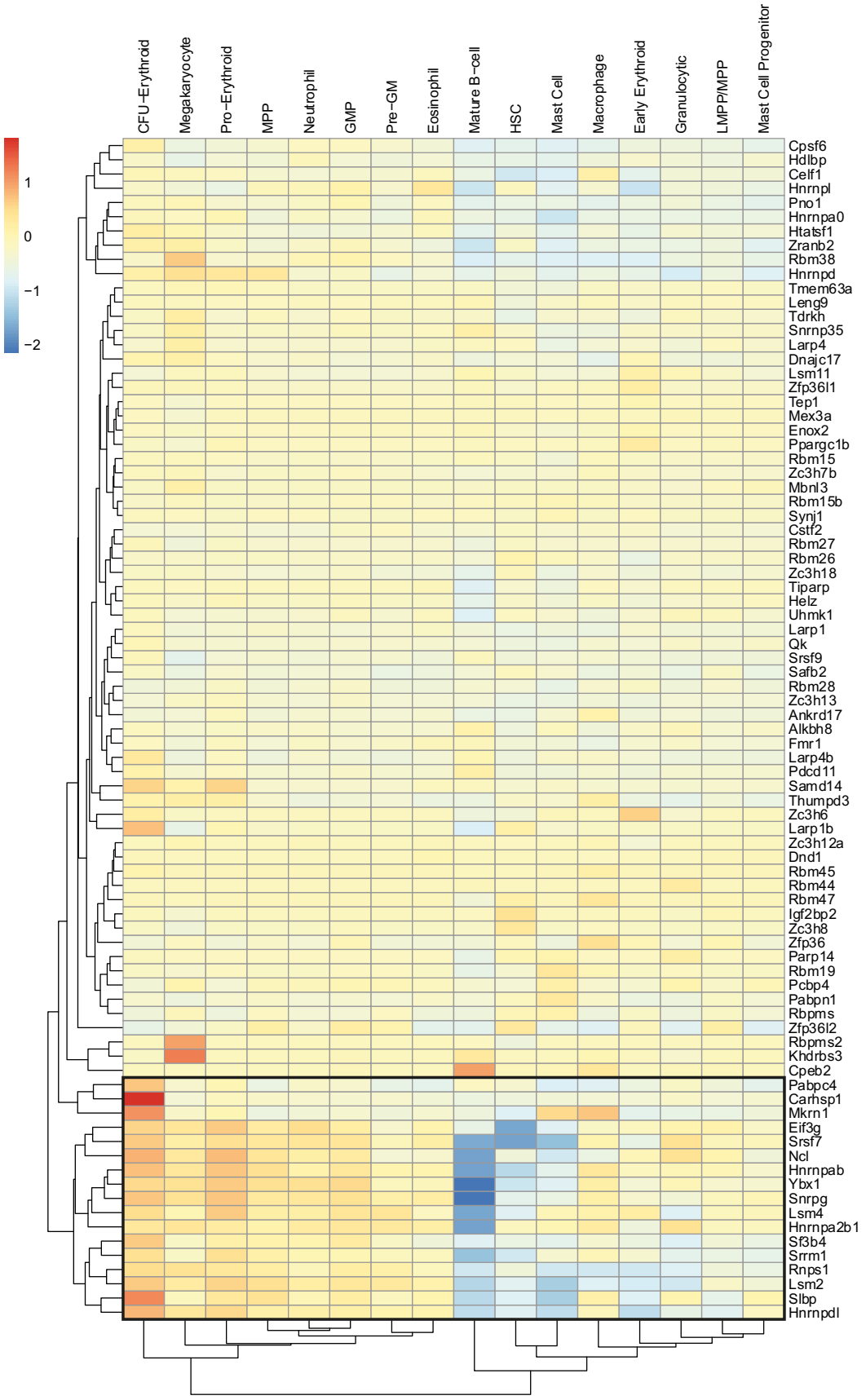


***Figure 33 - Heatmap of median gene expression per cluster.*** *Heatmap created using pheatmap displaying the median expression of RBP/SF genes within each of the clusters in the short-read data. The genes are clustered to highlight similar expression patterns between genes. A particular cluster of genes showing interesting expression patterns is outlined.*

### *3.4.2.2 Selection of KO target genes*

To narrow down the list of potential KO target genes, feature plots for each gene were created to map their expression patterns back onto the UMAP (Fig.34). This revealed that some genes were more widely expressed than the heatmaps made apparent. Genes which showed a clear lineage-specific or lineage-preferential expression pattern were identified, thus selecting a "shorterlist" of genes to investigate with higher priority: *Carhsp1* (erythroid), *Samd14* (erythroid/megakaryocytic), *Rbpms2* (megakaryocytic/HSCs), *Khdrbs3* (megakaryocytic) and *Larp1b* (erythroid).

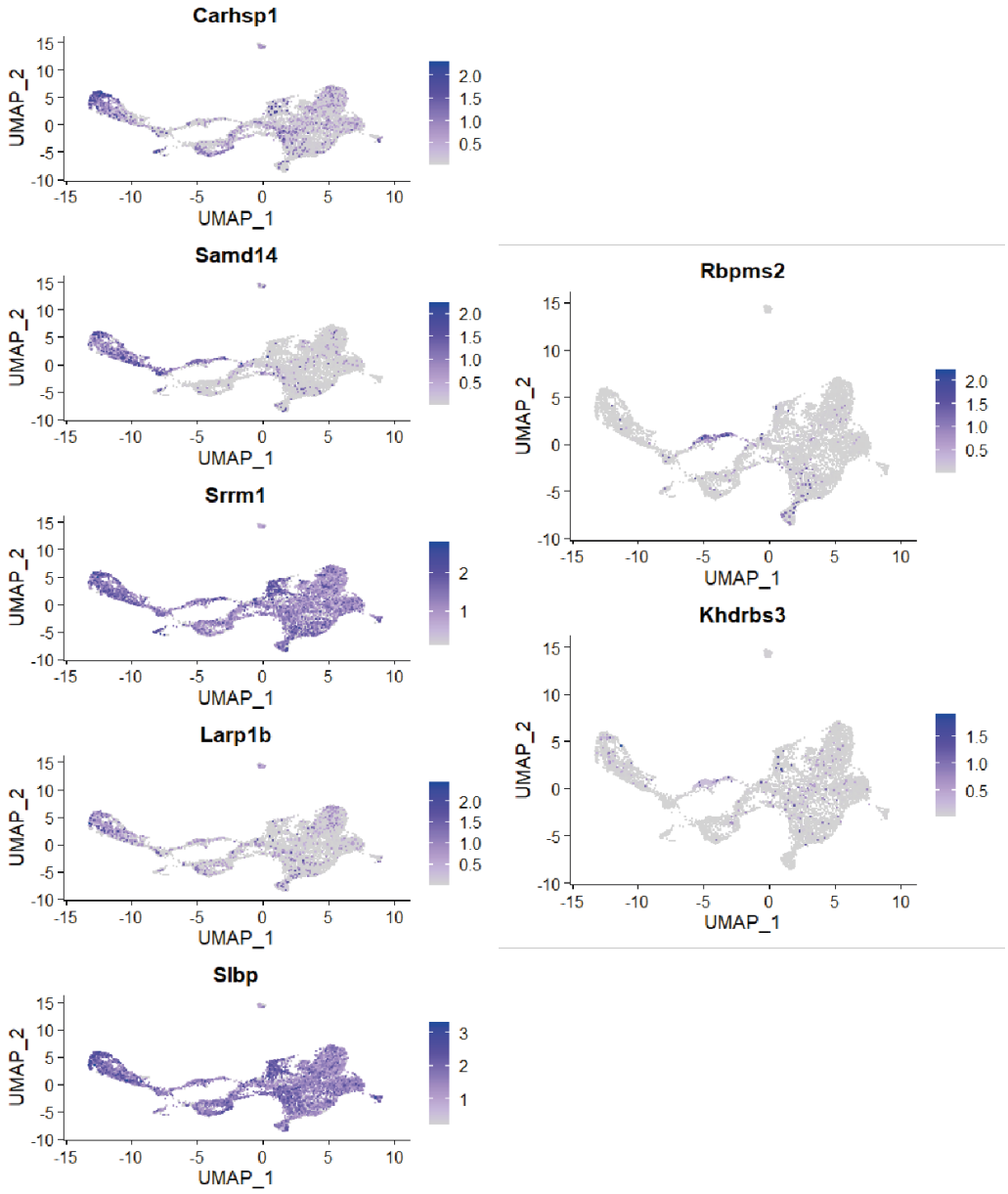


***Figure 34 - Cluster-specific expression patterns of lineage-specifically upregulated shortlist genes.*** *Overlaid onto the UMAP from Fig.27 to illustrate any cluster-specific patterns of expression (feature plots).* Srrm1 *and* Slbp *do not seem to show such patterns and are instead widely expressed.*

Subsequently, the expression levels of the genes in HAP1 and K562 cells were investigated using online resources (RNA-seq data for cell lines from The Human Protein Atlas (HPA; 2022; Carette Raaben and Wong *et al.*, 2011) and the wild-type expression values from the knockout cell line Horizon webpages (Horizon, no date). Table 13 lists the information for each gene. SAMD14 is not considered to be expressed in HAP1 or K562 in both data sources and therefore was eliminated from the list of KO target genes.

***Table 13 - Gene expression levels of the selected RBP/SF genes in HAP1 cells from The Human Protein Atlas (HPA), in K562 cells from HPA (blue) and in HAP1 from Horizon (orange).*** *HPA data is measured in normalised transcripts per million (nTPM); anything below 1 is considered to not be expressed. Horizon data is measured in transcripts per million (TPM); anything below 3 is considered to not be expressed. "Not expressed" values are highlighted in red.*

| | **Human Protein Atlas (<1 is "not expressed")** | | **Horizon (<3 is "not expressed")** |
|---|---|---|---|
| **Gene name** | **HAP1 (nTPM)** | **K562 (nTPM)** | **HAP1 (TPM)** |
| *RBM15* | 17.6 | 26.1 | 8.2 |
| *CARHSP1* | 79.3 | 94.5 | 97.7 |
| *KHDRBS3* | 19.5 | 2.6 | 16.0 |
| *RBPMS2* | 28.8 | 0.5 | NA |
| *SAMD14* | 0.3 | 0.3 | 0.5 |
| *LARP1B* | 32.7 | 37.1 | 27.1 |

*RBPMS2* is expressed in HAP1 but not in K562 according to HPA. The information was not available on the Horizon website because they do not offer a *RBPMS2* knockout cell line. This is because *RBPMS2* is one of the few genes with two copies in HAP1 cells. As described in the introduction to Chapter 2, HAP1 cells are not quite a fully haploid cell line as they have two copies of a portion of chromosome 15, one of which is fused to chromosome 19 (Essletzbichler *et al.*, 2014). This only encompasses around 330 genes in the HAP1 genome (Fig.33) but unfortunately does include *RBPMS2* (the only case from the selected KO target genes) which is located on Chr.15q22 (64,739,891-64,775,589; GRCh38.p14). Both alleles of *RBPMS2* would need to be knocked out in the HAP1 line and the double KO confirmed before proceeding further. Due to this and the lack of expression in K562, *RBPMS2* was also eliminated from the list of KO target genes. It is also worth noting that HAP1 cells also possess the Philadelphia chromosome (as in the parental KBM7 line), meaning that genes on chromosomes 9 or 22 at the sites of translocation might be affected. However, in this case none of the selected KO target genes are on either of these chromosomes.

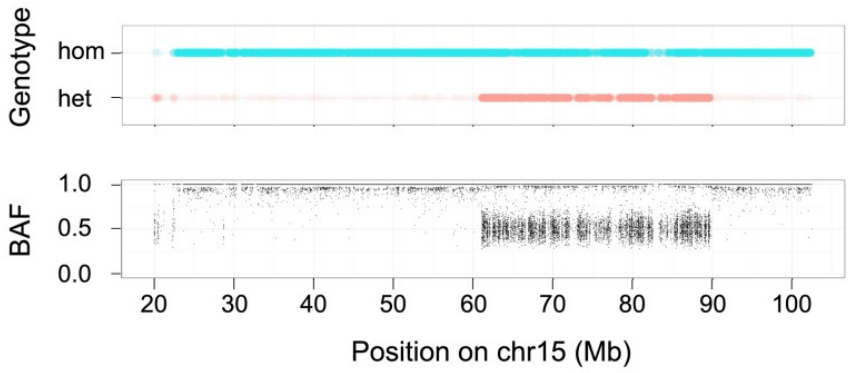


***Figure 35 - Disomy in the HAP1 parental cell line, KBM7.*** *Single-nucleotide polymorphism data from KBM-7 cells were analysed to identify chromosomal segments that are heterozygous and hence disomic. (Top) Heterozygous SNPs are depicted in red; homozygous SNPs, in turquoise. (Bottom) B-allele frequency (BAF) on Chromosome 15. A BAF of 0.5 is indicative of heterozygosity. Essletzbichler* et al*., 2014.*

### 3.4.3 *Rbm15* is expressed across haematopoietic lineages

RBM15 (RNA binding motif protein 15) has been highlighted in previous research for its known functions in m6A methylation of RNAs, splicing and haematopoietic homeostasis, as described in the Introduction and Chapter 2. As also previously mentioned, RBM15 has a role in the negative regulation of the thrombopoietin response in HSCs via the AS of the thrombopoietin receptor gene, *Mpl* (Xiao et al., 2015). This project focuses on exploring the functional consequences of deleting relevant SFs in the haematopoietic context and thus, because one MPL isoform inhibits production of the other with megakaryocytopoietic consequences, this instance of AS and its regulation is of significant interest. Given this, it was speculated whether *Rbm15* too exhibits haematopoietic lineage-specific expression. A UMAP feature plot was also generated for *Rbm15* in Seurat which shows that it is widely expressed across the clusters of cells, showing no exclusivity for any one lineage (Fig.34). *Rbm15* therefore provides a non-lineage-specific gene to compare with *Carhsp1*, *Khdrbs3* and *Larp1b*.

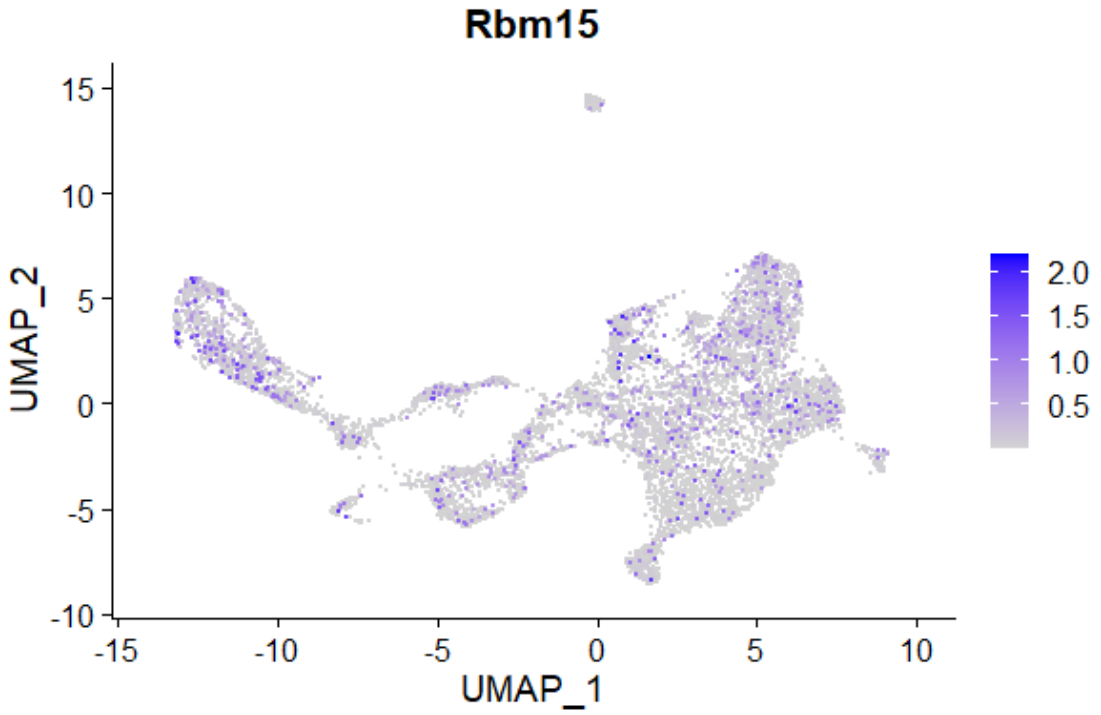


***Figure 36* - Rbm15 *expression displayed overlaying the UMAP from Fig.29 to illustrate that* Rbm15 *is expressed across all the present cell-types.***

### 3.4.4 Initial steps in generation of *CARHSP1*, *KHDRBS3* and *LARP1B* knockout lines

The single-cell short-read data analysis provided candidate RBP/SF genes with distinctly upregulated expression in discrete blood cell types. The next step was to investigate how depletion of these genes affected AS in haematologically relevant cell lines by use of CRISPR-KO and measurement of transcriptomic changes by bulk short-read and long-read sequencing. The genes selected for further investigation were *CARHSP1 and LARP1B*, genes strongly upregulated in CFU-Erythroid lineages, and *KHDRBS3*, the gene most strongly differentially expressed in the megakaryocyte lineage. The method of gene knockout for *CARHSP1, KHDRBS3*, and *LARP1B* were similar to that described in Chapter 2. As with the *RBM15*-KOs described in Chapter 2, three guides were tested for editing efficiency per gene. If time had allowed, these would have been single-cell sorted, and individual clones screened for evidence of frame-shifting indels.

#### *3.4.4.1 Confirmation of editing of KO target genes*

The workflow used for generating and validating CRISPR KOs was the same as for *RBM15-KO* generation described in Chapter 2. Briefly, HAP1 Cas9+ cells were

transfected with three different gRNAs, genomic DNA extracted, the region surrounding the Cas9 cut-site amplified by PCR, and amplicons sequenced by Sanger Sequencing.

*CARHSP1* was targeted by three guides within a 1kb region spanning exons 2 and 3 of the gene: gRNA_2 targeted exon 2 and gRNA_1 and gRNA_3 targeted exon 3. This allowed a single amplicon to be generated that encompassed all three Cas9 cut-sites, allowing easy screening of guide efficacy. Agarose gel electrophoresis revealed the *CARHSP1* amplicon generated to be large and present as a single clean band (Fig.37). Sanger sequencing of the amplicon revealed that gRNA_1 was effective in editing *CARHSP1* near to the gRNA_1-directed Cas9 cut-site (Fig.38).

*KHDRBS3* was also targeted by three gRNAs, but different amplicons had to be produced for each Cas9 cut-site due to the distance between each gRNA annealing site. The genomic PCR for gRNA_1 and gRNA_2 generated multiple bands when separated by agarose gel electrophoresis instead of a single amplicon, whilst the genomic PCR of gRNA_3 cut-site generated a single amplicon (Fig.37). Gel-extracted amplicons from gRNA_1, gRNA_2, and gRNA_3 of the appropriate size were sent for Sanger sequencing to determine the presence of indels. gRNA_1 and gRNA_3 amplicons did not produce a sequencing trace despite multiple attempts (not shown), however the gel-extracted gRNA_2 amplicon was sequenced successfully and showed the evidence of multiple overlapping traces after the gRNA_2-directed Cas9 cut-site (Fig.38).

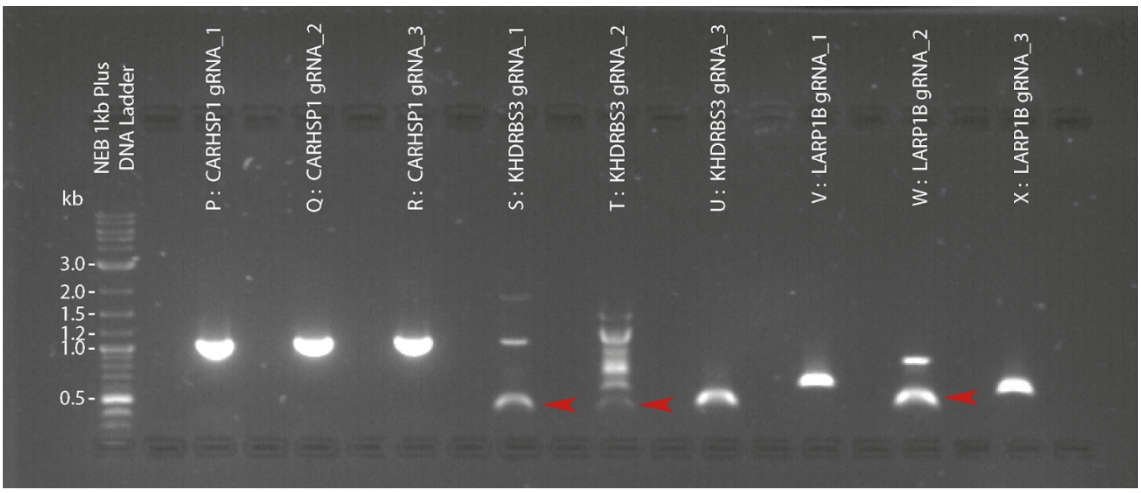


***Figure 37 - UV images from gel electrophoresis of samples P, Q, R, S, T, U, V, W, and X.*** *NEB 1kb Plus DNA Ladder was used as a size reference. Expected size of the amplicons: 992 bp for P/Q/R, 319 bp for S, 302 bp for T, 312 bp for U, 480 bp for V, 314 bp for W and 416 bp for X. The red arrowheads show the intended PCR product bands where multiple were produced; these were the bands extracted from the gel for Sanger sequencing analysis. The primers used in the PCR reactions to produce each amplicon are listed in Table 9.*

*LARP1B* also had three gRNA annealing regions sufficiently far apart to require the generation of three different amplicons. Two of the three amplicons produced single PCR products, but the amplicon for gRNA_2 produced some larger product in addition to the expected 314 bp band (Fig.37). Sanger sequencing of the gel-extracted amplicons revealed that gRNA_3 (Sample X) had multiple overlapping sequencing traces near the gRNA_3-directed Cas9 cut-site, implying indels had been generated (Fig.38).

To summarise, the most effective guides for each gene were: gRNA_1 for *CARHSP1*, gRNA_2 for *KHDRBS3*, and gRNA_3 for *LARP1B*. Due to a lack of time, these mixed populations containing edited and unedited cells were frozen down for future edit-verification by single-cell cloning and downstream analyses of any isolated knockout clones.

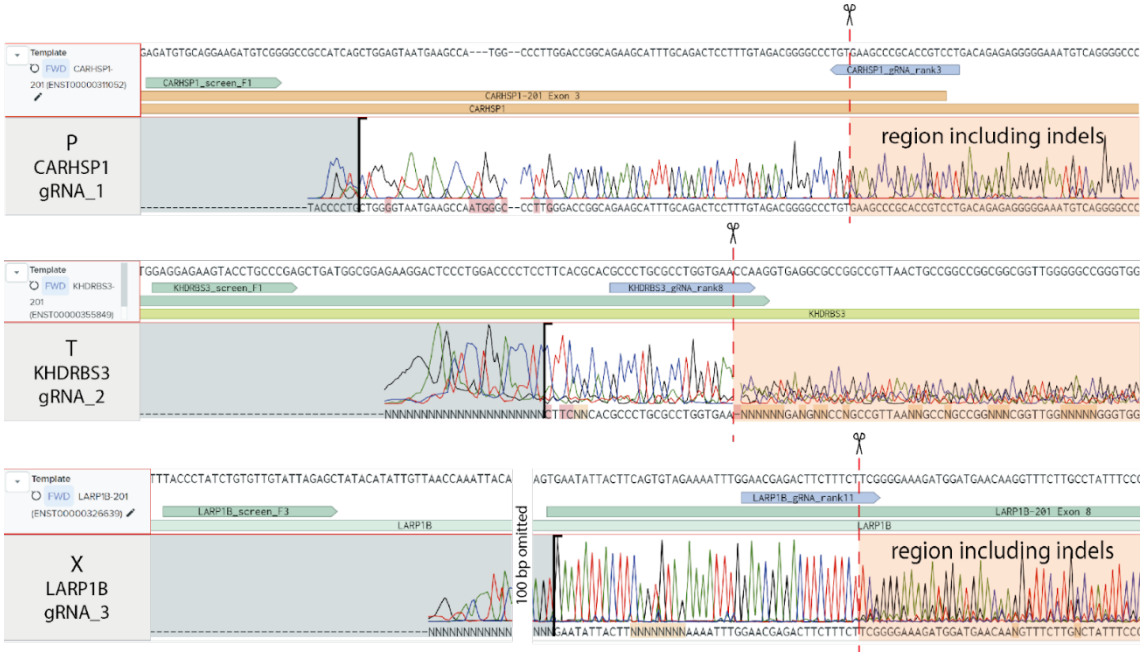


***Figure 38 - Sanger sequencing traces indicating signs of possible editing in samples P and X and an inconclusive trace for sample T.*** *Amplicons surrounding gRNA-directed Cas9 cut-sites in HAP1 Cas9+ cells were sequenced by Sanger sequencing and aligned to the genomic sequence of each gene using Benchling. Red dashed line indcates Cas9 cut-site for each gene, orange areas indicate regions where multiple overlapping sequencing traces become visible, blue large arrows indicate gRNA annealing location, and green large arrows show sequencing primer positions.*

## *3.5 Discussion*

The objectives of this Chapter of the project were centred on identifying target cell-type specific RBP/SF genes for KO experiments. Single-cell RNA-seq data from $Lin^-$ $cKit^+$ haemateopoetic stem and progenitor cells from six mice were analysed using Seurat to identify RBP/SF genes common to all six datasets. UMAP nearest neighbour dimensionality reduction analysis was then used to cluster cells by cell type, generate heatmaps clustering RBP/SF genes by median expression per cell within each cell type, and visualise the cell-type specificity of notable genes. The RBP/SF genes *CARHSP1 and LARP1B* were chosen, given their strong upregulation in CFU-Erythroid lineages, and *KHDRBS3,* given its high specific expression in megakaryocytes. These three genes were knocked out in HAP1 cells to generate bulk populations in preparation for future clone isolation and analysis.

### 3.5.1 Discussion of results

The heatmaps show comparatively little extreme up- or downregulation of RBP/SF genes; the majority of gene expression variability between clusters was apparently mild. This is perhaps surprising to see in stem and progenitor cells, considering that AS is a contributor to cellular differentiation (Vaquero-Garcia *et al.*, 2016). This limited expression could be an example of low read coverage per cell and exon-level data omitting biological information: the isoform ratios of these genes might have influence here but are not measured.

Alternatively, it could be due to unsuitable scaling of the data for plotting the heatmaps. Genes were identified from the heatmaps without too much difficulty but the expression scale of the individual cells heatmap highlighted only the very extreme expression patterns of individual cells within clusters. The majority of the heatmap therefore showed up- or downregulation of genes with less clarity. This also applied to the median expression heatmap but to a lesser extent. The scale of the data could be experimented with to see if new findings are revealed more discretely from an even distribution of gene expression.

Whilst a large portion of the average expression heatmap showed only mild gene expression changes between clusters, it did nicely visualise two groups of genes (top and bottom of Fig.33) which have stronger average expression changes during differentiation. Examples are *Slbp*, *Srsf7* and *Eif3g*: downregulated in HSCs but

increasingly expressed as a cell differentiates toward myeloid fates. Two more instances following the same trend are *Lsm2* and *Rnps1* which show almost identical expression patterns. In humans, LMS2 is a component of the U4/U6-U5 tri-snRNP complex and precatalytic spliceosome complex (Bertram *et al.*, 2017) and RNPS1 is a component of pre- and post-splicing multiprotein mRNP complexes (including the exon-junction complex formed after splicing) and also inhibits the formation of proapoptotic isoforms including Bcl-XS by AS regulation (Michelle *et al.*, 2012).

This trend of RBP/SF upregulation during differentiation is not reflected for all clusters; some genes appear to become more downregulated in the Mast Cell or Mature B-cell clusters (*Hnrnpdl*, *Srrm1*, *Sf3b4* etc.). These gene expression patterns suggest that splicing and AS may increase as a cell differentiates, but not for each cell type. These genes with stronger average expression changes would be suitable candidates for future experiments to explore these patterns further.

#### *3.5.1.1 Low-level expression of transcripts*

The mRNA transcripts in a cell comprise a small number of highly expressed transcripts and a large number of lowly expressed transcripts. Around 85% are present in 1-100 copies and many are present in just a few copies per cell (Macaulay and Voet, 2014). These are inherently missed in 10x Genomics and other scRNA-seq experiments analysing higher numbers of cells because the more cells that are included, the fewer less-abundant transcripts are detected.

It should be noted that more of the genes from the RBP/SF list could have in fact been expressed in the LK cells, however, since SF mRNAs are commonly amongst the lowly expressed transcripts, many may have not been detected in this experiment. This makes isoform analysis from the short-read dataset essentially impossible, particularly with the added limitations of scRNA-seq: low molecular capture rate, 3' targeting, low DNA conversion efficiency and uneven capturing of transcripts. Analysing the long-read data would work towards providing an insight into the isoform-level picture.

### 3.5.2 Functions of KO target genes

The functions of the shortlisted genes give some insight into the changes that might occur following their knockout.

*Carhsp1* (calcium-regulated heat stable protein 1) was found to be particularly upregulated in the CFU-Erythroid cluster as well as the Pro-Erythroid cluster to a lesser extent, implying expression of *Carhsp1* increases during erythroid lineage development. *CARHSP1* contains a cold-shock domain and two RNA-binding motifs, enabling its function of binding and regulating the stability of mRNA. Little is known about this gene as it does not feature much in the literature, however, Pfeiffer *et al.* (2011) did determine *Carhsp1* to enhance the stability of TNF-α mRNA. It has also been previously mentioned as a late erythroid gene (Hu, Yuan and Lodish, 2014). The Ser-53 residue of human CARHSP1 is phosphorylated via the PI3K-Akt signalling pathway, which regulates the cell cycle, cell growth and proliferation (Schäfer *et al.*, 2003); this pathway has been shown to be dysregulated in human cancers (Yuan and Cantley, 2008).

*Khdrbs3* (KH domain-containing, RNA-binding, signal transduction-associated protein 3) was upregulated mildly but specifically in the Megakaryocyte progenitor cluster. The protein encoded by this gene is thought to be involved in regulating AS by influencing mRNA splice site selection and exon inclusion. For example, *KHDRBS3* protein is involved in splice site selection of VEGF (Cohen *et al.*, 2005) and in regulating the AS of sarcomere protein titin (TTN), leading to intron retention (Boeckel *et al.*, 2021).

*Larp1b* (La Ribonucleoprotein 1B) is another gene showing Mk-Er lineage specificity, although to a lesser extent. The encoded protein is an RBP involved in the positive regulation of translation as its La motif and the neighbouring amino acids fold to form an RNA-recognition motif (RRM). Many isoforms are produced by AS of *LARP1B* (9 in human, 6 in mouse). It should be noted that there is less similarity between the human and mouse orthologues than for the other KO target genes.

The chromosomal locations and number of known protein-coding isoforms (human and mouse) along with the pattern of cluster-specific upregulation for each of the selected KO target genes are summarised in Table 14 below.

***Table 14 - Selected KO target gene information table.*** *Including chromosomal locus, exon count, strand and number of known protein-coding isoforms for both human and mouse as well as the clusters they are upregulated in according to the heatmaps (Fig.32, Fig.33).*

| | **Human – *Homo sapiens*** | | | | **Mouse – M*us musculus*** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Gene name** | **Locus** | **Exon count** | **Strand** | **Known Protein-coding isoforms** | **Locus** | **Exon count** | **Strand** | **Known protein-coding isoforms** | **Cluster-specific upregulation** |
| *RBM15* | **1**p13.3 | 3 | Forward | 9 | **3**; 3 F2.3 | 4 | Reverse | 1 | Non-specific |
| *CARHSP1* | **16**p13.2 | 12 | Reverse | 16 | **16** A1; 16 4.26 cM | 6 | Reverse | 1 | Er |
| *RBPMS2* | **15**q22.31 | 11 | Reverse | 2 | **9**; 9 C | 10 | Forward | 4 | Mk/HSCs |
| *KHDRBS3* | **8**q24.23 | 16 | Forward | 5 | **15** D3; 15 30.36 cM | 13 | Forward | 3 | Mk |
| *SAMD14* | **17**q21.33 | 12 | Reverse | 4 | **11**; 11D | 10 | Forward | 2 | Er/Mk |
| *LARP1B* | **4**q28.2 | 31 | Forward | 9 | **3**; 3 B | 17 | Forward | 6 | Er |

### 3.5.3 Guides verified for *CARHSP1* and *LARP1B*

Preliminary knockout experiments were undertaken which targeted the three haematopoietic lineage-specific genes identified in the Seurat analysis in this Chapter. For all three genes, one of the three guides tested was confirmed to edit their targets, although this was more ambiguous for *KHDRBS3* given the quality of the returned sequencing trace and the multiple unintended amplicon PCR products. Different primers should be designed and the PCRs optimised by altering the annealing temperature to obtain clean products of a single size for ease of purification. This should facilitate the generation of a neat Sanger sequencing trace and the subsequent confirmation of the efficacy of each of the three gRNAs targeting *KHDRBS3*.

The mixed populations of transfected cells were all frozen down for future experiments. The next step with these cells (if not re-transfecting at first) would be to put the samples transfected with the verified guides targeting *CARHSP1* and *LARP1B* back in culture and sort to subclone and isolate edited and wild-type clones, as in the *RBM15*-KO experiments in Chapter 2. Differential expression analysis can highlight the downstream consequences of editing the lineage-specific genes and these results can be compared against the results from editing non-lineage-specific *RBM15*.

# CHAPTER 4: DISCUSSION

This project aimed to identify genes encoding RBPs and SFs with potentially important roles in the haematopoietic system and investigate the transcriptomic consequences of knocking them out using CRISPR-Cas9. These experiments have yielded many interesting results pertaining to the function and mechanism of action of RBM15 as well as establishing a CRISPR-KO procedure in the lab and identifying optimisations that can be made to the process.

*RBM15* was edited in the HAP1 cell line but the expression of the transcript was found to be increased with respect to the wild-type control. Upon further investigation, it was discovered that the edited cell population had become diploid early into the experiment and that both alleles had been edited. Given the equal abundance of transcripts from both edited alleles and the nature of each edit, it appears that the clonal population generated (A2F7) expressed a near-wild-type RBM15 and a truncated RBM15 of just the first RRM. Measurement of the expression of RBM15 protein in the edited A2F7 clone and the wild-type A2E4 clone should be a priority. Wild-type verified haploid HAP1 cells should also be included in this assay for insurance. It would also be particularly interesting to compare RBM15 protein expression in cells with different levels of editing: haploid edited, diploid double-edited (A2F7) and diploid single-edited, for example.

Changes in ploidy caused complications in the process of deleting *RBM15*, although both alleles of the diploidised HAP1 cells were found to be edited. A key question to be answered is when did A2F7 become diploid? If diploid upon transfection, the gRNAs must have cut both *RBM15* alleles and different indels occurred upon repair. Yet it could have also become diploid upon transfection by a mechanism that is so far unclear. A ploidy-checking protocol must be established and integrated at regular intervals into the KO and single-cell sorting procedure, with optimisation tests performed to determine the best thresholds for confluency to avoid diploidisation. The ploidy of A2E4, A2F7, HAP1 wild-type and diploid HAP1 controls (the latter can be purchased on request from Horizon, as in Beigl, Kjosås, Seljeseth *et al.*, 2020) must be also confirmed.

An inverse relationship was found between the expression of *RBM15* and of a paralog *RBM15B* which seems to support literature documentation that these two genes share sequence and domain organisation as well as some functionality. There may be something significant underlying the relationship between RBM15 and RBM15B that is not yet known to the literature. Compensation between the genes is known in the context

of *XIST*-mediated inhibition of transcription but given the similarities between the two genes and the clear inverse relationship seen here, this could be true also for other functions of RBM15. When measuring RBM15 protein expression in clones isolated from this experiment, RBM15B expression should also be checked.

A large number of genes were shown to be differentially expressed between the edited diploid heterozygous cells and the wild-type control, with many of the genes encoding known RBM15 targets and binding sites. Gene ontology DAVID analysis revealed an interesting enrichment of differentially expressed genes annotated with the "Differentiation" ontology, and within this group, the gene *BMP7* was found to be substantially upregulated in *RBM15*-edited cells. The large number of differentially expressed genes might be due to RBM15 having several different crucial functions with different mechanisms of actions. AS could likely be a large factor, resulting in multiple isoforms of target genes and a subsequent change to their functions. Upcoming experiments will begin to test these hypotheses in more detail (described in Chapter 4.1).

Single-cell short-read analysis of a haematopoietic stem cell data set was performed and RBPs/SFs exhibiting haematopoietic lineage-specific expression were found, two preferentially expressed in the erythroid lineage and one in the megakaryocytic lineage. All three of the lineage-specific genes were confirmed to be edited following an experiment attempting to knock out the genes in the same manner as the *RBM15*-KO experiments. Mixed populations of these lines are frozen awaiting further experimentation.

## *4.1 Expanding the project further*

Splicing factors have a substantial degree of control over the fate of a cell. The expression of certain genes and isoforms of genes can be indicative of that fate. Relationships between the expression of splicing factors and the expression of cell fate-specific transcript isoforms have not yet been investigated on a large scale. Transcriptomic inferences have been made from the experiments in Chapter 2 on the bulk population level rather than at single-cell resolution and using short-read data rather than from long reads or a combined approach. From the long-read sequencing of the mouse cells in the dataset used in Chapter 3, differential isoform expression for the selected KO genes can be examined as well as any closely associated genes. Splicing factor genes with multiple isoforms which result in different protein sequences or different

untranslated regions (UTRs) should be identified. Utilising the power of long-read single-cell sequencing, comparisons should also be made into the differential expression of transcript isoforms with lineage-specific RBP/SF expression patterns (such as those uncovered in Chapter 3.3.2). These correlations could subsequently be tested at a fine resolution with single-cell transcriptomic analyses of models created with the RBP/SF genes knocked out.

# REFERENCES

Adusumalli, S. *et al.* (2019) 'Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease', *Aging Cell*, 18(3), p. e12928. Available at: https://doi.org/10.1111/acel.12928.

Akashi, K. *et al.* (1997) 'Bcl-2 Rescues T Lymphopoiesis in Interleukin-7 Receptor–Deficient Mice', *Cell*, 89(7), pp. 1033–1041. Available at: https://doi.org/10.1016/S0092-8674(00)80291-3.

Anczuków, O. *et al.* (2012) 'The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation', *Nature Structural & Molecular Biology*, 19(2), pp. 220–228. Available at: https://doi.org/10.1038/nsmb.2207.

Appasamy, P.M. (1999) 'Biological and clinical implications of interleukin-7 and lymphopoiesis', *Cytokines, Cellular & Molecular Therapy*, 5(1), pp. 25–39.

Appel, L.-M. *et al.* (2023) 'The SPOC domain is a phosphoserine binding module that bridges transcription machinery with co- and post-transcriptional regulators', *Nature Communications*, 14(1), p. 166. Available at: https://doi.org/10.1038/s41467-023-35853-1.

Babraham Bioinformatics (no date) *Trim Galore!* Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (Accessed: 30 September 2022).

Bai, H., Xu, P. and Chen, B. (2021) 'Gene signatures and prognostic values of m6A-related genes in multiple myeloma', *Curr Res Transl Med*, 69(2), p. 103288. Available at: https://doi.org/10.1016/j.retram.2021.103288.

Barash, Y. *et al.* (2010) 'Deciphering the splicing code', *Nature*, 465(7294), pp. 53–59. Available at: https://doi.org/10.1038/nature09000.

Baruchel, A. *et al.* (1991) 'Nonrandom t(1;22)(p12–p13;q13) in acute megakaryocytic malignant proliferation', *Cancer Genetics and Cytogenetics*, 54(2), pp. 239–243. Available at: https://doi.org/10.1016/0165-4608(91)90213-E.

Beigl, T.B. *et al.* (2020) 'Efficient and crucial quality control of HAP1 cell ploidy status', *Biology Open*, 9(11), p. bio057174. Available at: https://doi.org/10.1242/bio.057174.

*Benchling* (no date). Available at: https://benchling.com/ (Accessed: 24 October 2021).

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), pp. 289–300. Available at: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berget, S.M., Moore, C. and Sharp, P.A. (1977) 'Spliced segments at the 5' terminus of adenovirus 2 late mRNA', *Proc. Natl. Acad. Sci. U. S. A.*, 74(8), pp. 3171–3175. Available at: https://doi.org/10.1073/pnas.74.8.3171.

Berglund, A.-C. *et al.* (2008) 'InParanoid 6: eukaryotic ortholog clusters with inparalogs', *Nucleic Acids Res.*, 36(Database issue), pp. D263-6. Available at: https://doi.org/10.1093/nar/gkm1020.

Bertram, K. *et al.* (2017) 'Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation', *Cell*, 170(4), pp. 701-713.e11. Available at: https://doi.org/10.1016/j.cell.2017.07.011.

Bigas, A., Martin, D.I.K. and Milner, L.A. (1998) 'Notch1 and Notch2 Inhibit Myeloid Differentiation in Response to Different Cytokines', *Molecular and Cellular Biology*, 18(4), pp. 2324–2333. Available at: https://doi.org/10.1128/MCB.18.4.2324.

Blanchette, M. and Chabot, B. (1999) 'Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization', *EMBO J.*, 18(7), pp. 1939–1952. Available at: https://doi.org/10.1093/emboj/18.7.1939.

Boeckel, J.-N. *et al.* (2021) 'SLM2 Is A Novel Cardiac Splicing Factor Involved in Heart Failure due to Dilated Cardiomyopathy', *Genomics Proteomics Bioinformatics* [Preprint]. Available at: https://doi.org/10.1016/j.gpb.2021.01.006.

Boise, L.H. *et al.* (1993) 'bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death', *Cell*, 74(4), pp. 597–608. Available at: https://doi.org/10.1016/0092-8674(93)90508-N.

Brett, D. *et al.* (2000) 'EST comparison indicates 38% of human mRNAs contain possible alternative splice forms', *FEBS Letters*, 474(1), pp. 83–86. Available at: https://doi.org/10.1016/S0014-5793(00)01581-7.

Broad Institute (2022) *CRISPick*. Available at: https://portals.broadinstitute.org/gppx/crispick/public (Accessed: 30 April 2022).

Carette, J.E. *et al.* (2011) 'Ebola virus entry requires the cholesterol transporter Niemann-Pick C1', *Nature*, 477(7364), pp. 340–343. Available at: https://doi.org/10.1038/nature10348.

Carroll, A. *et al.* (1991) 'The t(l;22)(pl3;ql3) Is Nonrandom and Restricted to Infants With Acute Megakaryoblastic Leukemia: A Pediatric Oncology Group Study', *Blood*, 78(3), pp. 748–752. Available at: https://doi.org/10.1182/blood.V78.3.748.748.

Celik, H., Kramer, A. and Challen, G.A. (2016) 'DNA methylation in normal and malignant hematopoiesis', *International Journal of Hematology*, 103(6), pp. 617–626. Available at: https://doi.org/10.1007/s12185-016-1957-7.

Cendrowski, J. *et al.* (2020) 'Splicing variation of BMP2K balances abundance of COPII assemblies and autophagic degradation in erythroid cells', *eLife*. Edited by E.A. Miller, V. Malhotra, and E.A. Miller, 9, p. e58504. Available at: https://doi.org/10.7554/eLife.58504.

Chen, D. *et al.* (2021) 'Interleukin-7 Biology and Its Effects on Immune Cells: Mediator of Generation, Differentiation, Survival, and Homeostasis', *Frontiers in Immunology*, 12. Available at: https://www.frontiersin.org/articles/10.3389/fimmu.2021.747324 (Accessed: 30 May 2023).

Chen, L. *et al.* (2014) 'Transcriptional diversity during lineage commitment of human blood progenitors', *Science*, 345(6204), p. 1251033. Available at: https://doi.org/10.1126/science.1251033.

Chen, S. and Abdel-Wahab, O. (2021) 'Splicing regulation in hematopoiesis', *Current opinion in hematology*, 28(4), pp. 277–283. Available at: https://doi.org/10.1097/MOH.0000000000000661.

Chow, L.T. *et al.* (1977) 'An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA', *Cell*, 12(1), pp. 1–8. Available at: https://doi.org/10.1016/0092-8674(77)90180-5.

Cobb, B.S. *et al.* (2005) 'T cell lineage choice and differentiation in the absence of the RNase III enzyme Dicer', *Journal of Experimental Medicine*, 201(9), pp. 1367–1373. Available at: https://doi.org/10.1084/jem.20050572.

Coers, J., Ranft, C. and Skoda, R.C. (2004) 'A truncated isoform of c-Mpl with an essential C-terminal peptide targets the full-length receptor for degradation', *J. Biol. Chem.*, 279(35), pp. 36397–36404. Available at: https://doi.org/10.1074/jbc.M401386200.

Cohen, C.D. *et al.* (2005) 'Sam68-like mammalian protein 2, identified by digital differential display as expressed by podocytes, is induced in proteinuria and involved in splice site selection of vascular endothelial growth factor', *J. Am. Soc. Nephrol.*, 16(7), pp. 1958–1965. Available at: https://doi.org/10.1681/ASN.2005020204.

Conboy, J.G. *et al.* (1991) 'Tissue- and development-specific alternative RNA splicing regulates expression of multiple isoforms of erythroid membrane protein 4.1', *Journal of Biological Chemistry*, 266(13), pp. 8273–8280. Available at: https://doi.org/10.1016/S0021-9258(18)92973-X.

Cool, T. *et al.* (2020) 'Interleukin 7 receptor is required for myeloid cell homeostasis and reconstitution by hematopoietic stem cells', *Experimental Hematology*, 90, pp. 39-45.e3. Available at: https://doi.org/10.1016/j.exphem.2020.09.001.

Corbett, A.H. (2018) 'Post-transcriptional regulation of gene expression and human disease', *Current Opinion in Cell Biology*, 52, pp. 96–104. Available at: https://doi.org/10.1016/j.ceb.2018.02.011.

Cui, Y., Cai, M. and Stanley, H.E. (2017) 'Comparative Analysis and Classification of Cassette Exons and Constitutive Exons', *Biomed Res. Int.*, 2017, p. 7323508. Available at: https://doi.org/10.1155/2017/7323508.

Dahlin, J.S. *et al.* (2018) 'A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice', *Blood*, 131(21), pp. e1–e11. Available at: https://doi.org/10.1182/blood-2017-12-821413.

Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

Deguillien, M. *et al.* (2001) 'Multiple cis elements regulate an alternative splicing event at 4.1R pre-mRNA during erythroid differentiation', *Blood*, 98(13), pp. 3809–3816. Available at: https://doi.org/10.1182/blood.V98.13.3809.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. Available at: https://doi.org/10.1093/bioinformatics/bts635.

Doench, J.G. *et al.* (2014) 'Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation', *Nature Biotechnology*, 32(12), pp. 1262–1267. Available at: https://doi.org/10.1038/nbt.3026.

Dominissini, D. *et al.* (2012) 'Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq', *Nature*, 485(7397), pp. 201–206. Available at: https://doi.org/10.1038/nature11112.

Doulatov, S. *et al.* (2012) 'Hematopoiesis: A Human Perspective', *Cell Stem Cell*, 10(2), pp. 120–136. Available at: https://doi.org/10.1016/j.stem.2012.01.006.

Duncan, A.W. *et al.* (2005) 'Integration of Notch and Wnt signaling in hematopoietic stem cell maintenance', *Nature Immunology*, 6(3), pp. 314–322. Available at: https://doi.org/10.1038/ni1164.

Duran-Lozano, L. *et al.* (2021) 'Germline variants at SOHLH2 influence multiple myeloma risk', *Blood Cancer Journal*, 11(4), pp. 1–8. Available at: https://doi.org/10.1038/s41408-021-00468-6.

Dutta, A. *et al.* (2021) 'U2af1 is required for survival and function of hematopoietic stem/progenitor cells', *Leukemia*, 35(8), pp. 2382–2398. Available at: https://doi.org/10.1038/s41375-020-01116-x.

Dvinge, H. *et al.* (2016) 'RNA splicing factors as oncoproteins and tumour suppressors', *Nature Reviews Cancer*, 16(7), pp. 413–430. Available at: https://doi.org/10.1038/nrc.2016.51.

Dvinge, H. and Bradley, R.K. (2015) 'Widespread intron retention diversifies most cancer transcriptomes', *Genome Med.*, 7(1), p. 45. Available at: https://doi.org/10.1186/s13073-015-0168-9.

Edwards, C.R. *et al.* (2016) 'A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages', *Blood*, 127(17), pp. e24–e34. Available at: https://doi.org/10.1182/blood-2016-01-692764.

ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. Available at: https://doi.org/10.1038/nature11247.

Essletzbichler, P. *et al.* (2014) 'Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line', *Genome Res.*, 24(12), pp. 2059–2065. Available at: https://doi.org/10.1101/gr.177220.114.

Fabian, M.R., Sonenberg, N. and Filipowicz, W. (2010) 'Regulation of mRNA Translation and Stability by microRNAs', *Annual Review of Biochemistry*, 79(1), pp. 351–379. Available at: https://doi.org/10.1146/annurev-biochem-060308-103103.

Fairbrother, W.G. *et al.* (2002) 'Predictive Identification of Exonic Splicing Enhancers in Human Genes', *Science*, 297(5583), pp. 1007–1013. Available at: https://doi.org/10.1126/science.1073774.

Felli, N. *et al.* (2005) 'MicroRNAs 221 and 222 inhibit normal erythropoiesis and erythroleukemic cell growth via kit receptor down-modulation', *Proceedings of the National Academy of Sciences*, 102(50), pp. 18081–18086. Available at: https://doi.org/10.1073/pnas.0506216102.

Fiedler, K. and Brunner, C. (2012) 'The role of transcription factors in the guidance of granulopoiesis', *American Journal of Blood Research*, 2(1), pp. 57–65.

Fu, Y. *et al.* (2013) 'High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells', *Nature Biotechnology*, 31(9), pp. 822–826. Available at: https://doi.org/10.1038/nbt.2623.

Fujiwara, Y. *et al.* (1996) 'Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1.', *Proceedings of the National Academy of Sciences*, 93(22), pp. 12355–12358. Available at: https://doi.org/10.1073/pnas.93.22.12355.

Gao, Y., Vasic, R. and Halene, S. (2018) 'Role of alternative splicing in hematopoietic stem cells during development', *Stem Cell Investig*, 5, p. 26. Available at: https://doi.org/10.21037/sci.2018.08.02.

*GENCODE - Human Release 40* (no date). Available at: https://www.gencodegenes.org/human/ (Accessed: 11 May 2022).

Graveley, B.R. (2005) 'Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures', *Cell*, 123(1), pp. 65–73. Available at: https://doi.org/10.1016/j.cell.2005.07.028.

Greenberg, M.V.C. and Bourc'his, D. (2019) 'The diverse roles of DNA methylation in mammalian development and disease', *Nature Reviews Molecular Cell Biology*, 20(10), pp. 590–607. Available at: https://doi.org/10.1038/s41580-019-0159-6.

Gu, Z. (2021) *ComplexHeatmap Complete Reference*. Available at: https://jokergoo.github.io/ComplexHeatmap-reference/book/index.html (Accessed: 21 February 2022).

Haferlach, T. *et al.* (2014) 'Landscape of genetic lesions in 944 patients with myelodysplastic syndromes', *Leukemia*, 28(2), pp. 241–247. Available at: https://doi.org/10.1038/leu.2013.336.

Halperin, R.F. *et al.* (2021) 'Improved methods for RNAseq-based alternative splicing analysis', *Scientific Reports*, 11(1), p. 10740. Available at: https://doi.org/10.1038/s41598-021-89938-2.

Hao, Y. *et al.* (2021) 'Integrated analysis of multimodal single-cell data', *Cell*, 184(13), pp. 3573-3587.e29. Available at: https://doi.org/10.1016/j.cell.2021.04.048.

Heinrich, M.C. *et al.* (2003) 'PDGFRA activating mutations in gastrointestinal stromal tumors', *Science*, 299(5607), pp. 708–710. Available at: https://doi.org/10.1126/science.1079666.

Hicks, S.C. *et al.* (2018) 'Missing data and technical variability in single-cell RNA-sequencing experiments', *Biostatistics*, 19(4), pp. 562–578. Available at: https://doi.org/10.1093/biostatistics/kxx053.

Hiriart, E. *et al.* (2005) 'Interaction of the Epstein-Barr Virus mRNA Export Factor EB2 with Human Spen Proteins SHARP, OTT1, and a Novel Member of the Family, OTT3, Links Spen Proteins with Splicing Regulation and mRNA Export *', *Journal of Biological Chemistry*, 280(44), pp. 36935–36945. Available at: https://doi.org/10.1074/jbc.M501725200.

Hodges, C. *et al.* (2009) 'Nucleosomal Fluctuations Govern the Transcription Dynamics of RNA Polymerase II', *Science*, 325(5940), pp. 626–628. Available at: https://doi.org/10.1126/science.1172926.

Horizon (no date) *HAP1 knockout cell lines*. Available at: https://horizondiscovery.com/en/engineered-cell-lines/products/human-hap1-knockout-cell-lines (Accessed: 29 September 2022).

Hou, V.C. *et al.* (2002) 'Decrease in hnRNP A/B expression during erythropoiesis mediates a pre-mRNA splicing switch', *The EMBO Journal*, 21(22), pp. 6195–6204. Available at: https://doi.org/10.1093/emboj/cdf625.

Houck, K.A. *et al.* (1991) 'The Vascular Endothelial Growth Factor Family: Identification of a Fourth Molecular Species and Characterization of Alternative Splicing of RNA', *Molecular Endocrinology*, 5(12), pp. 1806–1814. Available at: https://doi.org/10.1210/mend-5-12-1806.

Hsu, H.L. *et al.* (1994) 'Positive and negative transcriptional control by the TAL1 helix-loop-helix protein.', *Proceedings of the National Academy of Sciences*, 91(13), pp. 5947–5951. Available at: https://doi.org/10.1073/pnas.91.13.5947.

Hsu, P.D. *et al.* (2013) 'DNA targeting specificity of RNA-guided Cas9 nucleases', *Nature Biotechnology*, 31(9), pp. 827–832. Available at: https://doi.org/10.1038/nbt.2647.

Hu, W., Yuan, B. and Lodish, H.F. (2014) 'Cpeb4-mediated translational regulatory circuitry controls terminal erythroid differentiation', *Dev. Cell*, 30(6), pp. 660–672. Available at: https://doi.org/10.1016/j.devcel.2014.07.008.

Huang, W. *et al.* (2021) 'N6-methyladenosine methyltransferases: functions, regulation, and clinical potential', *Journal of Hematology & Oncology*, 14(1), p. 117. Available at: https://doi.org/10.1186/s13045-021-01129-8.

Iwasaki, H. *et al.* (2003) 'GATA-1 Converts Lymphoid and Myelomonocytic Progenitors into the Megakaryocyte/Erythrocyte Lineages', *Immunity*, 19(3), pp. 451–462. Available at: https://doi.org/10.1016/S1074-7613(03)00242-5.

Jacobs, E., Mills, J.D. and Janitz, M. (2012) 'The role of RNA structure in posttranscriptional regulation of gene expression', *J. Genet. Genomics*, 39(10), pp. 535–543. Available at: https://doi.org/10.1016/j.jgg.2012.08.002.

Jin, S. *et al.* (2017) 'Splicing factor SF3B1K700E mutant dysregulates erythroid differentiation via aberrant alternative splicing of transcription factor TAL1', *PLOS ONE*, 12(5), p. e0175523. Available at: https://doi.org/10.1371/journal.pone.0175523.

Joshi, S.K. *et al.* (2019) 'Revisiting NTRKs as an emerging oncogene in hematological malignancies', *Leukemia*, 33(11), pp. 2563–2574. Available at: https://doi.org/10.1038/s41375-019-0576-8.

Kahles, A. *et al.* (2016) 'SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data', *Bioinformatics*, 32(12), pp. 1840–1847. Available at: https://doi.org/10.1093/bioinformatics/btw076.

Kikuchi, H. *et al.* (2011) 'GCN5 regulates the activation of PI3K/Akt survival pathway in B cells exposed to oxidative stress via controlling gene expressions of Syk and Btk', *Biochemical and Biophysical Research Communications*, 405(4), pp. 657–661. Available at: https://doi.org/10.1016/j.bbrc.2011.01.088.

Kim, E. *et al.* (2015) 'SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition', *Cancer Cell*, 27(5), pp. 617–630. Available at: https://doi.org/10.1016/j.ccell.2015.04.006.

Kim, E., Goren, A. and Ast, G. (2008) 'Alternative splicing: current perspectives', *Bioessays*, 30(1), pp. 38–47. Available at: https://doi.org/10.1002/bies.20692.

Kitamura, D. *et al.* (1989) 'Isolation and characterization of a novel human gene expressed specifically in the cells of hematopoietic lineage', *Nucleic Acids Res.*, 17(22), pp. 9367–9379.

Knuckles, P. *et al.* (2018) 'Zc3h13/Flacc is required for adenosine methylation by bridging the mRNA-binding factor Rbm15/Spenito to the m6A machinery component Wtap/Fl(2)d', *Genes & Development*, 32(5–6), pp. 415–429. Available at: https://doi.org/10.1101/gad.309146.117.

Kolasinska-Zwierz, P. *et al.* (2009) 'Differential chromatin marking of introns and expressed exons by H3K36me3', *Nature Genetics*, 41(3), pp. 376–381. Available at: https://doi.org/10.1038/ng.322.

Kolde, R. and Others (2019) 'Pheatmap: pretty heatmaps', *R package version*, 1(2), p. 747.

Kolodziejczyk, A.A. *et al.* (2015) 'Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation', *Cell Stem Cell*, 17(4), pp. 471–485. Available at: https://doi.org/10.1016/j.stem.2015.09.011.

König, J. *et al.* (2010) 'iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution', *Nat. Struct. Mol. Biol.*, 17(7), pp. 909–915. Available at: https://doi.org/10.1038/nsmb.1838.

Koralov, S.B. *et al.* (2008) 'Dicer Ablation Affects Antibody Diversity and Cell Survival in the B Lymphocyte Lineage', *Cell*, 132(5), pp. 860–874. Available at: https://doi.org/10.1016/j.cell.2008.02.020.

Kornblihtt, A.R. (2007) 'Coupling transcription and alternative splicing', *Adv. Exp. Med. Biol.*, 623, pp. 175–189. Available at: https://doi.org/10.1007/978-0-387-77374-2_11.

Kotecki, M., Reddy, P.S. and Cochran, B.H. (1999) 'Isolation and characterization of a near-haploid human cell line', *Exp. Cell Res.*, 252(2), pp. 273–280. Available at: https://doi.org/10.1006/excr.1999.4656.

Kuksin, M. *et al.* (2021) 'Applications of single-cell and bulk RNA sequencing in onco-immunology', *European Journal of Cancer*, 149, pp. 193–210. Available at: https://doi.org/10.1016/j.ejca.2021.03.005.

Landau, D.A. *et al.* (2013) 'Evolution and impact of subclonal mutations in chronic lymphocytic leukemia', *Cell*, 152(4), pp. 714–726. Available at: https://doi.org/10.1016/j.cell.2013.01.019.

Lander, E.S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. Available at: https://doi.org/10.1038/35057062.

Lee, J.-H. and Skalnik, D.G. (2012) 'Rbm15-Mkl1 Interacts with the Setd1b Histone H3-Lys4 Methyltransferase via a SPOC Domain That Is Required for Cytokine-Independent Proliferation', *PLOS ONE*, 7(8), p. e42965. Available at: https://doi.org/10.1371/journal.pone.0042965.

Leung, C.S. *et al.* (2023) 'Dysregulation of the chromatin environment leads to differential alternative splicing as a mechanism of disease in a human model of autism spectrum disorder', *Human Molecular Genetics*, p. ddad002. Available at: https://doi.org/10.1093/hmg/ddad002.

Li, D. and Tan, Y. (2021) 'Dysregulation of alternative splicing is associated with the pathogenesis of ulcerative colitis', *BioMedical Engineering OnLine*, 20(1), p. 121. Available at: https://doi.org/10.1186/s12938-021-00959-4.

Li, W. and Stanley, E.R. (1991) 'Role of dimerization and modification of the CSF-1 receptor in its activation and internalization during the CSF-1 response.', *The EMBO Journal*, 10(2), pp. 277–288. Available at: https://doi.org/10.1002/j.1460-2075.1991.tb07948.x.

Li, Yapu *et al.* (2021) 'A splicing factor switch controls hematopoietic lineage specification of pluripotent stem cells', *EMBO Rep.*, 22(1), p. e50535. Available at: https://doi.org/10.15252/embr.202050535.

Li, Yufeng *et al.* (2021) 'DRAK2 aggravates nonalcoholic fatty liver disease progression through SRSF6-associated RNA alternative splicing', *Cell Metabolism*, 33(10), pp. 2004-2020.e9. Available at: https://doi.org/10.1016/j.cmet.2021.09.008.

Lindtner, S. *et al.* (2006) 'RNA-binding Motif Protein 15 Binds to the RNA Transport Element RTE and Provides a Direct Link to the NXF1 Export Pathway*', *J. Biol. Chem.*, 281(48), pp. 36915–36928. Available at: https://doi.org/10.1074/jbc.M608745200.

Lion, T. *et al.* (1992) 'The Translocation t(l;22) (p13;q13) Is a Nonrandom Marker Specifically Associated With Acute Megakaryocytic Leukemia in Young Children', *Blood*, 79(12), pp. 3325–3330. Available at: https://doi.org/10.1182/blood.V79.12.3325.3325.

van Loosdregt, J. *et al.* (2010) 'Regulation of Treg functionality by acetylation-mediated Foxp3 protein stabilization', *Blood*, 115(5), pp. 965–974. Available at: https://doi.org/10.1182/blood-2009-02-207118.

Lozzio, C.B. and Lozzio, B.B. (1975) 'Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome', *Blood*, 45(3), pp. 321–334.

Lu, J. *et al.* (2008) 'MicroRNA-Mediated Control of Cell Fate in Megakaryocyte-Erythrocyte Progenitors', *Developmental Cell*, 14(6), pp. 843–853. Available at: https://doi.org/10.1016/j.devcel.2008.03.012.

Lu, X., Gross, A.W. and Lodish, H.F. (2006) 'Active Conformation of the Erythropoietin Receptor: RANDOM AND CYSTEINE-SCANNING MUTAGENESIS OF THE EXTRACELLULAR JUXTAMEMBRANE AND TRANSMEMBRANE DOMAINS *', *Journal of Biological Chemistry*, 281(11), pp. 7002–7011. Available at: https://doi.org/10.1074/jbc.M512638200.

Luco, R.F. *et al.* (2010) 'Regulation of Alternative Splicing by Histone Modifications', *Science*, 327(5968), pp. 996–1000. Available at: https://doi.org/10.1126/science.1184208.

Luco, R.F. *et al.* (2011) 'Epigenetics in alternative pre-mRNA splicing', *Cell*, 144(1), pp. 16–26. Available at: https://doi.org/10.1016/j.cell.2010.11.056.

Ma, X. *et al.* (2007) 'Rbm15 Modulates Notch-Induced Transcriptional Activation and Affects Myeloid Differentiation', *Molecular and Cellular Biology*, 27(8), pp. 3056–3064. Available at: https://doi.org/10.1128/MCB.01339-06.

Ma, Z. *et al.* (2001) 'Fusion of two novel genes, RBM15 and MKL1, in the t(1;22)(p13;q13) of acute megakaryoblastic leukemia', *Nature Genetics*, 28(3), pp. 220–221. Available at: https://doi.org/10.1038/90054.

Macaulay, I.C. and Voet, T. (2014) 'Single cell genomics: advances and future perspectives', *PLoS Genet.*, 10(1), p. e1004126. Available at: https://doi.org/10.1371/journal.pgen.1004126.

Maimon, A. *et al.* (2014) 'Mnk2 Alternative Splicing Modulates the p38-MAPK Pathway and Impacts Ras-Induced Transformation', *Cell Reports*, 7(2), pp. 501–513. Available at: https://doi.org/10.1016/j.celrep.2014.03.041.

Malcovati, L. *et al.* (2011) 'Clinical significance of SF3B1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms', *Blood*, 118(24), pp. 6239–6246. Available at: https://doi.org/10.1182/blood-2011-09-377275.

Mehmood, A. *et al.* (2020) 'Systematic evaluation of differential splicing tools for RNA-seq studies', *Brief. Bioinform.*, 21(6), pp. 2052–2065. Available at: https://doi.org/10.1093/bib/bbz126.

Michelle, L. *et al.* (2012) 'Proteins associated with the exon junction complex also control the alternative splicing of apoptotic regulators', *Mol. Cell. Biol.*, 32(5), pp. 954–967. Available at: https://doi.org/10.1128/MCB.06130-11.

Miller, R.M. *et al.* (2022) 'Enhanced protein isoform characterization through long-read proteogenomics', *Genome Biol.*, 23(1), p. 69. Available at: https://doi.org/10.1186/s13059-022-02624-y.

Milner, L.A. *et al.* (1996) 'Inhibition of granulocytic differentiation by mNotch1', *Proceedings of the National Academy of Sciences*, 93(23), pp. 13014–13019. Available at: https://doi.org/10.1073/pnas.93.23.13014.

Mincarelli, L. *et al.* (2023) 'Single-cell gene and isoform expression analysis reveals signatures of ageing in haematopoietic stem and progenitor cells', *Communications Biology*, 6(1), pp. 1–11. Available at: https://doi.org/10.1038/s42003-023-04936-6.

Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) 'Frequent Alternative Splicing of Human Genes', *Genome Research*, 9(12), pp. 1288–1293. Available at: https://doi.org/10.1101/gr.9.12.1288.

Motta-Mena, L.B., Heyd, F. and Lynch, K.W. (2010) 'Context-dependent regulatory mechanism of the splicing factor hnRNP L', *Mol. Cell*, 37(2), pp. 223–234. Available at: https://doi.org/10.1016/j.molcel.2009.12.027.

Naftelberg, S. *et al.* (2015) 'Regulation of alternative splicing through coupling with transcription and chromatin structure', *Annu. Rev. Biochem.*, 84, pp. 165–198. Available at: https://doi.org/10.1146/annurev-biochem-060614-034242.

NCBI (no date) *Primer-BLAST*. Available at: https://www.ncbi.nlm.nih.gov/tools/primer-blast/ (Accessed: 30 September 2022).

Ngian, Z.-K. *et al.* (2022) 'Truncated Tau caused by intron retention is enriched in Alzheimer's disease cortex and exhibits altered biochemical properties', *Proceedings of the National Academy of Sciences*, 119(37), p. e2204179119. Available at: https://doi.org/10.1073/pnas.2204179119.

Niu, C. *et al.* (2009) 'c-Myc is a target of RNA-binding motif protein 15 in the regulation of adult hematopoietic stem cell and megakaryocyte development', *Blood*, 114(10), pp. 2087–2096. Available at: https://doi.org/10.1182/blood-2009-01-197921.

Oh, H.K. *et al.* (2013) 'hnRNP A1 contacts exon 5 to promote exon 6 inclusion of apoptotic Fas gene', *Apoptosis*, 18(7), pp. 825–835. Available at: https://doi.org/10.1007/s10495-013-0824-8.

Ohlsson, E. *et al.* (2016) 'The multifaceted functions of C/EBPα in normal and malignant haematopoiesis', *Leukemia*, 30(4), pp. 767–775. Available at: https://doi.org/10.1038/leu.2015.324.

Okkenhaug, K. *et al.* (2002) 'Impaired B and T Cell Antigen Receptor Signaling in p110δ PI 3-Kinase Mutant Mice', *Science*, 297(5583), pp. 1031–1034. Available at: https://doi.org/10.1126/science.1073560.

Olbrich, T. *et al.* (2017) 'A p53-dependent response limits the viability of mammalian haploid cells', *Proc. Natl. Acad. Sci. U. S. A.*, 114(35), pp. 9367–9372. Available at: https://doi.org/10.1073/pnas.1705133114.

Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat. Genet.*, 40(12), pp. 1413–1415. Available at: https://doi.org/10.1038/ng.259.

Papaemmanuil, E. *et al.* (2011) 'Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts', *N. Engl. J. Med.*, 365(15), pp. 1384–1395. Available at: https://doi.org/10.1056/NEJMoa1103283.

Park, S.J. and Cochran, J.R. (2009) *Protein Engineering and Design*. CRC Press.

Patil, D.P. *et al.* (2016) 'm6A RNA methylation promotes XIST-mediated transcriptional repression', *Nature*, 537(7620), pp. 369–373. Available at: https://doi.org/10.1038/nature19342.

Pfeiffer, J.R. *et al.* (2011) 'CARHSP1 is required for effective tumor necrosis factor alpha mRNA stabilization and localizes to processing bodies and exosomes', *Mol. Cell. Biol.*, 31(2), pp. 277–286. Available at: https://doi.org/10.1128/MCB.00775-10.

Picelli, S. *et al.* (2013) 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nat. Methods*, 10(11), pp. 1096–1098. Available at: https://doi.org/10.1038/nmeth.2639.

Pronk, C.J.H. *et al.* (2007) 'Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy', *Cell Stem Cell*, 1(4), pp. 428–442. Available at: https://doi.org/10.1016/j.stem.2007.07.005.

Pucella, J.N., Upadhaya, S. and Reizis, B. (2020) 'The Source and Dynamics of Adult Hematopoiesis: Insights from Lineage Tracing', *Annual Review of Cell and Developmental Biology*, 36(1), pp. 529–550. Available at: https://doi.org/10.1146/annurev-cellbio-020520-114601.

Raffel, G.D. *et al.* (2007) 'Ott1(Rbm15) has pleiotropic roles in hematopoietic development', *Proc. Natl. Acad. Sci. U. S. A.*, 104(14), pp. 6001–6006. Available at: https://doi.org/10.1073/pnas.0609041104.

Rankin, E.B. and Sakamoto, K.M. (2018) 'The Cellular and Molecular Mechanisms of Hematopoiesis', in G.M. Kupfer, G.H. Reaman, and F.O. Smith (eds) *Bone Marrow Failure*. Cham: Springer International Publishing (Pediatric Oncology), pp. 1–23. Available at: https://doi.org/10.1007/978-3-319-61421-2_1.

Robinson, J.T. *et al.* (2011) 'Integrative genomics viewer', *Nat. Biotechnol.*, 29(1), pp. 24–26. Available at: https://doi.org/10.1038/nbt.1754.

Satija Lab, N.Y.G.C. (no date) *Tools for Single Cell Genomics*. Available at: https://satijalab.org/seurat/ (Accessed: 10 May 2022).

Schäfer, C. *et al.* (2003) 'CRHSP-24 phosphorylation is regulated by multiple signaling pathways in pancreatic acinar cells', *Am. J. Physiol. Gastrointest. Liver Physiol.*, 285(4), pp. G726-34. Available at: https://doi.org/10.1152/ajpgi.00111.2003.

Selheim, F. *et al.* (2000) 'Platelet-derived-growth-factor-induced signalling in human platelets: phosphoinositide-3-kinase-dependent inhibition of platelet activation', *Biochem. J*, 350 Pt 2, pp. 469–475.

Sévère, N., Miraoui, H. and Marie, P.J. (2011) 'The Casitas B lineage lymphoma (Cbl) mutant G306E enhances osteogenic differentiation in human mesenchymal stromal cells in part by decreased Cbl-mediated platelet-derived growth factor receptor alpha and fibroblast growth factor receptor 2 ubiquitination', *J. Biol. Chem.*, 286(27), pp. 24443–24450. Available at: https://doi.org/10.1074/jbc.M110.197525.

Sharp, P.A. (1994) 'Split genes and RNA splicing', *Cell*, 77(6), pp. 805–815. Available at: https://doi.org/10.1016/0092-8674(94)90130-9.

Shivdasani, R.A. *et al.* (1997) 'A lineage-selective knockout establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet development', *The EMBO Journal*, 16(13), pp. 3965–3973. Available at: https://doi.org/10.1093/emboj/16.13.3965.

Singh, R. (2018) 'RNA–Protein Interactions That Regulate Pre-mRNA Splicing', *Gene Expression*, 10(1–2), pp. 79–92.

Skokowa, J. *et al.* (2012) 'Interactions among HCLS1, HAX1 and LEF-1 proteins are essential for G-CSF-triggered granulopoiesis', *Nat. Med.*, 18(10), pp. 1550–1559. Available at: https://doi.org/10.1038/nm.2958.

Song, J.Y. *et al.* (2020) 'Hox genes maintain critical roles in the adult skeleton', *Proc. Natl. Acad. Sci. U. S. A.*, 117(13), pp. 7296–7304. Available at: https://doi.org/10.1073/pnas.1920860117.

Stanford University, E.P.C. (no date) *ENCODE*. Available at: https://www.encodeproject.org/ (Accessed: 23 September 2022).

Stegle, O., Teichmann, S.A. and Marioni, J.C. (2015) 'Computational and analytical challenges in single-cell transcriptomics', *Nature Reviews Genetics*, 16(3), pp. 133–145. Available at: https://doi.org/10.1038/nrg3833.

Tanaka, N. *et al.* (2014) 'Relative expression of hMena11a and hMenaINV splice isoforms is a useful biomarker in development and progression of human breast carcinoma', *International Journal of Oncology*, 45(5), pp. 1921–1928. Available at: https://doi.org/10.3892/ijo.2014.2591.

Tazi, J., Bakkour, N. and Stamm, S. (2009) 'Alternative splicing and disease', *Biochim. Biophys. Acta*, 1792(1), pp. 14–26. Available at: https://doi.org/10.1016/j.bbadis.2008.09.017.

*The Human Protein Atlas* (no date). Available at: https://www.proteinatlas.org/ (Accessed: 29 September 2022).

The Jackson Laboratory (2022) 'MGI-Mouse Genome Informatics-The international database resource for the laboratory mouse'. Available at: http://www.informatics.jax.org/ (Accessed: 22 March 2022).

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013) 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration', *Brief. Bioinform.*, 14(2), pp. 178–192. Available at: https://doi.org/10.1093/bib/bbs017.

Timchenko, L.T. *et al.* (1996) 'Identification of a (CUG) n Triplet Repeat RNA-Binding Protein and Its Expression in Myotonic Dystrophy', *Nucleic Acids Research*, 24(22), pp. 4407–4414. Available at: https://doi.org/10.1093/nar/24.22.4407.

Trowbridge, J.J. *et al.* (2009) 'DNA Methyltransferase 1 Is Essential for and Uniquely Regulates Hematopoietic Stem and Progenitor Cells', *Cell Stem Cell*, 5(4), pp. 442–449. Available at: https://doi.org/10.1016/j.stem.2009.08.016.

Uciechowski, P. and Dempke, W.C.M. (2020) 'Interleukin-6: A Masterplayer in the Cytokine Network', *Oncology*, 98(3), pp. 131–137. Available at: https://doi.org/10.1159/000505099.

Ullrich, S. and Guigó, R. (2020) 'Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development', *Nucleic Acids Research*, 48(3), pp. 1327–1340. Available at: https://doi.org/10.1093/nar/gkz1180.

Uranishi, H. *et al.* (2009) 'The RNA-binding Motif Protein 15B (RBM15B/OTT3) Acts as Cofactor of the Nuclear Export Receptor NXF1 *', *Journal of Biological Chemistry*, 284(38), pp. 26106–26116. Available at: https://doi.org/10.1074/jbc.M109.040113.

Urbanski, L.M., Leclair, N. and Anczuków, O. (2018) 'Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics', *WIREs RNA*, 9(4), p. e1476. Available at: https://doi.org/10.1002/wrna.1476.

Vaquero-Garcia, J. *et al.* (2016) 'A new view of transcriptome complexity and regulation through the lens of local splicing variations', *Elife*, 5, p. e11752. Available at: https://doi.org/10.7554/eLife.11752.

Vu, L.P., Luciani, L. and Nimer, S.D. (2013) 'Histone-modifying enzymes: their role in the pathogenesis of acute leukemia and their therapeutic potential', *International Journal of Hematology*, 97(2), pp. 198–209. Available at: https://doi.org/10.1007/s12185-012-1247-y.

Wadman, I.A. *et al.* (1997) 'The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins', *The EMBO Journal*, 16(11), pp. 3145–3157. Available at: https://doi.org/10.1093/emboj/16.11.3145.

Wahl, M.C. and Lührmann, R. (2015) 'SnapShot: Spliceosome Dynamics II', *Cell*, 162(2), pp. 456-456.e1. Available at: https://doi.org/10.1016/j.cell.2015.06.061.

Walter, K., Bonifer, C. and Tagoh, H. (2008) 'Stem cell–specific epigenetic priming and B cell–specific transcriptional activation at the mouse Cd19 locus', *Blood*, 112(5), pp. 1673–1682. Available at: https://doi.org/10.1182/blood-2008-02-142786.

Walter, M.J. *et al.* (2013) 'Clonal diversity of recurrently mutated genes in myelodysplastic syndromes', *Leukemia*, 27(6), pp. 1275–1282. Available at: https://doi.org/10.1038/leu.2013.58.

Wang, C. *et al.* (2014) 'Depletion of Sf3b1 impairs proliferative capacity of hematopoietic stem cells but is not sufficient to induce myelodysplasia', *Blood*, 123(21), pp. 3336–3343. Available at: https://doi.org/10.1182/blood-2013-12-544544.

Wang, H., La Russa, M. and Qi, L.S. (2016) 'CRISPR/Cas9 in Genome Editing and Beyond', *Annual Review of Biochemistry*, 85(1), pp. 227–264. Available at: https://doi.org/10.1146/annurev-biochem-060815-014607.

Wang, Q. *et al.* (2008) 'MicroRNA miR-24 inhibits erythropoiesis by targeting activin type I receptor ALK4', *Blood*, 111(2), pp. 588–595. Available at: https://doi.org/10.1182/blood-2007-05-092718.

Wang, Y. *et al.* (2015) 'Mechanism of alternative splicing and its regulation', *Biomedical Reports*, 3(2), pp. 152–158. Available at: https://doi.org/10.3892/br.2014.407.

Wang, Z. *et al.* (2004) 'Systematic Identification and Analysis of Exonic Splicing Silencers', *Cell*, 119(6), pp. 831–845. Available at: https://doi.org/10.1016/j.cell.2004.11.010.

Wiellette, E.L. *et al.* (1999) 'spen encodes an RNP motif protein that interacts with Hox pathways to repress the development of head-like sclerites in the Drosophila trunk',

*Development*, 126(23), pp. 5373–5385. Available at: https://doi.org/10.1242/dev.126.23.5373.

Woolthuis, C.M. and Park, C.Y. (2016) 'Hematopoietic stem/progenitor cell commitment to the megakaryocyte lineage', *Blood*, 127(10), pp. 1242–1248. Available at: https://doi.org/10.1182/blood-2015-07-607945.

Xiao, N. *et al.* (2015) 'Ott1 (Rbm15) regulates thrombopoietin response in hematopoietic stem cells through alternative splicing of c-Mpl', *Blood*, 125(6), pp. 941–948. Available at: https://doi.org/10.1182/blood-2014-08-593392.

Yang, G. *et al.* (2005) 'An erythroid differentiation–specific splicing switch in protein 4.1R mediated by the interaction of SF2/ASF with an exonic splicing enhancer', *Blood*, 105(5), pp. 2146–2153. Available at: https://doi.org/10.1182/blood-2004-05-1757.

Yasmin, N. *et al.* (2013) 'Identification of bone morphogenetic protein 7 (BMP7) as an instructive factor for human epidermal Langerhans cell differentiation', *Journal of Experimental Medicine*, 210(12), pp. 2597–2610. Available at: https://doi.org/10.1084/jem.20130275.

Yoshida, K. *et al.* (2011) 'Frequent pathway mutations of splicing machinery in myelodysplasia', *Nature*, 478(7367), pp. 64–69. Available at: https://doi.org/10.1038/nature10496.

Yuan, T.L. and Cantley, L.C. (2008) 'PI3K pathway alterations in cancer: variations on a theme', *Oncogene*, 27(41), pp. 5497–5510. Available at: https://doi.org/10.1038/onc.2008.245.

Zaccara, S., Ries, R.J. and Jaffrey, S.R. (2019) 'Reading, writing and erasing mRNA methylation', *Nature Reviews Molecular Cell Biology*, 20(10), pp. 608–624. Available at: https://doi.org/10.1038/s41580-019-0168-5.

Zhang, L. *et al.* (2015) 'Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing', *Elife*, 4. Available at: https://doi.org/10.7554/eLife.07938.

Zhao, Z. *et al.* (2022) 'N6-Methyladenosine Methylation Regulator RBM15 is a Potential Prognostic Biomarker and Promotes Cell Proliferation in Pancreatic Adenocarcinoma', *Front Mol Biosci*, 9, p. 842833. Available at: https://doi.org/10.3389/fmolb.2022.842833.

Zheng, G.X.Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nat. Commun.*, 8, p. 14049. Available at: https://doi.org/10.1038/ncomms14049.

Zolotukhin, A.S. *et al.* (2009) 'Nuclear export factor RBM15 facilitates the access of DBP5 to mRNA', *Nucleic Acids Res.*, 37(21), pp. 7151–7162. Available at: https://doi.org/10.1093/nar/gkp782.